АНАЛІЗ БІОМЕТРИЧНИХ ДАНИХ У РОЗВЕДЕННІ ТА СЕЛЕКЦІЇ ТВАРИН

НАВЧАЛЬНИЙ ПОСІБНИК

Миколаїв

МНАУ

2019

УДК 57.087.1: 519.2
А64

Авторський колектив:

С. С. Крамаренко
С. І. Луговий
А. В. Лихач
О. С. Крамаренко

Друкується за рішенням Вченої ради Миколаївського національного аграрного університету від «25» червня 2019 р., протокол № 11.

Рецензенти:

М. Д. Березовський – д-р с.-г. наук, професор, член-кореспондент НААН, головний науковий співробітник лабораторії селекції Інституту свинарства і агропромислового виробництва НААН України;

Р. В. Ставецька – д-р с.-г. наук, доцент, завідувач кафедри генетики, розведення та селекції тварин Білоцерківського національного аграрного університету;

Р. Л. Сусол – д-р с.-г. наук, доцент, завідувач кафедри технології виробництва і переробки продукції тваринництва Одеського державного аграрного університету.

Аналіз біометричних даних у розведенні та селекції тварин : навчальний посібник / С. С. Крамаренко, С. І. Луговий, А. В. Лихач, О. С. Крамаренко. – Миколаїв : МНАУ, 2019. – 211 с.
ISBN 978-617-7149-38-4

У навчальному посібнику наведено методи аналізу біометричних даних: генетичних, якісних і кількісних. Особливу увагу приділено одному із найсучасніших напрямків аналізу даних – ресамплінгу та його різним варіаціям.

Посібник розрахований на студентів, аспірантів, викладачів закладів вищої освіти аграрного та природничого напрямів та спеціалістів, які мають справу з біометричною обробкою результатів досліджень.

УДК 57.087.1: 519.2

ISBN 978-617-7149-38-4

ЗМІСТ

ВСТУП

Вже майже 40 років пройшло відтоді, як було видано посібник М. О. Плохинського «Руководство по биометрии для зоотехников». Однак, до сьогодні виконання практично кожної дисертації із розведення, селекції та генетики сільськогосподарських тварин не обходиться без використання цього видання. Й причин тут дві: досить високий рівень видання, з одного боку, а з іншого – відсутність інших, більш-менш вдалих посібників, які б могли замінити в роботі фахівця-селекціонера посібник М. О. Плохинського.

За роки, що минули, значно змінилися й підходи до аналізу біометричних даних. З'явилися й поширилися як ПЕОМ, так і програмне забезпечення, спрямоване на проведення статистико-математичного аналізу даних у різних галузях науки та техніки, в тому числі й біометричних. Тому змінився й принцип підготовки видань методичного спрямування – замість розробки методів прискорення та вдосконалення розрахункової частини аналізу (на що було зроблено особливий акцент у роботах по біометрії 1950-80-х рр.) зараз необхідно більше приділяти уваги новим методам та підходам до аналізу даних.

Вважаємо, що частково це завдання може вирішити дане видання. Воно складається із трьох частин. Частина І містить інформацію щодо методів аналізу генетичних даних. Цей напрямок зараз отримав значне поширення, особливо, після впровадження в селекційну роботу імуногенетичних та молекулярно-генетичних маркерів із кодомінантним (білковий поліморфізм крові чи молока, мікросателіти ДНК та ін.) чи домінантним (антигенні фактори крові, RAPD-маркери та ін.) типом успадкування.

Частина ІІ містить інформацію щодо методів аналізу якісних ознак (фенів). Причому особлива увага приділяється різноманітним варіантам дисперсійного аналізу якісних ознак та найсучаснішим методам аналізу просторово-структурованих популяцій (метапопуляційний підхід).

Частина ІІІ містить ряд базових біометричних підходів, але, поряд із цим, приведено аналітичні моделі, що використовуються при аналізі формування тваринницької продукції (моделі росту тварин, лактаційні криві, криві яєчної продуктивності). Особлива увага приділяється напрямку, який здобуває у останній час все більшого розповсюдження у генетико-селекційній роботі, а саме, інформаційно-ентропійному аналізу кількісних ознак.

У посібнику немає інформації стосовно методів непараметричної статистики. Для всіх бажаючих опанувати цими методами аналізу біометричних даних ми рекомендуємо наше попереднє видання: О. В. Шебаніна, С. С. Крамаренко, В. М. Ганганов. Практикум з біометрії. Методи непараметричної статистики. – Миколаїв: МДАУ, 2008. – 166 с.

Але при цьому, ми включили у дане видання ряд методик, які мають загальну назву resampling-процедури й базуються на багатократному відборі із вибірки вихідних даних псевдовибірок та їх наступний аналіз. Майже для кожного біометричного показника чи критерію ми наводимо приклади його

розрахунку з використанням різноманітних ресампінг-процедур: bootstrap- jackknife- або permutation-методів.

Всі відгуки, побажання чи зауваження стосовно даного видання просимо надсилати за електронною адресою: kssnail0108@gmail.com (Крамаренку С. С.).

ЧАСТИНА І
АНАЛІЗ ГЕНЕТИ ЧНИХ ДАНИХ

§ 1. Методи вивчення популяційних закономірностей

Генетика популяцій розглядає статистичні наслідки законів Г. Менделя, що виявляються в групі чи родині особин.

Предмет генетики популяцій – явище спадковості на популяційному рівні.

Основні завдання генетики популяцій:

- опис популяцій та їхнього генетичного складу;
- аналіз причин зміни генофонду.

У генетичному аспекті, **популяція** – *це просторово-часова група особин одного виду, які вільно схрещуються між собою*. Для неї характерно:

- вільне схрещування (панміксія);
- відсутність вибірковості при підборі чоловічих і жіночих організмів;
- відсутність вибірковості злиття гамет при заплідненні.

Розрізняють **природні популяції**, що формуються в природних умовах під впливом природного відбору, і **популяції, штучно сформовані людиною** в процесі штучного відбору і створення специфічних умов середовища.

У тваринництві під **популяцією** розуміють *досить велику групу тварин, пов'язаних спільністю походження, тривалістю розведення у певних зовнішніх умовах і подібними за напрямком розвитку ознаками продуктивності*.

Кожна популяція має визначену генетичну структуру і генофонд. **Генофондом** називається сукупність усіх генів, що мають члени сукупності. **Генетична структура** визначається концентрацією кожного гена (чи його алелів) у популяції, характером генотипів і частотою їхнього розподілу.

Алель – *одна із двох чи більше альтернативних форм гена, кожна з яких характеризується унікальною послідовністю нуклеотидів*; різні алелі даного гена звичайно відрізняються фенотипово.

Генотип – *вся генетична інформація організму, генетична структура організму за одним чи декількома досліджуваними локусами*.

Фенотип – *ознаки особини, що виявляються в результаті реалізації генотипу в певних умовах середовища*.

Локус – *місце розташування даного гена (чи мутації) на генетичній карті*.

Основним поняттям популяційної генетики є частота (насамперед, частота алеля). Для того, щоб однозначно охарактеризувати це поняття, необхідно спочатку звернутися до основ теорії ймовірності.

Імовірність (P) можна визначити, як *кількість сприятливих реалізацій будь-якої події, віднесене до загальної кількості можливих реалізацій*.

Наприклад, якщо тварина гетерозиготна за мастю (Rr), де червона масть обумовлена алелем R, а біла – r, то ймовірність того, що в гаметі виявиться алель R, дорівнює: $P(R) = \frac{1}{2}$.

У цьому випадку можливими результатами є два (*R* і *r*), а сприятливим ми вважаємо лише один (*R*).

Імовірність будь-якої події може набувати значення від 0 до 1. Імовірність, що дорівнює 1 означає, що подія обов'язково відбудеться і її називають *достовірною подією*; імовірність, що дорівнює 0 означає, що подія, навпаки, ніколи не відбудеться, і тоді її називають *неможливою подією*.

Як і з будь-якими числами, з ймовірностями можна робити арифметичні дії (додавати, множити, віднімати і ділити). Ці дії виконують на підставі двох нижченаведених законів.

Закон додавання ймовірностей. Імовірність того, що реалізується один із декількох взаємовиключних результатів даної події, дорівнює сумі ймовірностей кожного окремого результату.

Наприклад, при схрещуванні двох гетерозиготних тварин (*Rr*) імовірності появи нащадків із різними генотипами складають:

гомозигот *RR*: $P(RR) = 1/4$;
гетерозигот *Rr*: $P(Rr) = 2/4 = 1/2$;
гомозигот *rr*: $P(rr) = 1/4$.

Отже, ймовірність того, що серед нащадків з'явиться домінантний фенотип (тобто, що будь-яка тварина буде мати генотип або *RR*, або *Rr*) дорівнює:

$$P(R\_) = 1/4 + 1/2 = 3/4 .$$

Приклад. У стаді великої рогатої худоби породи шортгорн 250 тварин мали червону масть, 700 – чалу і 300 – білу.

Яка ймовірність того, що випадковим чином відібрана тварина буде не білої масті?

Скільки можна очікувати особин не білої масті в групі із 100 випадково відібраних тварин?

Усього в стаді: $N = 250 + 700 + 300 = 1250$ голів. Імовірність того, що одна випадковим чином відібрана тварина буде:

- червоної масті : $P(Че) = 250 : 1250 = 0{,}20$;
- чалої масті: $P(Ча) = 700 : 1250 = 0{,}56$;
- білої масті: $P(Б) = 300 : 1250 = 0{,}24$.

Зробимо перевірку, сума ймовірностей повинна становити 1,0:

$$0{,}20 + 0{,}56 + 0{,}24 = 1{,}00.$$

Отже, розрахунок імовірностей зроблено вірно. Імовірність того, що випадковим чином відібрана тварина буде мати або червону, або чалу масть дорівнює:

$$P(Че+Ча) = 0{,}2 + 0{,}56 = 0{,}76.$$

З іншого боку, оскільки сума ймовірностей взаємовиключних подій дорівнює 1, шукану імовірність можна знайти іншим способом. Оскільки P(Б) + P(не Б) = 1, то P(не Б) = 1 – P(Б), звідки P(не Б) = 1 – 0,24 = 0,76, де P(не Б) – імовірність того, що тварина має не білу масть.

На друге питання можна знайти відповідь, використовуючи формулу для розрахунку імовірності:

$$P(Б) = \frac{n_Б}{N}, \qquad (1.1)$$

де $n_Б$ – кількість тварин, що мають білу масть;

N – загальна кількість досліджуваних тварин.

Звідси можна оцінити величину $n_Б$ для будь-якого обсягу вибірки:

$$n_Б = P(Б)N. \qquad (1.2)$$

Отже, найбільш імовірна кількість особин білої масті серед будь-яких 100 тварин, відібраних з даної популяції, буде: $n_Б = 0{,}24 \times 100 = 24$.

Даний висновок можна узагальнити таким чином: кількість успіхів у n випробуваннях дорівнює кількості випробувань, помноженій на імовірність успіху при одному випробуванні.

Але необхідно пам’ятати, що на практиці у групі із 100 випадковим чином відібраних тварин із 1250, білих може виявитися будь-яка кількість – від 0 до 100. Але найбільш імовірна їх кількість становить 24. Більш докладно цю ситуацію буде розглянуто нижче.

Закон множення ймовірностей. Імовірність того, що декілька взаємно незалежних результатів реалізуються одночасно, дорівнює добутку ймовірностей кожного із цих результатів.

Наприклад, імовірність того, що при схрещуванні *Rr* × *Rr* певна тварина в потомстві одержить алель *r* від одного з батьків дорівнює 1/2; така ж імовірність того, що ця тварина одержить алель *r* від іншого із батьків. Отже, імовірність того, що ця тварина одержить алель *r* від обох батьків, дорівнює:

$$P(rr) = 1/2 \times 1/2 = 1/4.$$

Приклад. У стаді зі 150 голів великої рогатої худоби 30 тварин мали чорно-рябу масть, а 120 – червоно-рябу. При цьому 90 тварин були комолі, а інші 60 – рогаті.

Яка ймовірність того, що випадковим чином відібрана тварина буде мати чорно-рябу масть і буде комолою, якщо обидві ознаки успадковуються незалежно?

Спочатку розрахуємо ймовірності того, що дана тварина буде мати:

- чорно-рябу масть: P(ЧнР) = 30 : 150 = 0,2;
- червоно-рябу масть: P(ЧрР) = 120 : 150 = 0,8.

З іншого боку, ймовірність того, що дана тварина буде:

- комолою: P(К) = 90 : 150 = 0,6;
- рогатою: P(Р) = 60 : 150 = 0,4.

Тоді, ймовірність того, що будь-яка випадковим чином відібрана тварина буде одночасно чорно-рябої масті і комолою дорівнює:

$$P(\text{ЧнР} / \text{К}) = P(\text{ЧнР}) \times P(\text{К}) = 0{,}2 \times 0{,}6 = 0{,}12.$$

Таким чином, серед 150 тварин обидві ці ознаки будуть мати:

n (ЧнР / К) = 0,12 × 150 = 18 голів.

Часто ймовірність того чи іншого результату заздалегідь не відома. Тоді її можна визначити експериментально, спостерігаючи з якою частотою реалізується даний результат.

Частота (p_A) визначається як *відношення кількості реалізації даного результату (n) до загальної кількості проведених випробувань* (N):

$$p_A = \frac{n}{N}. \qquad (1.3)$$

Таким чином, частота події A розглядається як оцінка ймовірності цієї події $p \to P(A)$, якщо кількість реалізацій прагне до нескінченості ($N \to \infty$).

Оскільки на практиці оцінка частоти проводиться на основі вибіркової сукупності з популяції, то розглянуте значення частоти буде завжди відхилятися в той чи інший бік від значення імовірності даної події. Величина цього відхилення обернено пропорційна обсягу вибірки і називається *помилкою частоти* (SEp) і розраховується за формулою:

$$SEp = \sqrt{\frac{p(1-p)}{N}}. \qquad (1.4)$$

Таким чином, при $N \to \infty$ значення помилки частоти події A буде наближатися до нуля і в цьому випадку оцінки частоти та ймовірності розглянутої події будуть збігатися, тобто $p = P(A)$.

Однак є і ще одне ускладнення. Наприклад, імовірність того, що новонароджена дитина виявиться хлопчиком, дорівнює 0,5. У дійсності ж частота хлопчиків серед усіх немовлят загалом не завжди і не скрізь буде однакова. Так, Д. Сміт (1970) наводить наступні цифри: в Англії й Уельсі народжується приблизно 106 хлопчиків на кожні 100 дівчаток. Це можна пояснити різними причинами. Наприклад, можливо, що в одних жінок внутрішньоматкове середовище сприяє розвитку зародка чоловічої статі, а в інших – жіночої.

У багатьох випадках частота однієї з двох реалізацій у генеральній сукупності підказується сутністю досліджуваного явища. Наприклад, співвідношення 3:1 для домінантних і рецесивних форм при розщепленні гібридів при моногібридному схрещуванні (тобто p_0 = 0,75 для домінантної форми), чи p_0 = 0,5 при народженні тварин кожної статі і т.д.

Оскільки завжди вивчається вибірка, то зазвичай потрібно перевірити, чи можна розбіжність між фактичною (p) і очікуваною (p_0) частотою даної події пояснити впливом вибіркового варіювання.

У випадку досить великого обсягу вибірки це завдання може бути вирішене, використовуючи u-критерій:

$$u = \frac{|p_0 - p|}{SEp} = \frac{|p_0 - p|}{\sqrt{\frac{p(1-p)}{N}}}. \qquad (1.5)$$

Розходження вважаються вірогідними, якщо розраховане значення критерію перевищує 1,96.

Приклад. За матеріалами свинокомплексу «Калитянський» з 1976 р. по 2000 р. було отримано в опоросах 71 043 357 свинок і 75 984 558 кнурців.

Яка частота народження кнурців? Чи відповідає вона очікуваній (тобто, 0,5)?

Частота народження кнурців становить:

$$p = \frac{75984558}{71043357 + 75984558} = 0{,}517.$$

Таким чином, частота народження свинок становить, відповідно, 0,483. Вірогідність відхилення отриманих частот від очікуваних оцінимо, використовуючи *u*-критерій:

$$u = \frac{|0{,}5 - 0{,}517|}{\sqrt{\frac{0{,}517 \cdot 0{,}483}{147027915}}} = 412{,}51,$$

що набагато перевищує 1,96.

Таким чином, кнурці народжувалися набагато частіше, ніж можна очікувати при рівномірному співвідношенні статей. І гіпотеза, що p = 0,5 повинна бути відхилена.

У тих випадках, коли є однакова ймовірність прояву однієї з двох можливих ознак (наприклад, народження бичка чи телички, кнурця чи свинки і т. ін.) і доводиться працювати з малими вибірками, нормальна апроксимація неможлива. Тому доводиться використовувати непараметричні методи порівняння, а саме – **критерій знаків.**

Основою для порівняння служить фактична кількість реалізацій однієї з двох альтернатив (вибирається менше значення), що порівнюється з табличним значенням (для відповідного обсягу вибірки).

Табличні значення приведені в додатку А. Якщо фактичне значення виявляється більшим, ніж табличне, гіпотеза про рівну ймовірність появи альтернативних ознак приймається. У іншому випадку – гіпотеза відхиляється і вважається, що $p \neq 0{,}5$.

Приклад. У 10 виводках отримано 51 курочку і 37 півників. Чи суперечить це гіпотезі, що півники і курочки з'являються з рівною ймовірністю?

Розраховуємо загальний обсяг вибірки: n = 51 + 37 = 88. Згідно додатку А знаходимо, що критичне значення кількості півників (оскільки їх менше) складає 35. Оскільки фактичне значення (37) більше критичного, то гіпотеза про рівну ймовірність появи півників і курочок не відхиляється.

Контрольні питання:

1. Поняття про популяцію. Види популяцій.
2. Імовірність. Додавання та множення імовірностей.
3. Розрахунок частоти та її помилки.

§ 2. Закони розподілу дискретних імовірностей

Біноміальний розподіл. Дотепер ми говорили про одиничний експеримент, тобто, оцінювали ймовірність результату, що нас цікавить, при проведенні одного експерименту. Наприклад, якщо в стаді із 100 голів великої рогатої худоби 27 тварин були комолі, то ймовірність того, що одна випадковим чином відібрана тварина буде комолою становить $P(\text{К}) = 0{,}27$, а того, що рогатою – $P(\text{Р}) = 1{,}00 - 0{,}27 = 0{,}73$.

Але якщо з цього ж стада відібрати інші 100 тварин, то яка впевненість, що серед них комолими виявляться 27, а не 26 чи 28, чи навіть 20? А якщо відібрати 100 вибірок (із поверненням) по 100 тварин, то який характер розподілу будуть нести самі ймовірності, оцінені по кожному зі 100 експериментів?

З іншого боку, якщо відібрати з усього стада лише 5 тварин, яка ймовірність того, що всі п’ять з них будуть комолі, чи, навпаки, усі п’ять будуть мати роги?

Імовірність подібних подій описується біноміальним розподілом на підставі формули Бернуллі:

$$p_n(m) = \frac{n!}{m!(n-m)!} p^m q^{n-m}, \qquad (2.1)$$

де $p_n(m)$ – імовірність того, що з n випадковим чином відібраних з популяції організмів рівно m будуть мати ознаку, що нас цікавить;

p – імовірність появи даної ознаки в однієї особи;

$q = 1 - p$.

Наприклад, імовірність того, що всі п’ять тварин з розглянутого вище прикладу будуть комолими становить:

$$p_5(5) = \frac{5!}{5!(5-5)!} 0{,}27^5 0{,}73^{5-5} = 0{,}00143,$$

а імовірність того, що всі п’ять тварин будуть рогатими:

$$p_5(0) = \frac{5!}{0!(5-0)!} 0{,}27^0 0{,}73^{5-0} = 0{,}2073,$$

а того, що з п’яти випадковим чином відібраних тварин троє будуть комолими, а двоє – рогатими:

$$p_5(3) = \frac{5!}{3!(5-3)!} 0{,}27^3 0{,}73^{5-3} = 0{,}1049.$$

Повністю ймовірності всіх можливих варіантів для даного прикладу можна зобразити у вигляді таблиці (табл. 2.1).

Отже, найбільш імовірна поява серед п’яти випадковим чином відібраних тварин лише однієї комолої:

$$p_5(1) = 0{,}3834.$$

У цілому ж, імовірність появи серед п’яти відібраних не більше двох комолих складає:

$$p_5(0) + p_5(1) + p_5(2) = 0{,}2073 + 0{,}3834 + 0{,}2836 = 0{,}8743,$$

тобто, складає практично 90%. (Використано закон додавання ймовірностей; див. вище).

Таблиця 2.1 – Імовірності появи серед п’яти випадковим чином відібраних корів різної кількості комолих

m (кількість комолих)	C_n^m	p^m	q^{n-m}	$p_n(m)$	$mp_n(m)$
0	1	1	0,2073	0,2073	0
1	5	0,27	0,2840	0,3834	0,3834
2	10	0,0729	0,3890	0,2836	0,5672
3	10	0,01968	0,5329	0,1049	0,3147
4	5	0,00531	0,73	0,0194	0,0776
5	1	0,00143	1	0,0014	0,0070
Сума	32	×	×	1,0000	1,3499

Таким чином, якщо ми відберемо з розглянутої популяції сто разів по п’ять тварин, то в 87 випадках з них число комолих серед цих п’яти відібраних буде не більше двох (тобто, не зустрінеться жодної комолої, або зустрінеться одна чи дві комолі тварини).

Величини в другому стовпчику являють собою біноміальні коефіцієнти:

$$C_n^m = \frac{n!}{m!(n-m)!}. \qquad (2.2)$$

Для зручності ці коефіцієнти для обсягів вибірки від 1 до 12 приведені в додатку Б.

З іншого боку, якщо необхідно розрахувати очікувану кількість тварин, які мають ознаку, що нас цікавить, серед групи відібраних (у приведеному вище прикладі – середню кількість комолих з п’яти випадковим чином відібраних), то для цього необхідно знайти суму добутків імовірностей появи кожного варіанта результату (немає комолих, одна комола, дві комолих і т. ін.) на відповідне значення варіанта (тобто, кількість комолих з п’яти відібраних тварин). Ці значення приведено в останньому стовпчику таблиці 2.1.

Таким чином, найбільш очікувана кількість комолих з п’яти відібраних (при частоті комолих у всій популяції $p = 0{,}27$) становить:

$$N_m(n) = 0 \times 0{,}2073 + 1 \times 0{,}3834 + 2 \times 0{,}2836 + \ldots + 5 \times 0{,}0014 = 1{,}3499.$$

Однак простіше одержати це значення, скориставшись особливістю біноміального розподілу:

$$N_m(n) = pn, \qquad (2.3)$$

де p – імовірність появи даної події у однієї тварини;

n – обсяг вибірки.

У нашому випадку $p = 0{,}27$, $n = 5$, отже, очікувана (тобто найбільш імовірна) кількість комолих серед п’яти відібраних складатиме:

$$N_m(n) = 0{,}27 \times 5 = 1{,}35,$$

що близько до отриманого вище значення з точністю до округлення. Таким чином, серед десяти відібраних тварин найбільше ймовірно буде 2,7 комолих, а серед ста відібраних – 27 (як і дано за умовою прикладу; дивись вище).

Приклад. Припустимо, що схрещують мишу-альбіноса і мишу гомозиготного нормального типу (має забарвлене хутро). Яка ймовірність народження двох альбіносів із шести мишей у другому поколінні?

У першому поколінні усі миші будуть мати забарвлення. В другому ж поколінні забарвленими будуть 3/4 усіх мишей, а, відповідно, інші 1/4 будуть альбіносами. Таким чином, імовірність появи альбіноса в другому поколінні становить 1/4. Яка ж імовірність появи двох альбіносів із шести мишей у гнізді?

Скористаємося формулою Бернуллі, де $n = 6$, $m = 2$ і $p = 1/4$. Тоді,

$$p_6(2) = \frac{2!}{2!4!} 0{,}25^2 0{,}75^4 = 0{,}297.$$

Таким чином, імовірність появи двох альбіносів із шести мишей у другому поколінні складає близько 30%.

Оскільки найчастіше стать новонародженої особини не залежить від статі іншої особини певного гнізда, то імовірності появи визначеної кількості чоловічих чи жіночих нащадків у гнізді багатоплідних тварин (свиней, птиці, мишей, пацюків та ін.) також повинні відповідати біноміальному розподілу.

Приклад. У колонках 1 і 2 таблиці 2.2 наведено розподіл кількості півників у 80 виводках по 12 курчат у кожному. Чи можна цей розподіл вважати біноміальним?

Таблиця 2.2 – Розподіл 80 виводків по 12 курчат у кожному, що містять різну кількість півників

Кількість півників, m	Фактична кількість виводків	Біноміальні коефіцієнти, C_n^m	Теоретична кількість виводків
1	2	3	4
0	1	1	0,0
1	0	12	0,2
2	0	66	1,3
3	6	220	4,3
4	11	495	9,7
5	13	792	15,5
6	19	924	18,0
7	16	792	15,5
8	7	495	9,7
9	4	220	4,3
10	2	66	1,3
11	1	12	0,2

Продовження таблиці 2.2

1	2	3	4
12	0	1	0,0
Сума	80	4096	80,0

Насамперед, необхідно розрахувати частоту півників. Для цього розділимо сумарну кількість півників у 80 виводках на сумарну кількість усіх курчат (тобто 80 × 12):

$$p = \frac{0 \cdot 1 + 1 \cdot 0 + 2 \cdot 0 + 3 \cdot 6 + 4 \cdot 11 + ... + 12 \cdot 0}{12 \cdot 80} = 0{,}496.$$

Отже, частота курочок – q = 0,504. Як бачимо, ці значення мало відрізняються від очікуваних імовірностей появи особини різної статі (тобто, 0,5).

У тих випадках, коли імовірності двох альтернативних результатів однакові (і рівні, відповідно, 0,5), очікувані ймовірності можна одержати із формули:

$$p_n(m) = \frac{1}{2^n} C_n^m. \qquad (2.4)$$

Біноміальні коефіцієнти, приведені в третьому стовпчику, ми взяли з додатку Б. Тоді, ймовірність виводка, що містить, наприклад, чотири півника буде становити:

$$p_{12}(4) = \frac{1}{2^{12}} \cdot 495 = 0{,}1208.$$

Отже, теоретична кількість виводків, що містять чотири півники із 12 курчат у виводку, складає: $n_4^{12} = p_4^{12} N = 0{,}1208 \times 80 = 9{,}67.$ Ці значення і приведено в останній колонці таблиці 2.2.

Наскільки відповідає емпіричний розподіл біноміальному можна також проаналізувати на графіку (рис. 2.1).

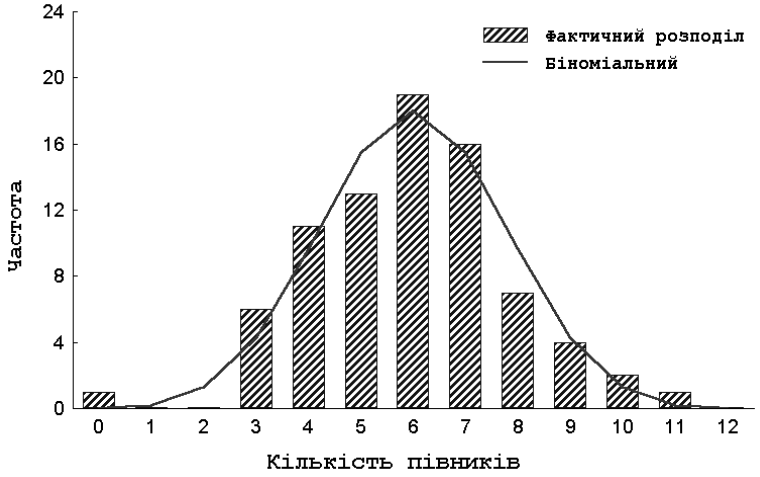

Рисунок 2.1 – Емпіричний розподіл виводків з різною кількістю півників та теоретична біноміальна крива

Біноміальний розподіл і розподіл Пуассона характеризуються двома параметрами: *максимальною імовірністю* (чи найімовірнішою кількістю

прояву очікуваного результату – μ) і *варіансою* (σ^2) частоти очікуваної події A в n незалежних випробуваннях.

Для біноміального розподілу ці параметри розраховуються за формулами:

$$\left.\begin{aligned} \mu &= np; \\ \sigma^2 &= pq. \end{aligned}\right\} \qquad (2.5)$$

Приклад. Із 550 корів симентальської породи 178 мали вим'я ванноподібної форми. Визначите дисперсію частоти особин з даною формою вим'я.

Частота ванноподібного вим'я у тварин даної групи складає:

$$p = \frac{178}{550} = 0{,}324.$$

Отже, дисперсія цієї частоти тоді становить:

$$\sigma^2 = 0{,}324 \times 0{,}676 = 120{,}46 \ .$$

Побудова довірчого інтервалу параметра біноміального розподілу (частоти ознаки) може бути проведена з визначеним рівнем точності на підставі нормальної апроксимації за формулою:

$$p_{1,2} = p \pm 1{,}96\sqrt{\frac{p(1-p)}{n}} \ . \qquad (2.6)$$

Однак у тих випадках, коли значення частоти нижче, ніж 0,2 чи перевищує 0,8, а також при малих обсягах вибірки (менше 100 особин), нормальна апроксимація може дати абсурдні значення границь довірчого інтервалу (нижче 0 чи вище 1). У цих випадках необхідно використовувати більш точний метод розрахунку границь довірчого інтервалу – вони знаходяться як корені квадратного рівняння:

$$(m - pn)^2 = 3{,}84np(1-p) \ ,$$

де m – кількість особин, що мають дану ознаку у вибірці тварин обсягом n.

У цьому випадку оцінки нижньої і верхньої границь 95% довірчого інтервалу можуть бути розраховані за формулою:

$$p_{1,2} = \frac{\left(p + \frac{3{,}84}{2n}\right) \pm 1{,}96\sqrt{\left(\frac{p(1-p)}{n}\right) + \left(\frac{3{,}84}{4n^2}\right)}}{\left(1 + \frac{3{,}84}{n}\right)}. \qquad (2.7)$$

Приклад. У вибірці із 50 корів 16 особин мали добовий надій від 17,5 до 20,5 кг. Яка частка таких тварин? Який буде її довірчий інтервал?

Частка особин з даним рівнем продуктивності становить:

$p = \frac{16}{50} = 0{,}32$, а її статистична помилка: $SEp = \sqrt{\frac{0{,}32 \times 0{,}68}{50}} = 0{,}07$.

Тоді довірчий інтервал даної оцінки буде складати:

$$0{,}32 - 1{,}96 \cdot 0{,}07 \le p \le 0{,}32 + 1{,}96 \cdot 0{,}07 \ ,$$

тобто

$$0{,}18 \le p \le 0{,}46 \ .$$

Розрахунок довірчого інтервалу на підставі формули 2.7 дасть більш точні оцінки, з урахуванням малого обсягу вибірки:

$$0{,}19 \le p \le 0{,}44 .$$

У тих випадках, коли є необхідність зробити порівняння параметрів двох біноміальних розподілів, можна використовувати наближений метод, заснований на нормальній апроксимації. Для цього використовується:

$$u = \frac{\left| (p_1 - p_2) - \dfrac{1}{2(n_1 + n_2)} \right|}{\sqrt{\dfrac{m_1 + m_2}{n_1 + n_2}\left(1 - \dfrac{m_1 + m_2}{n_1 + n_2}\right)\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}, \qquad (2.8)$$

де m_1 і m_2 – кількість особин з даною ознакою в двох вибірках обсягом n_1 і n_2;

p_1 і p_2 – їхні відповідні частоти.

Розходження визнаються достовірними, якщо розраховане значення *u-критерію* перевищує 1,96.

Розподіл Пуассона. У тих випадках, коли ймовірність появи успішного результату дуже мала навіть при великій кількості реалізацій експерименту (тобто p близько до 0 при $N \to \infty$), біноміальний розподіл ймовірностей може бути апроксимований розподілом Пуассона:

$$p_n(m) = e^{-\lambda} \frac{\lambda^m}{m!} , \qquad (2.9)$$

де $\lambda = kp$ – найімовірніша частота рідкісної події;

k – обсяг одиничної вибірки.

Приклад. Уявимо, що деяке рідкісне захворювання зустрічається в 0,1% особин даної популяції великої рогатої худоби. Якщо з цієї популяції випадково вибирають k = 5000 голів і перевіряють на захворювання, то яка імовірність, що рівно чотири з них будуть хворі.

Найбільш ймовірна кількість хворих тварин у вибірці з 5000 відібраних є λ = 5000 × 0,001 = 5 голів. Тоді, ймовірність знайти рівно чотири хворих тварини буде дорівнювати:

$$p_{5000}(4) = e^{-5} \frac{5^4}{4!} = 0{,}1755.$$

Графік розподілу ймовірностей для цього прикладу наведено на рисунку 2.2.

Як бачимо, він дуже асиметричний – імовірність знайти серед 5000 тварин більше, ніж 15 хворих особин практично дорівнює нулю.

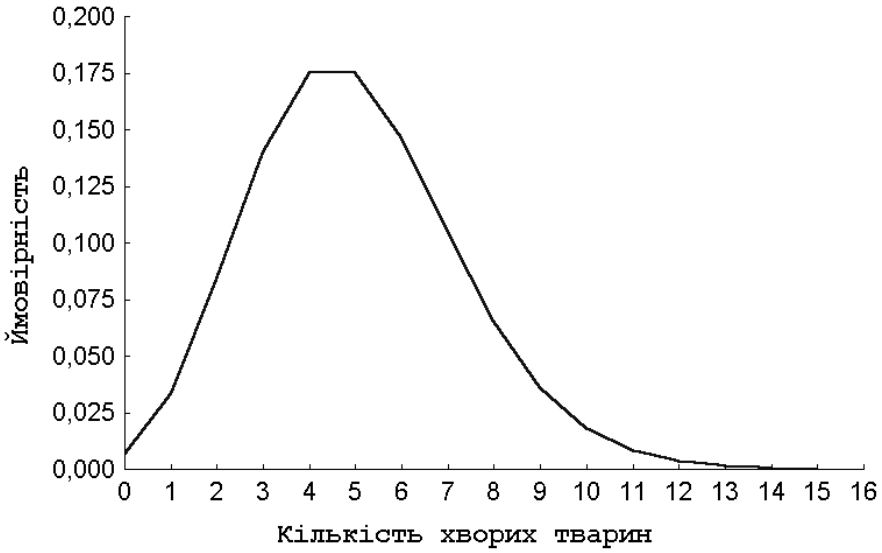

Рисунок 2.2 – Емпіричний розподіл імовірності знайти у вибірці з 5000 тварин різну кількість хворих (розподіл Пуассона)

Особливістю розподілу Пуассона є те, що його параметри виявляються рівними між собою, тобто:

$$\mu = \sigma^2 = \lambda. \qquad (2.10)$$

Ця особливість часто використовується при ідентифікації розподілу. Розподіл Пуассона характерний для тих випадків, коли наявність чи відсутність певної ознаки в організмі залежить від великої кількості випадкових факторів, що діють сумісно. Для сільськогосподарських тварин це кількість двоєн у корів, наявність глистяних інвазій, прояв альбінізму і т. ін.

Приклад. У групі із 100 корів кожна мала по чотири отелення. Серед них двійні були народжені лише 10 разів (табл. 2.3).

Таблиця 2.3 – Вихідні дані для розрахунку типу розподілу

Показник	Кількість отелень із двійнями				
	0	1	2	3	4
Фактичні частоти	90	6	3	1	0
Теоретичні частоти	86,1	12,9	0,95	0,05	0,00

Який тип розподілу має дана ознака? Розрахуйте параметри їхнього розподілу.

Спочатку розрахуємо частоту появи двійні:

$$p = \frac{0 \times 90 + 1 \times 6 + 2 \times 3 + 3 \times 1 + 4 \times 0}{100 \times 4} = 0{,}00375.$$

Оскільки це значення дуже мале, можна припустити, що ми маємо справу із розподілом Пуассона.

Оскільки $k = 4$, то найімовірніша кількість двоєн на одну тварину складає $\lambda = 4 \times 0{,}0375 = 0{,}15$, а її варіанса – $\sigma^2 = 0{,}25$.

Теоретична частота особин, що не мають жодної двійні з чотирьох отелень, тоді складе:

$$n_0 = np_4(0) = 100 \cdot e^{-0,15} \cdot \frac{0,15^0}{0!} = 86,1,$$

що мають одну двійню:

$$n_1 = np_4(1) = 100 \cdot e^{-0,15} \cdot \frac{0,15^1}{1!} = 12,9,$$

і т. д. Ці значення наведено у останньому рядку таблиці.

Оцінка вірогідності розходжень показників двох розподілів Пуассона λ_1 і λ_2 проводиться, використовуючи *u*-критерій:

$$u = \frac{\lambda_1 - \lambda_2}{\sqrt{\frac{\lambda_1}{n_1} + \frac{\lambda_2}{n_2}}}, \qquad (2.11)$$

де n_1 і n_2 – обсяги порівнюваних вибірок.

Розходження вважаються достовірними, якщо розраховане значення *u* перевищує 1,96.

Приклад. В одному стаді корів симентальської породи на 187 отелень зареєстровано 5 двоєн, а в іншому – з 560 отелень було 9 двоєн. Чи розрізняються тварини в цих стадах за частотою появи двоєн?

Найбільш імовірна кількість двоєн на одну тварину в першому стаді складає: $\lambda_1 = 5 : 187 = 0,027$, а в другому – $\lambda_2 = 9 : 560 = 0,016$. Тоді оцінка *u*-критерію буде становити:

$$u = \frac{0,027 - 0,016}{\sqrt{\frac{0,027}{187} + \frac{0,016}{560}}} = \frac{0,011}{0,013} = 0,85,$$

що набагато нижче критичного. Таким чином, тварини обох груп вірогідно не відрізняються за частотою народження двоєн.

Оскільки параметр розподілу Пуассона найчастіше дуже близький до нуля, його довірчий інтервал має дуже асиметричну форму і не може бути побудований по стандартній процедурі. Розрахунок його досить складний, тому для зручності в додатку В наведено нижню і верхню довірчі межі цього показника.

Наприклад, побудуємо 95% довірчий інтервал для частоти народження двоєн серед корів першої групи з прикладу. Для кількості двійневих отелень, що дорівнює 5, находимо із додатку В, що нижня межа становить 1,62, а верхня – 11,67. Таким чином, нижня межа 95% довірчого інтервалу частоти появи двоєн складатиме: $\lambda_H = 1,62 : 187 = 0,009$, а верхня – $\lambda_B = 11,67 : 187 = 0,062$.

Контрольні питання:

1. Біноміальний розподіл.
2. Розподіл Пуассона.

§ 3. Особливості генетичної структури панміктичних популяцій

Аналіз структури популяції за якісними ознаками містить у собі рішення ряду генетичних питань і обчислення основних генетико-статистичних параметрів. До основних елементів аналізу популяцій відноситься:

- оцінка структури популяції за частотами генотипів і фенотипів;
- визначення стану генотипової рівноваги популяції;
- оцінка частот алелів у популяції при різних типах успадкування ознак і залежно від впливу різних факторів (відбір, мутації, дрейф генів, міграції, тощо);
- визначення ступеня гетерозиготності популяції за окремими локусами та в цілому;
- визначення ступеня генетичної подібності за локусами і алелями поліморфних систем між тваринами різних груп у межах однієї популяції та між різними популяціями (генетична диференціація).

При генетичному аналізі популяцій використовується поняття **панміктична популяція**, маючи при цьому на увазі *досить велику групу особин певного виду, які можуть вільно схрещуватися одна з одною і на якій не відчувається тиску відбору, міграцій і мутацій.*

Панміктична популяція певною мірою є поняттям теоретичним, абстрактно-модельним. Така популяція не існує в конкретних умовах середовища. Але розкриті для неї закономірності у визначених межах можуть бути перенесені і на конкретні популяції, в яких порушені умови панміксії.

Спосіб визначення частот алелів чи генотипів змінюється залежно від характеру успадкування (кодомінантний чи домінантний), складності структури локусу (двохалельний, трьохалельний, поліалельний), а також від того, зчеплено чи роздільно успадковуються локуси.

3.1 Визначення частот алелів при кодомінантному типі успадкування і двоалельній системі локусу

У більшості поліморфних систем домінування одного алеля над іншим немає, а спостерігається **кодомінування**, при якому в фенотипі виявляються обидва алеля, при цьому фенотипово легко можна виділити як гомозиготні, так і гетерозиготні генотипи. Так, наприклад, у локусі гемоглобіну відомі два алельні типи, що зумовлюють синтез різних гемоглобінів трьох генотипів (Hb^A/Hb^A; Hb^A/Hb^B; Hb^B/Hb^B).

Позначимо кількість особин, що несуть домінантний гомозиготний генотип – D, гетерозиготний – H, а рецесивний гомозиготний – R. У цьому випадку частота домінантного гомозиготного генотипу буде: $d = \frac{D}{N}$, гетерозиготного: $h = \frac{H}{N}$, а рецесивного гомозиготного: $r = \frac{R}{N}$, де N – обсяг вибірки.

Для обчислення частот алелів найчастіше використовується метод максимальної правдоподібності, розроблений Р. Фішером. Відповідно до цього методу, частоту домінантного алеля можна визначити за формулою:

$$p = \frac{2D + H}{2N}, \tag{3.1}$$

а рецесивного:

$$q = \frac{2R + H}{2N}, \tag{3.2}$$

де $2N$ – загальна кількість алелів даного діалельного локусу у вибірці.

Якщо використовуються частоти генотипів, то оцінки частот алелів можна одержати за наступними формулами:

$$\begin{aligned} p &= d + \frac{h}{2}; \\ q &= r + \frac{h}{2}. \end{aligned} \tag{3.3}$$

Оскільки оцінка частот алелів проводиться на підставі вибірки, тому на неї впливають випадкові фактори. Вірогідність отриманих результатів можна визначити на підставі розрахунку помилок вибіркових частот алелів:

$$SEp = SEq = \sqrt{\frac{pq}{2N}}. \tag{3.4}$$

Отримані значення частот алелів вважаються достовірними (тобто, відрізняються від нуля) у тому випадку, якщо вони як мінімум втричі перевищують величини своїх помилок, тобто:

$$\frac{p}{SEp} \geq 3; \quad \frac{q}{SEq} \geq 3.$$

Приклад. У стаді великої рогатої худоби, кількістю 1000 тварин, було 300 особин генотипу Hb^A/Hb^A, 50 – генотипу Hb^A/Hb^B, і 650 – генотипу Hb^B/Hb^B. Які частоти алелів Hb^A і Hb^B?

Спочатку за формулами максимальної правдоподібності визначимо частоти алелів Hb^A і Hb^B.

$$p_{Hb^A} = \frac{2 \cdot 300 + 50}{2 \cdot 1000} = 0{,}325,$$

$$p_{Hb^B} = \frac{2 \cdot 650 + 50}{2 \cdot 1000} = 0{,}675.$$

Потім визначимо статистичну помилку для обох частот:

$$SEp_{Hb^A} = SEp_{Hb^B} = \sqrt{\frac{p_{Hb^A} p_{Hb^B}}{2N}} = \sqrt{\frac{0{,}325 \cdot 0{,}675}{2 \cdot 1000}} = 0{,}014.$$

Оскільки в обох випадках оцінки частот виявляються більше, ніж утричі більшими за свої помилки, їх можна розглядати як достовірні результати.

3.2 Визначення частот алелів і генотипів при двоалельній системі і домінуванні одного з алелів локусу

При домінуванні одного алеля локусу над іншим (*A* > *a*) формується два фенотипи: фенотип, що містить домінантний алель (*AA* і *Aa*) і фенотип, що містить обидва рецесивні алеля (*aa*). Рецесивний гомозиготний фенотип легко виділити серед інших домінантних фенотипів, тому його частоту використовують як основне джерело інформації про структуру популяції.

Для визначення частоти рецесивного алеля використовується формула:

$$q_a = \sqrt{q_{aa}^2} = \sqrt{\frac{n_{aa}}{N}}, \tag{3.5}$$

де n_{aa} – чисельність рецесивних гомозигот у вибірці обсягом N.

Потім за формулою: $p = 1 - q_a$ обчислюють частоту домінантного алеля.

Помилка вибіркових оцінок частот рецесивного і домінантного алелів визначається за формулами:

$$SEp_A = SEq_a = \sqrt{\frac{1-q_a^2}{4N}}. \tag{3.6}$$

Приклад. У стаді великої рогатої худоби, кількістю 200 голів виявлено двох сліпих телят. Цей дефект має рецесивну природу. Визначите частоту алелів *p* і *q*, а також очікувану кількість гетерозиготних особин.

За умовою прикладу, $n_{aa} = 2$ і $N = 200$, отже,

$$q = \sqrt{\frac{2}{200}} = \sqrt{0{,}01} = 0{,}1;$$

$$p = 1 - q = 0{,}9.$$

Помилки оцінок частот алелів складають:

$$SEp = SEq = \sqrt{\frac{1-0{,}1^2}{4 \cdot 200}} = 0{,}035.$$

Відношення оцінки частоти алеля *q* до його помилки виявляється менше трьох. Тому в таких випадках (коли обсяг вибірки невеликий чи зустрічаються поодинокі особини із рецесивним гомозиготним генотипом), необхідно використовувати модифіковані формули:

$$q_a = \sqrt{\frac{4n_{aa}+1}{4N+1}}, \tag{3.7}$$

$$SEq = \sqrt{\frac{1-q_a^2}{4N+1}}. \tag{3.8}$$

Для даного прикладу, розрахунок оцінки частоти рецесивного алеля по модифікованій формулі дає значення 0,106, а її помилки – 0,035.

Очікувана кількість тварин, що несуть рецесивний алель у гетерозиготному стані складає:

$$n_{Aa} = 2pqN = 2 \times 0{,}894 \times 0{,}106 \times 200 \approx 36 \text{ голів}.$$

3.3 Визначення частот генотипів і алелів при трьохалельній системі локусу і кодомінантному типі успадкування

Якщо локус, що детермінує будь-яку поліморфну ознаку складається з декількох алелів з кодомінантним типом успадкування, то для визначення частот генотипів і алелів використовується формула Бернштейна.

Якщо, наприклад, локус складається з трьох алелів, то формула структури популяції за частотами генотипів буде наступною:

$$p^2 + q^2 + z^2 + 2pq + 2pz + 2qz = 1,$$

де p, q і z – частоти алелів даного локусу.

Частоту кожного генотипу обчислюють за формулами:

$$p_{AA} = \frac{n_{AA}}{N};\ q_{BB} = \frac{n_{BB}}{N};\ z_{CC} = \frac{n_{CC}}{N};$$
$$x_{AC} = \frac{n_{AC}}{N};\ y_{AB} = \frac{n_{AB}}{N};\ r_{BC} = \frac{n_{BC}}{N}, \qquad (3.9)$$

а частоти алелів:

$$p_A = \frac{2n_{AA} + n_{AB} + n_{AC}}{2N};$$
$$q_B = \frac{2n_{BB} + n_{AB} + n_{BC}}{2N}; \qquad (3.10)$$
$$z_C = \frac{2n_{CC} + n_{AC} + n_{BC}}{2N},$$

де n_{AA}, n_{BB}, n_{CC} – кількість особин гомозиготного генотипу;

n_{AB}, n_{AC}, n_{BC} – кількість гетерозиготних особин;

$2N$ – загальна кількість алелів у вибірці.

При трьохалельній кодомінантній системі помилки частот різних алелів вже не рівні між собою й обчислюються за формулами:

$$SEp = \sqrt{\frac{p(1-p)}{2N}};$$
$$SEq = \sqrt{\frac{q(1-q)}{2N}}; \qquad (3.11)$$
$$SEz = \sqrt{\frac{z(1-z)}{2N}}.$$

Приклад. При дослідженні типів трансферину в стаді великої худоби бестужевської породи виявлено наступне співвідношення генотипів тварин:

Tf^ATf^A – 141, Tf^DTf^D – 183, Tf^ETf^E – 18, Tf^ATf^D – 117, Tf^ATf^E – 30, Tf^DTf^E – 6.

Необхідно визначити частоти алелів Tf^A, Tf^D та Tf^E .

Використовуючи приведені вище формули, розраховуємо:

$$p_A = \frac{2 \cdot 141 + 117 + 30}{2 \cdot 495} = 0{,}433;$$

$$q_D = \frac{2 \cdot 183 + 117 + 6}{2 \cdot 495} = 0{,}493;$$

$$z_E = \frac{2 \cdot 18 + 30 + 6}{2 \cdot 495} = 0{,}073.$$

Статистичні помилки оцінок частот алелів будуть наступними:

$$SEp_A = \sqrt{\frac{0{,}433 \cdot (1 - 0{,}433)}{2 \cdot 495}} = 0{,}016;$$

$$SEq_D = \sqrt{\frac{0{,}493 \cdot (1 - 0{,}493)}{2 \cdot 495}} = 0{,}016;$$

$$SEz_E = \sqrt{\frac{0{,}073 \cdot (1 - 0{,}073)}{2 \cdot 495}} = 0{,}008.$$

В усіх трьох випадках отримано оцінки частот алелів, що вірогідно відрізняються від 0, оскільки відношення оцінки частоти до її помилки перевищує 3.

При розрахунку частоти будь-якого алеля у випадку поліалельної кодомінантної системи успадкування, використовується наступний принцип: частота алеля дорівнюватиме відношенню суми подвійної кількості гомозиготного генотипу та частот гетерозиготних генотипів, що несуть даний алель, до $2N$ – загальної кількості алелів у вибірці.

Контрольні питання:

1. Визначення частот алелів при кодомінантному типі успадкування.
2. Визначення частот алелів і генотипів при двохалельній системі і домінуванні одного з алелів локусу.
3. Визначення частот генотипів і алелів при трьохалельній системі локусу і кодомінантному типі успадкування.

§ 4. Закон Гарді-Вайнберга

Для популяції, у якій алель *A* зустрічається з частотою *p* і алель *a* з частотою *q*, при випадковому схрещуванні ймовірність появи серед нащадків генотипів *AA*, *Aa* й *aa* можна розрахувати, використовуючи закони множення і додавання ймовірностей (див. § 1).

Наприклад, імовірність того, що особина одержить алель *A* від батька і від матері одночасно (а це події взаємно незалежні) являє собою добуток імовірностей одержання алеля *A*. Це, у свою чергу, дорівнює частоті алеля в генному пулі усієї популяції. Тоді, ймовірність того, що при формуванні зиготи зустрінуться два алеля *A* становитиме:

$$\mathrm{P}(AA) = \mathrm{P}(A) \times \mathrm{P}(A) = p \times p = p^2.$$

У свою чергу, ймовірність того, що нащадок одержить алель *a* від батька і від матері одночасно складає:

$$\mathrm{P}(aa) = \mathrm{P}(a) \times \mathrm{P}(a) = q \times q = q^2.$$

Генотип *Aa* формується, якщо батьківська гамета *A* зустрінеться з материнською гаметою *a*, чи батьківська гамета *a* зустрінеться з материнською *A*. Отже, ймовірністю появи генотипу *Aa* є сума наступних ймовірностей:

$$\mathrm{P}(Aa) = \mathrm{P}(A) \times \mathrm{P}(a) + \mathrm{P}(a) \times \mathrm{P}(A) = p \times q + q \times p = 2pq.$$

Таким чином, розподіл частот генотипів буде складати:

$$AA : Aa : aa = p^2 : 2pq : q^2.$$

Таке співвідношення генотипів при двохалельній системі локусу характерно для панміктичної популяції, тобто, без тиску на неї відбору, мутаційного процесу, міграцій, дрейфу генів і при вільному випадковому сполученні гамет при заплідненні (*закон Гарді-Вайнберга*).

Співвідношення частот генотипів $p^2 : 2pq : q^2$ буде зберігатися з покоління в покоління, тобто протягом поколінь буде спостерігатися **рівноважний стан** структури популяції.

Проте, під впливом відбору, мутацій, міграції структура популяції за частотою алелів і генотипів змінюється і популяція виходить зі стану генної рівноваги (тобто, змінюються величини частот кожного алеля і генотипів). При виникненні в популяції, що знаходиться в нерівноважному стані, умов панміксії, у першому ж поколінні нащадків настає зрівноважений стан і популяція знову набуває властивості панміксності (*закон К. Пірсона*).

Приклад. У певній популяції тварин частоти генотипів *AA*, *Aa* й *aa* становлять, відповідно, 256, 458 і 59 особин. Тоді, відповідно до методу максимальної правдоподібності (формули 3.1 та 3.2), частота цих алелів буде складати:

$$p_A = \frac{2 \cdot 256 + 458}{2 \cdot 773} = 0{,}627,$$

$$q_a = \frac{2 \cdot 59 + 458}{2 \cdot 773} = 0{,}373.$$

У наступному поколінні за умови повної панміксії співвідношення генотипів буде становити 0,393 : 0,468 : 0,139. І якщо чисельність популяції залишиться такою ж, то співвідношення генотипів в ній буде 304 : 362 : 107.

У другому поколінні (знову ж, за умови повної панміксії) частоти алелів також не зміняться:

$$p_A = \frac{2 \cdot 304 + 362}{2 \cdot 773} = 0{,}627,$$

$$q_a = \frac{2 \cdot 107 + 362}{2 \cdot 773} = 0{,}373.$$

Таким чином, як ми бачимо, **в умовах повної панміксії популяція в першому ж поколінні приходить у рівноважний стан і частоти алелів не змінюються з покоління в покоління.**

Закон Гарді-Вайнберга можна проілюструвати графічно (рис. 4.1).

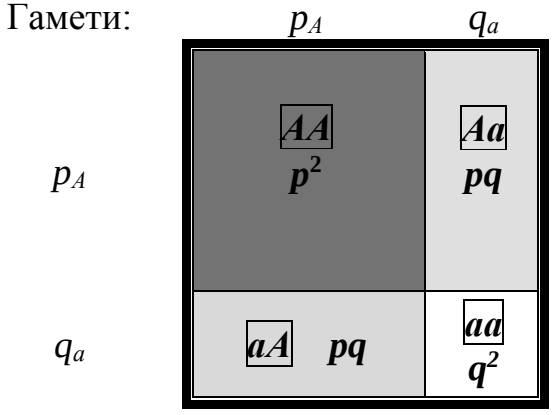

Рисунок 4.1 – Графічне представлення закону Гарді-Вайнберга

Частота алелів p і q пропорційна довжині грані одиничного квадрата (тобто, $p + q = 1$). Тоді, ймовірність появи генотипів пропорційна відповідним площам, при цьому площа квадрата залишається одиничною, тобто, $p^2 + 2pq + q^2 = 1$.

Як видно на рисунку 4.2, частка гетерозигот у популяції ніколи не може перевищити 0,5. З іншого боку, якщо відома ця частка (h), частоти алелів можна розрахувати на підставі формули:

$$p, q = \frac{1 \pm \sqrt{1 - 2h}}{2}. \qquad (4.1)$$

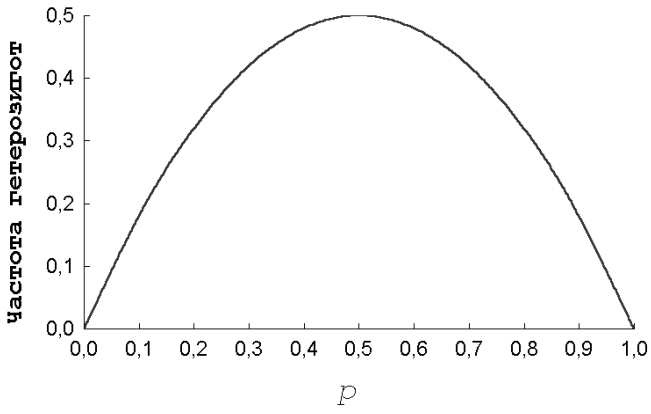

Рисунок 4.2 – Залежність між частотою гетерозигот і частотою алелів

Наслідки із закону Гарді-Вайнберга

Наслідок 1. У рівноважній популяції завжди будуть справедливими рівняння:

$$4DR = H^2,$$

чи

$$\frac{H}{\sqrt{DR}} = 2. \tag{4.2}$$

Наслідок 2. Для рівноважної популяції характерна наступна закономірність:

$$\sqrt{\frac{D}{N}} + \sqrt{\frac{R}{N}} = 1. \tag{4.3}$$

Ці наслідки дозволяють одержати тест на зрівноваженість генетичної структури популяції, тобто, оцінити чи не піддана популяція впливам відбору, мутацій, міграцій чи дрейфу генів. Крім того, другий наслідок із закону Гарді-Вайнберга дозволяє отримати оцінку частот алелів при домінуванні одного з них.

Оскільки оцінки частот алелів проводяться на підставі вибіркової сукупності, ці оцінки залежать від ряду випадкових факторів. Тому для одержання коректного висновку про зрівноваженість генетичної структури популяції необхідно використовувати статистичні критерії, що враховують особливості вибіркових сукупностей.

Перевірка відповідності розподілу частот генотипів у популяції проводиться з використанням критерію Хі-квадрат К. Пірсона:

$$\chi^2 = \frac{N(4DR - H^2)^2}{(2D + H)^2(2R + H)^2}, \tag{4.4}$$

де D – кількість особин, що мають домінантний гомозиготний генотип;

H – кількість гетерозигот у вибірці;

R – кількість рецесивних гомозигот;

N – обсяг вибірки.

Приклад. В стаді великої рогатої худоби бестужевської породи при дослідженні сироватки молока виявлено наступні частоти генотипів: Ig^AIg^A – 26; Ig^BIg^B – 271; Ig^AIg^B – 202 голів.

Необхідно визначити, чи знаходиться ця популяція в стані генетичної зрівноваженості.

Спочатку необхідно розрахувати частоти алелів (припускаючи, що популяція знаходиться в рівноважному стані) і їхні помилки:

$$p_{Ig^A} = \frac{2 \cdot 26 + 202}{2 \cdot 499} = 0{,}255,$$

$$q_{Ig^B} = \frac{2 \cdot 271 + 202}{2 \cdot 499} = 0{,}745.$$

$$SEp_{Ig^A} = SEq_{Ig^B} = \sqrt{\frac{0{,}255 \cdot 0{,}745}{2 \cdot 499}} = 0{,}014.$$

Потім, використовуючи закон Гарді-Вайнберга, розрахуємо теоретичні значення абсолютних частот генотипів:

$$\left(Ig^A Ig^A\right)^{teo} = Np^2 = 499 \cdot 0{,}255^2 = 32{,}4;$$

$$\left(Ig^A Ig^B\right)^{teo} = N2pq = 499 \cdot 2 \cdot 0{,}255 \cdot 0{,}745 = 189{,}6;$$

$$\left(Ig^B Ig^B\right)^{teo} = Nq^2 = 499 \cdot 0{,}745^2 = 277{,}0.$$

Як бачимо, є деяка розбіжність між фактичними і теоретичними частотами (розрахованими в припущенні того, що популяція знаходиться в рівноважному стані). Наскільки ці відхилення є випадковими, оцінимо, використовуючи критерій Хі-квадрат:

$$\chi^2 = \frac{499 \cdot (4 \cdot 26 \cdot 271 - 202^2)^2}{(2 \cdot 26 + 202)^2 \cdot (2 \cdot 271 + 202)^2} = 2{,}23.$$

Розраховане значення критерію тепер необхідно порівняти з табличним, що дорівнює 3,84. Якщо розраховане значення критерію буде менше, ніж 3,84 – вважається доведеним, що популяція знаходиться в рівноважному стані, а відхилення між фактичними і теоретичними частотами мають випадковий характер.

Якщо ж розраховане значення критерію Хі-квадрат дорівнює або більше, ніж 3,84, то робиться висновок, що дана популяція не знаходиться у зрівноваженому стані.

Оскільки в даному прикладі 2,23 < 3,84, то можна зробити висновок про те, що відхилення між фактичними і теоретичними значеннями частот генотипів обумовлені випадковими причинами й аналізована популяція знаходиться у рівноважному стані.

У випадку множинних алелів, оцінка відповідності фактичного розподілу генотипів зрівноваженому робиться на підставі наступної формули критерію Хі-квадрат К. Пірсона:

$$\chi^2 = \sum \frac{(Ф - T)^2}{T}, \qquad (4.5)$$

де $Ф$ – частоти генотипів, що спостерігаються фактично;

T – теоретичні частоти за умови дотримання закону Гарді-Вайнберга.

Розраховане значення критерію Хі-квадрат порівнюється з табличним для числа ступенів свободи: $df = \frac{m(m-1)}{2}$, де m – число алелів.

Приклад. При дослідженні типів трансферину в стаді великої худоби бестужевської породи виявлено наступне співвідношення генотипів тварин за локусом трансферину:

Tf^ATf^A – 141, Tf^DTf^D – 183, Tf^ETf^E – 18, Tf^ATf^D – 117, Tf^ATf^E – 30, Tf^DTf^E – 6.

Необхідно оцінити, чи знаходиться генетична структура цього стада у рівноважному стані.

Частоти алелів трансферину будуть дорівнювати:

$$p_A = \frac{2 \cdot 141 + 117 + 30}{2 \cdot 495} = 0{,}433;$$

$$q_D = \frac{2 \cdot 183 + 117 + 6}{2 \cdot 495} = 0{,}493;$$

$$z_E = \frac{2 \cdot 18 + 30 + 6}{2 \cdot 495} = 0{,}074.$$

Тоді, за умови, що популяція знаходиться у зрівноваженому стані, теоретична кількість особин, що мають генотип Tf^ATf^A повинна була б становити:

$$n_{AA}^T = Np_A^2 = 495 \cdot 0{,}433^2 = 92{,}8,$$

а генотип Tf^ATf^D:

$$n_{AD}^T = N2p_Aq_D = 495 \cdot 2 \cdot 0{,}433 \cdot 0{,}493 = 211{,}3,$$

і т. д.

Дані щодо фактичних і теоретичних частот генотипів тварин за локусом трансферину приведені в таблиці 4.1.

Таблиця 4.1 – Фактичні і теоретичні частоти генотипів тварин за локусом трансферину

Генотип	Фактична частота ($Ф$)	Теоретична частота (T)	$\frac{(Ф-T)^2}{T}$
Tf^ATf^A	141	92,8	25,03
Tf^DTf^D	183	120,3	32,68
Tf^ETf^E	18	2,7	86,70
Tf^ATf^D	117	211,3	42,08
Tf^ATf^E	30	31,7	0,09
Tf^DTf^E	6	36,1	25,10
Сума	495	494,9	211,68

Табличні значення критерію Хі-квадрат наведено в додатку Д для відповідного числа ступенів свободи.

У нашому випадку, табличне значення критерію при $df = 3$ становить 7,82. Оскільки розраховане значення критерію набагато перевищує табличне, можна зробити висновок, що дана популяція не знаходиться в стані генетичної зрівноваженості. Це добре видно і при порівнянні фактичних і теоретичних частот. Проглядається явний надлишок гомозиготних генотипів і нестача гетерозиготних.

Стандартний критерій Хі-квадрат може бути використаний лише в тих випадках, коли очікувані чисельності генотипів не менше, ніж 1-2. У протилежному випадку варто застосовувати його модифікацію – *G*-критерій.

При використанні даного критерію необхідно виконати наступні кроки. Спочатку проводиться розрахунок частот алелів, очікуваних чисельностей генотипів і використовується стандартний критерій Хі-квадрат (навіть якщо очікувані чисельності деяких генотипів невеликі), як це було показано в прикладі вище.

Потім обчислюється допоміжна величина *d*:

$$d = 1 + \frac{\sum_{i=1}^{m} \frac{1}{E_i}}{2(n-1)} - \frac{n^2 + 2n - 2}{2(n-1)N}, \tag{4.6}$$

де n – число генотипів;

N – сумарний обсяг вибірки;

T_i – теоретичні частоти генотипів.

На основі цієї величини розраховуються значення критерію і число його ступенів свободи:

$$\begin{aligned} G &= \frac{\chi^2}{d}; \\ df &= \frac{m(m-1)}{2d}, \end{aligned} \tag{4.7}$$

де χ^2 – значення стандартного критерію Хі-квадрат;

m – число алелів.

Оцінка рівня вірогідності критерію проводиться по таблиці стандартного критерію Хі-квадрат. Однак, оскільки розраховане число ступенів свободи найчастіше буває числом дробовим, а в таблиці наведено критичні значення тільки для цілих чисел, можна скористатися наступною апроксимацією:

$$\chi^2_{(df)} = 2{,}518 + 1{,}282 \cdot df - 0{,}0021 \cdot (df)^2 + 1{,}371 \cdot \ln df\ . \tag{4.8}$$

Приклад. У вибірці, що містить 100 тварин, чисельності генотипів трьохалельної системи трансферину були наступні:

AA = 81 голова, *DD* = 2, *EE* = 1, *AD* = 12, *AE* = 4 і *DE* = 0.

Необхідно оцінити, чи знаходиться ця популяція в стані генетичної зрівноваженості.

Спочатку розрахуємо частоти алелів на підставі формул Бернштейна (формула 3.10):

$$p_A = \frac{2 \cdot 81 + 12 + 4}{2 \cdot 100} = 0{,}89;$$

$$q_D = \frac{2 \cdot 2 + 12 + 0}{2 \cdot 100} = 0{,}08;$$

$$z_E = \frac{2 \cdot 1 + 4 + 0}{2 \cdot 100} = 0{,}03.$$

Потім розраховуємо очікувані чисельності генотипів, значення стандартного критерію Хі-квадрат і величини, обернені очікуваним частотам (табл. 4.2).

Таблиця 4.2 – Результати розрахунків очікуваної частоти генотипів, значення стандартного критерію Хі-квадрат і величин, обернених очікуваним частотам

Генотип	Фактична частота ($Ф$)	Теоретична частота (T)	$\frac{(Ф-T)^2}{T}$	$\frac{1}{T}$
AA	81	79,21	0,04	0,01
DD	2	0,64	2,89	1,56
EE	1	0,09	9,20	11,10
AD	12	14,24	0,35	0,07
AE	4	4,34	0,34	0,19
DE	0	0,48	0,56	2,08
Суми	100	100,00	χ^2 =13,40	15,00

Далі необхідно обчислити значення *G*-критерію і число його ступенів свободи, але спочатку розрахуємо допоміжну величину *d*:

$$d = 1 + \frac{15}{2 \cdot (6-1)} - \frac{6^2 + 2 \cdot 6 - 2}{2 \cdot (6-1) \cdot 100} = 2{,}45,$$

$$G = \frac{13{,}40}{2{,}45} = 5{,}47,$$

$$df = \frac{3 \cdot (3-1)}{2 \cdot 2{,}45} = 1{,}22.$$

Використовуючи апроксимуючий вираз (формула 4.8), розрахуємо критичне значення критерію для числа ступенів свободи, рівного 1,22:

$$\chi^2_{(df)} = 2{,}518 + 1{,}282 \cdot 1{,}22 - 0{,}0021 \cdot (1{,}22)^2 + 1{,}371 \cdot \ln 1{,}22 = 4{,}35.$$

Оскільки фактичне значення *G*-критерію перевищує табличне значення критерію Хі-квадрат (тобто, у нашому випадку 5,47 > 4,35), гіпотеза про те, що дана популяція знаходиться у стані генетичної зрівноваженості, повинна бути відхилена.

Контрольні питання:

1. Характеристика рівноважного стану структури популяції.
2. Особливості панміктичної популяції.
3. Як провести перевірку відповідності розподілу частот генотипів у популяції рівноважному стану?

§ 5. Фактори динаміки популяцій. Мутації і міграція

Будь-яка популяція може змінювати генетичне середовище під впливом зовнішніх і внутрішніх факторів і, отже, популяція має **генетичну пластичність**. Разом з тим, популяція здатна зберігати структуру в ряді поколінь чи протягом різних часових відрізків, що супроводжується формуванням її **генетичного гомеостазу**, тобто сталості.

Взаємодія цих двох властивостей популяції забезпечує її **генетичну динаміку**, на тлі якої формується пристосованість особин, що утворюють популяцію, до мінливих умов середовища і внутрішніх факторів.

Фактори, що здатні змінювати генетичну структуру популяції, різноманітні і кожен впливає на частоту алелів і генотипів. Основні з них:

- мутації;
- міграції;
- відбір;
- випадковий дрейф генів.

Перші три відносяться до систематичних, оскільки передбачуваним є не лише величина, але і напрямок змін алельних частот при їхньому впливі на генетичну структуру популяції. Водночас, для випадкового дрейфу генів можна оцінити тільки величину змін, але не їхній напрямок.

Мутації – *раптові і спонтанні зміни, які час від часу відбуваються із генами природнім шляхом*. Мутації є первинним джерелом нових алелів і, відповідно, генетичної мінливості особин та популяції в цілому.

Як показав М. П. Дубінін, процес виникнення мутацій і мутабільність організмів має адаптивне значення для популяції. Частка мутантного алеля в популяції залежить від його стану (домінантності чи рецесивності), від характеру його дії (летальна, напівлетальна, нейтральна), від характеру змін, викликаних мутантним алелем (морфологічних, біохімічних), від взаємодії з іншими алелями.

У великих популяціях мутантний рецесивний алель довше зберігається в гетерозиготному стані, а в нечисленних – він швидко переходить у гомозиготний стан і піддається впливу відбору, що або усуває гомозиготний рецесивний генотип, або сприяє його збереженню.

Припустимо, що існує два алеля одного локусу – A_1 і A_2, і що в результаті мутації A_1 перетворюється в A_2 з частотою u на одну гамету за одне покоління. Якщо в початковий момент часу частота алеля A_1 складає p_0, тоді в наступному поколінні його частота буде:

$$p_1 = p_0 - up_0 = p_0(1-u), \qquad (5.1)$$

де up_0 – частота алелів A_1, що за рахунок мутації перетворилися на алель A_2.

У наступному поколінні частота алеля A_1 складе:

$$p_2 = p_1 - up_1 = p_1(1-u) = p_0(1-u)(1-u) = p_0(1-u)^2,$$

з огляду на те, що $p_1 = p_0(1-u)$. Через t поколінь частота алеля A_1 буде дорівнювати:

$$p_t = p_0(1-u)^t. \tag{5.2}$$

Оскільки величина (1 – u) менше одиниці, зрозуміло, що p_t з часом зменшуватиметься. Якщо цей процес продовжується необмежено довго, частота алеля A_1 буде прямувати до нуля. При цьому, швидкість зміни частоти алеля дуже мала.

Наприклад, якщо $u = 10^{-5}$ на одну гамету за одне покоління, то для того, щоб змінити частоту алеля A_1 від 1 до 0,99, буде потрібно майже 1005 поколінь.

У загальному випадку, кількість поколінь, необхідна для зниження частоти алеля від p_0 до p_t, дорівнює:

$$t = \frac{\ln p_t}{\ln[p_0(1-u)]}. \tag{5.3}$$

Мутації генів часто бувають зворотними: алель A_1 перетворюється на A_2, а A_2 може, у свою чергу, перетворюватися на алель A_1.

Припустимо, що A_1 перетворюється на A_2 із частотою u, а зворотна мутація відбувається із частотою v. Якщо вихідні частоти алелів A_1 і A_2 дорівнюють p_0 і q_0, відповідно, то в наступному поколінні частота алеля A_1 буде складати:

$$p_1 = p_0 - up_0 + vq_0.$$

Якщо зміну частоти алеля за одне покоління позначити Δp, тобто, $\Delta p = p_1 - p_0$, то $\Delta p = (p_0 - up_0 + vq_0) - p_0 = up_0 + vq_0$. У тому випадку коли $\Delta p = 0$, настає рівновага між прямими і зворотними мутаціями.

Рівноважна частота алелів A_1 і A_2, тобто, коли кількість алелів A_1, що перетворюються за одне покоління на алелі A_2, дорівнює кількості алелів A_2, що перетворюються на алелі A_1, складає:

$$\begin{aligned} \hat{p}_{A1} &= \frac{v}{u+v}; \\ \hat{q}_{A2} &= \frac{u}{u+v}. \end{aligned} \tag{5.4}$$

Характерно, що ці частоти не залежать від вихідних частот алелів, а цілком визначаються тільки швидкістю прямих і зворотних мутацій. Наприклад, якщо припустити, що частота прямої і зворотної мутацій складають, відповідно, $u = 10^{-5}$ і $v = 10^{-6}$. Тоді рівноважні частоти будуть дорівнювати:

$$\hat{p} = \frac{10^{-6}}{10^{-5} + 10^{-6}} = 0{,}09;$$

$$\hat{q} = \frac{10^{-5}}{10^{-5} + 10^{-6}} = 0{,}91.$$

У тих випадках, коли виникає необхідність порівняти групи організмів за частотою виникнення мутацій, потрібно використовувати спеціальні критерії, що враховують, що частота виникнення мутацій – дуже рідкісна подія і

характер розподілу мутантних особин описується найчастіше розподілом Пуассона (див. § 2).

Для порівняння двох груп організмів (вибіркових сукупностей) у цьому випадку можна використовувати критерій, заснований на нормальній апроксимації розподілу Пуассона.

Якщо в першій вибірці обсягом N_1 виявлено m_1 мутацій, а в другій вибірці обсягом N_2 виявлено m_2 мутацій, то спочатку необхідно обчислити частоту мутацій у кожній вибірці за формулою:

$$p_i = \frac{m_i}{N_i}, \tag{5.5}$$

і середню частоту мутацій у всій сукупності організмів:

$$\overline{p} = \frac{m_1 + m_2}{N_1 + N_2}.$$

Потім необхідно розрахувати величину z_1 для вибірки, для якої $p_i < \overline{p}$:

$$z_1 = 2 \cdot \left(\sqrt{m_1 + 1} - \sqrt{N_1 \overline{p}} \right),$$

і величину z_2 для вибірки, для якої $p_i > \overline{p}$:

$$z_2 = 2 \cdot \left(\sqrt{m_2} - \sqrt{N_2 \overline{p}} \right).$$

Оцінка вірогідності розходжень за частотою виникнення мутацій на одну гамету у вибірках, що порівнюються, проводиться на підставі величини критерію:

$$Z = \sqrt{z_1^2 + z_2^2}\ . \tag{5.6}$$

Розходження визнаються достовірними (на рівні значущості α = 0,05), якщо значення критерію перевищує 1,96.

Приклад. У першому стаді, що налічує 34379 голів великої рогатої худоби, виявлено 39 тварин із трисомією XXX у статевій хромосомі, а в другому – 24 тварини із 20370 мали цю мутацію.

Необхідно обчислити, чи вірогідно розрізняються ці два стада за частотою виникнення даної геномної мутації.

Частота виникнення мутації в першому стаді становить:

$$p_1 = \frac{39}{2 \cdot 34379} = 5{,}67 \cdot 10^{-4}$$

на одну гамету, а в другому:

$$p_2 = \frac{24}{2 \cdot 20370} = 5{,}89 \cdot 10^{-4}$$

на одну гамету.

Середня частота виникнення мутації в сукупній групі становить:

$$\overline{p} = \frac{39 + 24}{68758 + 40740} = 5{,}75 \cdot 10^{-4}$$

на одну гамету.

Тоді для першої вибірки допоміжна величина становитиме:

$$z_1 = 2 \cdot \left(\sqrt{39+1} - \sqrt{68758 \cdot 5{,}75 \cdot 10^{-4}} \right) = 0{,}074,$$

а для другої:

$$z_2 = 2 \cdot \left(\sqrt{24} - \sqrt{40740 \cdot 5{,}75 \cdot 10^{-4}} \right) = 0{,}118 \ .$$

Остаточна величина критерію становитиме:

$$Z = \sqrt{0{,}074^2 + 0{,}118^2} = 0{,}019.$$

Оскільки розраховане значення критерію набагато менше 1,96, нульова гіпотеза про відсутність достовірних розходжень між двома порівнюваними групами стосовно частоти виникнення мутації даного типу визнається справедливою.

У тому випадку, якщо необхідно перевірити рівність частоти мутацій у декількох вибірках одночасно можна використовувати критерій Хі-квадрат:

$$\chi^2 = \sum_{i=1}^{k} \frac{(m_i - q_i m)^2}{q_i m}, \qquad (5.7)$$

із числом ступенів свободи: $df = k - 1$, де k – кількість вибірок.

У цій формулі m – сукупна кількість особин, які несуть мутацію, а q_i – частка особин i-тої вибірки серед усієї сукупності організмів.

Міграціями називають *або включення деякої кількості особин у дану популяцію (імміграція), або переселення з цієї популяції (еміграція)*. Цей процес може також викликати зміну частот алелів у популяції.

У практичному тваринництві імміграція здійснюється шляхом закупівель чи введення в стадо нових тварин, наприклад, для «прилиття крові» потрібних порід у місцеві популяції.

Припустимо, що після включення деякої кількості тварин чисельність популяції стала дорівнювати N. Кількість прибулих позначимо через I. Отже, частка іммігрантів складає: $i = \frac{I}{N}$. Тоді частка місцевих тварин буде: $\frac{N-I}{N} = 1 - i$.

Позначимо частоти алелів іммігрантів через p_i і q_i; частоти алелів популяції до міграції – p_1 і q_1. Тоді

$$p_1 = ip_i + (1-i)p_0 = i(p_i - p_0) + p_0. \qquad (5.8)$$

Різниця частот алелів у наступному поколінні й у вихідній популяції складає:

$$\Delta p = p_1 - p_0 = i(p_i - p_0) + p_0 - p_0 = i(p_i - p_0). \qquad (5.9)$$

Отже, зміна частот алелів у популяції залежить від їхнього первісного значення у вихідній популяції, частоти алелів у тварин, що іммігрують та від їхньої відносної частки в новій популяції.

Приклад. В стаді, що нараховує 1000 тварин відзначено наступну структуру генотипів: особин генотипу *AA* було 500, генотипу *Aa* – 400, і особин

генотипу *aa* – 100. У це стадо було введено 500 тварин, різних генотипів: *AA* – 100, *Aa* – 200 і *aa* – 200 особин. Які будуть частоти алелів *A* та *a* в новій змішаній популяції?

У вихідній популяції частоти алелів *A* та *a* були, відповідно:

$$p_A^0 = \frac{2 \cdot 500 + 400}{2 \cdot 1000} = 0{,}70,$$

$$p_a^0 = \frac{2 \cdot 100 + 400}{2 \cdot 1000} = 0{,}30.$$

Частоти алелів *A* та *a* серед іммігрантів:

$$p_A^i = \frac{2 \cdot 100 + 200}{2 \cdot 500} = 0{,}40,$$

$$p_a^i = \frac{2 \cdot 200 + 200}{2 \cdot 500} = 0{,}60.$$

Частка іммігрантів у змішаному стаді складає:

$$i = \frac{500}{500 + 1000} = \frac{1}{3}.$$

Тоді частоти алелів у змішаній популяції будуть:

$$p_A^1 = \frac{1}{3} \cdot (0{,}4 - 0{,}7) + 0{,}7 = 0{,}60,$$

$$p_a^1 = 1 - 0{,}6 = 0{,}4.$$

Зміна частот алелів при уведенні тварин складає:

$$\Delta p_A = \frac{1}{3} \cdot (0{,}4 - 0{,}7) = -0{,}1.$$

Якщо введення в популяцію чи виведення з неї тварин повторюється систематично, то частота алелів буде змінюватися в кожній генерації. Якщо тварин вводили (чи виводили з неї) лише однократно, то при першому ж стабілізуючому схрещуванні в змішаній популяції наступить генетична зрівноваженість.

Якщо частка іммігрантів і їхні частоти алелів залишаються постійними з покоління в покоління, то частота алелів у змішаній популяції через *t* генерацій буде складати:

$$p_t = (1 - i)^t (p_0 - p_i) + p_i. \qquad (5.10)$$

Якщо відомі частоти алелів у вихідній популяції і популяції, з якої іммігрують тварини, а також тривалість процесу імміграції (у поколіннях), то оцінку потоку мігрантів можна одержати за формулою:

$$i = 1 - \sqrt[t]{\frac{p_t - p_i}{p_0 - p_i}}. \qquad (5.11)$$

Приклад. Червоно-ряба худоба Чехословаччини створена у результаті прилиття крові червоної датської худоби місцевій червоній. Процес створення породи тривав близько 10 поколінь.

Яка була частка тварин, що вводилася (у середньому на одне покоління), якщо частота алеля Z у місцевої червоної худоби складала 0,6, у червоної датської – 0,14, а у тварин створеної чеської червоно-рябої породи – 0,56.

Отже, за умовою маємо, число генерацій $t = 10$, частоту алеля у вихідній популяції – $p_0 = 0{,}6$, частоту алеля у іммігрантів – $p_i = 0{,}14$, частоту алеля в змішаній групі через t поколінь – $p_t = 0{,}56$.

Тоді, середня частка тварин, що вводилися, дорівнює:

$$i = 1 - \sqrt[10]{\frac{0{,}56 - 0{,}14}{0{,}60 - 0{,}14}} = 1 - 0{,}991 = 0{,}009.$$

Іншими словами, потік генів *Z* від червоної датської худоби до червоно-рябої чеської йшов із середньою інтенсивністю 0,9% за одну генерацію.

Контрольні питання:

1. Які фактори здатні змінювати генетичну структуру популяції.
2. Поняття про мутації.
3. Поняття про міграції.
4. Як розрахувати частоту алелів у змішаній популяції через t генерацій, якщо частка іммігрантів і їхні частоти алелів залишаються постійними з покоління в покоління?

§ 6. Фактори динаміки популяцій. Випадковий дрейф генів

Випадковим дрейфом генів називається *зміна частот алелів у ряді поколінь, що викликана випадковими причинами, наприклад, нечисленністю популяції.*

Припустимо, що в даній популяції частоти двох алелів (*A* и *a*) дорівнюють 0,4 і 0,6, відповідно. Тоді в наступному поколінні частота алеля *A* може бути меншою (чи більшою), ніж 0,4 просто через те, що у вибірці гамет, що утворять зиготи цього покоління, частота алеля *A* в силу якихось причин виявилася меншою (чи більшою), ніж це можна було очікувати.

Дрейф генів – процес зовсім випадковий; він належить до особливого класу явищ, що мають назву *помилками вибірки*. Загальне правило полягає в тому, що величина «помилки» завжди перебуває у зворотній залежності від величини вибірки: чим менше величина вибірки, тим більше помилка (див. §§ 1, 2).

Таким чином, чим менша кількість особин, які схрещуються у популяції, тим більші зміни, зумовлені дрейфом генів, будуть спостерігатися на частотах алелів.

Однак, оскільки випадкові зміни частот алелів відбуваються в будь-яких напрямках, тенденція до підвищення чи, навпаки, зниження частоти алеля завжди може змінитися на зворотну, доки частота алеля не досягне нуля чи одиниці.

Якщо відома кількість батьків у вихідному поколінні і, відповідно, частоти алелів у ньому, то можна розрахувати імовірність одержати в наступному поколінні ті чи інші частоти алелів. Для цього необхідно оцінити варіансу (S^2) частот алелів у цьому поколінні:

$$S^2 = \frac{pq}{2N}, \tag{6.1}$$

де N – чисельність особин батьківського покоління;

p, q – частоти алелів *A* та *a* в батьківському поколінні.

Тоді із 95% імовірністю можна очікувати, що частоти алелів *A* та *a* в наступному поколінні (p_1 і q_1) будуть знаходиться в наступних межах:

$$\begin{aligned} p - 1{,}96\sqrt{\frac{pq}{2N}} \le p_1 \le p + 1{,}96\sqrt{\frac{pq}{2N}}; \\ q - 1{,}96\sqrt{\frac{pq}{2N}} \le q_1 \le q + 1{,}96\sqrt{\frac{pq}{2N}}. \end{aligned} \tag{6.2}$$

Це означає, що в 95 популяціях із 100 (для яких частоти алелів були p і q), частоти алелів у наступному поколінні – p_1 і q_1 будуть знаходитися в приведених вище інтервалах.

Ширина цього інтервалу значною мірою визначається чисельністю відповідної популяції; для значення $p = q = 0{,}5$ та різних чисельностях вихідної популяції розмах значень у наступному поколінні наведено у таблиці 6.1.

Таблиця 6.1 – Розмах значень p_1, очікуваний з 95% імовірністю за різної чисельності популяції

Чисельність популяції (N)	Кількість гамет ($2N$)	Варіанса $S^2 = \frac{pq}{2N}$	Розмах значень p_1, очікуваний з 95% імовірністю, тобто $p + 1{,}96\sqrt{\frac{pq}{2N}}$
5	10	0,025	0,18 - 0,82
50	100	0,0025	0,40 - 0,60
500	1000	0,00025	0,468 - 0,532

Таким чином, випадковий дрейф генів (при відсутності інших генетичних процесів) призводить, зрештою, до елімінації одного з алелів чи, навпаки, – його повного закріплення (тобто, $p \to 0$ чи $p \to 1$). Наслідком цього процесу є зниження гетерозиготності (оскільки максимальна частота гетерозигот у популяції має місце у випадку, коли $p = q = 0{,}5$).

Розрізняють *фактичну гетерозиготність*, що визначається як частота особин у популяції, гетерозиготних за визначеним локусом:

$$h = \frac{H}{N}, \tag{6.3}$$

і *очікувану гетерозиготність*, що визначається на підставі частот алелів у припущенні, що схрещування в популяції відбуваються випадковим чином:

$$\tilde{h} = \frac{2N}{2N-1}\left[1 - p^2 - q^2\right]. \tag{6.4}$$

Приклад. У популяції чисельності генотипів були наступні: $D = 250$, $H = 130$, $R = 15$. Необхідно оцінити фактичну і теоретичну гетерозиготність.

Оцінка фактичної гетерозиготність знаходиться за наведеною вище формулою і складає:

$$h = \frac{130}{250 + 130 + 15} = 0{,}329.$$

Для розрахунку очікуваної гетерозиготності спочатку необхідно розрахувати частоти алелів p і q, використовуючи метод максимальної правдоподібності (див. § 3):

$$p = \frac{2 \times 250 + 130}{2 \times 395} = 0{,}80,$$

$$q = \frac{2 \times 15 + 130}{2 \times 395} = 0{,}20.$$

Тоді, оцінка очікуваної гетерозиготності становитиме:

$$\tilde{h} = \frac{2 \cdot 395}{(2 \cdot 395) - 1} \times \left[1 - 0{,}8^2 - 0{,}2^2\right] = 0{,}320.$$

Навіть при цілком випадковому схрещуванні, гетерозиготність зменшиться за одне покоління в $\left(1-\frac{1}{2N}\right)$ рази, якщо N – чисельність популяції. Тоді в t-ому поколінні рівень гетерозиготності буде складати:

$$h_t = h_0 e^{\left(-\frac{t}{2N}\right)}, \tag{6.5}$$

де h_0 – вихідний рівень гетерозиготності.

Таким чином, якщо в ізольованій популяції (тобто, під час відсутності імміграції) протягом декількох поколінь зберігається лише невелика кількість особин, то генетична мінливість такої популяції поступово зменшується.

Але в реальних популяціях зустрічаються тварини різної статі в різній кількості і, крім того, їхня чисельність може щорічно змінюватися. Для таких популяцій С. Райт (1932) увів поняття *ефективної чисельності* (*Ne*).

Ефективною чисельністю *називається чисельність ідеальної популяції, в якій має місце такий же рівень дрейфу генів, що й у реальній, тобто в популяції, чисельністю Ne має місце таке зниження рівня гетерозиготності, що й у популяції чисельністю N = Nf + Nm.*

Оцінку значення *Ne* можна зробити за формулою:

$$Ne = \frac{4\ Nf\ Nm}{Nf + Nm}, \tag{6.6}$$

де *Nf* – кількість самок;

Nm – кількість самців.

Для *N* і *Ne* відмічено наступні залежності:

1. Якщо чисельності особин різної статі в популяції рівні (тобто *Nf* = *Nm*), то фактична та ефективна чисельності цієї популяції будуть співпадати (тобто *N* = *Ne*).

2. Якщо чисельність особин однієї статі більша чи менша чисельності особин іншої (тобто $Nf \neq Nm$), то ефективна чисельність буде завжди нижчою, ніж реальна (тобто *N* > *Ne*).

3. Чим менше в популяції відношення $\left(\frac{Ne}{N}\right)$, тим сильніше в цій популяції виражена тенденція до інбридингу і, відповідно, зниження рівня гетерозиготності (і, відповідно, збільшення гомозиготності).

Приклад. Яка буде ефективна чисельність популяції, що містить 50 корів і 10 бугаїв?

Використовуючи приведену вище формулу, знаходимо:

$$Ne = \frac{4 \cdot 50 \cdot 10}{50 + 10} = 33,$$

тобто, практично в два рази нижче, ніж реальна.

У крайньому випадку, коли спермою одного бугая запліднюється велика кількість корів (тобто $Nf \rightarrow \infty$, а *Nm* = 1), ефективна чисельність популяції дорівнює лише 4.

Інбридинг – *форма схрещування, при якому імовірність зустрічі гамет, що належать спорідненим тваринам, вища, ніж можна було б очікувати на підставі випадкового схрещування.*

Мірою генетичних наслідків інбридингу є коефіцієнт інбридингу (*F*) – імовірність того, що у будь-якої особини в даному локусі виявляться два алеля, ідентичні за походженням.

При інбридингу в популяції:

- знижується частота гетерозигот;
- виявляються рецесивні генотипи;
- підвищується фенотипова мінливість;
- відбувається зміна частот генотипів, при незмінності частот алелів.

У цьому випадку частоти генотипів виявляються наступними:

$$D = p^2 + pqF;$$

$$H = 2pq(1 - F);$$

$$R = q^2 + pqF.$$

Коефіцієнт інбридингу у випадку діалельної системи з повним домінуванням можна розрахувати, використовуючи наступні формули:

$$F = \frac{4DR - H^2}{(2D + H)(2R + H)};$$

$$F = \sqrt{\frac{\chi^2}{N}}; \qquad (6.7)$$

$$F = 1 - \frac{H}{\widetilde{H}},$$

де *D*, *H*, *R* – чисельність домінантних гомозигот, гетерозигот та рецесивних гомозигот, відповідно, у популяції чисельністю *N*;

χ^2 – значення критерію Хі-квадрат К. Пірсона, розрахованого для перевірки генної рівноваги в популяції (див. § 4);

$\widetilde{H}$ – очікувана чисельність гетерозигот, за умови панміксії в популяції.

Приклад. Розрахуйте коефіцієнт інбридингу для популяції, що характеризувалася наступним розподілом генотипів: *D* = 92, *H* = 56, *R* = 32.

Використовуючи першу з приведених вище формул, розраховуємо:

$$F = \frac{4 \cdot 92 \cdot 32 - 56^2}{(2 \cdot 92 + 56) \cdot (2 \cdot 32 + 56)} = 0{,}30.$$

При інбридингу реальна чисельність гетерозигот завжди нижча, ніж теоретична ($\widetilde{H}$). Для того, щоб розрахувати очікувану чисельність гетерозигот, спочатку необхідно знайти оцінки частот алелів:

$$p = \frac{2 \cdot 92 + 56}{2 \cdot 180} = 0{,}67,$$

$$q = \frac{2 \cdot 32 + 56}{2 \cdot 180} = 0{,}33.$$

Тоді очікувана чисельність гетерозигот становитиме:

$$\widetilde{H} = 2pqN = 2 \cdot 0{,}67 \cdot 0{,}33 \cdot 180 \approx 80 \text{ особин.}$$

Коефіцієнт інбридингу, отже, буде дорівнювати (використовуючи третю формулу):

$$F = 1 - \frac{56}{80} = 0{,}3.$$

У випадку поліалельної системи із повним домінуванням вибіркову оцінку коефіцієнта інбридингу можна одержати на підставі формули:

$$F = \sqrt{\frac{\chi^2}{N(k-1)}}, \qquad (6.8)$$

де χ^2 – значення критерію Хі-квадрат К. Пірсона, розрахованого для перевірки генної рівноваги в популяції;

k – кількість алелів.

Приклад. У стаді великої рогатої худоби бестужевської породи виявлено наступний розподіл генотипів за локусом трансферину: тварин із генотипом *AA* було 32 особини, *AD* – 36, *AE* – 60, *DD* – 57, *DE* – 90 і *EE* – 125 особин. Необхідно розрахувати коефіцієнт інбридингу для цієї популяції.

Спочатку розрахуємо частоти алелів на підставі формул Бернштейна:

$$p_A = \frac{2 \cdot 32 + 36 + 60}{2 \cdot 400} = 0{,}2;$$

$$q_D = \frac{2 \cdot 57 + 36 + 90}{2 \cdot 400} = 0{,}3;$$

$$z_E = \frac{2 \cdot 125 + 60 + 90}{2 \cdot 400} = 0{,}5.$$

Потім розраховуємо очікувані чисельності генотипів і значення критерію Хі-квадрат (табл. 6.2).

Таблиця 6.2 – Фактичні та теоретичні частоти генотипів за локусом трансферину і значення критерію Хі-квадрат

Генотип	Фактична частота (Φ)	Теоретична частота (T)	$\frac{(\Phi - T)^2}{T}$
AA	32	16	16,00
DD	57	36	12,25
EE	125	100	6,25
AD	36	48	3,00
AE	60	80	5,00
DE	90	120	7,50
Суми	400	400	χ^2 =50,00

Тоді коефіцієнт інбридингу в даній популяції буде становити:

$$F = \sqrt{\frac{50{,}0}{400 \cdot (3-1)}} = 0{,}25.$$

Проблематичнішою є оцінка коефіцієнта інбридингу у випадку двохалельної системи із повним домінуванням. У цьому випадку, як відомо, фенотипово можна виділити тільки дві групи особин. Завдання може бути вирішене лише в тому випадку, якщо серед особин із домінантним фенотипом можна точно виявити гомозигот і гетерозигот (не обов'язково всіх), наприклад, у результаті зворотних схрещувань з рецесивними гомозиготами.

Припустимо, у вибірці обсягом N виявлено d особин, що мають домінантний фенотип і r особин – рецесивних гомозигот. Із числа перших вдалося точно ідентифікувати a гомозиготних і b гетерозиготних особин. Сума $(a + b)$ звичайно менша, ніж d тому, що вірогідно розпізнаються не всі особини із домінантною ознакою. Тоді частоту домінантного алеля можна розрахувати за формулою:

$$p = \frac{d(2a+b)}{2N(a+b)}, \tag{6.9}$$

а оцінку коефіцієнта інбридингу можна одержати за формулою:

$$F = \frac{r - Nq^2}{Npq}. \tag{6.10}$$

Приклад. У вибірці, що містить 90 особин, 74 мали домінантний і 16 рецесивний фенотип. Особин із домінантною ознакою було проаналізовано шляхом зворотного схрещування і було встановлено, що 23 з них гомозиготні, а 14 гетерозиготні. Оцініть частоти алелів і коефіцієнт інбридингу.

Як бачимо, лише половину особин із домінантним фенотипом вдалося точно ідентифікувати, але і цього достатньо для того, щоб оцінити частоти алелів з високою точністю:

$$p = \frac{74 \cdot (2 \cdot 23 + 14)}{2 \cdot 90 \cdot (23 + 14)} = \frac{2}{3}.$$

Тепер скористаємося формулою 6.10 і розрахуємо оцінку коефіцієнта інбридингу:

$$F = \frac{16 - 90 \cdot \left(\frac{1}{3}\right)^2}{90 \cdot \frac{2}{3} \cdot \frac{1}{3}} = 0{,}30.$$

Контрольні питання:

1. Поняття про дрейф генів.
2. Розрахунок фактичної та очікуваної гетерозиготності.
3. Поняття про ефективну чисельність популяції та методика її оцінки.
4. Поняття про інбридинг та методика розрахунку коефіцієнта інбридингу.

§ 7. Основні форми відбору

У будь-якій конкретній популяції особини, які мають різний генотип, мають різну можливість залишити потомство. Більше того, особини, що навіть репродукують, залишають після себе різну кількість нащадків. Це зумовлено або їх низькою виживаністю, у порівнянні із особинами інших генотипів, або, з іншого боку, їх нижчою плодючістю. У результаті диференціального розмноження особин певного генотипу відбувається зміна частот алелів у популяції.

На відміну від мутагенезу, міграції і дрейфу генів, що випадковим чином (тобто, неспрямовано) змінюють вихідні частоти алелів і генотипів, диференціальне відтворення різних генетичних варіантів (тобто, *відбір*) сприяє підвищенню загальної пристосованості популяції і запобігає руйнівним наслідкам стохастичних процесів. Відбір підтримує і стабілізує сприятливий у даних умовах генетичний гомеостаз популяції.

У першому наближенні, залежно від характеру механізмів елімінації особин із популяції, виділяють дві форми відбору – *природний* і *штучний*.

При *природному відборі* перевага одних генотипів над іншими визначається на підставі рівня їхньої пристосованості до певних умов середовища; при штучному – ця перевага оцінюється суб’єктивно (у ході селекційної роботи).

Більш правильно було б вважати, що при штучному відборі селективна перевага тварин, що відбираються на плем’я, визначається спільно і тим, і іншим механізмом. Незважаючи на це, багато закономірностей, розкритих при вивченні механізмів дії природного відбору серед природних видів, можуть бути перенесені і на селекційний процес.

Як кількісну міру інтенсивності відбору звичайно використовують показник дарвінівської, чи відносної, *пристосованості* (*w*).

Пристосованість є мірою ефективності розмноження даного генотипу. До компонентів пристосованості входить цілий ряд біологічно важливих ознак:

- виживаність;
- плодючість;
- швидкість розвитку;
- тривалість репродуктивного віку, і т.д.

Тиск відбору на той чи інший алель локусу виражається *коефіцієнтом відбору* (*s*), що оцінює переважне відтворення спадкової ознаки в наступному поколінні і визначає швидкість зменшення частоти того чи іншого генотипу:

$$s = 1 - w. \qquad (7.1)$$

Розглянемо методику розрахунку показників пристосованості генотипів у випадку двохалельної системи.

Приклад. У популяції корів симентальської породи в батьківському і дочірньому поколіннях було відзначено наступний розподіл генотипів за системою *F-V* груп крові (табл 7.1):

Таблиця 7.1 – Розрахунок коефіцієнта відбору

Етапи розрахунку	Генотип		
	F/F	*F/V*	*V/V*
Кількість особин у батьківському поколінні	115	78	16
Кількість особин у дочірньому поколінні	120	60	20
Розрахунок:			
1. Середня кількість нащадків, що припадає на одну особину батьківського покоління	$\frac{120}{115}=1{,}043$	$\frac{60}{78}=0{,}769$	$\frac{20}{16}=1{,}250$
2. Пристосованість (*w*), щодо кращого генотипу	$\frac{1{,}043}{1{,}250}=0{,}83$	$\frac{0{,}769}{1{,}250}=0{,}62$	$\frac{1{,}250}{1{,}250}=1{,}0$
3. Коефіцієнт відбору (*s* = 1 – *w*)	1-0,83 = 0,17	1-0,62 = 0,38	0

Коефіцієнт відбору оцінює, таким чином, частку особин кожного генотипу, що **НЕ** вносять свого внеску у формування генного пула наступного покоління.

У розглянутому вище прикладі частоти алелів у батьківському поколінні будуть:

$$p_F^P=\frac{2\cdot 115+78}{2\cdot 209}=0{,}737;$$

$$q_V^P=\frac{2\cdot 16+78}{2\cdot 209}=0{,}263,$$

а в дочірньому:

$$p_F^O=\frac{2\cdot 120+60}{2\cdot 200}=0{,}75;$$

$$q_V^O=\frac{2\cdot 20+60}{2\cdot 200}=0{,}25.$$

Якщо коефіцієнт відбору для генотипу *FF* у батьківському поколінні складає 0,17, то із 115 особин даного генотипу алель *p* наступному поколінню передасть 115·(1-0,17) = 95,45, тобто, лише 95 тварин. Із 78 тварин генотипу *FV* лише 78·(1-0,38) = 48,36, тобто, 48 тварин передадуть наступному поколінню алелі *p* і *q*. Нарешті, оскільки для генотипу *VV* коефіцієнт відбору дорівнює 0, те всі 16 особин передадуть наступному поколінню алель *q*. Таким чином, реально, у формуванні генного пула наступного покоління будуть брати участь не 115, 78 і 16 особин різних генотипів, а тільки 95, 48 і 16, відповідно. Використовуючи ці частоти генотипів можна розрахувати частоти алелів у дочірньому поколінні. Вони виявляться точно такими, як показано вище.

Таким чином, у випадку дії штучного відбору, коефіцієнт відбору (*s*) – це частка тварин того чи іншого генотипу, що підлягає вибракуванню зі стада.

Розглянемо основні форми відбору.

7.1 Відбір проти рецесивних гомозигот

При відборі проти рецесивних гомозигот пристосованість генотипів (*w*) буде наступною: *AA* – 1; *Aa* – 1; *aa* – (1 – *s*).

Тоді частота рецесивного алеля (*a*) у наступному поколінні буде становити:

$$q_1 = \frac{q_0 - sq_0^2}{1 - sq_0^2}. \qquad (7.2)$$

Якщо рецесивні гомозиготи нежиттєздатні, тобто *s* = 1, то формула спрощується:

$$q_1 = \frac{q_0}{1 + q_0}. \qquad (7.3)$$

При відборі проти рецесивних леталей протягом *n* поколінь, вихідна частота алеля знизиться до рівня:

$$q_n = \frac{q_0}{1 + nq_0}, \qquad (7.4)$$

а частка гетерозигот складатиме:

$$h_n = \frac{2q_0[1 + (n-1)q_0]}{(1 + nq_0)^2}. \qquad (7.5)$$

Якщо заздалегідь вирішується завдання зниження частоти рецесивного (летального) алеля з рівня q_0 до рівня q_n, то необхідне число поколінь для цього становитиме:

$$n = \frac{q_0 - q_n}{q_0 \cdot q_n}. \qquad (7.6)$$

Приклад. У великої рогатої худоби мозкова грижа зумовлена аутосомним рецесивним алелем *c*. У стаді швіцької худоби серед 520 новонароджених телят виявилося дві особини із мозковою грижею. Особини із такою аномалією нежиттєздатні. Скільки необхідно поколінь проводити відбір проти алеля *c*, щоб його частота знизилася вдвічі?

Частота рецесивного алеля q_0 у даному стаді становить:

$$q_0 = \sqrt{\frac{2}{520}} = 0{,}062.$$

Для того щоб знизити його частоту вдвічі, тобто, до рівня q_n = 0,031, необхідно:

$$n = \frac{0{,}062 - 0{,}031}{0{,}062 \cdot 0{,}031} \approx 16 \text{ поколінь.}$$

Тоді частка гетерозигот, що несуть алель *c* у прихованій формі, у 16-му поколінні становитиме:

$$h_{16} = \frac{2 \cdot 0{,}062 \cdot [1 + (16-1) \cdot 0{,}062]}{(1 + 16 \cdot 0{,}062)^2} = 0{,}06.$$

Якщо рецесивні гомозиготи не мають летального прояву, то для того щоб знизити частоту алеля *a* з рівня q_0 до рівня q_n, коефіцієнт відбору і необхідна кількість поколінь пов'язані наступною залежністю:

$$s \cdot n = \frac{q_0 - q_n}{q_0 q_n} + \ln\left[\frac{q_0(1-q_n)}{q_n(1-q_0)}\right]. \quad (7.7)$$

У цілому, відбір проти рецесивних гомозигот (навіть у випадку їхньої летальності) – малоефективний, особливо при низькій вихідній частоті рецесивного алеля.

7.2 Відбір проти домінантного алеля у випадку повного домінування

У випадку повного домінування (тобто, коли домінантні гомозиготи і гетерозиготи фенотипово схожі), пристосованість генотипів (*w*) буде наступною: *AA* – (1 – *s*); *Aa* – (1 – *s*); *aa* – 1.

Відбір проти домінантних алелів йде більш ефективно, ніж відбір проти рецесивних, оскільки домінантні алелі виявляються не лише в гомозиготному, а й у гетерозиготному стані. При такому відборі частота домінантного алеля в наступному поколінні складе:

$$p_1 = \frac{p_0(1-s)}{1-s(1-q_0^2)}. \quad (7.8)$$

При повній елімінації генотипів *AA* і *Aa* (тобто, при *s* = 1), частота алеля *A* в наступному поколінні, природно, буде дорівнювати нулю.

Якщо заздалегідь відома частота домінантного алеля, що бажана в наступному поколінні, то інтенсивність відбору (тобто, коефіцієнт відбору) проти неї, можна розрахувати, використовуючи наступну формулу:

$$s = \frac{p_0 - p_1}{p_1 q_0^2 + p_0 - p_1}. \quad (7.9)$$

Приклад. Наявність куцого хвоста у деяких порід курей обумовлена домінантним аутосомним геном (*P*). У стаді із 200 курей 152 були із куцими хвостами. Яка частота алеля *P*? Яким повинен бути коефіцієнт відбору проти цього алеля, щоб у наступному поколінні його частота знизилася до 0,10?

Частота домінантного алеля у випадку повного домінування становить:

$$p_0 = 1 - \sqrt{\frac{200-152}{200}} = 0{,}51.$$

Для того щоб у наступному поколінні вона знизилася до рівня 0,1, коефіцієнт відбору повинен становити:

$$s = \frac{0{,}51 - 0{,}10}{0{,}1 \cdot 0{,}49^2 + 0{,}51 - 0{,}10} = 0{,}945.$$

Таким чином, із 152 курей із куцим хвостом повинно бути елімінovано:

$$152 \cdot (1 - 0{,}055) = 144 \text{ особини.}$$

7.3 Відбір проти домінантного алеля у випадку кодомінування

При даній формі відбору пристосованість генотипів (*w*) буде наступною: *AA* – (1 – *s*); *Aa* – 1; *aa* – 1.

У цьому випадку частота домінантного алеля в наступному поколінні складатиме:

$$p_1 = \frac{p_0 - sp_0^2}{1 - sp_0^2}. \tag{7.10}$$

Дана форма відбору (як і відбір проти рецесивних гомозигот) більшою мірою залежить від початкової частоти домінантного алеля – чим нижча початкова частота, тим менш ефективний відбір.

7.4 Відбір проти гетерозигот

Зниження пристосованості та зменшення плодючості часто зумовлюються транслокацією хромосом, наприклад, при транслокації Робертсона, що приводить до появи гетерозигот у результаті з'єднання різних хромосом в один агрегат. Тварини із таким хромосомним дефектом відрізняються зниженою виживаністю і плодючістю, чисельність їх у популяції зменшується під дією відбору.

При відборі проти гетерозигот, пристосованість генотипів (*w*) буде наступною: *AA* – 1; *Aa* – (1 – *s*); *aa* – 1.

Тоді частота домінантного алеля в наступному поколінні буде:

$$p_1 = \frac{p_0 - p_0 q_0 s}{1 - 2 p_0 q_0 s}. \tag{7.11}$$

При повній елімінації гетерозигот, частоти алелів у наступному поколінні становитимуть:

$$p_1 = \frac{p_0^2}{p_0^2 + q_0^2}; \qquad q_1 = \frac{q_0^2}{p_0^2 + q_0^2}. \tag{7.12}$$

У тому випадку, якщо $p_0 = q_0 = 0{,}5$, генотипова структура популяції буде знаходитися у стані рівноваги. Однак, ця рівновага дуже нестійка.

У тому випадку, коли початкова частота домінантного алеля перевищує частоту рецесивного, тобто, при $p_0 > q_0$, рецесивний алель буде поступово елімінуватися із популяції. У тому випадку, якщо $p_0 < q_0$ – поступовій елімінації буде піддаватися домінантний алель.

Приклад. Від схрещування білих норок із темними в першому поколінні одержують гетерозиготних особин, що мають світле забарвлення із темним хрестом на спині. На початку в популяції було 105 білих, 215 гетерозиготних і

80 темних норок. Норки якого кольору залишаться в популяції, якщо гетерозиготні форми цілком елімінуються в кожному поколінні?

Розраховуємо вихідну частоту алелів білого і темного кольору:

$$p_W^0 = \frac{2 \cdot 105 + 215}{2 \cdot 400} = 0{,}53;$$

$$q_B^0 = \frac{2 \cdot 80 + 215}{2 \cdot 400} = 0{,}47.$$

При повному усуненні гетерозигот із розмноження, частоти алелів у наступному поколінні будуть:

$$p_W^1 = \frac{0{,}53^2}{0{,}53^2 + 0{,}47^2} = 0{,}56;$$

$$q_B^1 = \frac{0{,}47^2}{0{,}53^2 + 0{,}47^2} = 0{,}44.$$

Як бачимо, відбір обумовив зниження частоти норок темного кольору. Таким чином, зрештою, у популяції залишаться тільки білі норки.

Частота алелів *A* та *a* в *n*-ному поколінні при повній елімінації гетерозигот буде дорівнювати:

$$p_n = \frac{p_0^{2^n}}{p_0^{2^n} + q_0^{2^n}};$$

$$q_n = \frac{q_0^{2^n}}{p_0^{2^n} + q_0^{2^n}}. \qquad (7.13)$$

При повній елімінації гетерозигот частота більш рідкісного алеля буде знижуватися швидше, ніж при елімінації гомозиготних генотипів цієї алелі. Наприклад, при летальності рецесивних гомозигот вихідна частота алеля *a* q_0 = 0,1 знизиться за три покоління відбору до рівня q_3 = 0,077, а при повній елімінації гетерозигот – до рівня $q_3 = 2{,}32 \cdot 10^{-8}$.

7.5 Відбір на користь гетерозигот

Відбір на користь гетерозигот, коли обидві гомозиготи мають знижену, порівняно із гетерозиготами пристосованість, називається також *наддомінуванням*. При такому відборі відбувається формування стійкої поліморфної рівноваги. Прикладом відбору на користь гетерозигот є селекція сірих каракульських овець, які в гетерозиготному стані мають більш високу виживаність.

Пристосованості генотипів (*w*) при відборі на користь гетерозигот будуть наступними: *AA* – $(1 - s_1)$; *Aa* – 1; *aa* – $(1 - s_2)$.

Частота домінантного і рецесивного алелів у наступному поколінні тоді становитиме:

$$p_1 = \frac{p_0 - p_0^2 s_1}{1 - p_0^2 s_1 - q_0^2 s_2};$$
$$q_1 = \frac{q_0 - q_0^2 s_2}{1 - p_0^2 s_1 - q_0^2 s_2}. \quad (7.14)$$

При повній елімінації гомозигот (тобто, якщо $s_1 = s_2 = 1$), у наступному поколінні частоти алелів *A* та *a* досягають рівноважного стану: $p_0 = q_0 = 0{,}5$.

При неповній елімінації гомозигот (тобто, якщо $s_1 \neq s_2 < 1$), частоти алелів будуть із кожним поколінням прагнути до рівноважного стану:

$$\hat{p} = \frac{s_2}{s_1 + s_2};$$
$$\hat{q} = \frac{s_1}{s_1 + s_2}. \quad (7.15)$$

Приклад. У норок домінантний ген у гетерозиготному стані зумовлює сріблясто-соболиний колір хутра («подих весни»), а в гомозиготному – має летальну дію. Рецесивний алель обумовлює темно-коричневий (стандартний) колір хутра, але життєздатність таких особин становить лише 90% від життєздатності гетерозигот.

Необхідно визначити стан рівноваги для частот алелів кольору хутра норок.

Для домінантних гомозигот, за умовою, пристосованість дорівнює нулю ($w_{AA} = 0$), тому коефіцієнт відбору для них становить – $s_1 = 1$. Для рецесивних гомозигот пристосованість становить $w_{aa} = 0{,}9$, а коефіцієнт відбору – $s_2 = 1 - 0{,}9 = 0{,}1$.

Тоді рівноважні частоти будуть:

$$\hat{p} = \frac{0{,}1}{1 + 0{,}1} = 0{,}09;$$
$$\hat{q} = \frac{1}{1 + 0{,}1} = 0{,}91.$$

7.6 Загальний випадок відбору

У загальному випадку, коли пристосованості генотипів (*w*) складають: *AA* – $(1 - s_1)$; *Aa* – $(1 - s_2)$; *aa* – $(1 - s_3)$, частота домінантного алеля в наступному поколінні становитиме:

$$p_1 = \frac{p_0^2(s_2 - s_1) + p_0(1 - s_2)}{p_0^2(2s_2 - s_1 - s_3) + 2p_0(s_3 - s_2) + 1 - s_3}. \quad (7.16)$$

Контрольні питання:

1. Поняття про відбір. Форми відбору
2. Якою буде пристосованість генотипів (*w*) за різних форм відбору?

§ 8. Генетична диференціація популяцій

Як було показано у § 3, при відсутності мутацій, міграцій, тиску відбору, майже безмежної чисельності та в умовах вільного схрещування (панміксії) популяція знаходиться у стані генетичної рівноваги й частоти генотипів обумовлюються законом Гарді-Вайнберга. Але у тих випадках, коли популяція не є цілісною, а складається з декількох субпопуляцій (наприклад, колоній), навіть при повній панміксії всередині цих субпопуляцій для популяції в цілому закон Гарді-Вайнберга вже не дотримується й частоти генотипів тоді для неї будуть дорівнювати:

$$\begin{aligned} AA &: p^2 + \sigma_q^2, \\ Aa &: 2pq - 2\sigma_q^2, \\ aa &: q^2 + \sigma_q^2, \end{aligned} \tag{8.1}$$

де

$$\sigma_q^2 = \frac{\sum_{i=1}^{s} (q_i - \bar{q})^2}{s}, \tag{8.2}$$

$$\bar{q} = \frac{\sum_{i=1}^{s} q_i}{s}, \tag{8.3}$$

де q_i – частота алеля в і-тій субпопуляції;
s – кількість субпопуляцій.

Вперше формули 8.1 отримав німецький вчений С. Валунд (Wahlund, 1928) тому явище впливу підрозділу популяції має назву *явища Валунда*.

Воно зумовлено тим, що частоти генотипів у популяції, що підрозділена на субпопуляції, змінюються на величину σ_q^2 (варіансу частот алелів), причому, частка гомозигот збільшується, а частка гетерозигот, навпаки, знижується. Таким чином, ефект структурування популяції подібний до прояву інбридингу (див. § 6).

Як ми пам’ятаємо, при інбридингу (інтенсивністю F) частоти генотипів у популяції мають наступний вигляд:

$$\begin{aligned} AA &: p^2 + pqF, \\ Aa &: 2pq(1-F), \\ aa &: q^2 + pqF, \end{aligned} \tag{8.4}$$

Звідси ми можемо отримати значення коефіцієнта інбридингу, виражене через варіансу частот алелів:

$$F = \frac{\sigma_q^2}{\bar{q} \cdot (1 - \bar{q})}. \quad (8.5)$$

Значення коефіцієнта інбридингу, отримане за формулою 8.5 має відношення до всієї популяції в цілому, але при цьому самі субпопуляції також можуть бути інбредованими в тій чи іншій мірі, тому С. Райт (Wright, 1951) запропонував декілька коефіцієнтів, які б відображували ступінь генетичної диференціації субпопуляцій:

- F_{IT} – коефіцієнт інбридингу особини (*I*) відносно цілої популяції (*T*);
- F_{IS} – коефіцієнт інбридингу особини (*I*) відносно субпопуляції (*S*), до якої вона відноситься;
- F_{ST} – коефіцієнт інбридингу субпопуляції (*S*) відносно всієї підрозділеної популяції (*T*).

Ці показники мають назву *індексів фіксації*, або *F-статистик*, і пов'язані між собою наступною залежністю:

$$F_{IS} = \frac{F_{IT} - F_{ST}}{1 - F_{ST}}, \quad (8.6)$$

або

$$1 - F_{IT} = (1 - F_{IS}) \cdot (1 - F_{ST}). \quad (8.7)$$

Розрахувати індекси фіксації можна за наступними формулами:

$$F_{IS} = 1 - \frac{\overline{H}_I}{\overline{H}_S}, \quad (8.8)$$

$$F_{IT} = 1 - \frac{\overline{H}_I}{H_T}, \quad (8.9)$$

$$F_{ST} = 1 - \frac{\overline{H}_S}{H_T}, \quad (8.10)$$

де $\overline{H}_I$, $\overline{H}_S$, H_T – середнє генне різноманіття особин у субпопуляціях, середнє генне різноманіття субпопуляцій та загальне генне різноманіття для популяції в цілому.

Продемонструвати розрахунок індексів фіксації можна на наступному прикладі.

Приклад. Популяція складається із трьох субпопуляцій, чисельності яких, а також співвідношення частот генотипів за аутосомним діалельним локусом наведено в таблиці 8.1. Необхідно визначити індекси фіксації для цієї популяції.

Таблиця 8.1 – Розміри субпопуляцій та частоти генотипів у них

Субпопуляція	N	Частота генотипу			Частота алеля		h_{teo}
		AA	*Aa*	*aa*	p_A	q_a	$(2 \cdot p_A \cdot q_a)$
№1	500	125	250	125	0,50	0,50	0,500
№2	100	50	30	20	0,65	0,35	0,455
№3	1000	100	500	400	0,35	0,65	0,455
У цілому	1600	275	780	545	0,42	0,58	

По-перше, необхідно визначити частоти алелів як для кожної із субпопуляцій, так і для популяції в цілому.

Для цього ми використовуємо формули 3.1 та 3.2. Отримані значення також заносимо в таблицю 8.1.

В останньому стовпчику таблиці 8.1 розрахуємо очікувану гетерозиготність.

Далі, розрахуємо показники генного різноманіття:

$$\overline{H}_I = \frac{\sum_{i=1}^{s}(N_i \cdot h_i)}{\sum_{i=1}^{s} N_i} = \frac{500 \cdot 0{,}5 + 100 \cdot 0{,}3 + 1000 \cdot 0{,}5}{500 + 100 + 1000} = 0{,}4875, \quad (8.11)$$

$$\overline{H}_S = \frac{\sum_{i=1}^{s}(N_i \cdot h_i^{teo})}{\sum_{i=1}^{s} N_i} = \frac{500 \cdot 0{,}5 + 100 \cdot 0{,}455 + 1000 \cdot 0{,}455}{500 + 100 + 1000} = 0{,}4691, \quad (8.12)$$

$$H_T = 2 \cdot \overline{p}_A \cdot \overline{q}_a = 2 \cdot 0{,}416 \cdot 0{,}584 = 0{,}4859. \quad (8.13)$$

Нарешті, використавши формули 8.8-8.10, розрахуємо індекси фіксації для цієї популяції:

$$F_{IS} = 1 - \frac{0{,}4875}{0{,}4691} = -0{,}0393,$$

$$F_{IT} = 1 - \frac{0{,}4875}{0{,}4859} = -0{,}0036,$$

$$F_{ST} = 1 - \frac{0{,}4691}{0{,}4859} = 0{,}0344.$$

При розрахунках індексів фіксації необхідно пам’ятати, що перші два з них (F_{IS} та F_{IT}) можуть набувати будь-яких значень у межах від –1 до +1, а останній (F_{ST}) – у межах від 0 до +1.

Контрольні питання:

1. Сутність явища Валунда
2. Сутність та методика розрахунку індексів фіксації, або F-статистик.

ЧАСТИНА ІІ
АНАЛІЗ ЯКІСНИХ ДАНИХ

§ 9. Фенетика популяцій. Оцінка частот фенів та рівня фенетичного розмаїття

Фенами (за О. В. Яблоковим) називаються *будь-які дискретні альтернативні варіації ознак і властивостей особин, що на всьому наявному матеріалі (обов'язково чисельному) далі не можуть бути поділені без втрати якостей*.

Фени завжди відображають генетичну конституцію даної особини, а своєю частотою – генетичну структуру популяції й інших груп особин даного виду.

Термін «фен» вперше було введено данським ученим В. Іоганнсеном у 1909 році, поряд із термінами «ген», «генотип», «алель».

Фенетика популяцій – це *поширення генетичних підходів і принципів на види і форми, власне генетичне вивчення яких ускладнене чи неможливе*.

Предмет фенетики – внутрішньовидова мінливість, доведена в остаточному підсумку до розгляду дискретних, альтернативних ознак-маркерів генотипового складу популяції – фенів.

Методи фенетики полягають у вичленовуванні різних фенів, характерних для мінливості досліджуваних форм, кількісне і якісне вивчення фенів.

Теоретичною основою фенетичного підходу є правило гомологічних рядів у спадковій мінливості, сформульоване М. І. Вавиловим у 1920 р. Він показав, що генетично і філогенетично близькі форми характеризуються подібними рядами спадкової мінливості.

Фенофонд – *сума фенів будь-якої сукупності особин.* Якщо кількість алелів – обмежена величина, що залежить від кількості генів, то кількість ознак практично нескінченна, її завжди можна збільшити – це залежить від бажання і наполегливості дослідника.

Феногеографія – *вивчення географічного розподілу окремих ознак (як правило, фенів та їхніх комплексів) у межах ареалу виду, проведене для дослідження проблем мікроеволюції та селекції*.

Велике значення для появи і становлення понять «фенофонд» і «феногеографія» мали проведені в 1927-1929 рр. дослідження О. С. Серебровського щодо частоти різних спадкових ознак у домашніх тварин (великої та дрібної рогатої худоби і курей) на великих територіях Радянського Союзу.

Використання методів фенетики і феногеографії дає можливість вирішувати цілий комплекс популяційно-генетичних завдань, як для природних, так і для штучних популяцій (табл. 9.1).

Вивчення і дослідження фенів у тваринництві розкриває основні особливості породного генофонду, алельні і неалельні взаємодії у тварин, а

також їх породний генотиповий склад. Фени дають можливість селекціонерам стежити за їх проявом (пенентрантністю) у породах і сімейних лініях. Наприклад, фени форми рогів, дерматогліфів, типів морфології, поведінки і т.ін. відрізняються коливаннями частоти не лише у близьких і далеких родичів, але, навіть, і в різних порід.

Таблиця 9.1 – Можливі сфери застосування фенетичного підходу

Завдання	Метод рішення
1. Виявлення меж між популяціями і їхніми групами	- за стійким, різким перепадом частот фенів; - на підставі збігу перепадів частот багатьох фенів
2. Установлення рівнів подібності й ієрархії популяційних угрупувань	- по відносній потужності фенетичних границь між багатьма популяційними вибірками досліджуваного виду
3. Виділення дрібних внутрішньопопуляційних структур	- на підставі виділення виявлених територіальних груп індивідуумів з унікальними фенами чи сполученнями фенів
4. Виявлення закономірностей географічної мінливості	- на підставі аналізу ізоліній і градієнтів частот фенів
5. Виявлення дії природного відбору	- на підставі збігу градієнта факторів середовища із градієнтом частот окремих фенів; - фенетичний аналіз особин, що вижили, при дії окремих факторів
6. Виявлення дії ізоляції	- на підставі збігу різких перепадів частот фенів із популяційними бар'єрами різного рангу (у межах однорідного біотопу)
7. Виявлення дії мутаційного процесу	- фенетичний аналіз популяції, що мешкає на території з підвищеним рівнем прояву мутагенних факторів
8. Виявлення тиску хвиль чисельності	- співставлення фенетичних характеристик однієї популяції на різних стадіях хвиль чисельності

Феноаналіз штучних популяцій дозволяє глибше охарактеризувати генотипи індивідуумів у стадах, визначити їх гетерогенність на різних рівнях і проаналізувати мікрофілогенез породи чи породної групи.

У породній феноселекції можна виділити дев'ять основних груп фенів:

1. Фени форми тіла (чи окремих його частин).
2. Фени забарвлення покриву (чи окремих його частин).
3. Фени будови волосяного покриву.
4. Морфологічні фени (типи дійок, копит, рогів, зубів, краніологічні фени).

5. Фени травної системи.
6. Фени дихальної системи.
7. Фени статевої системи.
8. Фени нервової системи.
9. Серологічні фени.

Усього в даний час нараховується близько 4000 фенів.

Використання в селекції фенетичного поліморфізму засноване на виявленні, в першу чергу, сімейних фенів, а також на підставі частот фенів – визначення «кращих» і «гірших» за продуктивністю тварин.

9.1 Оцінка частот фенів та побудова її довірчого інтервалу

Якщо у вибірці обсягом n відзначено n_A особин, що мають фен A, то частота його в даній вибірці складає:

$$p_A = \frac{n_A}{n}. \tag{9.1}$$

Статистична помилка цієї величини:

$$SE_{p_A} = \sqrt{\frac{p_A(1-p_A)}{n}}. \tag{9.2}$$

Якщо обсяг вибірки невеликий, то величина помилки може зрівнятися з оцінкою частоти. Практично можна враховувати оцінки лише в тих випадках, якщо SE_{PA} втричі менша, ніж p_A чи $1 - p_A$ (залежно від величини p_A – менша вона 0,5 чи більша).

Якщо у вибірці зустрічається m фенів (варіацій) однієї ознаки, то оцінки їхніх частот і помилок обчислюються за аналогічними формулами. Ці формули придатні, якщо частота будь-якого фена перебуває у межах від 0,2 до 0,8.

Величина помилки SE_{PA} часто використовується для розрахунку 95% довірчого інтервалу значення частоти: $p \pm 1{,}96 \times SE_{PA}$.

У випадках, якщо оцінки частот дуже великі або дуже малі (менше 0,2 чи більше 0,8), для більш точної оцінки частоти фена та її статистичної помилки необхідно користуватися φ-перетворенням Р. Фішера:

$$\varphi_A = 2 \cdot \arcsin\sqrt{\frac{n_A}{n}}, \tag{9.3}$$

$$SE\varphi_A = \frac{1}{\sqrt{n}}. \tag{9.4}$$

Приклад. У вибірці обсягом 78 особин виявлено 13 тварин, які мають фен *A*. Визначити частоту цього фена та його статистичну помилку.

Частота фена *A* в даній вибірці становить:

$$p_A = 13{:}78 = 0{,}167,$$

а її помилка:

$$SE_{p_A} = \sqrt{\frac{0{,}167 \cdot (1 - 0{,}167)}{78}} = 0{,}042 .$$

Довірчий інтервал для даної оцінки становить від: 0,167 – 1,96×0,042 до 0,167 + 1,96×0,042, тобто, від 0,085 до 0,249.

Однак, оскільки оцінка частоти фена *A* близька до 0, більш коректно використовувати φ-перетворення.

Тоді

$$\varphi_A = 2 \cdot \arcsin \sqrt{\frac{13}{78}} = 0{,}840 ;$$

$$SE\varphi_A = \sqrt{\frac{1}{78}} = 0{,}113 .$$

Межі 95% довірчого інтервалу будуть наступні:

$$\varphi_{A'} = 0{,}840 - 1{,}96 \times 0{,}113 = 0{,}619;$$

$$\varphi_{A''} = 0{,}840 + 1{,}96 \times 0{,}113 = 1{,}061.$$

Для того щоб повернутися від φ-перетворення до оцінок частот, використовується формула:

$$p = \sin^2 \frac{\varphi}{2} . \qquad (9.5)$$

Таким чином, для даного прикладу 95% довірчий інтервал частоти фена *A* буде становити $0{,}093 \le p_A \le 0{,}256$.

Для зручності в додатку Е наведено значення φ-перетворення Р. Фішера.

Значення φ для вибіркових оцінок частот, які відсутні в цій таблиці, можна розрахувати за допомогою лінійної інтерполяції за формулою:

$$\varphi = \varphi_н + u \cdot (\varphi_в - \varphi_н), \qquad (9.6)$$

де $u = \dfrac{p - p_н}{p_в - p_н}$;

$p_н$ та $p_в$ – найближчі до вибіркової оцінки значення частот, для яких у додатку Е наведено відповідні значення арксинус-перетворення Р. Фішера – $\varphi_н$ та $\varphi_в$, відповідно.

Наприклад, для значення p_A = 0,167 із прикладу в додатку Е немає точного значення. Найближчі до нього $p_н$ = 0,16 ($\varphi_н$ = 0,823) та $p_в$ = 0,17 ($\varphi_в$ = 0,850). Тоді, використовуємо формулу 9.6 і отримуємо значення:

$$\varphi = 0{,}823 + \left(\frac{0{,}167 - 0{,}16}{0{,}17 - 0{,}16} \right) \cdot (0{,}850 - 0{,}823) = 0{,}842 ,$$

яке відрізняється від точного лише третім знаком після коми.

Для зворотного перетворення використовується формула:

$$p = p_н + v \cdot (p_в - p_н), \qquad (9.7)$$

де $v = \dfrac{\varphi - \varphi_н}{\varphi_в - \varphi_н}$.

Наприклад, для значення $\varphi_{A'} = 0{,}619$ найближчі табличні складають $\varphi_н = 0{,}609$ (для нього $p_н = 0{,}09$) та $\varphi_в = 0{,}644$ (для нього $p_в = 0{,}10$). Тоді шукане значення нижньої межі 95% довірчого інтервалу становитиме:

$$p' = 0{,}09 + \left(\frac{0{,}619 - 0{,}609}{0{,}644 - 0{,}609}\right) \cdot (0{,}10 - 0{,}09) = 0{,}093,$$

яке не відрізняється від розрахованого за формулою 9.5.

Розрахована за формулою 9.7 верхня межа 95% довірчого інтервалу тоді складатиме 0,254, а точне значення – 0,256.

9.2 Оцінки фенетичного розмаїття

Мірою внутрішньопопуляційного фенетичного розмаїття стосовно якісних ознак, які мають *m* різних альтернативних варіацій (тобто фенів), є **середня кількість морф** (μ):

$$\mu = (\sum_{i=1}^{m} \sqrt{p_i})^2, \tag{9.8}$$

тобто квадрат суми квадратних коренів вибіркових оцінок частот фенів. Він показує, скільки у вибірці різних фенотипів (з урахуванням їх частот).

При рівності частот усіх фенів, що зустрічаються у вибірці, цей показник прагне до *m*; при нерівномірному розподілі частот – $\mu < m$; якщо популяція є *мономорфною* (тобто, зустрічається лише один фен) – $\mu = 1$.

Помилка середньої кількості морф розраховується за формулою:

$$SE_\mu = \sqrt{\frac{\mu(m - \mu)}{n}}. \tag{9.9}$$

На основі середньої кількості фенів обчислюється й інший показник – **частка рідкісних морф** (h_μ) та її статистична помилка:

$$h_\mu = 1 - \frac{\mu}{m}, \tag{9.10}$$

$$SE_{h_\mu} = \sqrt{\frac{h_\mu(1 - h_\mu)}{n}}. \tag{9.11}$$

Якщо розподіл частот морф у вибірці має рівномірний характер, то $h_\mu \approx 0$; при нерівномірності розподілу частот – $h_\mu > 0$.

Таким чином, μ оцінює ступінь розмаїття, а показник h_μ дає визначену характеристику цього розмаїття в розумінні співвідношення між частотами найменш численних та найбільш численних у цій вибірці (популяції) фенів.

Приклад. В стаді великої рогатої худоби симентальської породи племінного заводу «Дубов'язівський» виявлено 10 фенів типу лоба первісток (за 1976-1990 рр.). Чисельність тварин з різним феном лоба наведено в таблиці 9.2.

Таблиця 9.2 – Чисельність тварин з різним феном лоба

Фен типу лоба (i)	Кількість особин (n_i)	Частота фена (p_i)	$\sqrt{p_i}$
Широкий	852	0,127	0,356
Вузький	660	0,098	0,313
Плоский	880	0,131	0,362
Опуклий	758	0,113	0,336
Чотирикутний	410	0,061	0,247
Гостровершинний	700	0,104	0,323
Подовжений	1022	0,152	0,390
Високий	683	0,102	0,319
Середній	608	0,090	0,301
Низький	150	0,022	0,149
Сума	6723	1,000	3,095

Необхідно оцінити рівень внутрішньопопуляційного розмаїття корів за даною якісною ознакою.

Як видно із таблиці 9.2, чисельності і, відповідно, частоти особин з тим чи іншим феном значно коливаються. Найбільш розповсюдженим у даному племзаводі є «подовжений» фен лоба корів симентальської породи.

Для того щоб оцінити показники розмаїття (середню кількість морф і частку рідкісних морф), необхідно попередньо розрахувати частоту кожного фена (третій стовпчик таблиці) і корені квадратні з цих частот (четвертий стовпчик). Оцінка середньої кількості морф тоді розраховується як квадрат суми цих величин:

$$\mu = (0{,}356 + 0{,}313 + ... + 0{,}149)^2 = 3{,}095^2 = 9{,}579.$$

Помилка цього показника становить:

$$SE_\mu = \sqrt{\frac{9{,}58 \cdot (10 - 9{,}58)}{6723}} = 0{,}024.$$

Як бачимо, розраховане значення середньої кількості морф дуже близьке до кількості виділених фенів. Отже, популяція стосовно даної ознаки є високополіморфною.

Про це також свідчить і низьке значення частки рідкісних морф:

$$h_\mu = 1 - \frac{9{,}58}{10} = 0{,}042,$$

$$SEh_\mu = \sqrt{\frac{0{,}042 \cdot (1 - 0{,}042)}{6723}} = 0{,}002.$$

Показники фенетичної структури (мінливості та розмаїття) можуть бути також отримані з використанням методів рандомізації, або ресамплінгу.

Ресамплінг (від англ. resampling) – група статистичних методів, в основу яких покладено принцип формування великої кількості нових вибірок, що містять елементи вибірки вихідних даних, але розташованих кожен раз заново

у новому випадковому порядку – це так звані *псевдовибірки*. А подальший процес аналізу псевдовибірок залежить від гіпотези, що перевіряється.

Продемонструємо, як можна використати метод ресамплінгу для оцінки частоти фена та його статистичної помилки.

У нашому прикладі, що наведено вище, необхідно було розрахувати оцінку частоти фена *A*, його статистичної помилки та 95% довірчий інтервал, якщо у вибірці із 78 особин 13 мали фен *A*.

Використовуючи стандартні методи та підходи, які, до речі, базуються на екстраполяції біноміального розподілу нормальним, ми отримали: p_A = 13:78 = 0,167; $SE_{p_A} = \sqrt{\frac{0{,}167 \cdot (1-0{,}167)}{78}} = 0{,}042$; $0{,}093 \le p_A \le 0{,}256$.

Уявимо тепер наступний експеримент. Із вихідної вибірки, що містить 78 особин, почнемо обирати по одній особині, відмічати наявність або відсутність у неї фена *A*, та заново повертати її у вибірку. Зробимо це 78 разів та отримаємо першу псевдовибірку. Ймовірність того, що вона буде мати рівно 13 особин із феном *A* мала. Вона може мати і 14, і 12, і, навіть 9 чи 16 особин із даним феном. Зрештою, кількість особин із феном *A* у вибірці, обсягом 78 особин, може бути від 0 до 78.

Отримавши таким чином першу псевдовибірку, розрахуємо для неї значення p_A. Надалі, аналогічно, отримаємо другу псевдовибірку й розрахуємо для неї значення p_A. В кінцевому підсумку, сформуємо велику кількість таких псевдовибірок (наприклад, 1000 або навіть 10000) і для кожної розрахуємо відповідне значення p_A. В чисельному ресамплінгу така методика має назву *бутстрепа* (від англ. bootstrap).

Тоді, оцінку частоти фена *A* за допомогою бутстрепа можна отримати на підставі формули:

$$p_A^{boot} = \frac{\sum_{i=1}^{n} p_{Ai}^{boot}}{n}, \tag{9.12}$$

а її статистичну помилку:

$$SEp_A^{boot} = \sqrt{\frac{\sum_{i=1}^{n} (p_{Ai}^{boot} - p_A^{boot})^2}{n-1}}, \tag{9.13}$$

де p_{Ai}^{boot} – значення частоти фена *A* у *i*-тій псевдовибірці;

n – кількість сформованих псевдовибірок.

Для розрахунку нижньої та верхньої меж 95% довірчого інтервалу для бутстреп-оцінки (9.12) частоти фена необхідно на підставі ряду значень p_{Ai}^{boot} розрахувати 2,5%- та 97,5%-перцентилі, відповідно.

Наприклад, використавши генератор випадкових цифр на підставі біноміального розподілу, що є вбудованим у MS Excel, ми сформували 200 псевдовибірок для даних із нашого прикладу й отримали наступні бутстреп-

оцінки. Частота особин, що має фен *A* у даній вибірці становить 0,166 із статистичною помилкою 0,043. А 95% довірчий інтервал, оцінений за допомогою бутстреп-процедури для цих даних, складає: [0,090; 0,244].

Як бачимо, отримані на підставі бутстреп-процедури оцінки частоти, її помилки та 95% довірчого інтервалу дуже близькі до тих, що були отримані з використанням стандартних параметричних процедур.

Аналогічним чином можна отримати й відповідні бутстреп-оцінки для показників фенетичного розмаїття, що оцінюються за формулами 9.8-9.11. Для цього також необхідно, обравши випадковим чином із вихідної вибірки особину, відмітивши їх фен та повернувши у вихідну вибірку знову, сформувати нові псевдовибірки. І для кожної з них розрахувати середню кількість морф і частку рідкісних морф та їх відповідні статистичні помилки за формулами, аналогічними формулам 9.12 та 9.13, й 95% довірчі інтервали, як це було описано вище.

Наприклад, у вибірці обсягом 42 особини, фен *A* мали 12 особин, фен *B* – 5, фен *C* – 9, фен *D* – 11 та фен *E* – 5 особин. Показники фенетичного розмаїття, розраховані за формулами 9.8-9.11 складають: $\mu = 4{,}837 \pm 0{,}137$, $h_\mu = 0{,}033 \pm 0{,}027$.

Сформуємо 250 псевдовибірок знову використавши генератор випадкових цифр у табличному редакторі MS Excel та отримаємо, наприклад, наступні значення (табл. 9.3).

Таблиця 9.3 – Значення частот фенів у 250 псевдовибірках

Псевдовибірка	Частота фена					Показник	
	A	B	C	D	E	μ	h_μ
1	10	4	14	11	3	4,635	0,073
2	7	7	9	17	2	4,553	0,089
3	9	2	9	16	6	4,576	0,085
4	8	6	11	12	5	4,865	0,027
5	12	5	7	9	9	4,900	0,020
…	…	…	…	…	…	…	…
250	7	5	14	9	7	4,848	0,030

Тоді, відповідні бутстреп-оцінки для середньої кількості морф становитимуть: $\mu = 4{,}705 \pm 0{,}169$ (із 95% довірчим інтервалом [4,289; 4,938]), а для частки рідкісних морф: $h_\mu = 0{,}059 \pm 0{,}034$ (із 95% довірчим інтервалом [0,010; 0,143]).

Контрольні питання:

1. Дати визначення понять «фен», «фенофонд», «фенетика популяцій». Групи фенів.
2. Які завдання дозволяє виконувати вивчення і дослідження фенів?
3. За якими показниками проводять оцінку фенетичного розмаїття? Методика їх розрахунку.
4. Поняття про ресамплінг.

§ 10. Порівняння двох вибірок за частотами фенів. Асоціація фенів

Якщо оцінки частот фена, що аналізується знаходяться в межах від 0,2 до 0,8, а обсяги вибірок досить великі (кілька сотень особин), то для їхнього порівняння використовується стандартний *t*-критерій Ст'юдента, де вибіркові середні арифметичні замінюються оцінками частот ознаки в двох вибірках, а помилки середніх арифметичних – помилками оцінок частот:

$$t = \frac{|p_1 - p_2|}{\sqrt{SEp_1{}^2 + SEp_2{}^2}}. \qquad (10.1)$$

Розраховане значення критерію Ст'юдента порівнюється з табличним для числа ступенів свободи:

$$df = n_1 + n_2 - 2,$$

де n_1 і n_2 – обсяги першої і другої вибірок, відповідно.

Якщо $t > t_{крит}$, то частота ознаки у вибірці 1 вірогідно відрізняється від частоти цієї ж ознаки у вибірці 2. Табличні значення критерію Ст'юдента наведено в додатку Ж.

Якщо оцінки частот близькі до 1 чи 0, то використовується φ-перетворення (формула 9,3) і тоді оцінкою вірогідності розбіжностей між двома вибірками слугує величина *u*-критерію:

$$u = \frac{|\varphi_1 - \varphi_2|}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, \qquad (10.2)$$

де φ_1 і φ_2 – φ-перетворені значення частот ознаки, що аналізується у вибірках обсягом n_1 і n_2. Якщо $u > 1{,}96$, то вибірки вірогідно відрізняються за частотою даної ознаки.

Приклад. У першій вибірці, що містить 25 особин великої рогатої худоби виявлено п'ять тварин, що мають конічні дійки, а в другій вибірці, обсягом 15 особин – лише дві. Чи вірогідно відрізняються ці вибірки за частотою корів з конічними дійками?

Оскільки обсяги вибірок малі, використовуємо формулу 10.2. Частота корів з конічним типом дійок у першій вибірці становить 5 : 25 = 0,20, відповідно $\varphi_1 = 0{,}927$; частота тварин з аналогічною ознакою у другій вибірці – 2 : 15 = 0,133, і, відповідно $\varphi_2 = 0{,}747$.

Тоді значення u-критерію становить: $u = \frac{0{,}927 - 0{,}747}{\sqrt{\frac{1}{25} + \frac{1}{15}}} = 0{,}55$, що набагато нижче критичного.

Таким чином, нуль-гіпотеза щодо рівної частоти особин з даним типом дійок у двох вибірках підтверджується.

Асоціація ознак

Для статистичного підтвердження наявності зв'язку між двома якісними ознаками використовується критерій Хі-квадрат К. Пірсона:

$$\chi^2 = \sum \frac{(Ф - T)^2}{T}. \qquad (10.3)$$

Оцінка вірогідності відхилення фактичних частот сполучень ознак від теоретичних проводиться на підставі рівня значимості критерію Хі-квадрат при відповідному числі ступенів свободи $df = (v - 1) \times (c - 1)$, де v – кількість стовпців таблиці сполученості, а c – кількість її рядків.

Критичні значення критерію Хі-квадрат для різного числа ступенів свободи наведено в додатку Д.

Приклад. При аналізі прояву асоціації фенів крижів та постановки хвоста у 3300 корів симентальської породи, що належать племзаводу «Терезино» (1975-1985 рр.) виявлено наступні абсолютні частоти різних пар сполучень цих двох ознак (табл. 10.1).

Таблиця 10.1 – Частоти різних пар сполучень фенів крижів та постановки хвоста

Фен постановки хвоста	Фен крижів		Сума
	прямий	скошений	
Високопоставлений	700 / *474,0*	1140 / *1366,0*	1840
Низькопоставлений	150 / *376, 0*	1310 / *1084,0*	1460
Сума	850	2450	3300

Необхідно з'ясувати, чи має місце зв'язок між цими двома ознаками, тобто чи незалежно зустрічаються різні фени крижів і постановки хвоста у тварин даного господарства? Якщо ні, то які сполучення більш імовірні?

Для цього необхідно розрахувати теоретично очікувані абсолютні частоти всіх чотирьох можливих сполучень двох ознак і порівняти їх з фактичними з використанням критерію Хі-квадрат.

Для розрахунку теоретичних частот можна скористатися простим правилом: теоретична абсолютна частота для будь-якої клітинки таблиці дорівнює добутку суми частот по даному стовпцю і суми частот по даному рядку, поділеному на загальний обсяг вибірки. Причому це правило справедливе для будь-яких таблиць сполученості, незалежно від кількості рядків та стовпців.

Таким чином, розраховуємо теоретичні частоти для всіх чотирьох пар сполучень:

для сполучення «прямі крижі – високопоставлений хвіст»:

$$n_{ПВ} = \frac{850 \cdot 1840}{3300} = 474{,}0 \ ,$$

для сполучення «скошені крижі – високопоставлений хвіст»:

$$n_{ПН} = \frac{2450 \times 1840}{3300} = 1366{,}0 \text{ , і т. д.}$$

Усі розраховані теоретичні частоти приведені в таблиці 10.1 курсивом.

Як бачимо, навіть візуально видно, що фактичні і теоретичні частоти відрізняються одна від одної. Але наскільки вірогідна ця відмінність з'ясуємо за допомогою критерію Хі-квадрат:

$$\chi^2 = \frac{(700 - 474{,}0)^2}{474{,}0} + \ldots + \frac{(1310 - 1084{,}0)^2}{1084{,}0} = 328{,}1 \text{ .}$$

Число ступенів свободи в даному випадку дорівнює: *df* = (2-1)·(2-1) = 1. Як бачимо, розраховане значення критерію Хі-квадрат набагато перевищує табличне (3,84), що свідчить про те, що гіпотеза про незалежне зустрічання цієї пари ознак у даній групі тварин повинна бути відкинута. Тому приймається альтернативна їй гіпотеза про те, що має місце вірогідний зв'язок між появою у корови того чи іншого фенотипу крижів з різновидом постави хвоста.

Якщо проаналізувати дані таблиці 10.1 і порівняти фактичні частоти з теоретичними, то можна відзначити, що сполучення «прямі крижі – високопоставлений хвіст» і «скошені крижі – низькопоставлений хвіст» зустрічаються частіше, ніж очікувалося.

Після того, як доведено, що асоціація між парою ознак має місце, можна вирішувати друге завдання – оцінку сили цього зв'язку. Напрямок зв'язку може бути враховано лише у випадку таблиці 2×2 (тобто, із двома стовпцями та двома строчками). Для багатопільних таблиць зв'язок завжди має додатній знак.

Методи оцінки сили зв'язку для чотирипільної таблиці базуються на т.зв. «перехресному відношенні» (cross ratio) чи «***відношенні переваги***» (odd ratio).

Якщо ми маємо таблицю розмірності 2×2, де всі її клітинки позначено в такий спосіб:

a	***b***
c	***d***

то відношення переваги має вигляд: $\frac{ad}{bc}$.

Відомо, що у випадку незалежності ознак, це відношення дорівнює 1. На основі цієї закономірності було розроблено цілий ряд коефіцієнтів для оцінки сили зв'язку.

Найбільш простими і широко розповсюдженими є:

коефіцієнт асоціації Юла (Q_r):

$$Q_r = \frac{ad - bc}{ad + bc}, \tag{10.4}$$

$$SE_{Qr} = \frac{1 - Q_r^2}{2} \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}, \tag{10.5}$$

коефіцієнт контингенції Шарл'є (r_β):

$$r_\beta = \frac{|ad - bc| - \frac{n}{2}}{\sqrt{(a+b)(b+d)(a+c)(b+c)}} = \sqrt{\frac{\chi^2}{n}}, \tag{10.6}$$

$$SEr_\beta = \frac{1 - r_\beta^2}{\sqrt{n}}. \tag{10.7}$$

Коефіцієнти Юла та Шарл'є варіюють від –1 до +1. Чим ближче величина показника до 1, тим сильніше виражений зв'язок. Знак перед оцінкою коефіцієнта показує напрямок цього зв'язку – якщо відношення переваги більше одиниці, то зв'язок позитивний, якщо ж менше – негативний.

Необхідно відзначити, що коефіцієнт Юла завжди дає більш високе значення, ніж коефіцієнт Шарл'є (на тих самих даних).

Оцінки статистичної помилки коефіцієнтів сили зв'язку дозволяють оцінити вірогідність відхилення величини коефіцієнта від нуля і, у випадку потреби, порівняти два коефіцієнти, що розраховані для різних вибірок.

Приклад. Оцінимо силу зв'язку між ознаками, розглянутими в попередньому прикладі.

Використовуючи формули 10.4 та 10.5, розрахуємо значення коефіцієнта асоціації Юла та його статистичної помилки:

$$Q_r = \frac{700 \cdot 1310 - 150 \cdot 1140}{700 \cdot 1310 + 150 \cdot 1140} = +0{,}686,$$

$$SE_{Qr} = \frac{1 - 0{,}686^2}{2}\sqrt{\frac{1}{700} + \frac{1}{1140} + \frac{1}{150} + \frac{1}{1310}} = 0{,}026.$$

Аналогічно, за формулами 10.6 та 10.7 розрахуємо оцінки коефіцієнта контингенії Шарл'є та його статистичної помилки:

$$r_\beta = \frac{|700 \cdot 1310 - 150 \cdot 1140| - \frac{3300}{2}}{\sqrt{850 \cdot 2450 \cdot 1840 \cdot 1460}} = +0{,}315,$$

$$SEr_\beta = \frac{1 - 0{,}315^2}{\sqrt{3300}} = 0{,}016.$$

Як бачимо, в обох випадках зв'язок виявляється високовірогідним і позитивним.

У тих випадках, коли таблиця сполученості має більше рядків і стовпців, методи оцінки сили зв'язку між ознаками базуються безпосередньо на величині розрахованого критерію Хі-квадрат.

Сила такого зв'язку може бути оцінена за допомогою **міри Чупрова** (T) чи **міри Крамера** (V). Формули для обчислення цих показників та їхніх помилок наступні:

$$T = \sqrt{\frac{\chi^2}{n\sqrt{df}}}, \tag{10.8}$$

$$SE_T = \sqrt{\frac{1}{n\sqrt{df}}}; \tag{10.9}$$

$$V = \sqrt{\frac{\chi^2}{n \cdot \min\{v-1; c-1\}}}; \tag{10.10}$$

$$SE_V = \sqrt{\frac{1}{n \cdot \min\{v-1; c-1\}}}. \tag{10.11}$$

Вираз «min {v – 1; c – 1}» являє собою мінімальне значення з кількості стовпців таблиці сполученості мінус одиниця і кількості рядків таблиці мінус одиниця. Обидва показники варіюють від 0 до 1.

Приклад. В стаді корів симентальської породи було проаналізовано зв'язок між фенами вух та тривалістю тільності (виділено три градації – максимальна, середня і мінімальна). Всі вихідні дані наведено в таблиці 10.2.

Таблиця 10.2 – Фени вух та тривалість тільності корів

Тривалість тільності	Фен вух					Сума
	загострені	тупі	вузькі	широкі	м'ясисті	
Максимальна	44 / *64,6*	67 / *48,7*	45 / *57,0*	45 / *37,8*	54 / *46,8*	255
Середня	42 / *44,8*	31 / *33,8*	33 / *39,6*	35 / *26,2*	36 / *32,5*	177
Мінімальна	85 / *61,6*	31 / *46,5*	73 / *54,5*	20 / *36,0*	34 / *44,7*	243
Сума	171	129	151	100	124	675

Необхідно з'ясувати – чи має місце зв'язок між тривалістю тільності й особливостями будови вух тварин дослідної групи.

Перше за все, необхідно розрахувати теоретичні частоти для кожної пари сполучень і порівняти ці розраховані частоти з фактичними, використовуючи критерій Хі-квадрат. Розрахунок проводиться стандартним методом (див. вище). Теоретичні частоти наведено в таблиці 10.2 курсивом.

Далі, використовуючи формулу 10.3, визначаємо чи має місце вірогідний зв'язок між ознаками:

$$\chi^2 = \frac{(44-64{,}6)^2}{64{,}6} + \ldots + \frac{(34-44{,}7)^2}{44{,}7} = 53{,}38 \ .$$

У даному випадку число ступенів свободи дорівнює df = (5– 1)·(3 –1) = 8. Табличне значення критерію Хі-квадрат для даного числа ступенів свободи набагато нижче – 15,51 (додаток Д). Отже, можна вважати доведеним, що має місце вірогідний зв'язок між формою вух корів симентальської породи даного стада та тривалістю їх тільності.

Оцінку сили цього зв'язку розрахуємо, використовуючи міри Чупрова та Крамера:

$$T = \sqrt{\frac{53{,}38}{675 \cdot \sqrt{8}}} = 0{,}167; \quad SE_T = \sqrt{\frac{1}{675 \cdot \sqrt{8}}} = 0{,}023;$$

$$V = \sqrt{\frac{53{,}38}{675 \cdot 2}} = 0{,}199; \quad SE_V = \sqrt{\frac{1}{675 \cdot 2}} = 0{,}027.$$

Як бачимо, оцінка цього зв'язку хоча і вірогідна, але все-таки невелика.

Показники, що наведені в даному параграфі, також можуть бути розраховані з використанням одного з методів ресамплінгу, а саме – методу *перестановок*, або пермутацій (від англ. – permutation). Продемонструємо використання цього методу при порівнянні двох вибірок за частотою якісної ознаки (фена).

У прикладі, що було розібрано вище, необхідно було оцінити вірогідність відмінностей між двома вибірками. У першій вибірці, що містить 25 особин великої рогатої худоби виявлено п'ять корів, що мають конічні дійки, а в другій вибірці обсягом 15 особин – лише дві. Використовуючи φ-перетворення Р. Фішера, ми показали, що значення критерію u = 0,55, і є набагато нижче критичного. Таким чином, нуль-гіпотеза щодо рівної частоти особин з даним типом дійок у двох вибірках не може бути відкинута. Рівень значущості отриманої оцінки критерію дорівнює p = 0,582.

Перевіримо тепер ту ж саму нуль-гіпотезу, використовуючи метод перестановок. Для цього уявимо наступний експеримент. Розташуємо всі наші вихідні дані в один ряд. Тоді він буде складатися із нулів (для корів, що не мають конічних дійок) та одиниць (для корів, що мають конічні дійки). Всього такий ряд буде складатися із 40 цифр (перші 25 – для першої вибірки, та останні 15 – для другої), серед яких буде 7 одиниць (п'ять особин із конічними дійками із першої вибірки та дві – із другої) і, відповідно, 33 нуля. Далі випадковим чином перетасуємо ці 40 цифр. Розрахуємо для перших 25 цифр (наша перша вибірка) частоту особин із конічними дійками та для наступних 15 цифр (наша друга вибірка).

Для цих двох пар псевдооцінок частот розрахуємо за формулою 10.2 значення псевдооцінки критерію u. Знову перетасуємо вибірку з вихідними даними та знову розрахуємо псевдооцінку критерію u. Нарешті, проведемо таку процедуру ще багато-багато разів, наприклад M (бажано, щоб M було 1000 або, навіть, 10000).

З отриманої кількості псевдооцінок критерію u підрахуємо скільки разів ці оцінки дорівнювали, або були більші, ніж та, що ми отримали для вихідних даних (тобто, u = 0,55). Нехай такі псевдооцінки зустрілися m разів. Тоді, рівень значущості отриманої нами оцінки u-критерію, буде дорівнювати:

$$p = \frac{m}{M+1}. \qquad (10.12)$$

Сенс даного підходу в наступному. Чим вище отримана оцінка критерію (тобто, чим більші відмінності за частотами фена у двох вибірках, що

порівнюються), тим менше шансів отримати таке ж (або навіть більше) значення для випадковим чином сформованої вибірки із такими ж вихідними умовами.

Для нашого прикладу із 100 перестановок було отримано 72 псевдооцінки *u*-критерію, що дорівнювали або переважали значення 0,55. Таким чином, рівень значущості для ресамплінг-критерію, використаного для даних із нашого прикладу, становить:

$$p = 72 : (100+1) = 0{,}713.$$

Таке високе отримане значення також підтверджує, що нуль-гіпотеза не може бути відкинута.

Аналогічним чином цей підхід може бути використаний і при аналізі асоціації між ознаками.

Наприклад, в таблиці 10.3 наведено дані щодо одночасного зустрічання фенів (кожен з трьома морфами) у однієї особини.

Таблиця 10.3 – Дані щодо кількості особин із різними фенами

Морфа фена *B*	Морфа фена *A*			Сума
	*A*1	*A*2	*A*3	
*B*1	11	4	3	18
*B*2	6	12	4	22
*B*3	5	7	9	21
Сума	22	23	16	61

Тобто, фенотип *A*1*B*1 мають 11 особин, а фенотип *A*3*B*2 – 4 особини.

Якщо ми використаємо для цих даних стандартний критерій Хі-квадрат К. Пірсона, то отримаємо оцінку χ^2 = 10,767, що при чотирьох ступенях свободи (див. формулу 10.3) дасть нам підставу для відхилення нуль-гіпотези про те, що ці два фени успадковуються незалежно (із рівнем значущості p = 0,029).

Але таблиця 10.3 містить деякі значення частот, що не переважають 5 і тому висновок, отриманий на підставі використання стандартного критерію Хі-квадрат К. Пірсона, може мати деяку помилку.

Тому нам необхідно використати таку методику перевірки даної нуль-гіпотези, яка б не мала ніяких припущень щодо обсягів вихідних даних. Найбільш придатний для цих цілей також критерій перестановок. Його використання в даному випадку майже аналогічне тому, як це було описано вище. Тобто, вихідні дані формуються наступним чином. Для першої вибірки записуємо наступні коди: 11 одиниць, 6 двійок та п’ять трійок. За ними зразу ж записуємо коди для другої вибірки – чотири одиниці, дванадцять двійок та сім трійок. За ними зразу ж записуємо коди для третьої вибірки: три одиниці, чотири двійки та дев’ять трійок. Всього наша вихідна база даних містить 61 цифру. Перші 22 – це коди для морфи *A*1, наступні 23 – це коди для морфи *A*2, а останні 16 – це коди для морфи *A*3. Коди означають наступне. Одиниці – це коди для морфи *B*1, двійки – для морфи *B*2 та трійки – для морфи *B*3.

Тепер випадковим чином перетасуємо всі цифри та отримаємо першу псевдовибірку значень. Наприклад, вона буде мати наступний вигляд (табл. 10.4).

Таблиця 10.4 – Дані щодо кількості особин із різними фенами у псевдовибірці

Морфа фена *B*	Морфа фена *A*			Сума
	*A*1	*A*2	*A*3	
*B*1	7	7	4	18
*B*2	6	8	8	22
*B*3	9	8	4	21
Сума	22	23	16	61

Як бачимо, змінилися лише значення частот у клітинах таблиці 10.3, тоді як маргінальні частоти (суми по стовпчиках та строчках) залишилися без змін. Для цієї псевдовібирки розраховується псевдооцінка критерію Хі-квадрат К. Пірсона. Вона складає: $\chi^2 = 2{,}183$.

Знову перетасовуємо вихідну вибірку та отримуємо наступну псевдооцінку критерію Хі-квадрат К. Пірсона. І виконуємо цю дію багато разів. Рівень значущості для ресамплінг-критерію отримуємо, знову ж, за формулою 10.12.

Для даних із нашого прикладу після 500 перестановок даних із таблиці 10.3 було отримано 13 оцінок критерію Хі-квадрат К. Пірсона, що дорівнювали або перевищували значення 10,767.

Таким чином, рівень значущості для ресамплінг-критерію, використаного для даних із нашого прикладу, становить:

$$p = 13 : (500+1) = 0{,}026.$$

Таке низьке значення свідчить про те, що нульова гіпотеза повинна, все ж таки, бути відхилена.

Аналогічним чином ресамплінг-процедура (а саме, метод перестановок) може бути використана для оцінки рівня значущості будь-якого критерію рівня асоціації, що описані у даному параграфі.

Контрольні питання:

1. Як провести порівняння двох вибірок за частотами фенів, якщо оцінки частот фена, що аналізується знаходяться в межах від 0,2 до 0,8, а обсяги вибірок досить великі?

2. Які коефіцієнти використовуються для оцінки сили зв’язку між двома якісними ознаками?

§ 11. Дисперсійний аналіз якісних ознак. Однофакторний дисперсійний аналіз

11.1 Однофакторний дисперсійний аналіз диморфних ознак

В основі однофакторного дисперсійного аналізу якісних ознак із двома альтернативними варіантами лежить наступне правило розкладання сумарної варіанси (σ_T^2) на факторіальну (σ_X^2) і залишкову (σ_Z^2):

$$\sigma_T^2 = \sigma_X^2 + \sigma_Z^2 . \tag{11.1}$$

Отже, припустимо, що у нас є *s* груп (вибірок), чисельність кожної з яких становить n_i (де $1 \le i \le s$), кількість організмів з певною ознакою в кожній групі дорівнює m_i. Тоді сумарна варіанса (σ_T^2) такого дисперсійного комплексу буде дорівнювати:

$$\sigma_T^2 = N \cdot \overline{p} \cdot (1 - \overline{p}), \tag{11.2}$$

де N – сумарна кількість всіх об'єктів дисперсійного комплексу: $N = \sum n_i$;

$\overline{p}$ – середня зважена частота ознаки для всієї вибірки:

$$\overline{p} = \frac{\sum_{i=1}^{s} m_i}{N} . \tag{11.3}$$

Використовуючи прості перетворення, можна одержати інший вираз для розрахунку сумарної варіанси:

$$\sigma_T^2 = \sum_{i=1}^{s} m_i - \frac{\left(\sum_{i=1}^{s} m_i \right)^2}{N} . \tag{11.4}$$

Факторіальна варіанса (σ_X^2), що відображає мінливість **між** вибірками, визначається мірою нерівності частот ознаки, що аналізується в різних вибірках і може бути розрахована за формулою:

$$\sigma_X^2 = \sum_{i=1}^{s} \frac{m_i^2}{n_i} - \frac{\left(\sum_{i=1}^{s} m_i \right)^2}{N} . \tag{11.5}$$

Нарешті, залишкова дисперсія (σ_Z^2), що відображає розходження **всередині** кожної із вибірок (у цілому), розраховується за формулою:

$$\sigma_Z^2 = \sum_{i=1}^{s} n_i \cdot p_i \cdot (1 - p_i) , \tag{11.6}$$

де p_i – частота ознаки в *i*-тій вибірці.

Формулу 11.6 можна подати в іншому вигляді:

$$\sigma_Z^2 = \sum_{i=1}^{s} m_i - \sum_{i=1}^{s} \frac{m_i^2}{n_i} . \tag{11.7}$$

Вірогідність впливу фактора тоді можна оцінити за допомогою дисперсійного відношення:

$$F = \frac{\sigma_X^2 \cdot (N - s)}{\sigma_Z^2 \cdot (s - 1)}, \qquad (11.8)$$

яке має F-розподіл Фішера-Снедекора із числом ступенів свободи $df_1 = s - 1$ і $df_2 = N - s$.

Приклад. У п’яти популяціях корів червоної степової породи було визначено кількість особин, що мають фен «прямі крижі» (табл. 11.1). Необхідно перевірити гіпотезу про рівність частот особин із даним феном у всіх п’яти проаналізованих вибірках.

Таблиця 11.1 – Кількість особин, що мають фен «прямі крижі» у п’яти популяціях корів

	Популяція					Суми
	1	2	3	4	5	
n	57	77	117	113	80	$N = 444$
m	16	34	33	93	53	$\Sigma\, m = 229$
$p = m/n$	0,281	0,442	0,282	0,823	0,663	
C_Z	11,509	18,987	23,692	16,460	17,888	$\sigma_Z^2 = 88{,}536$

Спочатку необхідно розрахувати частоти даного фена як у кожній із вибірок, так і у всій сумарній вибірці в цілому. Наприклад, для вибірки 1 частота особин із даним феном становить:

$$p_1 = 16/57 = 0{,}281.$$

А для всієї вибірки в цілому:

$$\overline{p} = 229/444 = 0{,}516.$$

Тепер, маючи всі необхідні дані, можна розрахувати оцінки сумарної і залишкової варіанси на підставі формул 11.2 та 11.6:

$$\sigma_T^2 = 444 \cdot 0{,}516 \cdot (1 - 0{,}516) = 110{,}886;$$

$$\sigma_Z^2 = 57 \cdot 0{,}281 \cdot (1 - 0{,}281) + \ldots + 80 \cdot 0{,}663 \cdot (1 - 0{,}663) = 88{,}536.$$

Факторіальну варіансу простіше визначити як різницю між сумарною і залишковою:

$$\sigma_X^2 = \sigma_T^2 - \sigma_Z^2 = 110{,}886 - 88{,}536 = 22{,}350.$$

Визначимо тепер число ступенів свободи для факторіальної, залишкової і сумарної варіанс:

$$df_X = s - 1 = 5 - 1 = 4;$$
$$df_Z = N - s = 444 - 5 = 439;$$
$$df_T = N - 1 = 444 - 1 = 443.$$

Розрахуємо середні квадрати для факторіальної і залишкової варіанс:

$$MS_X = \frac{\sigma_X^2}{df_X} = \frac{22{,}350}{4} = 5{,}588;$$

$$MS_Z = \frac{\sigma_Z^2}{df_Z} = \frac{88{,}536}{439} = 0{,}202\,.$$

Значення дисперсійного відношення розрахуємо як відношення середнього квадрата для факторіальної варіанси до відповідного значення для залишкової:

$$F = \frac{5{,}588}{0{,}202} = 27{,}66\,.$$

Це розраховане значення набагато перевищує табличне значення критерію Фішера для числа ступенів свободи: $df_1 = 4$ і $df_2 = 439$ ($F_{\alpha=0{,}05} = 2{,}39$; див. додаток Ж). Отже, нульова гіпотеза про рівність частот даного фена в п'яти вивчених популяціях повинна бути відкинута.

Всі отримані результати заносимо в стандартну таблицю дисперсійного аналізу (табл. 11.2).

Таблиця 11.2 – Результати дисперсійного аналізу

Джерело мінливості	σ^2	*df*	*MS*	*F*	*p*
X	22,350	4	5,588	27,66	< 0,001
Z	88,536	439	0,202		
T	110,886	443			

Необхідно відзначити, що рівень значущості критерію Фішера-Снедекора для результатів дисперсійного аналізу якісних ознак даного типу буде відповідати рівню значимості критерію Хі-квадрат К. Пірсона, використаного для перевірки тієї ж самої нуль-гіпотези лише у випадку досить великого сумарного обсягу вибірок (порядку декількох тисяч).

У цілому, якщо $df \to \infty$, то F-критерій Фішера-Снедекора і критерій Хі-квадрат К. Пірсона поєднуються рівнянням:

$$F(df_X) = \frac{\chi^2}{df_X}\,. \qquad (11.9)$$

Таким чином, у випадку аналізу двох груп (вибірок, субпопуляцій і т. ін.), значення цих критеріїв (а, відповідно, і рівень їхньої значущості) збігається.

Оцінку сили впливу фактора (у тому випадку, звичайно, якщо доведено вірогідний його вплив на мінливість частот у різних градаціях) може бути зроблено двома способами.

Перший спосіб. При використанні цього способу оцінка сили впливу розраховується як відношення факторіальної варіанси до загальної:

$$\eta^2 = \frac{\sigma_X^2}{\sigma_T^2}\,, \qquad (11.10)$$

Даний показник якого може варіювати від 0 до 1. Нулю він може дорівнювати у випадку, коли частота особин з даною ознакою є однаковою у всіх порівнюваних групах (тобто, градаціях фактора). Досягти одиниці

показник сили впливу може тільки в особливому випадку – за наявності тільки двох градацій фактора (двох вибірок), причому всі особини однієї з вибірок мають ознаку, а всі особини з іншої – не мають її. У випадку *s* груп (*s* > 2) оцінка $\acute{\eta}^2$ з формули 11.10 ніколи не досягає одиниці.

Оцінка сили впливу має безпосереднє відношення до критерію Хі-квадрат К. Пірсона, оскільки, використовуючи формулу Брандта-Снедекора (Бейли, 1962), можна показати, що

$$\eta^2 = \frac{\chi^2}{N}, \tag{11.11}$$

де χ^2 – оцінка критерію Хі-квадрат К. Пірсона, використана для перевірки тієї ж гіпотези, тобто гіпотези про рівність частот у всіх аналізованих вибірках (тобто відсутність впливу фактора).

Для розглянутого вище прикладу ця величина буде дорівнювати, відповідно: $\eta^2 = \frac{22{,}350}{110{,}886} = 0{,}202$.

Оцінка рівня значущості цієї величини може бути зроблена з огляду на те, що величина

$$\chi^2 = \eta^2 \cdot N \tag{11.12}$$

має розподіл Хі-квадрат з числом ступенів свободи:

$$df = s - 1. \tag{11.13}$$

У нашому прикладі, розраховане значення χ^2 становить 0,202 × 444 = 89,69, що набагато більше, ніж табличне значення критерію Хі-квадрат з числом ступенів свободи *df* = 5 – 1 = 4 ($\chi^2_{\alpha=0{,}05} = 9{,}49$).

Другий спосіб. Якщо використовувати другий спосіб, то оцінка сили впливу фактора розраховується за формулою:

$$\eta^2 = \frac{S_A^2}{S_A^2 + S_Z^2}, \tag{11.14}$$

де

$$S_Z^2 = MS_Z; \tag{11.15}$$

$$S_A^2 = \frac{MS_A - MS_Z}{n^*}; \tag{11.16}$$

$$n^* = \frac{1}{s-1} \cdot \left(N - \frac{\sum_{i=1}^{s} n_i}{N} \right). \tag{11.17}$$

Формула 11.17 використовується в тому випадку, якщо обсяги вибірок нерівні. У випадку, якщо обсяги всіх вибірок рівні між собою і дорівнюють, наприклад, *n*, то *n** = *n*.

Продемонструємо використання другого способу оцінки сили впливу фактора на тих же даних із нашого прикладу.

Оскільки обсяги вибірок не рівні, спочатку використовуємо формулу 11.17 для розрахунку усередненого показника *n**:

$$n^* = \frac{1}{5-1} \cdot \left(444 - \frac{57^2 + 77^2 + ... + 80^2}{444} \right) = 87{,}3 \ .$$

Далі, використовуємо формулу 11.16 для розрахунку S_A^2:

$$S_A^2 = \frac{5{,}588 - 0{,}202}{87{,}3} = 0{,}062 \ .$$

Таким чином, оцінка сили впливу фактора для нашого прикладу буде дорівнювати:

$$\eta^2 = \frac{0{,}062}{0{,}062 + 0{,}202} = 0{,}235 \ .$$

Г. Шеффе (1963) приводить наступну методику для розрахунку меж 95% довірчого інтервалу отриманої оцінки.

Спочатку розраховуються величини:

$$L = \frac{1}{n^*} \cdot \left(\frac{F}{F^2} - 1 \right); \qquad (11.18)$$

$$R = \frac{1}{n^*} \cdot \left(\frac{F}{F_1} - 1 \right), \qquad (11.19)$$

де F – оцінка дисперсійного відношення, отримана в результаті проведення дисперсійного аналізу;

F_1 – табличне значення критерію Фішера-Снедекора для α = 0,025; $df_1 = s - 1$; $df_2 = N - s$;

F_2 – табличне значення критерію Фішера-Снедекора для α = 0,975; $df_1 = s - 1$; $df_2 = N - s$.

Тоді, верхня і нижня межі 95% довірчого інтервалу для оцінки η^2 будуть дорівнювати, відповідно:

$$\eta_U^2 = \frac{1}{1 + \frac{1}{L}}; \qquad (11.20)$$

$$\eta_L^2 = \frac{1}{1 + \frac{1}{R}} \ . \qquad (11.21)$$

Таким чином, для даних із нашого прикладу межі 95% довірчого інтервалу буде дорівнювати:

$$L = \frac{1}{87{,}3} \cdot \left(\frac{27{,}66}{0{,}121} - 1 \right) = 2{,}607 \ ; \ R = \frac{1}{87{,}3} \cdot \left(\frac{27{,}66}{2{,}815} - 1 \right) = 0{,}101 \ ;$$

$$\eta_U^2 = \frac{1}{1+\frac{1}{2{,}607}} = 0{,}723\,;\ \eta_L^2 = \frac{1}{1+\frac{1}{0{,}101}} = 0{,}092\,.$$

Отже, оцінка сили впливу фактора дорівнює 0,235 із 95% довірчим інтервалом: [0,092; 0,723].

Оскільки цей інтервал не містить 0, можна говорити про вірогідний вплив фактора (тобто, про вірогідність розходжень частот ознаки в аналізованих вибірках).

Відзначимо, що при використанні другого способу в деяких випадках може бути отримане значення оцінки сили впливу фактора, що має від'ємну величину (що, з першого погляду, не має сенсу). У деяких випадках нижня межа довірчого інтервалу також може мати від'ємний знак. Усе це свідчить про те, що фактор не має ніякого суттєвого впливу на мінливість ознаки (вірніше, його частот у різних вибірках) і що вибірки за частотою даного фену навіть більш подібні, ніж це можна було б очікувати при випадковому відборі особин при проведенні аналізу.

Використання першого способу ніколи не дає таких безглуздих значень, однак його оцінка є зміщеною; і це зміщення тим більше, чим сильніше відрізняються вибірки за чисельністю. Однак, на таке зміщення можна не зважати (особливо, при використанні вибірок одного порядку), якщо врахувати, що рівень значущості оцінки сили впливу фактора *(ή²)* визначається на підставі критерію Хі-квадрат К. Пірсона, для якого рівність обсягів вибірок при перевірці нульової гіпотези не є обов'язковою умовою застосування. Єдина вимога, у цьому випадку, щоб мінімальна частота певної ознаки в одній із вибірок була не менше 3-5 (Справочник по прикладной статистике, 1990).

У нашому прикладі оцінки сили впливу фактора, отримані першим і другим способом, виявилися досить близькими.

Крім того, як було показано Н. А. Плохинским (1964: С. 82), оцінки сили впливу фактора, отримані цими двома способами, пов'язані між собою простою залежністю.

11.2 Однофакторний дисперсійний аналіз поліморфних ознак

В основі однофакторного дисперсійного аналізу якісних ознак, що характеризуються декількома альтернативними варіаціями (морфами), також лежить правило розкладання загальної (сумарної) варіанси на дві частини – факторіальну (σ_X^2) і залишкову (σ_Z^2). Однак, у цьому випадку формули для їхнього розрахунку мають інший вигляд.

Отже, припустимо, що ми маємо справу з *s* окремими вибірками й у кожній вибірці відзначається наявність *k* морф. Тоді вихідну таблицю даних можна представити в наступному вигляді (табл. 11.3).

Таблиця 11.3 – Вихідні дані із *s* окремими вибірками коли у кожній вибірці відзначається наявність k морф

Морфи	Вибірки					Суми
	1	2	3	…	*s*	
1	m_{11}	m_{21}	m_{31}	…	m_{s1}	n_1
2	m_{12}	m_{22}	m_{32}	…	m_{s2}	n_2
3	m_{13}	m_{23}	m_{33}	…	m_{s3}	n_3
…	…	…	…	…	…	…
k	m_{1k}	m_{2k}	m_{3k}	…	m_{sk}	n_k
Суми	N_1	N_2	N_3	…	N_s	N

У цій таблиці чисельності особин у кожній вибірці являють собою суму частот окремих морф:

$$N_i = \sum_{j=1}^{k} m_{ij} \ , \tag{11.22}$$

а суми по рядках являють загальну кількість особин, що характеризуються даним феном:

$$n_j = \sum_{i=1}^{s} m_{ij} \ . \tag{11.23}$$

Загальна кількість особин являє собою суму частот по рядках і стовпцях:

$$N = \sum_{i=1}^{s} \sum_{j=1}^{k} m_{ij} \ . \tag{11.24}$$

Першим етапом аналізу є розрахунок частот всіх фенів для кожної вибірки окремо:

$$p_{ij} = \frac{m_{ij}}{N_i} \ . \tag{11.25}$$

Крім цього, розраховуються середні (узагальнено для усіх вибірок у цілому) частоти фенів:

$$\overline{p}_j = \frac{n_j}{N} \ . \tag{11.26}$$

Тоді наша таблиця з вихідними даними набуває наступного вигляду (табл. 11.4).

Таблиця 11.4 – Розраховані частоти фенів

Морфи	Вибірки					Середні частоти
	1	2	3	…	*s*	
1	p_{11}	p_{21}	p_{31}	…	p_{s1}	$\overline{p}_1$
2	p_{12}	p_{22}	p_{32}	…	p_{s2}	$\overline{p}_2$
3	p_{13}	p_{23}	p_{33}	…	p_{s3}	$\overline{p}_3$
…	…	…	…	…	…	…
k	p_{1k}	p_{2k}	p_{3k}	…	p_{sk}	$\overline{p}_k$

Тепер можна розрахувати сумарну, залишкову і факторіальну варіанси:

$$\sigma_T^2 = N \cdot \left[1 - \sum_{j=1}^{k} {}^{-2} p_j \right]; \tag{11.27}$$

$$\sigma_Z^2 = \sum_{i=1}^{s} \left[N_i \cdot \left[1 - \sum_{j=1}^{k} p_{ij}^2 \right] \right]; \tag{11.28}$$

$$\sigma_X^2 = \sum_{i=1}^{s} \left[\frac{\sum_{j=1}^{k} m_{ij}^2}{N_i} \right] - \frac{\sum_{j=1}^{k} n_j^2}{N}. \tag{11.29}$$

Для кожної з варіанс надалі розраховуються відповідні значення числа ступенів свободи:

$$df_{\mathrm{T}} = N - 1; \tag{11.30}$$

$$df_{\mathrm{Z}} = N - s; \tag{11.31}$$

$$df_{\mathrm{X}} = s - 1. \tag{11.32}$$

Середні квадрати для факторіальної та залишкової компонентів розраховуються як відношення відповідних варіанс до числа ступенів свободи:

$$MS_X = \frac{\sigma_X^2}{s-1}; \tag{11.33}$$

$$MS_Z = \frac{\sigma_Z^2}{N-s}. \tag{11.34}$$

Нарешті, дисперсійне відношення розраховується як відношення факторіального середнього квадрата до залишкового:

$$F = \frac{MS_X}{MS_Z}. \tag{11.35}$$

Оцінка значущості отриманого дисперсійного відношення проводиться на підставі порівняння розрахункового F із табличним значенням критерію Фішера-Снедекора для числа ступенів свободи: $df_1 = df$ і $df_2 = df$ (див. додаток Ж).

Приклад. У трьох вибірках корів червоної степової породи було зареєстровано наявність трьох морф із частотами, що приведені в таблиці 11.5. Необхідно з'ясувати, чи вірогідно розрізняються ці вибірки за структурою фенетичної мінливості.

Таблиця 11.5 – Кількість морф у трьох вибірках корів

Фен	Вибірка			Суми
	1	2	3	
A	15	18	11	$n_1 = 44$
B	17	27	23	$n_2 = 67$
C	25	30	15	$n_3 = 70$
Суми	$N_1 = 57$	$N_2 = 75$	$N_3 = 49$	$N = 181$

Спочатку розрахуємо відповідні частоти – по кожній клітинці таблиці та маргінальні (тобто, середні для всіх трьох вибірок) та занесемо їх в нову таблицю (табл. 11.6).

$$p_{11} = \frac{15}{57} = 0{,}263;$$

$$p_{12} = \frac{17}{57} = 0{,}298;$$

$$p_{21} = \frac{18}{75} = 0{,}240;$$

$$\overline{p}_1 = \frac{44}{181} = 0{,}243$$

і т. д.

Таблиця 11.6 – Частоти фенів у вибірках

Фен	Вибірка			Середні частоти
	1	2	3	
A	0,263	0,240	0,224	0,243
B	0,298	0,360	0,469	0,370
C	0,439	0,400	0,307	0,387
$N_i \cdot \left[1 - \sum_{j=1}^{k} p_{ij}^2\right]$	36,993	48,975	31,164	

Тепер ми можемо розрахувати сумарну варіансу:

$$\sigma_T^2 = 181 \cdot \left[1 - (0{,}243^2 + 0{,}370^2 + 0{,}387^2)\right] = 118{,}425.$$

Залишкова варіанса являє собою суму трьох (по числу вибірок) доданків, кожен з яких розраховується за аналогічною формулою, де замість сумарної чисельності підставляється обсяг відповідної вибірки, а замість середніх частот – відповідні для вибірки частоти фенів. Ці значення приведені в таблиці 11.6 в останньому рядку.

Таким чином, залишкова варіанса буде дорівнювати:

$$\sigma_Z^2 = 36{,}993 + 48{,}975 + 31{,}164 = 117{,}132 \ .$$

Факторіальну варіансу можна не розраховувати, оскільки вона може бути знайдена як різниця сумарної та залишкової, тобто, $\sigma_X^2 = 118{,}425 - 117{,}132 = 1{,}293$.

В принципі, формула 11.29 дасть ту ж величину, з точністю до округлення.

Тепер, використовуючи формули 11.30-11.35, ми можемо заповнити таблицю дисперсійного аналізу (табл. 11.7).

Як бачимо, розрахункове значення дисперсійного відношення набагато менше, ніж табличне значення критерію Фішера-Снедекора із числом ступенів свободи: $df_1 = 2$ і $df_2 = 178$ ($F_{\alpha = 0{,}05} = 3{,}05$).

Таблиця 11.7 – Результати дисперсійного аналізу

Джерело мінливості	σ^2	*df*	*MS*	*F*	*p*
X	1,293	2	0,647	0,98	> 0,05
Z	117,132	178	0,658		
T	118,425	180			

Отже можна зробити висновок, що нульова гіпотеза не може бути відкинута. Таким чином, частоти відповідних трьох фенів у трьох порівнюваних вибірках вірогідно не відрізняються між собою.

Оцінку сили впливу фактора можна розрахувати двома способами.

Перший спосіб. Оцінка сили впливу фактора розраховується як відношення факторіальної варіанси до сумарної:

$$\eta^2 = \frac{\sigma_X^2}{\sigma_T^2}. \tag{11.36}$$

Перевірка нульової гіпотези (у даному випадку – щодо рівності частот відповідних морф у всіх вибірках) перевіряється порівнянням величини

$$\chi^2 = \eta^2 \cdot N \cdot (k-1) \tag{11.37}$$

із табличним значенням критерію Хі-квадрат К. Пірсона із числом ступенів свободи:

$$df = (s-1)\cdot(k-1). \tag{11.38}$$

У нашому прикладі сила впливу фактора, розрахована цим способом, складає всього $\acute{\eta}^2 = 1{,}293 : 118{,}425 = 0{,}011$.

Ця оцінка (як і очікується) має низьку значущість, оскільки величина $\chi^2 = 0{,}011\times181\times2 = 3{,}98$ менше табличного значення критерію Хі-квадрат із числом ступенів свободи: $df = 2 \times 2 = 4$ ($\chi^2_{\alpha=0{,}05} = 9{,}49$).

Другий спосіб. Як і для диморфних ознак, можна використати наступну формулу:

$$\eta^2 = \frac{S_A^2}{S_A^2 + S_Z^2}, \tag{11.39}$$

де

$$S_Z^2 = MS_Z; \tag{11.40}$$

$$S_A^2 = \frac{MS_A - MS_Z}{n^*}; \tag{11.41}$$

$$n^* = \frac{1}{s-1}\cdot\left(N - \frac{\sum_{i=1}^{s} n_i}{N}\right). \tag{11.42}$$

Тоді, оцінка сили впливу фактора буде складати: η^2 = -0,00029, тобто практично дорівнює 0.

95% довірчий інтервал для оцінки η^2, розрахований по формулах 11.18-11.21, складає: [-0,013; 0,388]. Оскільки цей інтервал включає 0, приймається нуль-гіпотеза про відсутність вірогідних розходжень за структурою фенетичної мінливості у трьох досліджуваних вибірках тварин.

Контрольні питання:

1. Яким чином проводиться оцінка сили впливу фактора?
2. Методика розрахунку меж 95% довірчого інтервалу за Г. Шеффе (1963).

§ 12. Двофакторний дисперсійний аналіз якісних ознак

12.1 Двофакторний дисперсійний аналіз диморфних ознак

В основі двохфакторного дисперсійного аналізу (ДДА) якісних ознак, що представлені лише двома варіантами, лежить вже знайомий нам закон розкладання сумарної мінливості, однак з деякими доповненнями. У випадку, коли одночасно розглядається вплив на залежну змінну відразу двох (незалежних) факторів (наприклад, місце походження тварин і рік відбору матеріалу) факторіальна варіанса (σ_X^2) сама вже представлена сумою трьох компонентів:

$$\sigma_X^2 = \sigma_A^2 + \sigma_B^2 + \sigma_{A\times B}^2, \qquad (12.1)$$

де σ_A^2 – варіанса, що визначається впливом фактора *A*;

σ_B^2 – варіанса, що визначається впливом фактора *B;*

$\sigma_{A\times B}^2$ – варіанса, зумовлена одночасним впливом фактора *A* та фактора *B*.

Таким чином, у випадку проведення ДДА необхідно розрахувати чотири оцінки варіанси (загальну, залишкову і двох головних факторів), і ще дві оцінки варіанси (факторіальну та спільного впливу факторів *A×B)* можна розрахувати арифметично. Приведемо послідовність виконання розрахунків при проведенні ДДА якісних ознак із двома альтернативними станами на наступному прикладі.

Приклад. У двох різних стадах корів червоної степової породи (фактор *A*; дві градації) за три послідовних роки (фактор *B*; три градації) були зібрані вибірки (*n*), серед яких підраховано кількість тварин, що мали певний фен (*m*).

Необхідно з’ясувати, чи має місце розходження в частоті цієї морфи між окремими популяціями і в різні роки дослідження, оцінити силу впливу просторового фактора, часового фактора та їхнього спільного впливу (якщо вони мають місце).

Усі вихідні дані наведено в таблиці 12.1.

Таблиця 12.1 – Дані щодо частки тварин, що мають певний фен у різних стадах у різні роки

	A1			*A2*			Суми
	B1	*B2*	*B3*	*B1*	*B2*	*B3*	
m	35	55	20	40	85	35	M = 270
n	120	150	125	250	230	175	N = 1050
p	0,292	0,367	0,160	0,160	0,370	0,200	0,257
n·p·(1 – p)	24,808	34,847	16,800	33,600	53,613	28,000	

Розрахуємо спочатку частоти аналізованої ознаки для всіх сполучень фактора *A* та фактора *B* (всього отримуємо шість груп). Наприклад, для першої

градації фактора *A* в сполученні з першою градацією фактора *B* це значення становитиме: $p_{1\times B1} = 35 : 120 = 0{,}292$ і т. п.

Далі, розрахуємо середню зважену частоту аналізованої ознаки по всіх можливих сполученнях ознаки. Для цього сумарна кількість зареєстрованих морф в обох популяціях за три роки дослідження поділимо на загальну кількість досліджених особин:

$$\overline{p} = 270 : 1050 = 0{,}257.$$

Розрахуємо сумарну варіансу:

$$\sigma_T^2 = N \cdot \overline{p} \cdot (1 - \overline{p}) = 1050 \cdot 0{,}257 \cdot 0{,}743 = 200{,}499. \qquad (12.2)$$

Для кожного сполучення градацій обох факторів розрахуємо аналогічні величини. Наприклад, для першої градації фактора *A* в сполученні із першою градацією фактора *B* це значення складатиме:

$$C_{A1B1} = 120 \cdot 0{,}292 \cdot 0{,}708 = 24{,}808.$$

Ці значення приведені в таблиці 12.1 у самому нижньому рядку.

Сума цих величин дає нам оцінку залишкової варіанси (σ_Z^2). Таким чином, для нашого прикладу, значення залишкової варіанси дорівнюватиме:

$$\sigma_Z^2 = 24{,}808 + 34{,}847 + ... + 28{,}000 = 191{,}668.$$

Різниця між сумарною варіансою і залишковою дадуть нам оцінку факторіальної варіанси (тобто, суму всіх трьох її компонентів в формулі (12.1)):

$$\sigma_X^2 = 200{,}499 - 191{,}668 = 8{,}831.$$

Для того, щоб тепер вичленувати з цієї загальної факторіальної варіанси варіансу, зумовлену територіальним фактором (*A*), побудуємо нову допоміжну таблицю (табл. 12.2):

Таблиця 12.2 – Допоміжні дані для проведення ДДА

Показник	*A1*	*A2*
m_A	110	160
n_A	395	655
p_A	0,278	0,244
$n \cdot p_A \cdot (1 - p_A)$	79,283	120,824

Заповнюється ця таблиця таким чином: спочатку необхідно знайти кількість усіх тварин з даною морфою сумарно для всіх градацій фактора *B* усередині градації *A1*: $m_1 = 35 + 55 + 20 = 110$. Аналогічно знайдемо сумарну кількість усіх тварин з даною морфою для другої градації фактора *A*: $m_2 = 160$.

Далі, знайдемо загальну кількість аналізованих особин для всіх градацій фактора *B* всередині градації *A1*: $n_1 = 120 + 150 + 125 = 395$. І таку ж величину для другої градації фактора *A*: $n_2 = 655$.

Потім, розрахуємо відносні частоти особин, що мають дану морфу для градацій фактора *A* без врахування градацій фактора *B*:

$p_1 = 110 : 395 = 0{,}278$ і $p_2 = 160 : 655 = 0{,}244$.

За аналогією з формулою (12.2), знайдемо для кожної градації фактора *A* варіансу частот (останній рядок таблиці) та їхню суму:

$$C_A = 79{,}283 + 120{,}824 = 200{,}107.$$

Тоді, варіанса, викликана впливом тільки фактора *A* (без врахування впливу фактора *B*) буде дорівнювати:

$$\sigma_A^2 = \sigma_T^2 - C_A = 200{,}499 - 200{,}107 = 0{,}392. \quad (12.3)$$

Для того, щоб тепер вичленувати із загальної факторіальної варіанси варіансу, зумовлену часовим фактором (*B*), побудуємо ще одну допоміжну таблицю (табл. 12.3).

Таблиця 12.3 – Допоміжні дані для вичленування із загальної факторіальної варіанси варіанси, зумовленої часовим фактором (*B*)

Показник	*B1*	*B2*	*B3*
m_B	75	140	55
n_B	370	380	300
p_B	0,203	0,368	0,183
$n \cdot p_B \cdot (1 - p_B)$	59,863	88,379	44,853

Всі клітинки цієї таблиці заповнюються аналогічно тому, як ми це робили вище. Таким чином, проводиться аналіз впливу тільки фактора *B* (точніше, його градацій) без урахування впливу фактора *A*.

У підсумку одержуємо, що сума значень в останньому рядку таблиці дорівнює: C_B = 59,863 + 88,379 + 44,853 = 193,095. Тоді, за аналогією із формулою (12.3), варіанса, викликана впливом лише фактора *B* дорівнюватиме: $\sigma_B^2 = 200{,}499 - 193{,}095 = 7{,}404$.

Нарешті, останню компоненту (варіансу, що зумовлена одночасним впливом факторів *A* та *B*) знаходимо як різницю факторіальної варіанси і варіанс, пов'язаних із факторами *A* та *B*, тобто, використовуємо формулу (12.1):

$$\sigma_{A\times B}^2 = 8{,}831 - (0{,}392 + 7{,}404) = 1{,}035.$$

Наступним етапом нашого аналізу буде розрахунок числа ступенів свободи для всіх шести розрахованих нами варіанс:

$$df_T = N - 1; \quad (12.4)$$
$$df_A = a - 1; \quad (12.5)$$
$$df_B = b - 1; \quad (12.6)$$
$$df_{A\times B} = (a - 1)\cdot(b - 1); \quad (12.7)$$
$$df_X = a \cdot b - 1; \quad (12.8)$$
$$df_Z = N - a \cdot b, \quad (12.9)$$

де *a* – число градацій фактора *A*;

b – число градацій фактора *B*.

Для нашого прикладу, відповідні значення числа ступенів свободи будуть наступні:

df_T = 1049; df_A = 1; df_B = 2; $df_{A\times B}$ = 2; df_X = 5 і df_Z = 1044.

Середні квадрати розраховуються (як і при будь-якому типі дисперсійного аналізу) як відношення варіанс до відповідного значенням їх

числа ступенів свободи. Наприклад, середній квадрат для фактора *A* буде дорівнювати:

$$MS_A = \frac{\sigma_A^2}{df_A} = \frac{0{,}392}{1} = 0{,}392 \text{ і т. д.} \qquad (12.10)$$

Завершальним етапом ДДА є розрахунок дисперсійних відношень для трьох основних джерел мінливості вихідних даних дисперсійного комплексу (фактора *A*, фактора *B* та спільної дії факторів *A* і *B*). Однак, при цьому необхідно враховувати тип фактора. Як відомо, фактори бувають фіксованими та випадковими.

Якщо дослідника цікавлять розходження між визначеними градаціями фактора (наприклад, визначеними видами, визначеними популяціями, визначеними роками і т. п.), то така модель дисперсійного аналізу називається моделлю I типу або **моделлю із фіксованими (fixed) факторами**.

І навпаки, якщо градації фактора обрані випадково із великого (нескінченно) числа можливих його станів, то така модель дисперсійного аналізу називається моделлю II типу або **моделлю із випадковими (random) факторами**.

Якщо один із факторів має випадково обрані градації, а другий – фіксовані, то така модель дисперсійного аналізу називається **змішаною моделлю.**

Залежно від того, який із двох факторів ДДА є випадковим чи фіксованим, мають місце чотири різні ситуації і, відповідно, чотири схеми розрахунку дисперсійних відношень та оцінки рівня їхньої значущості, що представлені в таблиці 12.4.

Таблиця 12.4 – Схеми розрахунку дисперсійних відношень та оцінки рівня їхньої значущості

Градації фактора *B*	Градації фактора *A*	
	фіксовані	випадкові
Фіксовані	$F_A = \frac{MS_A}{MS_Z};$ $F_B = \frac{MS_B}{MS_Z};$ $F_{A\times B} = \frac{MS_{A\times B}}{MS_Z};$	$F_A = \frac{MS_A}{MS_Z};$ $F_B = \frac{MS_B}{MS_{A\times B}};$ $F_{A\times B} = \frac{MS_{A\times B}}{MS_Z};$
Випадкові	$F_A = \frac{MS_A}{MS_{A\times B}};$ $F_B = \frac{MS_B}{MS_Z};$ $F_{A\times B} = \frac{MS_{A\times B}}{MS_Z};$	$F_A = \frac{MS_A}{MS_{A\times B}};$ $F_B = \frac{MS_B}{MS_{A\times B}};$ $F_{A\times B} = \frac{MS_{A\times B}}{MS_Z};$

Якщо тепер прийняти, що в нашому прикладі градації фактора *A* та фактора *B* є фіксованими (тобто, нас цікавить мінливість частоти особин з певною морфою саме в цих двох популяціях і саме за ці три роки), то таблиця з результатами дисперсійного аналізу буде мати наступний вигляд (табл. 12.5).

Таблиця 12.5 – Результати дисперсійного аналізу

Джерело мінливості	σ^2	*df*	*MS*	*F*	*p*
A	0,392	1	0,392	2,13	0,145
B	7,404	2	3,702	20,12	*<0,001*
A×B	1,035	2	0,518	2,82	0,060
X	8,831	5	1,766	9,60	*<0,001*
Z	191,668	1044	0,184		
Y	200,499	1049			

Таким чином, нами доведено тільки вірогідний вплив фактора *B*, тобто, року проведення дослідження. Часова компонента мінливості дуже значима, хоча при цьому спостерігається і деяка тенденція до прояву спільного впливу територіального і часового аспектів дослідження.

Оцінку сили впливу кожного з факторів можна провести двома способами.

Перший спосіб. Оцінити силу впливу кожного з факторів (чи їхнього спільного впливу) можна, використовуючи формулу:

$$\chi^2 = \eta^2 \cdot N \qquad (12.11)$$

Рівень значущості цієї оцінки перевіряється шляхом порівняння розрахованого відповідного значення величини (12.11) із табличним значенням критерію Хі-квадрат із відповідним числом ступенів свободи (формули 12.5-12.7).

Для нашого прикладу, сила впливу часового фактора дорівнює:

$$\eta^2_B = \frac{\sigma^2_B}{\sigma^2_T} = \frac{7{,}404}{200{,}499} = 0{,}0369 \ .$$

Рівень значущості цієї оцінки можна визначити, розрахувавши за формулою 12.11 величину: $N \cdot \acute{\eta}^2 = 1050 \cdot 0{,}0369 = 38{,}75$.

Ця величина значно перевершує табличне значення критерію Хі-квадрат із числом ступенів свободи: $df_B = 2$ ($\chi^2_{\alpha=0{,}05} = 5{,}99$).

Отже, нульова гіпотеза для цього фактора повинна бути відкинута.

Відзначимо, що отримані оцінки при проведенні ДДА мають дещо зміщений характер і ступінь цього зміщення залежить від розходження в кількості досліджених об’єктів для кожного сполучення градацій кожного фактора. При використанні вибірок одного порядку цим зміщенням можна знехтувати.

Розрахунок сили впливу фактора (факторів) таким способом правомочний лише у випадку використання моделі із фіксованими факторами.

Другий спосіб базується на розкладанні факторіальних середніх квадратів. У випадку проведення ДДА, формули для такого розкладання залежать від типу використаної моделі (див. вище); у підсумковому вигляді вони представлені в таблиці 12.6.

Таблиця 12.6 – Формули для розкладання факторіальних середніх квадратів

Градації фактора *B*	Градації фактора *A*	
	фіксовані	випадкові
Фіксовані	$S_A^2 = \frac{MS_A - MS_Z}{b \cdot n^*};$ $S_B^2 = \frac{MS_B - MS_Z}{a \cdot n^*};$ $S_{A\times B}^2 = \frac{MS_{A\times B} - MS_Z}{n^*}$	$S_A^2 = \frac{MS_A - MS_Z}{b \cdot n^*};$ $S_B^2 = \frac{MS_B - MS_{A\times B}}{a \cdot n^*};$ $S_{A\times B}^2 = \frac{MS_{A\times B} - MS_Z}{n^*};$
Випадкові	$S_A^2 = \frac{MS_A - MS_{A\times B}}{b \cdot n^*};$ $S_B^2 = \frac{MS_B - MS_Z}{a \cdot n^*};$ $S_{A\times B}^2 = \frac{MS_{A\times B} - MS_Z}{n^*};$	$S_A^2 = \frac{MS_A - MS_{A\times B}}{b \cdot n^*};$ $S_B^2 = \frac{MS_B - MS_{A\times B}}{a \cdot n^*};$ $S_{A\times B}^2 = \frac{MS_{A\times B} - MS_Z}{n^*};$

де n^* – усереднений по всіх градаціях усіх факторів обсяг вибірки:

$$n^* = \frac{1}{(a \cdot b) - 1} \cdot \left(N - \frac{\sum_{i=1}^{a \cdot b} n_i^2}{N} \right). \qquad (12.12)$$

Якщо припустити (як і раніше), що ми маємо справу з двома фіксованими факторами, то використовуючи формули, приведені у верхній лівій клітинці таблиці 12.6 і формулу 12.12 одержимо наступні проміжні значення:

$$n^* = \frac{1}{(2 \cdot 3) - 1} \cdot \left[1050 - \frac{120^2 + 150^2 + ... + 175^2}{1050} \right] = 172{,}2;$$

$$S_A^2 = \frac{0{,}392 - 0{,}184}{3 \cdot 172{,}2} = 0{,}00040;$$

$$S_B^2 = \frac{3{,}702 - 0{,}184}{2 \cdot 172{,}2} = 0{,}01002;$$

$$S_{A\times B}^2 = \frac{0{,}518 - 0{,}184}{172{,}2} = 0{,}00194;$$

$$S_Z^2 = MS_Z = 0{,}18359\ .$$

Тоді, силу впливу факторів *A*, *B* та їхнього сполучення *A*×*B* можна знайти за формулами:

$$\eta_A^2 = \frac{S_A^2}{S_T^2}; \tag{12.13}$$

$$\eta_B^2 = \frac{S_B^2}{S_T^2}; \tag{12.14}$$

$$\eta_{A\times B}^2 = \frac{S_{A\times B}^2}{S_T^2}; \tag{12.15}$$

де

$$S_T^2 = S_A^2 + S_B^2 + S_{A\times B}^2 + S_Z^2. \tag{12.16}$$

При порівнянні оцінок сили впливу головних факторів та їхнього сполучення, отриманих двома представленими способами (табл. 12.7) для даних нашого прикладу можна помітити, що вони досить близькі, що дає деяку перевагу при використанні першого способу (без розкладання середніх квадратів), однак, тільки у випадках використання моделі ДДА із фіксованими факторами *A* та *B*.

Таблиця 12.7 – Результати оцінки сили впливу головних факторів та їхнього сполучення

Показник	Спосіб оцінки сили впливу:		
	перший	другий	
η_A^2	0,0020	0,0020	[-0,0012; 0,815]
η_B^2	0,0369	0,0519	[0,0128; 0,695]
$\eta_{A\times B}^2$	0,0052	0,0099	[-0,0014; 0,391]

Крім того, ми використовували формули 11.18-11.21 і розрахували довірчий інтервал для оцінок, отриманих другим способом. При розрахунках, ми використовували для оцінки *n** у формулах 11.18 і 11.19, формулу 12.12 для взаємодії факторів, і формулу 11.17 для головних факторів *A* чи *B*, попередньо замінивши *s* на *a* чи *b*, відповідно.

Як і очікувалося на підставі результатів безпосереднього впливу факторів та їхнього сполучення, на вихідну ознаку (див. табл. 12.5) тільки для фактора *B* у 95% довірчий інтервал не потрапляє значення нуль і, відповідно, лише стосовно цього фактора нульова гіпотеза може бути відкинута.

Деякі автори, наприклад Г. Ф. Лакін (1980), пропонують не включати в суму (12.16) компоненти середніх квадратів фактора (факторів чи їхнього сполучення), якщо під час ДДА не доведено його (їх) впливу на ознаку, і розраховувати оцінку сили впливу фактора без його (їх) компоненти.

12.2 Двофакторний дисперсійний аналіз поліморфних ознак

ДДА у випадку присутності декількох альтернативних варіацій аналізованої ознаки проводиться за тим же принципом, що і для диморфних ознак. Лише змінюється формула для розрахунку варіанс і розрахунку рівня значущості впливу фактора (факторів чи їхнього сполучення).

Продемонструємо алгоритм ДДА для поліморфних ознак на наступному прикладі.

Приклад. Також, як і раніше, ми маємо справу із двома популяціями, що були проаналізовані протягом трьох послідовних років. Частоти (абсолютні) трьох різних варіацій представлені нижче в таблиці 12.8.

Таблиця 12.8 – Абсолютні значення трьох різних варіацій

Фени	*A1*			*A2*			Суми
	B1	*B2*	*B3*	*B1*	*B2*	*B3*	
A	15	15	15	120	15	50	M1 = 230
B	35	40	45	50	40	35	M2 = 245
C	40	50	60	30	50	20	M3 = 250
Обсяг вибірок (*n*)	90	105	120	200	105	105	N = 725

Розрахуємо частоти окремих фенів у всіх вибірках і середні частоти фенів по всій сукупності даних і занесемо ці дані в наступну таблицю 12.9.

Таблиця 12.9 – Частоти окремих фенів у всіх вибірках і середні частоти фенів по всій сукупності даних

Фени	*A1*			*A2*			Середні частоти
	B1	*B2*	*B3*	*B1*	*B2*	*B3*	
A	0,167	0,143	0,125	0,600	0,143	0,476	0,317
B	0,389	0,381	0,375	0,250	0,381	0,333	0,338
C	0,444	0,476	0,500	0,150	0,476	0,191	0,345
$C = n \cdot \left(1 - \sum_{i=1}^{3} p_i^2\right)$	56,129	63,820	71,250	111,000	63,820	65,736	

Розрахуємо сумарну варіансу дисперсійного комплексу за наступною формулою:

$$\sigma_T^2 = N \cdot \left(1 - \sum_{i=1}^{k} \overline{p}_i^{-2}\right). \qquad (12.17)$$

Таким чином, сумарна варіанса дорівнює:

$$\sigma_T^2 = 725 \cdot \left[1 - (0{,}317^2 + 0{,}338^2 + 0{,}345^2)\right] = 483{,}025$$

Залишкова варіанса розраховується як сума значень, приведених в останньому рядку таблиці 12.9. Її значення, відповідно, становитиме:

$$\sigma_Z^2 = 56{,}129 + 63{,}820 + ... + 65{,}736 = 431{,}755 \ .$$

Факторіальна варіанса (у даному випадку вона включає вплив фактора *A*, фактора *B* і їхнього спільного впливу *A*×*B*; див. формулу (12.1) являє собою різницю між сумарною і залишковою варіансами:

$$\sigma_X^2 = \sigma_T^2 - \sigma_Z^2 = 483{,}025 - 431{,}755 = 51{,}270 \ .$$

Для того, щоб розрахувати частку факторіальної варіанси, зумовлену впливом фактора *A*, необхідно скласти нову допоміжну таблицю (табл. 12.10).

Таблиця 12.10 – Допоміжні дані для розрахунку частки факторіальної варіанси, зумовленої впливом фактора *A*

Фен	*A1*	*A2*
	Абсолютні частоти	
A	45	185
B	120	125
C	150	100
n	315	410
	Відносні частоти	
A	0,143	0,451
B	0,381	0,305
C	0,476	0,244
$C_A = n \cdot \left(1 - \sum_{i=1}^{3} p_i^2\right)$	191,461	264,056

Тут у кожній клітинці градації фактора *A* ми підсумовуємо всі значення по градаціях фактора *B*. Аналогічно, підраховуємо суму по кожному стовпцю, розраховуємо частоти і частки варіанси.

Використовуючи отримані з цієї допоміжної таблиці дані, розраховуємо оцінку варіанси, зумовлену впливом фактора *A* за формулою:

$$\sigma_A^2 = \sigma_T^2 - \sum_{i=1}^{a} C_{Ai} \ . \qquad (12.18)$$

Таким чином, її значення становитиме:

$$\sigma_A^2 = 483{,}025 - (191{,}461 + 264{,}054) = 27{,}508 .$$

Аналогічно розраховуємо оцінку варіанси, зумовлену впливом фактора *B* (табл. 12.11).

Таким чином, значення варіанси, що зумовлена впливом фактора *B* становитиме:

$$\sigma_B^2 = 483{,}025 - (185{,}285 + 127{,}641 + 149{,}337) = 20{,}762 .$$

Таблиця 12.11 – Допоміжні дані для розрахунку частки факторіальної варіанси, зумовленої впливом фактора *B*

Фен	*B1*	*B2*	*B3*
	Абсолютні частоти		
A	135	30	65
B	85	80	80
C	70	100	80
n	290	210	225
	Відносні частоти		
A	0,466	0,143	0,289
B	0,293	0,381	0,356
C	0,241	0,476	0,355
$C_B = n \cdot \left(1 - \sum_{i=1}^{3} p_i^2\right)$	185,285	127,641	149,337

Нарешті, значення варіанси спільного впливу факторів *A*×*B* розраховується як різниця між факторіальною варіансою (σ_X^2) і сумою оцінок варіанс факторів *A* та *B*:

$$\sigma_{A\times B}^2 = \sigma_X^2 - (\sigma_A^2 + \sigma_B^2) = 51{,}270 - (27{,}508 + 20{,}762) = 3{,}000. \quad (12.19)$$

Після того, як розраховано всі компоненти варіації, розрахуємо число ступенів свободи для кожного фактора та їх сполучення. Для цього використовуємо формули 12.4-12.9. Тоді відповідні величини будуть наступні:

df_T = 724; df_A = 1; df_B = 2; $df_{A\times B}$ = 2; df_X = 5 и df_Z = 719.

Значення середніх квадратів розраховуємо як відношення значень варіанс до відповідного значення числа ступенів свободи (див. формулу 12.10).

Всі отримані результати заносимо в підсумкову таблицю дисперсійного аналізу (табл. 12.12).

Таблиця 12.12 – Результати дисперсійного аналізу

Джерело мінливості	σ^2	*df*	*MS*	*F*	*p*
A	27,508	1	27,508	45,85	< 0,001
B	20,762	2	10,381	17,30	< 0,001
A×B	3,000	2	1,500	2,50	0,083
X	51,270	5	10,254	17,09	< 0,001
Z	431,755	719	0,600		
T	483,025	724			

Відповідні факторіальні відношення розраховуємо, вважаючи, що наші фактори (територіальний і часовий) мають фіксовані градації (див. табл. 12.4).

Отже, нульова гіпотеза (щодо відсутності впливу) повинна бути спростована нами як для просторового фактора (фактор *A*), так і часового (фактор *B*). Однак, спільного впливу факторів *A* та *B* не виявлено, хоча і було

відмічено деяку тенденцію до непропорційної зміни частот фенів у різних популяціях у різні роки дослідження.

Сила впливу кожного з факторів (і їхнього спільного впливу) розраховується як відношення відповідної варіанси до сумарної. Наприклад, для фактора *A* це значення складає:

$$\eta_A^2 = \frac{\sigma_A^2}{\sigma_T^2} = \frac{27{,}508}{483{,}025} = 0{,}0569\,. \qquad (12.20)$$

Оцінка рівня значущості показників сили впливу проводиться із використання розподілу Хі-квадрат К. Пірсона. Величини:

$$\chi_A^2 = \eta_A^2 \cdot N \cdot (k-1)\,, \qquad (12.21)$$

$$\chi_B^2 = \eta_B^2 \cdot N \cdot (k-1)\,, \qquad (12.22)$$

$$\chi_{A\times B}^2 = \eta_{A\times B}^2 \cdot N \cdot (k-1) \qquad (12.23)$$

мають розподіл Хі-квадрат зі ступенями свободи, відповідно:

$$df_A = (a-1)\cdot(k-1), \qquad (12.24)$$

$$df_B = (b-1)\cdot(k-1), \qquad (12.25)$$

$$df_{A\times B} = (a-1)\cdot(b-1)\cdot(k-1), \qquad (12.26)$$

де *a* – число градацій фактора *A*;

b – число градацій фактора *B*;

k – число використаних в аналізі фенів.

У нашому прикладі, оцінка величини Хі-квадрат для ознаки *A* становитиме: $\chi_A^2 = 0{,}0569 \cdot 725 \cdot (3-1) = 82{,}505$, що значно вище, ніж табличне значення критерію Хі-квадрат із числом ступенів свободи: $df = (2-1)\cdot(3-1) = 2$ ($\chi_{\alpha=0{,}05}^2 = 5{,}99$).

Аналогічно можна оцінити рівень значущості для сили впливу фактора *B* та спільного впливу *A*×*B*.

Крім того, оцінки сили впливу факторів чи їхнього сполучення, а також їхній 95% довірчий інтервал, можна розрахувати, використовуючи формули, приведені в таблиці 12.7, формули 12.13-12.16 і матеріали, викладені наприкінці попереднього розділу.

Контрольні питання:

1. Особливості застосування моделей із фіксованими (fixed) та випадковими (random) факторами.
2. Які відмінності у проведенні двохфакторного дисперсійного аналізу диморфних та поліморфних ознак?

§ 13. Ієрархічний двофакторний дисперсійний аналіз якісних ознак

13.1 Ієрархічний двофакторний дисперсійний аналіз диморфних ознак

Ієрархічна структура дисперсійного аналізу виникає в тому випадку, коли немає повного сполучення всіх градацій фактора *A* із усіма градаціями фактора *B*. Структуру повного двофакторного дисперсійного аналізу (**crossed 2-way ANOVA**) й ієрархічного дисперсійного аналізу у випадку двох факторів (**nested 2-way ANOVA**) можна схематично представити в такий спосіб (рис. 13.1).

ДДА

	B1	*B2*	*B3*	*B4*	*B5*
A1	X	X	X	X	X
A2	X	X	X	X	X

ІДДА

	B1	*B2*	*B3*	*B4*	*B5*	*B6*
A1	X	X	X			
A2				X	X	X

Рисунок 13.1 – Схематична структура повного двофакторного дисперсійного аналізу та ієрархічного дисперсійного аналізу у випадку двох факторів

У представленій вище схемі ієрархічного дисперсійного аналізу перша градація фактора *A* сполучається тільки з градаціями фактора *B* – *B1, B2* і *B3*, а друга градація фактора *A* – тільки з градаціями фактора *B* – *B4, B5* і *B6*. Сполучень типу *A1*×*B5* не може існувати взагалі в природі через, наприклад, їхню просторову та/чи часову роз'єднаність.

Подібна схема структури дисперсійного аналізу виникає, наприклад, у тому випадку, коли проводяться дослідження в двох регіонах. В одному регіоні аналізується частота даної ознаки в трьох популяціях (1-3), а в іншому – у трьох інших популяціях (4-6). Відповідно, просто фізично не може бути сполучення першого регіону та п'ятої популяції.

Більш складна ієрархічна система може містити кілька рівнів організації, але з обов'язковим підпорядкуванням нижче розташованих рівнів розташованим вище. Наприклад, окремі деми можуть входити до складу популяцій, що, у свою чергу, входять до складу локалітів, що у свою чергу розташовані в різних регіонах і т. д. і т. п.

З інших термінів, що зустрічаються в російсько- та україномовній літературі для позначення ієрархічного дисперсійного аналізу, можна привести наступні визначення – **гніздовий план дисперсійного аналізу** чи **план дисперсійного аналізу з угрупуванням**.

При проведенні ієрархічного ДДА (ІДДА) для якісних ознак із двома альтернативними варіаціями основна послідовність розрахунків залишається

без зміни, деякі модифікації використовуються тільки для розрахунку спеціальних факторіальних варіанс і дисперсійних відношень.

У цьому випадку факторіальна варіанса складається лише із двох компонентів – варіанси **між градаціями фактора *A*** (σ_A^2) та варіанси **між градаціями фактора *B* у межах градацій фактора *A*** ($\sigma_{B(A)}^2$).

Таким чином, сума цих двох компонентів (σ_X^2) являє собою частку мінливості ознаки **між градаціями фактора *B* та між градаціями фактора *A***.

Розглянемо алгоритм розрахунку ІДДА на наступному прикладі.

Приклад. У двох регіонах було проаналізовано частоту фена *A* серед корів англерської породи. При цьому, у межах кожного регіону досліджувалося по чотири окремі популяції.

Нам необхідно виявити, чи мають місце розходження частоти даної ознаки для тварин, що утримуються в різних регіонах і між популяціями в межах різних регіонів?

Усі вихідні дані приведені в таблиці 13.1.

Таблиця 13.1 – Частота фена А серед корів англерської породи

	A1				*A2*				Суми
	B1	*B2*	*B3*	*B4*	*B5*	*B6*	*B7*	*B8*	
m	13	17	15	10	45	50	35	15	$M = 200$
n	150	125	225	250	275	300	275	150	$N = 1750$
p	0,087	0,136	0,067	0,040	0,164	0,167	0,127	0,100	$\overline{p} = 0{,}114$
C_Z	11,915	14,688	14,065	9,600	37,704	41,733	30,490	13,500	

Після того, як розраховано частоти фена для кожної вибірки і середню частоту даної ознаки у всій сукупності ($\overline{p} = 0{,}114$) ми можемо перейти до оцінки сумарної і залишкової варіанс:

$$\sigma_T^2 = N \cdot \overline{p} \cdot (1 - \overline{p}) = 1750 \cdot 0{,}114 \cdot 0{,}886 = 176{,}757, \quad (13.1)$$

$$\sigma_Z^2 = \sum_{i=1}^{b} C_{Z\,i} = 11{,}915 + 14{,}688 + \ldots + 13{,}500 = 173{,}695, \quad (13.2)$$

де *b* – число градацій фактора *B* (тобто, сукупна кількість досліджуваних популяцій і в першому, і в другому регіонах).

Тоді факторіальна варіанса (σ_X^2) може бути розрахована як різниця між сумарною і залишковою:

$$\sigma_X^2 = \sigma_T^2 - \sigma_Z^2 = 176{,}757 - 173{,}695 = 3{,}062. \quad (13.3)$$

Як ми вже вказували вище, цей компонент являє собою наступну суму:

$$\sigma_X^2 = \sigma_A^2 + \sigma_{B(A)}^2, \quad (13.4)$$

тому, розрахувавши одне з цих значень, друге можна одержати простим розрахунком.

Компоненту факторіальної варіанси (σ_A^2), що пов'язана із впливом фактора *A* (тобто, розходженнями між регіонами) обчислюємо на підставі даних, підсумованих для всіх популяцій у межах кожного регіону (табл. 13.2).

Таблиця 13.2 – Дані для розрахунку компоненти факторіальної варіанси (σ_A^2), що пов'язана із впливом фактора *A*

Показник	*A1*	*A2*
m_A	55	145
n_A	750	1000
p_A	0,073	0,145
$C_A = n \cdot p_A \cdot (1 - p_A)$	50,753	123,975

Тоді, шукану величину може бути розраховано за формулою:

$$\sigma_A^2 = \sigma_T^2 - \sum_{i=1}^{a} C_{Ai} = 176{,}757 - (50{,}753 + 123{,}975) = 2{,}029, \quad (13.5)$$

де *a* – число градацій фактора *A*.

Відповідно, компонента факторіальної варіанси ($\sigma_{B(A)}^2$), що пов'язана із впливом фактора *B* у межах фактора *A*, буде дорівнювати:

$$\sigma_{B(A)}^2 = \sigma_X^2 - \sigma_A^2 = 3{,}062 - 2{,}029 = 1{,}033. \quad (13.6)$$

Число ступенів свободи для кожної компоненти мінливості дисперсійного комплексу розраховується за наступними формулами:

$$df_T = N - 1; \quad (13.7)$$

$$df_A = a - 1; \quad (13.8)$$

$$df_{B(A)} = b - a; \quad (13.9)$$

$$df_X = b - 1; \quad (13.10)$$

$$df_Z = N - b. \quad (13.11)$$

Таким чином, відповідні для нашого прикладу числа ступенів свободи будуть наступні:

$df_T = 1749$; $df_A = 1$; $df_{B(A)} = 6$; $df_X = 7$ та $df_Z = 1742$.

Середні квадрати розраховуються стандартно, як відношення варіанс до відповідного значення числа ступенів свободи; наприклад, для фактора *A* значення середнього квадрата буде дорівнювати:

$$MS_A = \frac{\sigma_A^2}{df_A} = \frac{2{,}029}{1} = 2{,}029 \text{ і т. п.} \quad (13.12)$$

Як вказувалося вище, у випадку проведення ІДДА принципово змінюється правило розрахунку дисперсійних відношень.

Для нашого прикладу вони розраховуються за наступними формулами:

$$F_A = \frac{MS_A}{MS_{B(A)}} = \frac{2{,}029}{0{,}172} = 11{,}80; \quad (13.13)$$

$$F_{B(A)} = \frac{MS_{B(A)}}{MS_Z} = \frac{0{,}172}{0{,}100} = 1{,}72. \quad (13.14)$$

Тоді підсумкова таблиця ІДДА буде мати наступний вигляд (табл. 13.3).

Таблиця 13.3 – Результати дисперсійного аналізу

Джерело мінливості	σ^2	*df*	*MS*	*F*	*p*
A	2,029	1	2,029	11,80	0,014
B(A)	1,033	6	0,172	1,72	0,113
X	3,062	7	0,437	4,37	<0,001
Z	173,695	1742	0,100		
T	176,757	1749			

Таким чином, відхиляється нуль-гіпотеза лише стосовно регіональної мінливості за частотою ознаки (із рівнем значущості p = 0,014). Водночас, у межах своїх регіонів популяції виявляються гомогенними.

Оцінку сили впливу факторів, використаних в аналізі, можна провести, як звичайно, двома способами.

Перший спосіб. У цьому випадку оцінки сили впливу фактора розраховуються як відношення відповідної факторіальної варіанси до сумарної:

$$\eta_A^2 = \frac{\sigma_A^2}{\sigma_T^2}; \tag{13.15}$$

$$\eta_{B(A)}^2 = \frac{\sigma_{B(A)}^2}{\sigma_T^2}\ . \tag{13.16}$$

Рівень значущості цих оцінок визначається на підставі порівняння величин:

$$\chi_A^2 = \frac{\eta_A^2 \cdot N}{F_{B(A)}}, \tag{13.17}$$

$$\chi_{B(A)}^2 = \eta_{B(A)}^2 \cdot N \tag{13.18}$$

із табличними значеннями критерію Хі-квадрат К. Пірсона для відповідних чисел ступенів свободи (формули 13.7 і 13.8).

Формули 13.15-13.18 можуть бути застосовані у випадку, коли $b \geq$ 12-16 (особливо, перша з них).

Другий спосіб ґрунтується на розкладанні оцінок факторіальних середніх квадратів.

Спочатку необхідно розрахувати величини:

$$n^* = \frac{N - n_a}{b - a}; \tag{13.19}$$

$$n^{**} = \frac{n_a - \frac{\sum_{i=1}^{b} n_i^2}{N}}{a - 1}; \tag{13.20}$$

$$n^{***} = \frac{N - \frac{\sum_{j=1}^{a} n_j^2}{N}}{a-1}, \tag{13.21}$$

де

$$n_a = \sum_{i=1}^{a} \frac{\sum_{j=1}^{b} n_j^2}{N_i}. \tag{13.22}$$

Тоді компоненти відповідних середніх квадратів можна знайти за формулами:

$$S_Z^2 = MS_Z; \tag{13.23}$$

$$S_{B(A)}^2 = \frac{MS_{B(A)} - MS_Z}{n^*}; \tag{13.24}$$

$$S_A^2 = \frac{n^* \cdot MS_A - n^{**} \cdot MS_{B(A)} + (n^{**} - n^*) \cdot MS_Z}{n^* \cdot n^{***}}. \tag{13.25}$$

А оцінки сили впливу факторів *A* та *B(A)*:

$$\eta_A^2 = \frac{S_A^2}{S_T^2}; \tag{13.26}$$

$$\eta_{B(A)}^2 = \frac{S_{B(A)}^2}{S_T^2}, \tag{13.27}$$

де

$$S_T^2 = S_A^2 + S_{B(A)}^2 + S_Z^2. \tag{13.28}$$

Для даних з нашого прикладу відповідні величини будуть наступні:

$$n_a = \frac{150^2 + 125^2 + 225^2 + 250^2}{750} + \frac{275^2 + 300^2 + 275^2 + 150^2}{1000} = 465{,}5;$$

$$n^* = \frac{1750 - 465{,}5}{8-2} = 214{,}1;$$

$$n^{**} = \frac{465{,}5 - \frac{150^2 + 125^2 + \ldots + 150^2}{1750}}{2-1} = 228{,}4;$$

$$n^{***} = \frac{1750 - \frac{750^2 + 1000^2}{1750}}{2-1} = 857{,}1;$$

$$S_Z^2 = 0{,}100; \quad S_{B(A)}^2 = \frac{0{,}172 - 0{,}100}{214{,}1} = 0{,}00034;$$

$$S_A^2 = \frac{214{,}1 \cdot 2{,}029 - 228{,}4 \cdot 0{,}172 + 0{,}100 \cdot (228{,}4 - 214{,}1)}{214{,}1 \cdot 857{,}1} = 0{,}00216.$$

Тоді оцінки сили впливу факторів *A* та *B(A)* дорівнюватимуть:

$$\eta_A^2 = \frac{0{,}00034}{0{,}00034 + 0{,}00216 + 0{,}100} = 0{,}0033;$$

$$\eta_{B(A)}^2 = \frac{0{,}00216}{0{,}00034 + 0{,}00216 + 0{,}100} = 0{,}0211.$$

13.2 Ієрархічний двофакторний дисперсійний аналіз поліморфних ознак

У випадку, коли ознака має кілька альтернативних варіацій ($k \geq 3$) загальний алгоритм ієрархічного ДДА змінюється слабко; зміни стосуються лише формул для розрахунку варіанс. Продемонструємо застосування ІДДА для аналізу поліморфної ознаки на наступному прикладі.

Приклад. У двох регіонах було проаналізовано частоту трьох фенів корів червоної степової породи. При цьому, в межах кожного регіону досліджувалося по чотири окремі популяції. Необхідно оцінити рівень фенетичної диференціації виду між регіонами і між окремими популяціями в межах регіонів. Усі вихідні дані занесені в таблицю 13.4.

Таблиця 13.4 – Кількість фенів у корів червоної степової породи в різних регіонах та популяціях

Фен	*A1*				*A2*				Суми
	B1	*B2*	*B3*	*B4*	*B5*	*B6*	*B7*	*B8*	
M	15	15	17	25	35	25	45	40	*K*1 = 217
N	35	45	35	30	55	10	60	25	*K*2 = 295
S	40	50	68	100	60	50	100	45	*K*3 = 513
n	90	110	120	155	150	85	205	110	*N* = 1025

В таблиці 13.5 приведено частоти окремих фенів і середня зважена частота фенів у цілому по всім даним.

Таблиця 13.5 – Частота окремих фенів у корів червоної степової породи в різних регіонах та популяціях

Фен	*A1*				*A2*				Середні частоти
	B1	*B2*	*B3*	*B4*	*B5*	*B6*	*B7*	*B8*	
M	0,167	0,136	0,142	0,161	0,233	0,294	0,220	0,364	$p_1 = 0{,}212$
N	0,389	0,409	0,292	0,194	0,367	0,118	0,293	0,227	$p_2 = 0{,}288$
S	0,444	0,455	0,567	0,645	0,400	0,588	0,488	0,409	$p_3 = 0{,}500$
C	56,129	66,792	67,770	80,665	97,653	47,081	128,659	71,356	

Використовуючи формулу 13.1, розраховуємо сумарну варіансу:

$$\sigma_T^2 = 1025 \cdot \left[1 - (0{,}212^2 + 0{,}288^2 + 0{,}500^2)\right] = 637{,}665.$$

За аналогічним принципом розраховуємо і окремі компоненти залишкової варіанси для кожної вибірки окремо. Ці значення наведені в останньому рядку таблиці 13.5. Тоді, величина залишкової варіанси буде дорівнювати:

$$\sigma_Z^2 = 56{,}129 + 66{,}792 + ... + 71{,}356 = 617{,}106.$$

Факторіальна варіанса (у даному випадку вона являє собою суму $\sigma_A^2 + \sigma_{B(A)}^2$) тоді дорівнюватиме:

$$\sigma_X^2 = 637{,}665 - 617{,}106 = 20{,}559.$$

Наступним етапом аналізу буде розкладання факторіальної варіанси на окремі її компоненти. Для цього побудуємо допоміжну таблицю, в яку занесемо вихідні дані, об'єднавши при цьому всі популяції для кожного регіону (фактор *A*), розрахуємо частоти по кожній ознаці і частки компонента варіанси σ_A^2 (табл. 13.6).

Таблиця 13.6 – Частоти по кожній ознаці і частки компонента варіанси σ_A^2

	A1	*A2*
	Абсолютні частоти	
k1	72	145
k2	145	150
k3	258	255
n	475	550
	Відносні частоти	
p_1	0,152	0,264
p_2	0,305	0,273
p_3	0,543	0,464
C_A	279,785	352,263

Тоді оцінку варіанси, що зумовлена відмінностями між регіонами можна одержати за формулою:

$$\sigma_A^2 = \sigma_T^2 - \sum_{i=1}^{a} C_{Ai} \qquad (13.29)$$

$$\sigma_A^2 = 637{,}665 - (279{,}785 + 352{,}263) = 5{,}616$$

Варіанса, що зумовлена відмінностями між популяціями всередині регіонів тоді буде дорівнювати:

$$\sigma_{B(A)}^2 = \sigma_X^2 - \sigma_A^2 \qquad (13.30)$$

$$\sigma_{B(A)}^2 = 20{,}559 - 5{,}616 = 14{,}943.$$

Число ступенів свободи для кожного компонента розраховується за формулами 13.7-13.11:

$df_T = 1024$; $df_A = 1$; $df_{B(A)} = 6$; $df_X = 7$ та $df_Z = 1017$.

Середні квадрати розраховуються за стандартним принципом – як відношення варіанси до відповідного числа ступенів свободи. Дисперсійні відношення розраховуємо за формулами 13.13 та 13.14.

Отримані результати заносимо в таблицю дисперсійного аналізу (табл. 13.7).

Таблиця 13.7 – Результати дисперсійного аналізу

Джерело мінливості	σ^2	*df*	*MS*	*F*	*p*
A	5,616	1	5,616	2,25	0,184
B(A)	14,943	6	2,491	4,10	<0,001
X	20,559	7	2,937	4,84	<0,001
Z	617,106	1017	0,607		
T	637,665	1024			

Таким чином, нульова гіпотеза приймається тільки стосовно фактора *A* (тобто, відмінності між регіонами відсутні). Водночас, стосовно окремих популяцій всередині регіонів вона повинна бути відкинута із рівнем значущості $p < 0{,}001$.

Оцінка сили впливу факторів далі проводиться із використанням формул 13.19-13.28.

Контрольні питання:

1. Умови застосування ієрархічного дисперсійного аналізу.
2. Розрахунок сили впливу фактора.

§ 14. Фенетичний аналіз структурованих популяцій

1973 рік вважається відправною точкою сучасного етапу фенетики популяцій. Саме цього року в журналі «Природа» було надруковано статтю М. В. Тимофєєва-Ресовського та О. В. Яблокова по фенетиці популяцій і вийшла друком книга «Нарис вчення про популяцію» під авторством М. В. Тимофєєва-Ресовського зі співавторами, в якій був розділ, присвячений фенетиці популяцій.

А вже у 1976 р. у м. Саратові на базі Саратовського державного університету пройшла Перша Всесоюзна нарада по фенетиці популяцій. Міжвузівський збірник «Фізіологічна та популяційна екологія тварин» за матеріалами цієї наради було опубліковано лише в 1978 р.

У 1979 р. у м. Москві на базі Інституту біології розвитку пройшла Друга Всесоюзна нарада по фенетиці популяцій. Нарешті, у 1980 р. відбулася публікація книги О. В. Яблокова «Фенетика: еволюція, популяція, ознака», що являє собою перший спеціальний збірник по фенетиці популяцій.

Перший навчальний посібник по фенетиці популяцій було видано через п’ять років під авторством О. В. Яблокова та Н. І. Ларіної і він має назву «Введення у фенетику популяцій: Новий підхід до вивчення природних популяцій».

Починаючи із 1982 р. практично кожні 2-3 роки виходили матеріали нарад по фенетиці популяцій (1985, 1988, 1990 р.). Останнє видання подібного плану («Популяційна фенетика») вийшло в 1997 р., хоча у вигляді окремих публікацій роботи із фенетики популяцій продовжують з’являтися майже безперервно.

Якщо проаналізувати питання, що найчастіше аналізуються в дослідженнях із фенетики популяцій, то на перше місце, звичайно ж, виходять роботи, присвячені каталогізації фенів у межах окремих таксономічних груп.

Друге питання, що найбільш часто аналізують автори – це поширення тих чи інших фенів у межах досліджуваного регіону (тобто, стосується питань феногеографії) і пошук можливих зв’язків із факторами зовнішнього середовища.

Особливий клас представляють роботи порівняльного плану, в яких проводиться аналіз фенетичної структури фонових видів в умовно чистих місцях існування і в забруднених (найчастіше, урбаноценозах).

Практично весь математико-статистичний апарат, що використовується у дослідженнях із фенетики популяцій, базується на методиках оцінки середнього числа морф і частки рідкісних морф, розроблених Л. А. Животовським (1982; 1991) і використанні критерія Хі-квадрат К. Пірсона (чи його модифікацій для нечисленних частот). Інколи використовується формула К. Шеннона чи інші методи (класичного) статистичного аналізу.

Можливо, це стало однією з причин зниження уваги до досліджень у сфері фенетики популяцій.

Паралельно з цим, в останні 10-15 років, навпаки, великого розвитку досягли методи генетичного аналізу популяцій, насамперед, стосовно характеру та ступеня їхньої структурованості.

У 1951 р. у своїй (тепер вже класичній) роботі С. Райт (Wright, 1951) вводить поняття F-статистик, що можуть бути використані як для оцінки рівня інбридингу в популяції, так і для оцінки міри генетичної диференціації субпопуляцій у межах однієї підрозділеної популяції. При цьому, всі розрахунки С. Райта базуються на моделі одного локусу з двома алелями.

Бурхливий розвиток у 60-х роках минулого сторіччя методів електрофорезу білків призвів до необхідності подальшого розвитку уявлень С. Райта і це було зроблено М. Неєм (Nei, 1973; 1978). Він розробив і запропонував аналоги F-статистик, що можуть бути використані у випадку поліалельних систем (якими, наприклад, є майже всі ізоферменти).

Крім того, Б. Вейр і С. Кокерхем (Weir, Cockerham, 1984) розробили методики оцінки F-статистик, ґрунтуючись на алгоритмі дисперсійного аналізу частот алелів, генотипів і фенотипів у популяціях чи їхніх структурних одиницях.

Подальша еволюція F-статистик відбувалася на фоні переходу досліджень від рівня макромолекул до рівня структури ДНК і аналізу послідовності нуклеотидів. Для аналізу молекулярної мінливості Екскоффієром та співавторами (Excoffier et al., 1992) було розроблено принципово новий метод, що за аналогією з Дисперсійним аналізом (ANOVA в англомовній літературі) одержує назву AMOVA (**A**nalysis of **Mo**lecular **Va**riance), тобто Аналіз Молекулярної Мінливості.

Крім співзвучної назви, AMOVA має ще багато спільного із класичним дисперсійним аналізом і, насамперед – це той же принцип розкладання сумарної мінливості на окремі компоненти, що був запропонований Р. Фішером ще в 1925 р.

Таким чином, загальна методологія дисперсійного аналізу в останні 20 років зазнала значного розширення, в результаті чого суттєво збільшилася сфера його використання.

Фенетична мінливість (та її елементарна одиниця – фен із рядом морф) залишалася незадіяною на фоні бурхливого росту уваги до проблем структурованості природних та штучних популяцій.

Однак, фен являє собою елементарну, дискретну, генетично зумовлену ознаку й у цьому сенсі має всі якості, що необхідні для його використання у якості об’єкта в дослідженнях подібного плану.

Єдиною проблемою виходу фена і фенетики на принципово новий рівень (із описового на аналітичний) була відсутність необхідного математико-статистичного апарату.

Очевидно, вперше метод оцінки географічної мінливості фенетичного складу групи природних популяцій із застосуванням дисперсійного аналізу якісних ознак було використано в роботі Є. М. Панова зі співавторами (1993). Принаймні, інших прикладів нами не знайдено. При цьому автори використали

алгоритм однофакторного дисперсійного аналізу якісної ознаки (фена), що описаний у підручнику «Біометрія» (с. 302) Г. Ф. Лакіна, опублікованому в 1973 р. Однак, вже наступні перевидання цього підручника в 1980 р. та 1990 р. методики дисперсійного аналізу якісних ознак не містять. Немає їх і в інших класичних та базових підручниках і посібниках із біометрії для біологів.

Однак, при цьому необхідно віддати належне М. О. Плохинському (1969; 1970), який практично в кожній своїй книзі приводить алгоритми різних варіантів дисперсійного аналізу якісних ознак (одно-, двох- та навіть трифакторного).

Швидше за все, таке відношення до дисперсійного аналізу якісних ознак можна пояснити «конкуренцією» із боку критерію Хі-квадрат (та його аналогів), що дуже часто вирішує ті ж завдання, але з меншими витратами часу. Використання критерію Мантеля-Хейзеля, а також бурхливий розвиток (на фоні зростання досконалості та поширеності комп'ютерної техніки) лог-лінійних моделей також «відтісняють» дисперсійний аналіз якісних ознак, і цей метод практично не використовується в сучасних наукових дослідженнях. (Ми знайшли тільки два згадування про використання цієї процедури при пошуку в мережі Інтернет, проведеному за допомогою пошукової системи Google. Обидва вони мали відношення до медичних досліджень. Варто окремо згадати, що в медичних дослідженнях значно вагомішу роль відіграють методи некількісного та непараметричного аналізу через специфічність самого об'єкта вивчення).

Запропонований алгоритм дисперсійного аналізу якісних ознак і його різноманітні варіації, викладені в §§ 11-13, на нашу думку, можуть надати другий подих дослідженням із фенетики популяцій (особливо у випадку їх складної ієрархічної структурованості, як це часто буває на практиці).

Нижче ми приводимо методики і приклади використання фенетичного аналізу структурованих популяцій. Оскільки деякі методи розрахунку можуть бути реалізовані лише за допомогою спеціальних статистичних комп'ютерних програм, ми також наводимо їхні специфікації (описи та адреси сайтів, де їх можна знайти і завантажити free-версії).

Загальна послідовність аналізу має наступний вигляд:

1. Оцінка показника фенетичної диференціації (P_{ST}) популяції, використовуючи методику дисперсійного аналізу.

2. Оцінка довірчого інтервалу отриманого показника P_{ST} із використанням розподілу Хі-квадрат, *F*-розподілу Фішера-Снедекора і методів ресамплінгу.

3. Оцінка варіанси і статистичної помилки показника P_{ST} із використанням методів ресамплінгу (jackknife-процедура).

4. Розрахунок парних оцінок P_{ST} між кожною парою субпопуляцій (підготовка матриці фенетичних дистанцій).

5. Перевірка моделі «ізоляції відстанню» (Isolation-by-distance; IBD) шляхом оцінки коефіцієнта кореляції між матрицею фенетичних дистанцій і

матрицею географічних відстаней між субпопуляціями, використовуючи тест Мантеля (Mantel test).

6. Розрахунок коефіцієнтів (і оцінка рівня їхньої значущості) лінійної регресії між величинами $P_{ST}/(1-P_{ST})$ матриці фенетичних дистанцій і логарифмами відповідних географічних відстаней між субпопуляціями.

7. Оцінка коефіцієнтів просторової автокореляції (*I* Морана та *C* Джирі).

Аналіз просторової фенетичної структури наземного молюска *Brephulopsis bidens*

У 1989 р. нами було вивчено одну популяцію наземних молюсків *B.bidens*, розташовану в передмісті м. Сімферополя. Молюски *B.bidens* формували більш-менш численні скупчення в заростях густої трав'янистої рослинності. У червні 1989 р. нами було зібрано (на ділянках площею 1,0-1,5 м^2) 12 вибірок молюсків *B.bidens* у визначеному порядку (рис. 14.1). Відстань між вибірками складало близько 4 метрів. Але одна вибірка (№ 1) не збереглася, тому в аналіз ми включили лише 11, що залишилися.

Рисунок 14.1 – Розташування вибірок *B.bidens* у просторі дослідної популяції

Кожна вибірка містила 19-93 статевозрілих особини *B.bidens*, дані стосовно яких було проаналізовано. Для кожної вибірки було підраховано кількість особин, що мають на черепашках радіальні темні пігментні смужки (Пигм$^+$). На основі цих матеріалів і проводилися подальші дослідження фенетичної структури даної локальної популяції й оцінка рівня її фенетичної диференціації (табл. 14.1).

Таблиця 14.1 – Розподіл особин, що мають пігментні смужки на черепашці

Показник	Субпопуляція										
	2	3	4	5	6	7	8	9	10	11	12
Кількість особин зі смужками	58	43	37	35	13	27	46	17	31	12	36
Обсяг вибірки	93	67	60	49	24	59	61	25	58	19	44
Частка особин зі смужками	0,624	0,642	0,617	0,714	0,542	0,458	0,754	0,680	0,534	0,632	0,818

Відносні частоти даної ознаки свідчать про її значну просторову мінливість у межах вивченої популяції, що може бути виражено в числовій формі у вигляді оцінки P_{ST}.

Розрахунок даного показника проводиться на підставі алгоритму дисперсійного аналізу якісних ознак (табл. 14.2).

Таблиця 14.2 – Алгоритм дисперсійного аналізу якісних ознак

Джерело мінливості	Варіанса (H)	Число ступенів свободи (df)	Середній квадрат (MS)	Компоненти середнього квадрата $E(MS)$
За рахунок мінливості між субпопуляціями	H_B	$df_B = s - 1$	MS_B	$MS_S + n \cdot MS_B$
За рахунок мінливості всередині субпопуляцій	H_S	$df_S = N - s$	MS_S	MS_S
Загальна мінливість	H_T	$df_T = N - 1$		

У випадку двох альтернативних варіацій ознаки, оцінки загальної та залишкової варіанс знаходяться за відповідними формулаами:

$$H_T = N \cdot \overline{p} \cdot (1 - \overline{p})$$
$$H_S = \sum_{i=1}^{s} n_i \cdot p_i \cdot (1 - p_i) \qquad (14.1)$$

де N – сумарний обсяг всіх об’єктів дисперсійного комплексу: $N = \sum n_i$;

$\overline{p}$ – середня зважена частота ознаки для всієї вибірки;

s – число груп (вибірок), чисельність кожної з яких становить n_i (де $1 \le i \le s$);

p_i – відносна частота шуканої ознаки в i-тій вибірці.

Тоді оцінка показника фенетичної диференціації проводиться за наступною формулою:

$$P_{ST} = \frac{MS_B - MS_S}{MS_B + (n-1) \cdot MS_S}, \qquad (14.2)$$

де n – обсяг кожної вибірки, у випадку їхньої рівності.

Якщо обсяги вибірок не однакові, то використовують наступну величину:

$$n^* = \frac{1}{s-1} \cdot \left[N - \frac{\sum_{i=1}^{s} n_i^2}{N} \right]. \qquad (14.3)$$

При розрахунку варіанси H_S, зручніше спочатку розрахувати для кожної вибірки значення $n_i \times p_i \cdot (1 - p_i)$, а потім їх додати. Оцінку H_B простіше знайти як різницю $H_T - H_S$.

Таким чином, для наших даних по 11 вибірках наземного молюска *B.bidens* заповнена таблиця дисперсійного аналізу має наступний вигляд (табл. 14.3).

Таблиця 14.3 – Результати дисперсійного аналізу

Джерело мінливості	*H*	*df*	*MS*	*F*	*p*
B	5,387	10	0,5387	2,378	0,0092
S	124,166	548	0,2266		
T	129,553	558			

Вірогідне значення критерію Фішера-Снедекора свідчить про те, що нульова гіпотеза (щодо рівності частот ознаки у всіх 11 вибірках) повинна бути відхилена з імовірністю прийняття помилкового рішення $p = 0{,}0092$, тобто, в 9 випадках із 1000 (статистично дуже мала величина ймовірності).

Оскільки обсяги вибірок неоднакові, спочатку за формулою (14.3) необхідно знайти значення *n*:*

$$n^* = \frac{1}{11-1} \cdot \left[559 - \frac{(93^2 + 67^2 + ... + 44^2)}{559} \right] = 49{,}96 \ .$$

Тоді оцінка індексу фенетичної диференціації молюсків *B.bidens* у межах 11 вибірок вивченої популяції буде дорівнювати:

$$P_{ST} = \frac{0{,}5387 - 0{,}2266}{0{,}5387 + (49{,}96 - 1) \cdot 0{,}2266} = 0{,}0268 \ .$$

У графічному вигляді фенетична структурованість даної популяції представлена на рисунку 14.2.

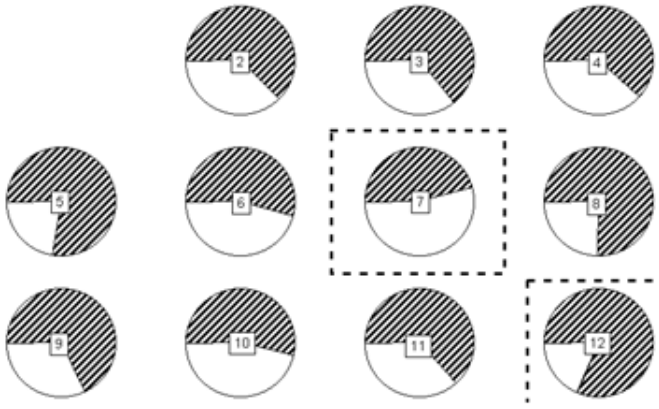

Рисунок 14.2 – Частота зустрічання черепашок із пігментними смужками (заштрихований сектор) у локальній популяції молюска *B.bidens*

Як бачимо, вся територія популяції досить гомогенна, однак у вибірці № 7 частота особин зі смужками на черепашці менша, а у вибірці № 12 – більша, ніж в інших вибірках із цієї популяції.

Оцінка 95% довірчого інтервалу для отриманого показника фенетичної диференціації дослідженої популяції *B.bidens* P_{ST} може бути зроблена трьома різними способами. Вони дають різні оцінки нижньої і верхньої меж.

Оцінка 95% довірчого інтервалу для P_{ST}, використовуючи розподіл Хі-квадрат.

Нижня $P_{ST}(L)$ і верхня $P_{ST}(U)$ довірчі межі в цьому випадку розраховуються за формулами:

$$P_{ST}(U) = \frac{(s-1)\cdot P_{ST}}{\chi^2_U \cdot (1-P_{ST}) + (s-1)\cdot P_{ST}}, \quad (14.4)$$

$$P_{ST}(L) = \frac{(s-1)\cdot P_{ST}}{\chi^2_L \cdot (1-P_{ST}) + (s-1)\cdot P_{ST}}, \quad (14.5)$$

де χ^2_U – табличне значення розподілу Хі-квадрат для α = 0,975 і числа ступенів свободи $df = s - 1$;

χ^2_L – табличне значення розподілу Хі-квадрат для α = 0,025 і числа ступенів свободи $df = s - 1$.

Для даної популяції наземного молюска *B.bidens* довірчі межі показника фенетичної диференціації, відповідно, будуть дорівнювати:

$$P_{ST}(U) = \frac{(11-1)\cdot 0{,}0268}{3{,}247\cdot(1-0{,}0268)+(11-1)\cdot 0{,}0268} = 0{,}0782,$$

$$P_{ST}(L) = \frac{(11-1)\cdot 0{,}0268}{20{,}483\cdot(1-0{,}0268)+(11-1)\cdot 0{,}0268} = 0{,}0133.$$

Таким чином, інтервальна оцінка шуканого показника становить: [0,0133; 0,0782].

Оцінка 95% довірчого інтервалу для P_{ST}, використовуючи F-розподіл Фішера-Снедекора.

Нижня $P_{ST}(L)$ і верхня $P_{ST}(U)$ довірчі межі в цьому випадку розраховуються за формулами:

$$P_{ST}(U) = \frac{1}{1+\frac{1}{R}}, \quad (14.6)$$

$$P_{ST}(L) = \frac{1}{1+\frac{1}{L}}, \quad (14.7)$$

де

$$L = \frac{1}{n^*}\cdot\left(\frac{F}{F_2}-1\right); \quad (14.8)$$

$$R = \frac{1}{n^*}\cdot\left(\frac{F}{F_1}-1\right), \quad (14.9)$$

де F – оцінка дисперсійного відношення, отримана в результаті проведення дисперсійного аналізу;

F_1 – табличне значення критерію Фішера-Снедекора для $\alpha = 0{,}025$; $df_1 = s - 1$; $df_2 = N - s$;

F_2 – табличне значення критерію Фішера-Снедекора для $\alpha = 0{,}975$; $df_1 = s - 1$; $df_2 = N - s$.

Значення n^* розраховується за формулою (14.3).

Для даної популяції наземного молюска *B.bidens* довірчі межі показника фенетичної диференціації, відповідно, будуть наступні:

$$L = \frac{1}{49{,}96} \cdot \left[\frac{2{,}378}{2{,}072} - 1 \right] = 0{,}0030 \; ; \; R = \frac{1}{49{,}96} \cdot \left[\frac{2{,}378}{0{,}323} - 1 \right] = 0{,}1273 ,$$

$$P_{ST}(L) = \frac{1}{1 + \frac{1}{0{,}0030}} = 0{,}0030 \; ; \; P_{ST}(U) = \frac{1}{1 + \frac{1}{0{,}1273}} = 0{,}1129 .$$

Таким чином, інтервальна оцінка шуканого показника становить: [0,0030; 0,1129].

Оцінка 95% довірчого інтервалу для P_{ST}, використовуючи бутстреп-процедуру

Межі 95% довірчого інтервалу для оцінки P_{ST} можуть бути отримані, використовуючи один із методів ресамплінгу, а саме бутстреп-процедуру. Як ми вже вказували, основою цього методу є принцип відбору вибірок із поверненням та формування багатьох штучних вибірок (псевдовибірок) обсягом n із однієї і тієї ж емпіричної сукупності такого ж обсягу.

Наприклад, у вибірку із субпопуляції № 2 потрапило 93 особини равликів, з яких 58 мали пігментні смужки на черепашці. Нам необхідно сформувати псевдовибірку, що буде також містити 93 особини. В ідеальному випадку, нам необхідно із цих вибраних 93 особин обрати першу, яка трапиться, відмітити її фен (має чи не має смужки на черепашці) та повернути назад у вибірку. Далі відібрати другу особину також випадковим чином, відмітити її фен і також повернути назад у вибірку. Аналогічним чином зробити ще 91 раз. У підсумку, ми отримаємо першу псевдовибірку для вибірки субпопуляції № 2.

Аналогічним чином необхідно сформувати і другу, і третю, і четверту, і інші псевдовибірки для цієї популяції, а також для кожної із решти 10 субпопуляцій.

На практиці це робиться наступним чином. Оскільки ми маємо вибіркову частоту (наприклад, для субпопуляції № 2 вона становить 0,624; див. табл. 14.1), а також обсяг вибірки (93 особини), то можна скористатися генератором псевдовипадкових чисел, наприклад, що вбудований у табличний редактор MS Excel. За допомогою цього генератора можна дуже легко генерувати подібні псевдовибірки, що будуть мати біноміальний розподіл (див. §2) із заданими нами параметрами (наприклад, для субпопуляції № 2 ці параметри

становитимуть: p = 0,624; n = 93). Краще, щоб таких псевдовибірок для кожної субпопуляції було декілька сотень чи навіть тисяч. Наприклад, ми згенерували по 1000 таких псевдовибірок для кожної субпопуляції (табл. 14.4).

Таблиця 14.4 – Розподіл особин у 1000 псевдовибірках, що мають пігментні смужки на черепашці

Псевдо-вибірки	Номер субпопуляцій											P_{ST}
	2	3	4	5	6	7	8	9	10	11	12	
1	61	44	39	35	17	28	48	19	25	11	37	0,0513
2	60	51	35	36	12	32	43	13	32	11	31	0,0145
3	49	42	41	33	15	28	47	18	36	14	33	0,0228
4	66	40	38	33	14	35	47	16	29	7	38	0,0381
5	61	51	36	34	12	28	45	18	21	13	37	0,0702
6	57	51	33	39	15	27	49	16	33	10	34	0,0472
7	61	48	30	35	12	23	43	14	32	10	34	0,0434
8	53	42	34	36	12	34	45	17	37	12	38	0,0218
9	51	46	42	33	13	30	37	17	38	10	38	0,0235
10	56	38	42	33	15	31	42	20	35	15	33	0,0093
…												
1000	60	42	32	38	12	25	48	19	35	12	33	0,0420

В останньому стовпці таблиці 14.4 наведено відповідні оцінки (псевдооцінки) показника фенетичної диференціації P_{ST}, що було розраховано аналогічно тому, як це було зроблено вище для емпіричних даних.

В якості нижньої та верхньої межі 95% довірчого інтервалу оцінки P_{ST}, можна тепер використати 2,5% та 97,5% перцентилі для розподілу псевдооцінок P_{ST}. У нашому випадку 95% довірчий інтервал має наступний вигляд: [0,0121; 0,0895].

Порівняльний аналіз процедур оцінювання меж довірчого інтервалу оцінки P_{ST} показує, що найбільш стійка оцінка та, що отримана при використанні бутстреп-процедури, а найменш ефективний метод – той, що базується на використанні розподілу Хі-квадрат, який у випадку негативних значень вибіркової оцінки P_{ST} дає абсурдні значення.

Оцінювання варіанси та статистичної помилки вибіркової оцінки P_{ST}

Методи ресамплінгу також незамінні і в тому випадку, коли нам необхідно оцінити варіансу та статистичну помилку показника фенетичної диференціації.

Продемонструємо, як можна провести необхідне оцінювання з використанням одного з методів ресамплінгу, а саме jackknife-методу чи методу «складного ножа» (від англ. jackknife – складний ніж).

Для цього по-перше необхідно одержати вибірку зі значень jackknifing-оцінок P_{ST}. Це можна зробити таким способом. Видаляємо із таблиці вихідних значень (табл. 14.1) перше значення (для субпопуляції № 2, тобто, 58 особин із

пігментними смужками із 93 проаналізованих особин) і розраховуємо значення P_{ST}, ґрунтуючись на 10 значеннях, що залишилися. Далі видаляємо значення для субпопуляції № 3 (43 із 67) і знову розраховуємо значення P_{ST}. Так повторюємо 11 разів (стільки, скільки маємо субпопуляцій) і одержуємо 11 псевдооцінок P_{ST} (табл. 14.5).

Таблиця 14.5 – Псевдооцінки P_{ST}, отримані при видаленні значень для кожної субпопуляції послідовно

Субпопуляція										
2	3	4	5	6	7	8	9	10	11	12
0,0349	0,0331	0,0327	0,0283	0,0286	0,0130	0,0225	0,0298	0,0268	0,0300	0,0157

Середнє значення $\overline{P}_{ST}$ і варіансу $Var(P_{ST})$ можна тепер оцінити, ґрунтуючись на цій вибірці jackknife-оцінок P_{ST}:

$$\overline{P}_{ST} = s \cdot P_{ST} - \frac{s-1}{s} \cdot \sum_{i=1}^{s} P_{ST}^{i}, \qquad (14.10)$$

$$Var(P_{ST}) = \frac{s-1}{s} \cdot \sum_{i=1}^{s} \left[P_{ST}^{i} - \frac{\sum_{i=1}^{s} P_{ST}^{i}}{s} \right]^2, \qquad (14.11)$$

де P_{ST}^{i} – jackknife-оцінка P_{ST}, отримана в результаті видалення значень для *i*-тої субпопуляції (тобто значення, приведені в табл. 14.5).

Для аналізованої популяції *B.bidens* відповідні значення будуть наступні:

$$\overline{P}_{ST} = 11 \cdot 0{,}0268 - \frac{11-1}{11} \cdot 0{,}2954 = 0{,}0263,$$

$$Var(P_{ST}) = \frac{11-1}{11} \cdot \left[\left(0{,}0349 - \frac{0{,}2954}{11} \right)^2 + \ldots + \left(0{,}0157 - \frac{0{,}2954}{11} \right)^2 \right] = 0{,}000452.$$

Статистична помилка отриманої оцінки P_{ST} розраховується як корінь квадратний з варіанси; у нашому випадку ця величина дорівнює:

$$SE_{\overline{P}_{ST}} = \sqrt{Var(P_{ST})} = \sqrt{0{,}000452} = 0{,}0213. \qquad (14.12)$$

Таким чином, результати дисперсійного аналізу (табл. 14.3) свідчать про те, що молюски із 11 субпопуляцій, включених до аналізу, вірогідно відрізняються за частотою даної ознаки (отримане значення дисперсійного відношення F = 2,378 із рівнем значущості p = 0,0092). Вибіркова оцінка показника фенетичної диференціації досліджуваної популяції *B.bidens* стосовно ознаки наявність/відсутність на черепашці пігментних смужок $P_{ST} \pm SEP_{ST}$ = 0,0268±0,0213 із довірчим інтервалом [0,0121; 0,0895]. Оскільки нуль не потрапляє в межі 95% довірчого інтервалу, відповідно нуль-гіпотеза щодо відсутності фенетичної структурованості повинна бути відхилена.

Отже, доведено наявність її фенетичної диференціації.

Як ми бачимо на рис.14.2, дана гетерогенність, насамперед, обумовлена «випаданням» із загальної закономірності субпопуляцій № 7 і № 12.

Однак, якщо поставити перед собою більш широкі цілі, то можна задатися питанням про можливу присутність просторової структурованості популяції.

Природно, нульовою гіпотезою (яку нам і варто перевірити) буде твердження про те, що наша популяція не має ніякої просторової структури, тобто, значення частот ознаки в різних її частинах (субпопуляціях) є результатом випадкових процесів.

Альтернативна гіпотеза, навпаки, буде стверджувати, що популяція просторово структурована, тобто, більш близькі субпопуляції будуть більш подібні одна до одної стосовно частот фена, ніж більш віддалені. Визначається це, насамперед, більш високою частотою обміну генетичною інформацією між сусідніми субпопуляціями (тобто, має місце потік генів – gene flow). Дана модель популяційної структури має назву моделі «ізоляції відстанню» (isolation-by-distance; IBD). Її ми і будемо перевіряти.

У найбільш загальному випадку, модель IBD можна перевірити на підставі рівня значущості коефіцієнта кореляції між відповідними елементами двох матриць – матриці попарних значень P_{ST} і матриці географічної дистанції між кожною парою субпопуляцій.

Однак використовувати в цьому випадку коефіцієнт парної лінійної кореляції Пірсона-Браве (або навіть непараметричні коефіцієнти кореляції Кендалла і Спірмена) не можна, оскільки елементи матриць не є незалежними випадковими величинами, як того вимагає теорія і сфера застосування даних коефіцієнтів.

Єдиним прийнятним у даній ситуації статистичним критерієм є тест Мантеля, що саме і призначений для оцінки рівня зв’язку між елементами двох матриць (звичайно, однієї розмірності).

Матриця значень P_{ST} між кожною парою субпопуляцій представлена в таблиці 14.6. (Жирним курсивом у ній виділено вірогідні значення P_{ST}).

Таблиця 14.6 – Матриця значень P_{ST} між кожною парою субпопуляцій

Субпопуляція	Субпопуляція									
	2	3	4	5	6	7	8	9	10	11
3	-0,0123	X								
4	-0,0138	-0,0147	X							
5	0,0025	-0,0059	0,0025	X						
6	-0,0125	-0,0075	-0,0180	0,0340	X					
7	***0,0411***	***0,0513***	0,0334	***0,1095***	-0,0157	X				
8	0,0248	0,0139	0,0268	-0,0146	0,0736	***0,1548***	X			
9	-0,0190	-0,0249	-0,0203	-0,0282	-0,0014	0,0676	-0,0146	X		
10	0,0023	0,0076	-0,0033	0,0482	-0,0302	-0,0054	***0,0849***	0,0143	X	
11	-0,0326	-0,0347	-0,0354	-0,0213	-0,0322	0,0249	0,0027	-0,0431	-0,0614	X
12	***0,0683***	***0,0552***	***0,0743***	0,0081	***0,1450***	***0,2248***	-0,0078	0,0215	***0,1467***	0,0559

Оцінка рівня значущості показника P_{ST} у випадку використання тільки двох вибірок (субпопуляцій) може бути отримана шляхом розрахунку величини:

$$\chi^2 = N \cdot \left(1 - \frac{H_S}{H_T}\right), \qquad (14.13)$$

яка має розподіл Хі-квадрат з числом ступенів свободи $df = 1$.

Наприклад, при порівнянні субпопуляцій № 2 і № 7 відповідна оцінка дорівнює: $\chi^2 = 152 \cdot \left(1 - \frac{36{,}472}{37{,}467}\right) = 4{,}037$, а при порівнянні субпопуляцій № 7 і № 9: $\chi^2 = 84 \cdot \left(1 - \frac{20{,}084}{20{,}952}\right) = 3{,}480$.

Тому в першому випадку відзначається вірогідна відмінність між субпопуляціями стосовно частоти ознаки, що аналізується (оскільки для $df = 1$ табличне значення $\chi^2_{\alpha=0{,}05} = 3{,}84$), а в другому – ні, хоча за абсолютним значенням ступінь фенетичної диференціації між субпопуляціями № 7 і № 9 більший, ніж між № 2 і № 7 (табл. 14.6).

Матриця географічної відстані між кожною парою субпопуляцій по прямій лінії може бути побудована після безпосереднього виміру відстані між місцем розташування кожної пари субпопуляцій на картосхемі (із дотриманням масштабу).

Якщо ж, з іншого боку, є координати всіх субпопуляцій на тій же картосхемі (звичайно, у прямокутній системі координат; географічні широту і довготу використовувати не можна), то відстані можуть бути розраховані і дотримуючись закону Піфагора.

У таблиці 14.7 наведено відстані між кожною парою субпопуляцій *B.bidens* у просторі популяції, що аналізується.

Таблиця 14.7 – Відстані між кожною парою субпопуляцій *B.bidens* у просторі популяції, що аналізується

Субпопуляція	Субпопуляція									
	2	3	4	5	6	7	8	9	10	11
3	4									
4	8	4								
5	5,5	8,5	12,5							
6	4	5,5	9	4						
7	5,5	4	5,5	8	4					
8	8,5	5,5	4	12	8	4				
9	9	11,5	14	4	5,5	9	12,5			
10	8	9	11,5	5,5	4	5,5	9	4		
11	9	8	9	9	5,5	4	5,5	8	4	
12	11,5	9	8	12	9	5,5	4	12	8	4

Значення коефіцієнта кореляції Мантеля між відповідними елементами даних матриць (трикутних) можна одержати за формулою:

$$R_M = \frac{1}{n-1} \cdot \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{(x_{ij} - \bar{x})}{\sigma_x} \cdot \frac{(y_{ij} - \bar{y})}{\sigma_y} \qquad (14.14)$$

де x_{ij}, y_{ij} – відповідні елементи матриць *X* і *Y*;

$\bar{x}$, $\bar{y}$ – середні значення для всіх елементів матриць *X* і *Y*, відповідно;

σ_x, σ_y – середні квадратичні значення для всіх елементів матриць X і Y, відповідно;

n – кількість елементів у кожній матриці (= $s \cdot (s - 1)/2$).

Таким чином, значення даного показника одержати дуже просто, навіть не маючи спеціального програмного забезпечення, а навіть у MS Excel. Для цього необхідно всі елементи матриць виписати в два стовпці (кількість елементів у цих вибірках буде для нашого прикладу $n = 11(11 – 1) / 2 = 55$).

Потім від кожного елемента кожної вибірки необхідно відняти середнє значення по відповідній вибірці і поділити на середнє квадратичне відхилення відповідної вибірки (тобто, в такий спосіб ми переходимо до стандартизованих величин). Знайти суму добутків пар цих величин і поділити цю суму на $n – 1$ (у нашому випадку, на 54).

Для даних із аналізованої популяції *B.bidens* коефіцієнт кореляції Мантеля буде дорівнювати:

$$R_M = \frac{1}{55-1} \cdot (-2{,}94738) = -0{,}055.$$

Однак оцінити рівень значущості даної величини вже набагато складніше, оскільки відсутня формула для розрахунку статистичної помилки коефіцієнта кореляції Мантеля і, відповідно, не може бути використана стандартна формула, що використовує розподіл Ст’юдента.

Оцінка рівня значущості отриманої величини коефіцієнта Мантеля проводиться таким способом. Один із рядків чи один зі стовпців матриці географічних відстаней міняється місцями з іншим рядком (чи стовпцем). Для отриманої таким способом матриці географічних відстаней і матриці парних значень P_{ST} розраховується коефіцієнт Мантеля.

Потім проводиться перестановка ще двох випадковим чином обраних рядків (чи стовпців) і знову розраховується коефіцієнт Мантеля. Дана операція проводиться 1000, 5000 чи 10000 разів.

Серед отриманих таким чином випадкових оцінок коефіцієнта Мантеля підраховують кількість таких, котрі є більшими чи дорівнюють вибірковій оцінці даного коефіцієнта для матриць вихідних даних.

Відношення цього числа до числа перестановок і є оцінкою рівня значущості вибіркового коефіцієнта Мантеля. Чим менше це число – тим нижча вірогідність того, що при випадковому розміщенні елементів матриць буде отримано таке ж саме (чи навіть вище) значення оцінки коефіцієнта кореляції.

Даний метод оцінювання рівня значущості коефіцієнта Мантеля називається permutation-процедурою.

Виконати її вручну практично неможливо, тому для оцінки коефіцієнта Мантеля необхідно спеціальне програмне забезпечення. Ми рекомендуємо для цього програму **GenAIEx** v.6 (Genetic Analysis in Excel). Однією з її переваг є те, що вона вбудовується в MS Excel і може використовувати ті ж вихідні дані. Тобто, їй не потрібний спеціальний формат файлів з даними (як у багатьох інших програмах, що мають ті ж можливості). Іншою перевагою даної програми є її вільне поширення. Її free-версію можна завантажити із сайту авторів: www.anu.edu.au/BoZo/GenAIEx.

Ми скористалися даною програмою і здійснили 9999 перестановок. У підсумку ми одержали рівень значущості для коефіцієнта Мантеля: p = 0,378.

Таким чином, 3780 оцінок коефіцієнта Мантеля, отриманих для випадковим чином сформованих матриць, перевищували чи дорівнювали вибірковій оцінці (-0,055). Це свідчить про те, що нульова гіпотеза про відсутність зв'язку між відповідними елементами двох аналізованих матриць не може бути відхилена.

Іншими словами, ми не одержали доказів моделі IBD і розподіл частот черепашок молюска *B.bidens* із пігментними смужками в межах досліджуваної популяції носить випадковий характер.

З іншого боку, модель IBD може бути перевірена, використовуючи модель лінійної регресії, де в якості залежної змінної виступають величини $P_{ST}/(1-P_{ST})$, розраховані на підставі значень відповідної матриці (табл. 14.6), а в якості незалежної – натуральні логарифми попарних відстаней між субпопуляціями, отриманих з таблиці 14.7.

На рисунку 14.3 наведено графік лінійної регресії ($Y = a + b \times X$) між відповідними показниками, а в таблиці 14.8 – коефіцієнти даної лінії регресії, отримані стандартним методом (методом найменших квадратів; використано програму **STATISTICA** v.5.5) і за допомогою bootstrap-процедури (використовуючи 1000 повторних вибірок; використано програму **S-PLUS**).

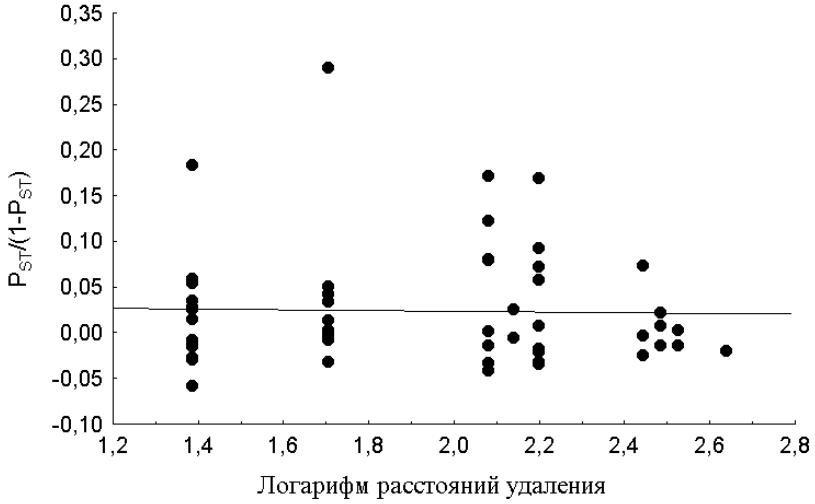

Рисунок 14.3 – Графік лінійної регресії, використаної для перевірки гіпотези IBD

Таблиця 14.8 – Коефіцієнти лінії регресії, отримані різними методами

Коефіцієнти регресії	Метод найменших квадратів	Bootstrap-процедура	
		оцінки	довірчий інтервал
a	0,0314 ± 0,0427	0,0335 ± 0,0348	[-0,0248; 0,1087]
b	-0,0040 ±0,0219	-0,0048 ± 0,0170	[-0,0370; 0,0280]

Оцінки, отримані двома способами досить близькі й вірогідно не перевищують нуль. На підставі даних оцінок ми також змушені відхилити нуль-гіпотезу щодо придатності моделі IBD у випадку даної популяції *B.bidens*.

Тест Мантеля, однак, не дає повної інформації про характер структурованості популяції, а лише служить для перевірки гіпотези, що близько розташовані регіони більш подібні за феноструктурою, а більш віддалені – менш подібні. Таким чином, він дає адекватне відображення структури популяції у випадку наявності більш-менш вираженого градієнта частот ознаки. У випадку складної, «плямистої» структури, коли відносини між сусідніми ділянками мають «нелінійний» характер, більш точну картину можна одержати, використовуючи показники просторової автокореляції (Spatial autocorrelation), наприклад, **коефіцієнт *I* Морана.**

Просторова автокореляція подібна коефіцієнту автокореляції, що використовується в аналізі часових рядів. Тільки у випадку просторової автокореляції проводиться оцінка ступеня зв'язку між різними значеннями тієї ж самої ознаки, що просторово зміщена одна від одної на попередньо задану величину. Ця величина (так само як і в аналізі часових рядів) зветься «лаг» (від англійського «lag» – зсув) і має своє вираження в одиницях довжини (наприклад, метрах).

У загальному вигляді формула для розрахунку коефіцієнта Морана має наступну структуру:

$$I = \left(\frac{n}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}} \right) \cdot \left(\frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij} \cdot (x_i - \bar{x}) \cdot (x_j - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right), \qquad (14.15)$$

де n – число точок чи просторових одиниць (у нашому випадку, субпопуляцій);

x – значення змінної (у нашому випадку, це – частота фена);

$\bar{x}$ – середнє значення ознаки по всій сукупності значень;

w_{ij} – «вага», що відбиває ступінь близькості (чи далекості) між точками i і j у просторі.

У найбільш простому варіанті в якості вагової змінної може бути використано відстань між кожною парою точок у прямокутній системі координат (див., наприклад, табл. 14.7).

У випадку відсутності будь-якої просторової автокореляції, значення коефіцієнта Морана буде близьким до величини:

$$E(I) = -\frac{1}{n-1}. \qquad (14.16)$$

Вибіркові значення коефіцієнта Морана можуть набувати будь-яких значень у інтервалі від –1 до +1 (як і коефіцієнти кореляції Пірсона-Браве, Спірмена чи Кендалла). Додатні значення *I* свідчать про наявність позитивної просторової автокореляції (тобто, просторово більш близькі райони більш подібні стосовно частоти ознаки, що аналізується), тоді як від'ємні значення – про наявність негативної просторової автокореляції (тобто, просторово більш віддалені райони більш подібні стосовно частоти ознаки, що аналізується).

Оцінка рівня значущості вибіркових оцінок коефіцієнта Морана може бути зроблена двома способами. При використанні першого способу спочатку розраховується варіанса цього показника – *Var(I)*, а потім проводиться оцінка величини:

$$z = \frac{I - E(I)}{\sqrt{Var(I)}}, \qquad (14.17)$$

яка має стандартний нормальний розподіл.

Таким чином, значення $|z| > 1{,}96$ буде свідчити про необхідність відхилити нульову гіпотезу з імовірністю $p < 0{,}05$.

Ми в даному розділі не будемо наводити формул для оцінки варіанси коефіцієнта Морана. По-перше, вони досить громіздкі. По-друге, у них немає ніякої необхідності, оскільки всі необхідні розрахунки проводяться автоматично, використовуючи спеціалізовані програми (ми нижче зупинимося на них).

Другий спосіб використовує процес рандомізації. При цьому оцінки ознаки, що аналізується (частоти фена) розподіляються випадковим чином по окремих точках (тобто, субпопуляціях) і розраховується оцінка коефіцієнта Морана. Так проводиться багато разів (наприклад, 1000, 5000 чи 10000; *Nruns*). Потім підраховується, скільки псевдооцінок, отриманих у результаті перестановок значень, були більшими чи дорівнювали вибірковій оцінці *I* (*NGE*). Тоді двосторонній рівень значущості вибіркової оцінки коефіцієнта Морана буде дорівнювати:

$$p = 2 \cdot \left(\frac{NGE + 1}{Nruns + 1} \right). \qquad (14.18)$$

Ми у своєму аналізі використовували програму **ROOKCASE**, розроблену доктором M. Sawada (University of Ottawa), що вбудовується в MS Excel. Це дуже зручно, оскільки дозволяє проводити обробку даних в одному середовищі. (Нагадаємо, описана вище програма для оцінки коефіцієнта Мантеля також має вид Add-in.)

Програма є free-версією і може бути отримана від автора за запитом на адресу: msawada@uottawa.ca.

Формат вихідних даних для розрахунку коефіцієнта просторової автокореляції Морана (також, як і інших показників, наприклад, коефіцієнта *C* Джирі) простий. Він являє собою три стовпці цифр: перший – це координати

точок по осі *X*, другий – по осі *Y*, третій – власні значення ознаки, що аналізується (частоти ознаки, розміри, і т. п.).

Результати можуть бути отримані як у цілому для всієї сукупності значень, так і окремо для точок, відстань між якими задається дослідником.

Наприклад, нас цікавило як змінюється значення коефіцієнта Морана при розгляді частот ознаки в різних масштабах.

Ми обрали чотири інтервали відстаней: від 0 до 4 м, від 4 до 8 м, від 8 до 12 м і від 12 до 16 м. На рисунку 14.4 наведено корелограму для ознаки, що аналізується. Для кожного інтервалу лага наведено значення коефіцієнта Морана плюс-мінус одна статистична помилка, а також рівень значущості даних величин, отриманий у результаті 1000 пермутацій.

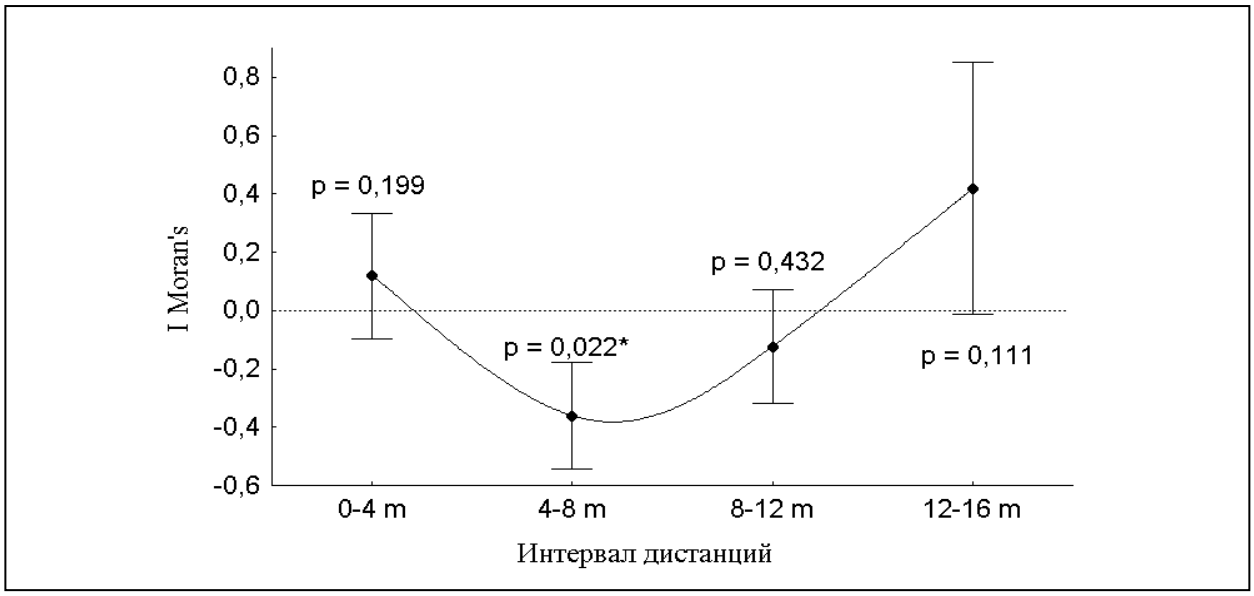

Рисунок 14.4 – Корелограма частоти черепашок із пігментними смужками в аналізованій популяції *B.bidens*

Лише одне значення коефіцієнта Морана виявилося вірогідним; для величини лага від 4 до 8 м (I = -0,361±0,182). Це свідчить про те, що вибірки, віддалені одна від одної на 4-8 м виявляються більш гетерогенними за частотою даного фена, ніж це можна було очікувати при випадковому формуванні фенетичної структури популяції молюска *B.bidens*.

Контрольні питання:

1. Загальна послідовність фенетичного аналізу структурованих популяцій.
2. Яким чином проводиться оцінка рівня значущості величини коефіцієнта Мантеля?
3. У яких випадках має сенс застосування показників просторової автокореляції?

ЧАСТИНА ІІІ
АНАЛІЗ КІЛЬКІСНИХ ДАНИХ

§ 15. Варіаційний ряд та аналіз вибіркових даних

15.1 Побудова варіаційного ряду

Варіаційний ряд розподілу – це подвійний ряд, в якому для ранжованих значень варіант вибірки наведено їх частоти. Форму запису сукупності у вигляді варіаційного ряду часто використовують для статистичного аналізу великих вибірок.

Розглянемо на одному прикладі особливості побудови варіаційного ряду та розрахунку вибіркових показників за його допомогою. Проаналізуємо дані щодо вмісту жиру в молоці корів (табл. 15.1).

Таблиця 15.1 – Вміст жиру в молоці корів, %

4,01	4,12	3,64	3,50	3,70	3,80
3,67	4,20	3,90	3,59	3,60	3,83
3,94	3,85	3,56	3,94	3,79	3,73
4,17	3,93	3,62	3,38	3,45	3,98
3,95	3,63	3,73	3,51	3,71	3,76
3,93	3,93	3,40	3,80	3,56	3,95
4,06	3,70	3,94	3,95	3,54	3,86
3,93	4,07	3,70	4,00	3,57	3,58
3,65	4,07	3,83	3,85	3,69	3,99
4,05	3,94	3,65	3,71	3,66	3,62
3,70	4,08	3,98	3,82	4,13	3,51
4,46	3,75	3,53	3,70	3,70	3,78
3,76	3,95	3,64	3,62	3,79	4,14
4,04	3,96	3,42	3,78	3,70	
4,04	3,96	3,25	3,64	3,36	
3,94	3,78	3,44	3,71	3,82	
3,60	3,70	3,53	3,60	4,02	
4,42	3,71	3,97	3,56	3,72	
4,08	3,58	3,91	3,80	3,53	
4,49	3,67	3,62	3,92	3,57	

Для побудови інтервального варіаційного ряду, по-перше, необхідно визначити кількість інтервалів та його ширину.

Існує велика кількість правил щодо вибору необхідної кількості інтервалів. Найвідоміше з них – це правило Стерджеса:

$$i = 1 + 3{,}32 \cdot \lg n, \qquad (15.1)$$

де n – обсяг вибірки.

Але доцільніше для виборі кількості інтервалів варіаційного ряду обирати непарне число, найближче до значення, що дає правило Стерджеса.

У нашому прикладі значення, що дає правило Стерджеса, дорівнює: $i = 1 + 3{,}32 \cdot \lg\ 113 = 7{,}82.$

Отже, доцільно обрати найближче непарне число, тобто, 7.

Ширина інтервалу визначається за формулою:

$$h = \frac{x_{\max} - x_{\min}}{i}. \qquad (15.2)$$

Але, для полегшення подальших розрахунків, ширина інтервалу повинна бути кратною 10 (наприклад, 0,1, 20 або 250 і т. п.). Для цього мінімальне значення вибірки ($x_{\min}$) у формулі 15.2 можна дещо зменшити, а максимальне ($x_{\max}$) – дещо збільшити. Але ніяк не навпаки!

У нашому прикладі: $x_{\max}$ = 4,49%, $x_{\min}$ = 3,25%. Різниця між ними становить 1,24. Найближче число, яке кратне до 7 та 10 – це 1,40. Тому ми трохи зменшимо мінімальне значення (до 3,20) й підвищимо максимальне (до 4,60). Тоді ширина інтервалу складатиме:

$$h = \frac{x_{\max} - x_{\min}}{i}.$$

Межами нашого варіаційного ряду будуть значення:

3,20 – 3,40 – 3,60 – 3,80 – 4,00 – 4,20 – 4,40 – 4,60.

Але для того, щоб значення вибірки, що потрапляють на межу між класами однозначно трактувалися, підвищимо нижні межі класових інтервалів (починаючи із другого) на одну одиницю мінімального розряду (у нашому випадку – на 0,01). Таким чином вихідна матриця для нашого варіаційного ряду матиме наступний вигляд (табл. 15.2).

Таблиця 15.2 – Вихідна матриця варіаційного ряду

i	*Інтервал*					
1	3,20-3,40					
2	3,41-3,60					
3	3,61-3,80					
4	3,81-4,00					
5	4,01-4,20					
6	4,21-4,40					
7	4,41-4,60					
Всього						

Після того, як класові інтервали визначено, підраховуються абсолютні частоти, з якими варіанти вибірки потрапляють у той чи інший класовий інтервал (табл. 15.3).

Абсолютні частоти наведено в стовпчику n_i. Крім того необхідно розрахувати накопичені частоти (стовпчик s_i в табл. 15.3). Накопичена частота класового інтервалу – це кількість значень у вибірці, що є меншими чи дорівнюють верхній межі класового інтервалу.

Таблиця 15.3 – Абсолютні на накопичені частоти

i	*Інтервал*	n_i	s_i			
1	3,20-3,40	4	4			
2	3,41-3,60	21	25			
3	3,61-3,80	40	65			
4	3,81-4,00	30	95			
5	4,01-4,20	15	110			
6	4,21-4,40	0	110			
7	4,41-4,60	3	113			
Всього		113	-			

Для першого інтервалу: $s_1 = n_1$; для другого: $s_2 = s_1 + n_2$; для третього: $s_3 = s_2 + n_3$ і т. д.

Накопичені частоти потрібні нам для розрахунку медіани вибірки та для перевірки відповідності розподілу вибірки нормальному закону.

Необхідно визначити середини класових інтервалів (ось для чого ширину інтервалу було обрано кратною 10). Для першого інтервалу середина становить:

$$x_1 = \frac{3{,}20 + 3{,}40}{2} = 3{,}30, \text{ для другого:}$$

$$x_2 = \frac{3{,}40 + 3{,}60}{2} = 3{,}50, \text{ і т. п.}$$

Одиницю, на яку ми збільшили нижню межу кожного інтервалу, до уваги не беремо.

Розраховуємо вибіркове середнє арифметичне значення за формулою:

$$\bar{x} = \frac{\sum n_i x_i}{n}. \qquad (15.3)$$

Для цього в межах кожного інтервалу необхідно помножити його середину на абсолютну частоту. Тобто, для першого інтервалу:

$$n_1 \cdot x_2 = 4 \cdot 3{,}30 = 13{,}20 \text{ і т.п .}$$

Розраховані дані наведено в таблиці 15.4 (стовпчик $n_i \cdot x_i$). Всі необхідні для цього значення розраховуємо та знаходимо суму значень цього стовпчика.

Таблиця 15.4 – Проміжні результати визначення вибіркового середнього арифметичного значення

i	*Інтервал*	n_i	s_i	x_i	$n_i \cdot x_i$	
1	3,20-3,40	4	4	3,30	13,20	
2	3,41-3,60	21	25	3,50	73,50	
3	3,61-3,80	40	65	3,70	148,00	
4	3,81-4,00	30	95	3,90	117,00	
5	4,01-4,20	15	110	4,10	61,50	
6	4,21-4,40	0	110	4,30	0,00	
7	4,41-4,60	3	113	4,50	13,50	
Всього		113	-	-	426,70	

Таким чином, вибіркове середнє арифметичне значення становитиме:

$$\bar{x} = \frac{426{,}70}{113} = 3{,}776 \approx 3{,}78.$$

Значення середнього квадратичного відхилення (с.к.в.) розраховується за формулою:

$$\sigma = \sqrt{\frac{\sum n_i (x_i - \bar{x})^2}{n-1}}. \qquad (15.4)$$

Для цього необхідно в межах кожного інтервалу помножити його абсолютну частоту на квадрат різниці між серединою цього інтервалу та вибірковим середнім арифметичним значенням. Для першого інтервалу ця величина становитиме:

$$n_1 \cdot (x_1 - \bar{x})^2 = 4 \cdot (3{,}30 - 3{,}78)^2 = 0{,}9216 \text{ і т. п.}$$

Результати розрахунків наведено в табл. 15.5.

Таблиця 15.5 – Вихідні дані для розрахунку середнього квадратичного відхилення

i	*Інтервал*	n_i	s_i	x_i	$n_i \cdot x_i$	$n_i \cdot (x_i - \bar{x})^2$
1	3,20-3,40	4	4	3,30	13,20	0,9216
2	3,41-3,60	21	25	3,50	73,50	1,6464
3	3,61-3,80	40	65	3,70	148,00	0,2560
4	3,81-4,00	30	95	3,90	117,00	0,4320
5	4,01-4,20	15	110	4,10	61,50	1,5360
6	4,21-4,40	0	110	4,30	0,00	0,0000
7	4,41-4,60	3	113	4,50	13,50	1,5552
Всього		113	-	-	426,70	6,3472

Таким чином, вибіркове значення с.к.в. становитиме:

$$\sigma = \sqrt{\frac{6{,}3472}{113-1}} = 0{,}238.$$

Але, оскільки розрахунок с.к.в. проводився не на підставі окремих значень, а на підставі варіаційного ряду, необхідно врахувати поправку Шепарда.

Для цього від розрахованого значення необхідно відняти величину (h/12). Таким чином, остаточне значення с.к.в. становитиме:

$$\sigma = 0{,}238 - \frac{0{,}20}{12} = 0{,}221 \approx 0{,}22.$$

Коефіцієнт варіації розраховується за формулою:

$$CV = \frac{\sigma}{\bar{x}} \cdot 100\%, \qquad (15.5)$$

а його статистична помилка:

$$S_{CV} = \frac{CV}{\sqrt{2n}}. \qquad (15.6)$$

Для даних із нашого прикладу значення коефіцієнта варіації складатиме: $CV = \frac{0,22}{3,78} \cdot 100\% = 5,82\%$ із статистичною помилкою: $S_{CV} = \frac{5,82}{\sqrt{2 \cdot 113}} = 0,39\%$.

Альтернативним показником, який, як і вибіркове середнє арифметичне, використовується для оцінки центру розподілу, є вибіркова медіана. Для інтервального варіаційного ряду значення медіани розраховується за формулою:

$$Me = x_0 + h \cdot \left(\frac{\frac{n}{2} - s_{Me-1}}{n_{Me}} \right), \qquad (15.7)$$

де x_0 – нижня межа медіанного інтервалу;

s_{Me-1} – накопичені частоти інтервалу, який передує медіанному;

n_{Me} – частота медіанного інтервалу.

Медіанним є інтервал 3,61-3,80, оскільки на цей інтервал припадає перша накопичена частота, що перевищує половину всього обсягу вибірки (65 перевищує 113/2 = 56,5). Таким чином значення медіани становить:

$$Me = 3,61 + 0,2 \cdot \left(\frac{\frac{113}{2} - 25}{40} \right) = 3,77 .$$

15.2 Помилки вибіркових показників та їх довірчі інтервали

Інтервальною називають оцінку, яка характеризується двома числами – границями інтервалу, який охоплює оцінюваний параметр. Така оцінка являє собою деякий інтервал, в якому із заданою ймовірністю знаходиться шуканий параметр. За центр інтервалу беруть вибіркову точкову оцінку.

Довірча ймовірність – це достатньо висока ймовірність практично вірогідної події, яка гарантує отримання надійних статистичних висновків. Вона позначається P, а ймовірність перевищити цей рівень – α. Відповідно, $\alpha = 1 - P$. Ймовірність α називають *рівнем значущості*, або *рівнем істотності*, який характеризує відносну кількість помилкових висновків від загальної кількості всіх статистичних висновків.

Довірчим інтервалом для параметра θ називається такий інтервал, відносно якого можна із близькою до одиниці довірчою ймовірністю $P = 1 - \alpha$ стверджувати, що він містить шукане значення параметра θ.

Для більшості вибіркових показників інтервальне оцінювання безпосередньо пов’язано із розрахунками їх відповідних статистичних помилок. Проілюструємо методику розрахунку довірчих інтервалів для показника вмісту жиру в молоці матерів корів дослідного стада (див. вище наведений приклад).

Розрахунок довірчого інтервалу для вибіркового середнього арифметичного

Статистична помилка вибіркового середнього арифметичного розраховується за формулою:

$$S_{\bar{x}} = \frac{\sigma}{\sqrt{n}}, \tag{15.8}$$

де σ – вибіркове середнє квадратичне відхилення (с.к.в.);

n – обсяг вибірки.

Таким чином, для даних з нашого прикладу помилка середнього арифметичного становитиме: $S_{\bar{x}} = \frac{0{,}22}{\sqrt{113}} = 0{,}02$.

Тоді із 95% довірчою ймовірністю значення вибіркового середнього арифметичного знаходиться в інтервалі:

$$\bar{x} - t \cdot S_{\bar{x}} \le \bar{x} \le \bar{x} + t \cdot S_{\bar{x}}, \tag{15.9}$$

де *t* – значення критерію Ст'юдента для числа ступенів свободи *df* = *n* – 1 (додаток И).

Обсяг вибірки становить 113 голів. Тоді число ступенів свободи для критерію Ст'юдента буде: *df* = 113 – 1 = 112. У Додатку И немає точного значення для такого числа ступенів свободи. Але, оскільки, для 110 та 120 ступенів свободи табличне значення не змінюється (1,98), його можна прийняти за шукане.

Таким чином, для нашого прикладу 95% довірчий інтервал для вибіркового середнього арифметичного становитиме:

$$3{,}78 - 1{,}98 \cdot 0{,}02 \le \bar{x} \le 3{,}78 + 1{,}98 \cdot 0{,}02,$$

або

$$3{,}74 \le \bar{x} \le 3{,}82.$$

У тих випадках, коли обсяг генеральної сукупності, з якої відібрана вибірка, має фіксоване значення (*N*), у формулу (15.8) необхідно вносити виправлення:

$$S_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \cdot \sqrt{1 - \frac{n}{N}}. \tag{15.10}$$

Розрахунок довірчого інтервалу для вибіркового с.к.в.

Статистична помилка вибіркового с.к.в. розраховується за формулою:

$$S_{\sigma} = \frac{\sigma}{\sqrt{2n}}. \tag{15.11}$$

Таким чином, для даних із нашого прикладу помилка вибіркового с.к.в. становитиме: $S_{\sigma} = \frac{0{,}22}{\sqrt{2 \cdot 113}} = 0{,}015.$

Але при побудові довірчого інтервалу для вибіркового с.к.в. використовується не розподіл Ст'юдента, а розподіл Хі-квадрат Пірсона.

З 95% довірчою ймовірністю можна стверджувати, що істинне значення вибіркового с.к.в. лежить у межах:

$$\sigma_н = \sqrt{\frac{(n-1)\cdot\sigma^2}{\chi_1^2}}; \tag{15.12}$$

$$\sigma_в = \sqrt{\frac{(n-1)\cdot\sigma^2}{\chi_2^2}}, \tag{15.13}$$

де χ_1^2 – значення критерію Хі-квадрат К. Пірсона для числа ступенів свободи $df = n - 1$ та рівня значущості $\alpha = 0{,}025$;

χ_2^2 – значення критерію Хі-квадрат К. Пірсона для числа ступенів свободи $df = n - 1$ та рівня значущості $\alpha = 0{,}975$.

Для зручності розрахунки довірчого інтервалу для вибіркового с.к.в. можна проводити за формулами:

$$\sigma_н = c_1 \cdot \sigma; \tag{15.14}$$

$$\sigma_в = c_2 \cdot \sigma, \tag{15.15}$$

де коефіцієнти c_1 та c_2 знаходяться в таблиці 15.6.

Таблиця 15.6 – Коефіцієнти для розрахунку довірчого інтервалу для вибіркового с.к.в.

df	c_1	c_2	*df*	c_1	c_2
20	0,765	1,444	120	0,888	1,145
30	0,799	1,337	130	0,892	1,138
40	0,821	1,280	140	0,895	1,133
50	0,837	1,243	150	0,899	1,128
60	0,849	1,217	160	0,901	1,123
70	0,858	1,198	170	0,904	1,119
80	0,866	1,183	180	0,907	1,115
90	0,873	1,171	190	0,909	1,112
100	0,879	1,161	200	0,911	1,109
110	0,884	1,152	250	0,920	1,096

У тих випадках, коли число ступенів свободи для вибірки відсутнє в таблиці 15.6, шукані значення коефіцієнтів c_1 та c_2 можна отримати лінійною інтерполяцією.

Наприклад, для даних із нашого прикладу ми отримали вибіркове значення с.к.в. 0,22. Як розрахувати 95% довірчий інтервал цього показника?

Число ступенів свободи для вибірки складає $df = 113 - 1 = 112$. В таблиці 15.6 найближчі значення – це 110 та 120. Тоді методом лінійної інтерполяції коефіцієнти для розрахунків нижньої та верхньої довірчих меж для вибіркового с.к.в. можна розрахувати за формулою:

$$c = c' + \frac{(c'' - c') \cdot (df - df')}{(df'' - df')}, \qquad (15.16)$$

де *df'* – табличне число ступенів свободи, що менше вибіркового, а *c'* – відповідне значення допоміжного коефіцієнта;

df'' – табличне число ступенів свободи, що більше вибіркового, а *c''* – відповідне значення допоміжного коефіцієнта.

Нижче вибіркового числа ступенів свободи в таблиці наведені значення для *df* = 110, а вище – *df* = 120. Тоді на підставі лінійної інтерполяції можна визначити значення допоміжних коефіцієнтів:

$$c_1 = 0{,}884 + \frac{(0{,}888 - 0{,}884) \cdot (112 - 110)}{120 - 110} = 0{,}885;$$

$$c_2 = 1{,}152 + \frac{(1{,}145 - 1{,}152) \cdot (112 - 110)}{120 - 110} = 1{,}151.$$

Відповідно, нижня та верхня межі для вибіркового с.к.в. становитимуть:

$$\sigma_н = 0{,}885 \cdot 0{,}22 = 0{,}195;$$

$$\sigma_в = 1{,}151 \cdot 0{,}22 = 0{,}253.$$

Розрахунок довірчого інтервалу для вибіркової медіани

95% довірчий інтервал для вибіркової медіани можна також розрахувати за формулою 15.9, де замість вибіркового середнього арифметичного та його помилки підставляється вибіркова медіана та її помилка, відповідно:

$$Me - t \cdot S_{Me} \le Me \le Me + t \cdot S_{Me}, \qquad (15.17)$$

де статистична помилка вибіркової медіани розраховується за формулою:

$$S_{Me} = \frac{Q_3 - Q_1}{\sqrt{n}}. \qquad (15.18)$$

У цій формулі використовуються значення першого (Q_1) та третього квартилів (Q_3). *Квартилі* розподіляють весь вибірковий ранжований ряд на чотири рівні частини. Таким чином, 25% вибіркових варіант мають значення нижчі, ніж перша квартиль, 50% значень знаходяться між першою та третьою квартиллю, і ще 25% значень переважають третю квартиль. Медіана вважається другою квартиллю.

Значення першої та третьої квартилів для даних, що згруповані у варіаційний ряд можна розрахувати за формулами:

$$Q_1 = x_{Q_1} + h \cdot \left(\frac{\frac{n}{4} - s_{Q_1 - 1}}{n_{Q_1}} \right) \qquad (15.19)$$

та

$$Q_3 = x_{Q_3} + h \cdot \left(\frac{\frac{3 \cdot n}{4} - s_{Q_3 - 1}}{n_{Q_3}} \right), \qquad (15.20)$$

де x_{Q_1} та x_{Q_3} – нижні межі інтервалів, що містять першу та треті квартилі, відповідно;

$s_{Q_1 - 1}$ та $s_{Q_3 - 1}$ – накопичені частоти інтервалів, що передують інтервалам, що містять першу та третю квартилі, відповідно;

n_{Q_1} та n_{Q_3} – частоти інтервалів, що містять першу та третю квартилі, відповідно.

Оскільки у вибірці 113 значень, нижче першої квартилі знаходяться (113/4) = 28,25 варіант. Таким чином, перша квартиль знаходиться в інтервалі, накопичена частота якої вперше переважає 28,25. Це інтервал 3,61-3,80. Менше третьої квартилі знаходиться (3·113/4) = 84,75 варіант. Таким чином, третя квартиль знаходиться у інтервалі, накопичена частота якої вперше переважає 84,75. Це інтервал 3,81-4,00. Таким чином значення першої і третьої квартилів будуть дорівнювати:

$$Q_1 = 3{,}61 + 0{,}2 \cdot \left(\frac{\frac{113}{4} - 25}{40} \right) = 3{,}63,$$

$$Q_3 = 3{,}81 + 0{,}2 \cdot \left(\frac{\frac{3 \cdot 113}{4} - 65}{30} \right) = 3{,}94.$$

Таким чином, помилка вибіркової медіани дорівнюватиме:

$$S_{Me} = \frac{3{,}94 - 3{,}63}{\sqrt{113}} = 0{,}03.$$

Тоді з 95% ймовірністю можна стверджувати, що вибіркова медіана знаходиться в межах:

$$3{,}77 - 1{,}98 \cdot 0{,}03 \le Me \le 3{,}77 + 1{,}98 \cdot 0{,}03,$$

або

$$3{,}71 \le Me \le 3{,}83.$$

Оцінка вибіркових показників, а також їх довірчих інтервалів, особливо у разі нечисельних вибірок, а також вибірок, розподіл яких не відповідає нормальному (див. § 16) через наявність значень або надмірно малих, або надмірно великих, ніж решта вибіркових значень, може бути здійснена з використанням методів чисельного ресамплінгу і, насамперед, бутстреп-методу.

При використанні цього методу із вибірки, що складається із фактичних значень, випадковим чином відбираються значення у нову, штучним чином створену вибірку такого ж обсягу (вона має назву *псевдовибірка*) шляхом відбору із поверненням. Таким чином, у псевдовибірку одне значення із фактичної вибірки може зовсім не потрапити, а інше – потрапити двічі, тричі і більше разів.

Наприклад, якщо наша вибірка із вихідними даними містить значення 1, 2, 3, 4 та 5, то перша псевдовибірка може складатися із наступних елементів: 2, 3, 3, 4, 4, друга псевдовибірка – 1, 1, 3, 5, 5 і т. п. Таких штучних псевдовибірок повинно бути сформовано досить багато (100, 500, 1000 або навіть 5000), наприклад, *M*.

Для кожної із них розраховується значення середнього арифметичного. Тоді бутстреп-оцінка середнього арифметичного може бути знайдена за формулою:

$$\overline{X}_{boot} = \frac{\sum_{i=1}^{M} \overline{X}i}{M}, \tag{15.21}$$

а її статистична помилка:

$$SE_{\overline{X}_{boot}} = \sqrt{\frac{\sum_{i=1}^{M} (\overline{X}_i - \overline{X}_{boot})^2}{M-1}}, \tag{15.22}$$

де $\overline{X}_i$ – оцінка середнього арифметичного, отримана для *i*-тої псевдовибірки.

Нижня та верхня межі 95% довірчого інтервалу для бутстреп-оцінки середнього арифметичного знаходиться як 2,5% та 97,5% перцентилі для сукупності оцінок $\overline{X}_i$.

Приклад. Було отримано десять значень рівня молочної продуктивності корів червоної степової породи експериментальної групи за 305 днів лактації: 4972, 3440, 4484, 3705, 6283, 4661, 3962, 5069, 10144 та 3922 кг. Необхідно розрахувати середнє арифметичне та його статистичну помилку.

Відразу ж звернемо увагу на те, що вибірка наших вихідних даних відносно нечисельна і, по-друге, значення 10144 кг значно виділяється серед усіх інших показників. Але, з іншого боку, ми не можемо не врахувати це отримане під час експерименту значення. (Вважається, що сам експеримент виконано без помилок і всі отримані результати валідні.)

Тому використаємо бутстреп-процедуру для оцінки середнього арифметичного та його помилки.

Для цього, використовуючи генератор випадкових чисел, що є, наприклад, в MS Excel, створимо псевдовибірки обсягом 10 величин із вихідних даних і розрахуємо для них середні арифметичні значення. А далі, на

підставі формул 15.21 та 15.22, розрахуємо бутстреп-оцінку середнього арифметичного та її помилку.

Для 50 псевдовибірок, сформованих для ілюстрації цього прикладу, бутстреп-оцінка середнього арифметичного склала 4893,8 кг (фактичне середнє арифметичне значення складає для цієї вибірки – 5064,2кг) із статистичною помилкою – 488,9 кг (статистична помилка для фактичного середнього арифметичного значення – 621,9 кг).

Як бачимо, якщо оцінки середнього арифметичного, отримані класичними методами та методами ресамплінгу, значно не відрізняються, то помилка середнього у другому варіанті майже на 20% нижче. Завищене значення помилки середнього арифметичного, розрахованого класичними методами, зумовлено якраз наявністю т. зв. «викиду», тобто значення 10144.

Таким чином, методи ресамплінгу є більш стійкими до характеру розподілу елементів вибірки та наявності в ній «викидів», тому можуть давати більш надійні результати у селекційній роботі.

Контрольні питання:

1. Послідовність побудови варіаційного ряду.
2. Методика розрахунку довірчого інтервалу для вибіркового середнього арифметичного.
3. Яким чином здійснюється оцінка вибіркових показників, а також їх довірчих інтервалів, у разі нечисельних вибірок, а також вибірок, розподіл яких не відповідає нормальному?

§ 16. Нормальний розподіл та його використання в селекційній роботі

Нормальний розподіл (або *розподіл Гауса-Лапласа*) має сукупність, розподіл варіант якої відповідає формулі:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\bar{x})^2}{2\sigma^2}}. \qquad (16.1)$$

Для перевірки відповідності нормальному закону розподілу вибірки, що задана у вигляді варіаційного ряду, необхідно виконати наступні розрахунки.

Розглянемо цю методику на підставі варіаційного ряду, що було нами отримано у таблиці 15.1.

Необхідно побудувати нову таблицю, що повинна мати наступні дані щодо варіаційного ряду (табл. 16.1).

Таблиця 16.1 – Дані варіаційного ряду

i	*Інтервал*	n_i	s_i	x_i					
1	3,20-3,40	4	4	3,30					
2	3,41-3,60	21	25	3,50					
3	3,61-3,80	40	65	3,70					
4	3,81-4,00	30	95	3,90					
5	4,01-4,20	15	110	4,10					
6	4,21-4,40	0	110	4,30					
7	4,41-4,60	3	113	4,50					
Всього		113	-	-					

Для середини кожного інтервалу необхідно розрахувати його нормоване відхилення за формулою:

$$t_i = \frac{x_i - \bar{x}}{\sigma}. \qquad (16.2)$$

Наприклад, для середини першого інтервалу нормоване відхилення має значення: $t_1 = \frac{3{,}30 - 3{,}78}{0{,}22} = -2{,}182$.

Результати розрахунків наведено в таблиці 16.2.

Таблиця 16.2 – Проміжні дані із урахуванням нормованого відхилення

i	*Інтервал*	n_i	s_i	x_i	t_i				
1	3,20-3,40	4	4	3,30	-2,182				
2	3,41-3,60	21	25	3,50	-1,273				
3	3,61-3,80	40	65	3,70	-0,364				
4	3,81-4,00	30	95	3,90	0,545				
5	4,01-4,20	15	110	4,10	1,455				
6	4,21-4,40	0	110	4,30	2,364				
7	4,41-4,60	3	113	4,50	3,273				
Всього		113	-	-	-				

Далі для значень нормованого відхилення необхідно визначити ординати нормальної кривої (теоретичні). Це можна зробити двома способами.

По-перше, знайти відповідні значення в таблиці ординат нормальної нормованої кривої, що наведена в Додатку К. (Необхідно пам'ятати, що *f(-t) = f(t)*, тобто знак, що знаходиться перед значенням нормованого відхилення, можна не враховувати).

По-друге, ординати нормальної кривої можна розрахувати на підставі формули:

$$f(t) \approx 0{,}4 \cdot e^{-\frac{t^2}{2}}. \qquad (16.3)$$

Отримані значення наведено в таблиці 16.3.

Таблиця 16.3 – Розраховані значення ординат нормальної кривої

i	*Інтервал*	n_i	s_i	x_i	t_i	*f(t)*			
1	3,20-3,40	4	4	3,30	-2,182	0,0370			
2	3,41-3,60	21	25	3,50	-1,273	0,1780			
3	3,61-3,80	40	65	3,70	-0,364	0,3744			
4	3,81-4,00	30	95	3,90	0,545	0,3447			
5	4,01-4,20	15	110	4,10	1,455	0,1389			
6	4,21-4,40	0	110	4,30	2,364	0,0245			
7	4,41-4,60	3	113	4,50	3,273	0,0019			
Всього		113	-	-	-	1,0993			

Правильність розрахунків можна перевірити наступним чином. Сума значень ординат повинна дорівнювати:

$$\sum f(t) = \frac{\sigma}{h}, \qquad (16.4)$$

де *h* – ширина класового інтервалу варіаційного ряду.

У нашому випадку: $\frac{\sigma}{h} = \frac{0{,}22}{0{,}2} = 1{,}10$, що з точністю до третього знаку після коми відповідає сумі значень ординат нормальної кривої (1,0993).

Розраховуються теоретичні значення частот кожного інтервалу за формулою:

$$n^{teo} = f(t) \cdot \frac{nh}{\sigma} \qquad (16.5)$$

Наприклад, теоретична частота для першого інтервалу становитиме: $n^{teo} = 0{,}0370 \cdot \frac{113 \cdot 0{,}20}{0{,}22} = 3{,}80$.

Результати розрахунку всіх теоретичних частот наведено в таблиці 16.4.

В ідеалі сума фактичних частот повинна дорівнювати сумі теоретичних, тобто $\sum n_i = \sum n^{teo}$. (У нашому випадку різниця у другому знаку після коми, що зумовлено округленням значень при проведенні проміжних розрахунків).

Наступний стовпчик – це накопичені теоретичні значення.

Таблиця 16.4 – Теоретичні значення частот кожного інтервалу

i	*Інтервал*	n_i	s_i	x_i	t_i	*f(t)*	n^{teo}	s^{teo}	
1	3,20-3,40	4	4	3,30	-2,182	0,0370	3,80	3,80	
2	3,41-3,60	21	25	3,50	-1,273	0,1780	18,28	22,08	
3	3,61-3,80	40	65	3,70	-0,364	0,3744	38,46	60,55	
4	3,81-4,00	30	95	3,90	0,545	0,3447	35,41	95,96	
5	4,01-4,20	15	110	4,10	1,455	0,1389	17,27	110,22	
6	4,21-4,40	0	110	4,30	2,364	0,0245	2,52	112,74	
7	4,41-4,60	3	113	4,50	3,273	0,0019	0,19	112,93	
Всього		113	-	-	-	1,0993	112,93	-	

Графічним зображенням відповідності розподілу вибірки нормальному закону є гістограма (рис. 16.1).

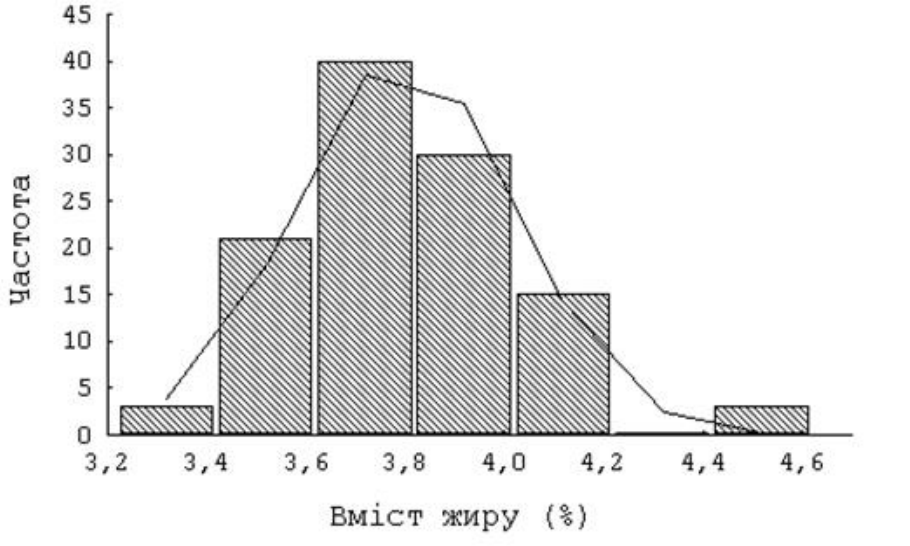

Рисунок 16.1 – Гістограма розподілу за вмістом жиру в молоці корів дослідного стада. Наведено теоретичну лінію нормального розподілу

Крім того, відповідність вибіркового розподілу нормальному закону можна перевірити на підставі *кумуляти* (рис. 16.2).

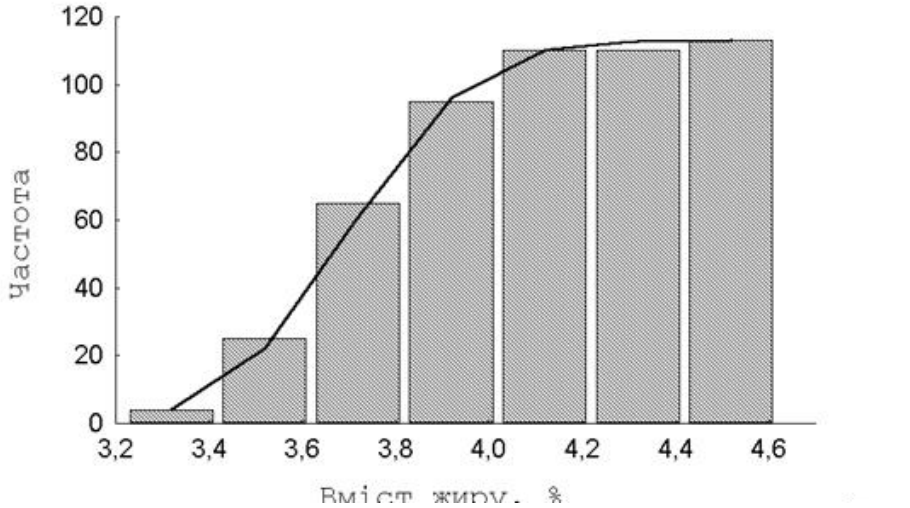

Рисунок 16.2 – Кумулята розподілу за вмістом жиру в молоці корів дослідного стада. Наведено теоретичну лінію нормального розподілу

Для перевірки відповідності вибіркового розподілу нормальному закону можна використовувати критерій *Колмогорова-Смирнова*, який являє собою модуль максимальної різниці між фактичними та теоретичними накопиченими частотами:

$$D = max \mid s_i - s^{teo} \mid. \qquad (16.6)$$

Для розрахунку критерію Колмогорова-Смирнова необхідно в межах кожного інтервалу розрахувати модуль різниці між фактичними та теоретичними накопиченими частотами.

Наприклад, для першого інтервалу це значення складатиме:

$$D = max \mid 4 - 3{,}80 \mid = 0{,}20.$$

Отримані результати наведено в таблиці 16.5.

Таблиця 16.5 – Розрахований модуль різниці між фактичними та теоретичними накопиченими частотами

i	*Інтервал*	n_i	s_i	x_i	t_i	$f(t)$	n^{teo}	s^{teo}	
1	3,20-3,40	4	4	3,30	-2,182	0,0370	3,80	3,80	0,20
2	3,41-3,60	21	25	3,50	-1,273	0,1780	18,28	22,08	2,92
3	3,61-3,80	40	65	3,70	-0,364	0,3744	38,46	60,55	**4,45**
4	3,81-4,00	30	95	3,90	0,545	0,3447	35,41	95,96	0,96
5	4,01-4,20	15	110	4,10	1,455	0,1389	17,27	110,22	0,22
6	4,21-4,40	0	110	4,30	2,364	0,0245	2,52	112,74	2,74
7	4,41-4,60	3	113	4,50	3,273	0,0019	0,19	112,93	0,07
Всього		113	-	-	-	1,0993	112,93	-	-

Знаходимо максимальне значення – в нашому випадку воно дорівнює 4,45 (за модулем) для третього інтервалу.

Оцінка вірогідності отриманого значення критерію Колмогорова-Смирнова перевіряється на підставі значення λ:

$$\lambda = \frac{D}{\sqrt{n}}. \qquad (16.7)$$

Якщо розраховане значення λ не перевищує 1,36, вибірковий розподіл відповідає нормальному закону; якщо λ більше, ніж 1,36 – вибірковий розподіл вірогідно відхиляється від нормального.

Для нашого прикладу: $\lambda = \frac{4{,}45}{\sqrt{113}} = 0{,}42$; таким чином розподіл матерів корів за вмістом жиру відповідає нормальному закону.

Крім того, відповідність вибіркового розподілу нормальному закону можна перевірити, розрахувавши показники форми кривої розподілу, а саме – коефіцієнт асиметрії (*As*) та коефіцієнт ексцесу (*Ex*):

$$As = \frac{\sum (x_i - \bar{x})^3}{n\sigma^3}, \qquad (16.8)$$

$$Ex = \frac{\sum (x_i - \bar{x})^4}{n\sigma^4} - 3. \qquad (16.9)$$

Перший із них дає оцінку скошеності кривої розподілу відносно його центру, а другий – визначає наявність піку або плато у центрі розподілу.

Формули 16.8 та 16.9 використовуються для малих вибірок.

Якщо ми маємо справу із варіаційним рядом, оцінки коефіцієнтів асиметрії та ексцесу розраховуються наступним чином.

Будується нова таблиця, яка має такі стовпчики (табл. 16.6).

Таблиця 16.6 – Вихідні дані для розрахунку коефіцієнтів асиметрії та ексцесу

i	*Інтервал*	n_i	*a*	$n_i \cdot a$	$n_i \cdot a^2$	$n_i \cdot a^3$	$n_i \cdot a^4$
1	3,20-3,40	4	-2	-8	16	-32	64
2	3,41-3,60	21	-1	-21	21	-21	21
3	3,61-3,80	40	0	0	0	0	0
4	3,81-4,00	30	+1	30	30	30	30
5	4,01-4,20	15	+2	30	60	120	240
6	4,21-4,40	0	+3	0	0	0	0
7	4,41-4,60	3	+4	12	48	192	768
Суми		113	-	43	175	289	1123

Показник *a* беруть таким чином, що для інтервального класу з найбільшою частотою він дорівнює нулю, для значень, що менше модального класу – зменшується на одиницю, а для значень, що більше модального класу – збільшується на одиницю. (До речі, вибір значень *a* не має суттєвої ролі).

На підставі розрахованих сум чотирьох останніх стовпчиків розраховуються проміжні величини:

$$m_1 = \frac{\sum n_i a}{n}, \tag{16.10}$$

$$m_2 = \frac{\sum n_i a^2}{n}, \tag{16.11}$$

$$m_3 = \frac{\sum n_i a^3}{n}, \tag{16.12}$$

$$m_4 = \frac{\sum n_i a^4}{n}. \tag{16.13}$$

Тоді оцінки коефіцієнтів асиметрії та ексцесу можна отримати за формулами:

$$As = \frac{m_3 - 3m_1 m_2 + 2m_1^3}{\left(\sqrt{m_2 - m_1^2}\right)^3}, \tag{16.14}$$

та

$$Ex = \frac{m_4 - 4m_1 m_3 + 6m_1^2 m_2 - 3m_1^4}{\left(\sqrt{m_2 - m_1^2}\right)^4} - 3. \tag{16.15}$$

Для даних із нашого прикладу: $m_1 = \frac{43}{113} = 0{,}381$; $m_2 = \frac{175}{113} = 1{,}549$; $m_3 = \frac{289}{113} = 2{,}558$; $m_4 = \frac{1123}{113} = 9{,}938$.

Тоді значення коефіцієнту асиметрії становитиме:

$$As = \frac{2{,}558 - 3 \cdot 0{,}381 \cdot 1{,}549 + 2 \cdot 0{,}381^3}{\left(\sqrt{1{,}549 - 0{,}381^2}\right)^3} = 0{,}541,$$

а коефіцієнта ексцесу:

$$Ex = \frac{9{,}938 - 4 \cdot 0{,}381 \cdot 2{,}558 + 6 \cdot 0{,}381^2 \cdot 1{,}549 - 3 \cdot 0{,}381^4}{\left(\sqrt{1{,}549 - 0{,}381^2}\right)^4} - 3 = 0{,}718.$$

Оцінка вірогідності отриманих значень проводиться на підставі значень їх статистичних помилок. Для великих вибірок статистичні помилки коефіцієнтів асиметрії та ексцесу розраховуються за формулами:

$$S_{As} = \sqrt{\frac{6}{n+3}}; \qquad (16.16)$$

$$S_{Ex} = \sqrt{\frac{24}{n+5}}. \qquad (16.17)$$

Для даних із нашого прикладу ці значення становлять: S_{As} = 0,227; S_{Ex} = 0,451.

Якщо відношення коефіцієнта асиметрії чи ексцесу до своїх помилок перевищуватиме 3, то вважається, що вибірковий розподіл вірогідно відхиляється від нормального закону. В інших випадках – вибірковий розподіл відповідає нормальному закону.

У нашому випадку відношення коефіцієнта асиметрії до своєї помилки становить: $\frac{As}{S_{As}} = \frac{0{,}541}{0{,}227} = 2{,}38$, а відношення коефіцієнта ексцесу до своєї помилки: $\frac{Ex}{S_{Ex}} = \frac{0{,}718}{0{,}451} = 1{,}59$.

В обох випадках ці значення не перевищують 3.

Таким чином, ми ще раз підтвердили, що розподіл корів за вмістом жиру в молоці вірогідно не відхиляється від нормального.

Використання нормального розподілу в селекційній роботі

Вважається, що випадкова величина *x* розподілена нормально, якщо функція її щільності *f(x)* має вигляд:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{(x-\bar{x})^2}{2\sigma^2}}, \qquad (16.18)$$

де $\bar{x}$ – середнє арифметичне значення;

σ – середнє квадратичне відхилення (с.к.в.).

Якщо значення вибіркової сукупності стандартизувати, тобто для кожного значення розрахувати його нормоване відхилення:

$$t_i = \frac{x_i - \bar{x}}{\sigma}, \tag{16.19}$$

то рівняння для функції щільності для *f(t)* буде мати наступний вигляд:

$$f(t) = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{t^2}{2}}. \tag{16.20}$$

Цей розподіл називають *одиничним (стандартизованим) нормальним розподілом*. Він характеризується тим, що його вибіркові параметри дорівнюють, відповідно, $\bar{x} = 0$ і $\sigma = 1$.

Графік функції щільності розподілу одиничного нормального розподілу приведено на рисунку 16.3.

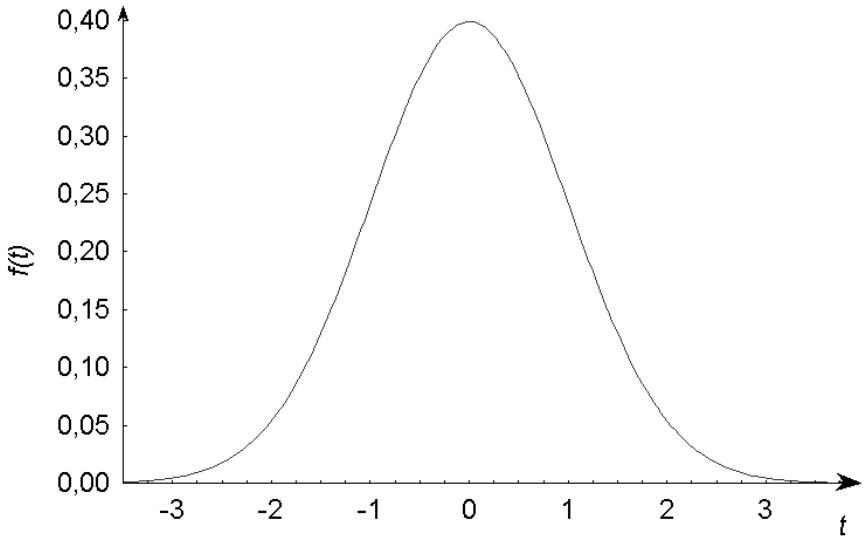

Рисунок 16.3 – Графік функції щільності одиничного нормального розподілу

Дана крива називається кривою Гауса-Лапласа, а її рівняння (16.20) виражає функціональний зв'язок між імовірністю $P(t_i)$ і нормованим відхиленням t_i.

Ординати одиничної нормальної кривої (тобто значення *f(t)*) наведено в Додатку К для будь-яких *t* від 0 до 3,59. У випадку якщо величина *t* від'ємна, її ординату знаходять, використовуючи наступну особливість нормального розподілу:

$$f(-t) = f(t). \tag{16.21}$$

Інтегруючи вираз 16.20, одержуємо інтегральну функцію нормального розподілу *Φ(t)*:

$$\Phi(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{-\frac{t^2}{2}} dt. \tag{16.22}$$

Вираз 16.22 визначає частку варіант у вибірці, значення яких не перевершують величину *t*.

Графік інтегральної функції нормального розподілу наведено на рисунку 16.4, а значення для будь-яких *t* від 0 до 3,59 – у Додатку Л.

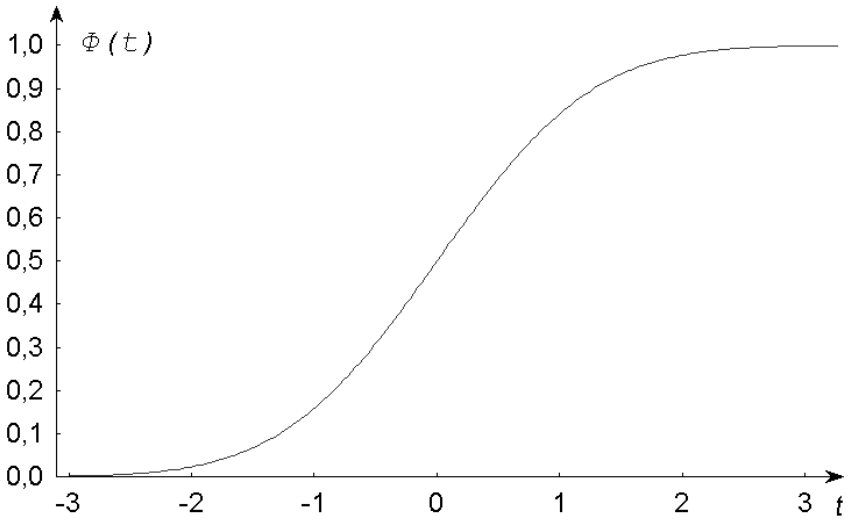

Рисунок 16.4 – Графік інтегральної функції одиничного нормального розподілу

Для від'ємних значень *t* користуються тією властивістю нормального розподілу, що:

$$\Phi(-t) = 1 - \Phi(t). \qquad (16.23)$$

Особливості нормального розподілу дуже часто використовуються в селекційній роботі.

Приклад. В стаді великої рогатої худоби, що містить 1000 голів, середній надій становить 3500 кг, а середнє квадратичне відхилення – 500 кг.

Скільки корів у даному стаді має надій 4000 кг?

Спочатку необхідно перейти від вихідних значень до стандартизованих величин. Для значення $x = 4000$ нормоване відхилення буде становити:

$$t = \frac{4000 - 3500}{500} = +1{,}0 \ .$$

Згідно Додатку К знаходимо ординату для значення $t = 1{,}0$. Ця величина складає $f(1{,}0) = 0{,}242$. Отже, з 1000 голів, надій 4000 кг буде мати $0{,}242 \times 1000 = 242$ корови.

Приклад. В стаді великої рогатої худоби, що містить 1000 голів, середній надій становить 3500 кг, а середнє квадратичне відхилення – 500 кг.

Скільки корів даного стада варто записати в ДПК, якщо стандарт I класу за надоєм для досліджуваної породи 4500 кг?

Спочатку знову перейдемо від вихідних значень до стандартизованих величин. Для значення $x = 4500$ нормоване відхилення буде становити:

$$t = \frac{4500 - 3500}{500} = +2{,}0 \ .$$

В Додатку Л знаходимо для величини t = 2,0 значення інтегральної функції нормального розподілу. Воно складає $\Phi(2,0)$ = 0,9772. Отже, з 1000 голів даного стада 1000 × 0,9772 = 977 корів мають надій, що не перевищує 4500 кг і, отже, лише 23 корови можуть бути занесені в ДПК, оскільки їхній надій перевищує 4500 кг.

Приклад. У стаді великої рогатої худоби, що містить 1000 голів, середній надій становить 3500 кг, а середнє квадратичне відхилення – 500 кг.

Скільки корів із даного стада мають надій від 3000 до 4000 кг?

Графічне зображення рішення даної задачі наведено на рисунку 16.5.

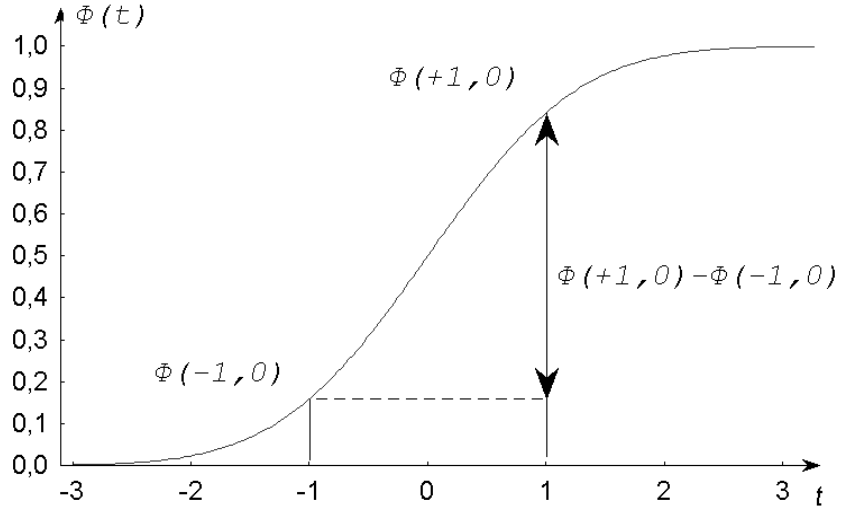

Рисунок 16.5 – Графічне рішення задачі з прикладу

По-перше, замінимо обидва граничні значення їх стандартизованими величинами:

$$t_{x=3000} = \frac{3000 - 3500}{500} = -1{,}0;$$

$$t_{x=4000} = \frac{4000 - 3500}{500} = +1{,}0.$$

Таким чином, для того, щоб обчислити частку тварин, що мають рівень продуктивності від 3000 до 4000 кг, необхідно, використовуючи таблицю інтегральної функції (Додаток Л), обчислити, скільки тварин мали рівень продуктивності не більше 4000 кг, а потім відняти від цієї величини частку тварин, що мають рівень молочної продуктивності не більше 3000 кг.

Таким чином, шукана величина дорівнює різниці $\Phi(+1,0) - \Phi(-1,0)$.

У свою чергу, використовуючи формулу (16.23), можна показати, що:

$$\Phi(-1,0) = 1 - \Phi(+1,0).$$

У Додатку Л знаходимо, що $\Phi(+1,0)$ = 0,8413. Тоді шукана величина дорівнює:

$$\Phi(+1,0) - (1 - \Phi(+1,0)) = 0{,}8413 - (1 - 0{,}8413) = 0{,}6826.$$

Отже, серед 1000 голів рівень продуктивності від 3000 кг до 4000 кг будуть мати 1000 × 0,6826 = 683 корови.

Обидві функції одиничного нормального розподілу знаходять широке застосування при прогнозуванні ефекту селекції. Знаючи точку відсікання за рівнем продуктивності тварин у племінне ядро (*t*), можна розрахувати інтенсивність селекції:

$$i = \frac{f(t)}{1 - \Phi(t)}, \tag{16.24}$$

де $t = \frac{x - M}{\sigma}$;

x – точка відсікання тварин вибірки в племядро;

M – середня продуктивність тварин вибірки;

σ – середнє квадратичне значення.

Середнє арифметичне значення рівня продуктивності тварин племядра тоді можна розрахувати за формулою:

$$M_{PP} = M + i\sigma, \tag{16.25}$$

а прогнозований рівень продуктивності нащадків тварин із племядра:

$$M_{OFF} = M + i \cdot \sigma \cdot h^2, \tag{16.26}$$

де h^2 – коефіцієнт успадкування ознаки, за якою ведеться селекція.

Графічне зображення селекційного диференціала (*Sd*) та ефекту селекції (*SE*) наведено на рисунку 16.6.

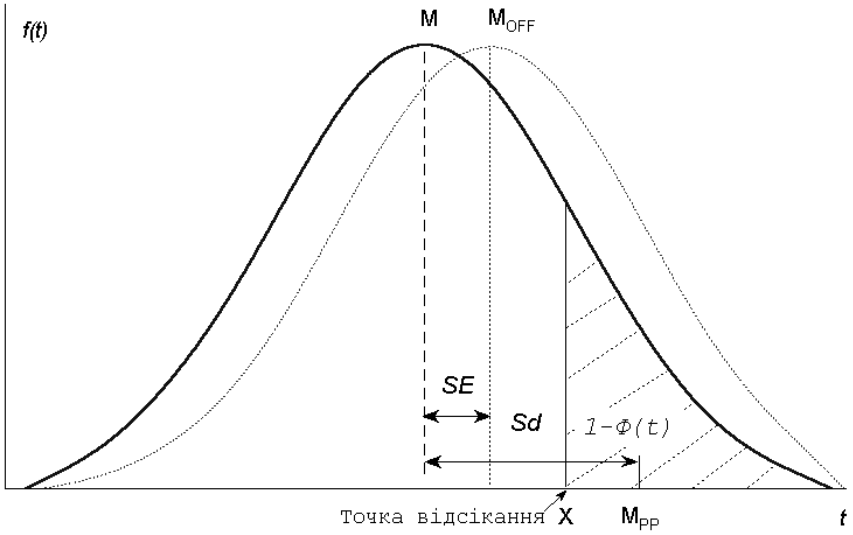

Рисунок 16.6 – Модельний приклад, що демонструє принцип розрахунку селекційного диференціала та ефекту селекції

Приклад. Середня жива маса овець при народженні становить 4000 г із середнім квадратичним відхиленням 900 г. Яку можна очікувати середню живу масу при народженні ягнят, отриманих від тварин, відібраних у племядро, якщо границя відбору становить 4900 г? Коефіцієнт успадкування даної ознаки становить: h^2 = 0,3.

Переведемо величину границі відсікання тварин стада в племядро в нормоване відхилення:

$$t = \frac{4900 - 4000}{900} = +1{,}0 .$$

Далі, використовуючи Додаток К, знайдемо ординату нормальної кривої для даного значення нормованого відхилення границі відсікання: $f(+1{,}0) = 0{,}242$, а по Додатку Л – значення інтеграла її ймовірності: $\Phi(+1{,}0) = 0{,}8413$.

Далі підставляємо ці значення у формулу 16.24 для розрахунку інтенсивності селекції:

$$i = \frac{0{,}2420}{1 - 0{,}8413} = 1{,}522 .$$

Останню величину, а також значення середнього квадратичного відхилення і коефіцієнта успадкування підставляємо у формулу 16.26 і розраховуємо прогнозовану середню живу масу при народженні ягнят наступного покоління:

$$M_{OFF} = 4000 + 1{,}522 \cdot 900 \cdot 0{,}3 = 4410{,}9 \text{ г.}$$

Як легко розрахувати, ефект селекції в цьому випадку можна очікувати приблизно 411 г.

Аналогічно, використовуючи формулу 16.25, можна розрахувати середнє значення маси при народженні ягнят, народжених від тварин, відібраних у племядро. Воно становить 5370 г.

Аналогічні розрахунки можна провести і знаючи лише частку відібраних у племядро тварин (p), оскільки вона пов’язана з інтегралом імовірності одиничного нормального розподілу простою залежністю:

$$p = 1 - \Phi(t). \qquad (16.27)$$

Приклад. Середня жива маса овець при народженні складає 4000 г із середнім квадратичним відхиленням 900 г. Який можна очікувати ефект селекції, якщо в племядро відібрано 25% найбільш крупних при народженні овець? Коефіцієнт успадкування даної ознаки становить: $h^2 = 0{,}3$.

За умовою: $p = 0{,}25$. Тоді, використавши формулу 16.27, одержуємо, що $\Phi(t) = 1 - 0{,}25 = 0{,}75$. Потім, в Додатку Л знаходимо значення t, для якого значення $\Phi(t)$ складає 0,75. Це значення дорівнює 0,67.

Отже, інтенсивність селекції становитиме:

$$i = \frac{f(+0{,}67)}{1 - \Phi(+0{,}67)} = \frac{0{,}3187}{0{,}2500} = 1{,}275,$$

а ефект селекції, відповідно, $SE = 1{,}275 \times 900 \times 0{,}3 = 344$ г.

Контрольні питання:

1. Принцип застосування критерію Колмогорова-Смирнова.
2. Яким чином визначається відхилення вибіркового розподілу від нормального закону на основі використання коефіцієнта асиметрії чи ексцесу?

§ 17. Перевірка статистичних гіпотез. Параметричні методи

Статистичною гіпотезою (*H*) називається припущення щодо параметрів чи форми розподілу генеральної сукупності (чи сукупностей), що перевіряється на підставі даних вибіркового розподілу. З параметрів розподілу найчастіше виникає необхідність перевірити рівність середніх арифметичних у двох вибірках (перевірка у відношенні центральної тенденції) чи вибіркових варіанс (перевірка у відношенні рівня мінливості). В обох випадках мається на увазі, що обидві вибірки взяті з генеральних сукупностей із нормальним типом розподілу (див. § 16). Якщо характер розподілу вибіркових варіант відрізняється від нормального, використовуються непараметричні методи.

При перевірці статистичних гіпотез у відношенні центральної тенденції найчастіше виникає необхідність оцінити вірогідність розходжень між двома вибірковими значеннями середніх арифметичних, розрахованих на підставі двох вибірок. Ця оцінка завжди проводиться з використанням *критерію Ст'юдента* (його ще називають *t*-тест). Однак, цей критерій має кілька модифікацій і вибір відповідної для кожного випадку визначається властивостями самих вибірок.

Насамперед, необхідно визначити чи є порівнювані вибірки *залежними* (наприклад, коли оцінюється рівень молочної продуктивності у матерів та їхніх дочок із одного стада) чи *незалежними* (наприклад, коли проводиться порівняння рівня продуктивності двох груп тварин, що належать різним господарствам). У першому випадку проводиться оцінка рівня значущості вірогідності відмінності від нуля *середньої арифметичної різниці пар значень* у двох вибірках, а в другому – оцінка вірогідності розходжень *різниці двох вибіркових середніх арифметичних*.

Залежні вибірки

Як вже вказувалося вище, у випадку залежних вибірок нульова й альтернативна гіпотези формулюються таким способом:

$$H_0: \bar{d} = 0;$$

$$H_A: \bar{d} \neq 0,$$

де $\bar{d}$ – середня різниця між парними варіантами першої (x) і другої вибірок (y), тобто, $\bar{d} = \overline{x - y}$.

Перевірка нульової гіпотези в цьому випадку проводиться, використовуючи наступну формулу критерію Ст'юдента:

$$t = \frac{\left|\bar{d}\right|}{\sqrt{\frac{\sum d_i^2 - n\bar{d}^2}{n(n-1)}}}, \qquad (17.1)$$

де n – обсяг вибірок, а $d_i = x_i - y_i$.

Перевірка рівня значущості отриманого значення критерію Ст'юдента проводиться, використовуючи табульовані значення критерію, із числом

ступенів свободи: $df = n - 1$. Табличні значення критерію Ст'юдента для різних рівнів значущості наведено в Додатку И.

Приклад. Було визначено вміст жиру в молоці у 10 корів та їхніх дочок. Чи можна сказати, що дочки більш жирномолочні, ніж матері? Усі вихідні дані і проміжні величини приведені в таблиці 17.1.

Таблиця 17.1 – Вихідні дані та проміжні величини для визначення вірогідності різниці між дочками та матерями

№ п/п	Вміст жиру в молоці матерів, % (x)	Вміст жиру в молоці дочок, % (y)	d_i	d_i^2
1	3,6	3,5	0,1	0,01
2	3,3	3,6	-0,3	0,09
3	3,7	3,6	0,1	0,01
4	3,5	3,8	-0,3	0,09
5	3,0	3,5	-0,5	0,25
6	3,1	3,4	-0,3	0,09
7	4,0	3,9	0,1	0,01
8	3,4	3,7	-0,3	0,09
9	3,8	4,0	-0,2	0,04
10	3,7	4,1	-0,4	0,16
Суми	35,1	37,1	-2,0	0,84

Середня різниця тоді буде дорівнювати: $\bar{d} = \frac{-2,0}{10} = -0,2$. Отже, оцінка критерію Ст'юдента становитиме:

$$t = \frac{0,2}{\sqrt{\frac{0,84 - 10 \cdot 0,2^2}{10 \cdot 9}}} = 2,86.$$

Число ступенів свободи дорівнює: $df = 10 - 1 = 9$. Табличні значення для цього числа ступенів свободи складають:

$$t_{\alpha=0,05} = 2,26;$$
$$t_{\alpha=0,01} = 3,25.$$

Отже, можна зробити висновок, що нульова гіпотеза повинна бути відхилена із рівнем значущості $0,01 < p < 0,05$. (Точний рівень значущості $p = 0,019$).Дочки виявляються вірогідно більш жирномолочними, ніж їх матері.

Незалежні вибірки

Для незалежних вибірок нульова й альтернативна гіпотези формулюються таким способом:

$$H_0: \ \bar{x} - \bar{y} = 0;$$
$$H_A: \ \bar{x} - \bar{y} \neq 0.$$

У цьому випадку вибір формули для критерію Ст'юдента залежить від рівності варіанс двох вибіркових сукупностей. Ця перевірка проводиться, використовуючи F-критерій Фішера-Снедекора:

$$F = \frac{\sigma_1^2}{\sigma_2^2}. \qquad (17.2)$$

Значення критерію Фішера-Снедекора завжди перевищує одиницю, оскільки являє собою відношення більшої варінси до меншої. Отримане в такий спосіб значення критерію порівнюється із табличним для відповідних чисел ступенів свободи: $df_1 = n_1 - 1$ і $df_2 = n_2 - 1$. Відповідно, n_1 і n_2 – обсяги порівнюваних вибірок.

Табличні значення критерію Фішера-Снедекора наведено в Додатку Ж. Однак, якщо обсяги вибірок досить великі (більше 15-20 варіант), рівень значущості *F*-критерію можна оцінити, використовуючи *u*-апроксимацію:

$$u = \frac{(1-k_2)\cdot\sqrt[3]{F} - (1-k_1)}{\sqrt{k_2\sqrt[3]{F^2} + k_1}}, \qquad (17.3)$$

де $k_1 = 2/(9\cdot df_1)$, а $k_2 = 2/(9\cdot df_2)$.

Критичне значення *u*-критерію, при якому визнається вірогідним розходження між обома порівнюваними вибірковими варіансами, становить 1,645.

При рівності варіанс, значення критерію Ст'юдента розраховується за формулою:

$$t = \frac{|\bar{x} - \bar{y}|\sqrt{\frac{n_x \cdot n_y}{n_x + n_y}}}{\sqrt{\frac{\sigma_x^2(n_x - 1) + \sigma_y^2(n_y - 1)}{n_x + n_y - 2}}}. \qquad (17.4)$$

Цей вираз значно спрощується, якщо обсяги вибірок рівні між собою:

$$t = \frac{|\bar{x} - \bar{y}|}{\sqrt{\frac{\sigma_x^2 + \sigma_y^2}{n}}}. \qquad (17.5)$$

Розраховане значення критерію порівнюється із табличним для числа ступенів свободи: $df = n_x + n_y - 2$. (У випадку рівності вибірок ця формула перетвориться в $df = 2\cdot(n - 1)$).

Приклад. У двох групах корів червоної степової породи було проведено оцінку жирномолочності. В першій групі, що містила 50 тварин, середній вміст жиру в молоці був 3,0% із середнім квадратичним відхиленням – 0,15%. В другій групі, що містила 75 корів, середній вміст жиру в молоці був 3,8% із середнім квадратичним відхиленням – 0,18%. Чи можна стверджувати, що обидві групи вірогідно відрізняються за жирномолочністю?

Спочатку, використовуючи критерій Фішера-Снедекора, з'ясуємо чи вірогідно відрізняються варіанси порівнюваних вибірок чи ні. Підставляємо вихідні значення у формулу 17.2, причому в чисельнику записуємо значення варіанси для другої вибірки (оскільки воно більше):

$$F = \frac{0{,}18^2}{0{,}15^2} = 1{,}44.$$

Оскільки обсяги вибірок великі, оцінку значущості цієї величини можна розрахувати, використовуючи апроксимацію за формулою 17.3:

$$u = \frac{(1-0{,}0045)\cdot\sqrt[3]{1{,}44} - (1-0{,}003)}{\sqrt{0{,}0045\cdot\sqrt[3]{1{,}44^2} + 0{,}003}} = 1{,}36 \ .$$

Оскільки ця величина не перевищує граничну (тобто 1,645), робимо висновок, що оцінки варіанс у порівнюваних вибірках вірогідно не відрізняються. Отже, можна використовувати формулу 17.4 для розрахунку величини критерію Ст'юдента:

$$t = \frac{\left|3{,}0 - 3{,}8\right|\sqrt{\frac{50\cdot 75}{50+75}}}{\sqrt{\frac{0{,}15^2\cdot(50-1) + 0{,}18^2\cdot(75-1)}{50+75-2}}} = 26{,}3.$$

Число ступенів свободи для цього значення складає:

$$df = 50 + 75 - 2 = 123.$$

Безпосередня інтерпретація отриманого значення критерію Ст'юдента є неможливою, оскільки в Додатку И не наведено табличне значення для числа ступенів свободи рівного 123. Але, як можна помітити, при збільшенні числа ступенів свободи від 120 до 200 табличні значення критерію Ст'юдента практично не змінюються. Тому можна впевнено стверджувати, що розраховане для прикладу значення критерію Стьюдента (26,3) має високий рівень значущості ($p < 0{,}001$). (Точне значення рівня значущості для критерію Ст'юдента становить $p = 2{,}48\cdot 10^{-52}$.)

Таким чином, можна стверджувати, що рівень жирномолочності корів другої групи вірогідно вищий, ніж у корів першої групи.

У тому випадку, якщо оцінки вибіркових варіанс вірогідно відрізняються між собою, оцінку вірогідності розходжень між двома середніми проводять на підставі формули:

$$t = \frac{\left|\bar{x} - \bar{y}\right|}{\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}} \qquad (17.6)$$

із числом ступенів свободи:

$$df = (n_x + n_y - 2)\cdot\left(\frac{1}{2} + \frac{\sigma_x^2\sigma_y^2}{\sigma_x^4 + \sigma_y^4}\right). \qquad (17.7)$$

Приклад. Середня жива маса корів герефордської породи в стаді чисельністю 150 голів (при аутбридингу) становила 600 кг із середнім квадратичним відхиленням 20 кг, а середня жива маса корів цієї ж породи в іншому стаді чисельністю 200 голів (при тісному інбридингу) – 580 кг із середнім квадратичним відхиленням 30 кг. Чи впливає інбридинг на живу масу корів?

Спочатку знову, використовуючи критерій Фішера-Снедекора, з'ясуємо, чи вірогідно розрізняються варіанси порівнюваних вибірок чи ні. Підставляємо вихідні значення у формулу 17.2:

$$F = \frac{30^2}{20^2} = 2{,}25.$$

Розраховане за формулою 17.3 значення *u*-критерію для даного значення критерію Фішера-Снедекора становить 5,10, що набагато перевищує критичне (1,645). Таким чином, вибіркові варіанси відрізняються вірогідно і для розрахунку оцінки критерію Ст'юдента потрібно використовувати формулу 17.6:

$$t = \frac{|600 - 580|}{\sqrt{\frac{20^2}{150} + \frac{30^2}{200}}} = 7{,}46.$$

Використовуючи формулу 17.7, розрахуємо число ступенів свободи для даного значення критерію Ст'юдента:

$$df = (150 + 200 - 2) \cdot \left(\frac{1}{2} + \frac{20^2 \cdot 30^2}{20^4 + 30^4} \right) \approx 303.$$

У Додатку И не наведено критичних значень для такого значення числа ступенів свободи, однак при $df > 200$ можна в якості табличних використовувати значення 1,96; 2,59 і 3,31 для трьох рівнів значущості, відповідно. Оскільки розраховане значення критерію Ст'юдента набагато перевищує граничне, навіть для третього рівня значущості, можна зробити висновок, що тісний інбридинг призводить до вірогідного зниження живої маси корів із рівнем значущості: $p < 0{,}001$. (Точне значення рівня значущості становить $p = 9{,}2 \cdot 10^{-13}$.)

Середнє арифметичне значення не може служити єдиною характеристикою вибіркового розподілу, оскільки рівність у відношенні центральної тенденції не гарантує рівності двох вибірок за рівнем мінливості чи характером розподілу. Наприклад, дві вибірки, що мають тип розподілу, представлений на рисунку 17.1 не розрізняються у відношенні оцінок середніх арифметичних значень ($M_1 = M_2$), однак у відношенні рівня мінливості розрізняються значно (σ_1 майже в два рази нижче, ніж σ_2).

Тому детальний аналіз вибіркових матеріалів вимагає перевірки не тільки рівності середніх, але і рівності рівня мінливості.

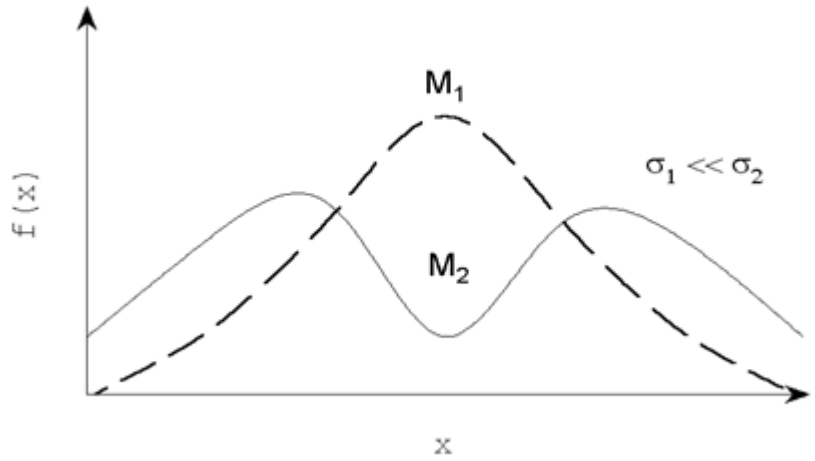

Рисунок 17.1 – Приклад розподілу двох гіпотетичних вибірок з рівними середніми, але що розрізняються у відношенні середніх квадратичних відхилень

Крім того, від рівності чи нерівності вибіркових варіанс залежить і вибір формули критерію Ст'юдента, яка використовується при порівнянні двох середніх значень.

Крім того, перевірка рівності варіанс – обов'язковий попередній етап при проведенні дисперсійного аналізу Р. Фішера, оскільки в його основу покладено умову нормальності розподілів, з яких відібрані вибірки, і рівності вибіркових варіанс (див. § 18).

Вибір того чи іншого статистичного критерію для перевірки рівності вибіркових варіанс залежить від особливостей самих порівнюваних вибіркових сукупностей. Схематично це представлено в таблиці 17.2.

Таблиця 17.2 – Статистичні критерії для різних вибірок

<table>
<tr><td colspan="3">Вибірки мають нормальний розподіл</td><td rowspan="3">Вибірки не мають нормальний розподіл</td></tr>
<tr><td rowspan="2">дві вибірки</td><td colspan="2">більше, ніж дві вибірки</td></tr>
<tr><td>обсяги вибірок однакові</td><td>обсяги вибірок відрізняються</td></tr>
<tr><td>Критерій Фішера-Снедекора</td><td>Критерій Кохрена</td><td>Критерій Бартлетта</td><td>Критерій Левене
Критерій Брауна-Форсайта</td></tr>
</table>

Критерій Фішера-Снедекора

Критерій Фішера-Снедекора (F-критерій) використовується у випадку перевірки рівності двох варіанс. Особливе значення цей критерій має у дисперсійному аналізі Р. Фішера, коли проводиться порівняння факторіальної і випадкової варіанс (див. § 18).

У найпростішому випадку, при досить великому обсязі двох порівнюваних вибірок, перевірка рівності рівня мінливості (насамперед, вибіркових середніх квадратичних відхилень) може бути зроблена, використовуючи *и*-критерій:

$$u = \frac{|\sigma_1 - \sigma_2|}{\sqrt{\frac{\sigma_1^2}{2(n_1 - 1)} + \frac{\sigma_2^2}{2(n_2 - 1)}}}. \qquad (17.8)$$

Розходження між двома вибірковими середніми квадратичними відхиленнями вважаються вірогідними, якщо оцінка *u*-критерію перевищує 1,96. Однак, у випадку нечисленних вибірок ($n_1 = n_2 < 15$) цей критерій не прийнятний.

Застосувавши нескладні перетворення можна показати, що значення, одержане при використанні формули 17.8 залежить, насамперед, від співвідношення $\frac{\sigma_1}{\sigma_2}$, тобто, співвідношення вибіркових варіанс. Остання величина і зветься критерієм Фішера-Снедекора:

$$F = \frac{\sigma_1^2}{\sigma_2^2}. \qquad (17.9)$$

Детальний опис використання цього критерію (а також його нормальної апроксимації у випадку великого обсягу вибірок) дано вище.

Критерій Кохрена

Критерій Кохрена (G_C) використовується в тих випадках, коли необхідно перевірити гіпотезу про рівність декількох вибіркових варіанс одночасно, тобто ***гомоскедастичності*** вибірок (як, наприклад, у випадку проведення дисперсійного аналізу) при рівності обсягів усіх вибірок (тобто, $n_1 = n_2 = n_3 = \ldots = n_k$)... Оцінка цього критерію проводиться за формулою:

$$G_C = \frac{\sigma_{\max}^2}{\sigma_1^2 + \sigma_2^2 + \ldots + \sigma_k^2}, \qquad (17.10)$$

де $\sigma_{\max}^2$ – максимальна за значенням вибіркова варіанса;

k – число порівнюваних вибірок.

Нуль-гіпотеза про рівність вибіркових варіанс відкидається, якщо розраховане значення критерію Кохрена перевищує табличне при заданому числі вибірок (*k*) і числі ступенів свободи ($df = n - 1$). Критичні значення критерію Кохрена наведено в Додатку М.

Приклад. Було вивчено вплив ефекту гетерозису на відтворювальні ознаки (масу гнізда при відлученні) при різних поєднаннях двох порід свиней. Статичні показники даної ознаки наведено в таблиці 17.3.

Таблиця 17.3 – Статистичні показники маси гнізда при відлученні за різних поєднань

Поєднання	$\bar{X}$, кг	σ, кг
♀ Велика біла × ♂ Велика біла	65,3	4,70
♀ Дюрок × ♂ Дюрок	63,1	2,42
♀ Дюрок × ♂ Велика біла	65,9	4,09
♀ Велика біла × ♂ Дюрок	65,0	6,11

Необхідно перевірити гіпотезу, що тип поєднання не впливає на рівень мінливості даної ознаки. У кожному випадку було проаналізовано по 20 свиноматок.

Як видно із даних, що наведені в таблиці 17.3, тип поєднання практично не впливав на середнє значення маси гнізда при відлученні свиноматок. Однак рівень мінливості цього показника варіював досить значно. Використовуємо критерій Кохрена для перевірки значущості цього варіювання:

$$G_C = \frac{6{,}11^2}{4{,}70^2 + 2{,}42^2 + 4{,}09^2 + 6{,}11^2} = 0{,}455.$$

У Додатку М, на жаль, немає точного значення для числа ступенів свободи: $df = 20 - 1 = 19$. Однак, можна помітити, що при чотирьох вибірках ($k = 4$) табличне значення критерію Кохрена навіть при $df = 18$ виявляється нижчим, ніж розраховане: $G_C = 0{,}425$. Тому нульову гіпотезу варто відхилити (на рівні значущості $p = 0{,}05$) і вважати доведеним, що вид поєднання впливає на рівень мінливості маси гнізда при відлученні досліджених свиноматок. (Точне значення критерію Кохрена в даному випадку для $df = 19$ складає: $G_C = 0{,}421$).

Критерій Бартлетта

У тих випадках, коли обсяги порівнюваних вибірок не однакові, для перевірки гіпотези про рівність декількох вибіркових варіанс необхідно використовувати більш складний критерій Бартлетта (M).

Єдина вимога до даних, що аналізуються, при використанні цього критерію – це нормальність вибіркового розподілу. Обсяги вибірок можуть варіювати від 3 до $+\infty$.

При використанні критерію Бартлетта розраховується його значення за формулою:

$$M = N\left(\ln \frac{1}{N} \sum (n_i - 1)\sigma_i^2 \right) - \sum (n_i - 1) \ln \sigma_i^2, \qquad (17.11)$$

де $N = \sum (n_i - 1).$

Показано, що якщо нульова гіпотеза про рівність усіх вибіркових варіанс справедлива, то величина:

$$\chi^2 = \frac{M}{1 + \frac{\left(\sum \frac{1}{n_i - 1} \right) - \frac{1}{N}}{3(k-1)}} \qquad (17.12)$$

має розподіл Хі-квадрат із числом ступенів свободи: $df = k - 1$, де k – кількість проаналізованих вибірок.

У тому випадку, якщо значення критерію Бартлетта перевищує табличне значення критерію Хі-квадрат (див. Додаток Д), то нуль-гіпотеза відхиляється і, відповідно, вважається статистично доведеною нерівність вибіркових варіанс.

Приклад. Проаналізуємо дані з таблиці 17.3, використовуючи критерій Бартлетта. Для наочності будемо вважати, що одна з вибірок містила дані про 10 свиноматок. Усі вихідні і проміжні дані занесено в таблицю 17.4.

Таблиця 17.4 – Вихідні дані для розрахунку критерію Бартлетта

Поєднання	σ^2	$(n_i - 1)$	$1/(n_i - 1)$	$(n_i - 1)\,\sigma^2$	$\ln \sigma^2$	$(n_i - 1)\ln\sigma^2$
♀ Велика біла × ♂ Велика біла	22,09	9	0,111	198,81	3,095	27,855
♀ Дюрок × ♂ Дюрок	5,86	19	0,053	111,34	1,768	33,592
♀ Дюрок × ♂ Велика біла	16,73	19	0,053	317,87	2,817	53,523
♀ Велика біла × ♂ Дюрок	37,33	19	0,053	709,27	3,620	68,780
Суми	-	66	0,270	1337,29	-	183,750

Тоді значення критерію Бартлетта буде дорівнювати:

$$M = 66 \cdot \left[\ln \frac{1337,29}{66} \right] - 183,750 = 14,827.$$

Далі, оцінимо рівень значущості критерію Бартлетта, а для цього розрахуємо значення критерію Хі-квадрат за формулою 17.12:

$$\chi^2 = \frac{14,827}{1 + \dfrac{0,270 - \dfrac{1}{66}}{3 \times (4-1)}} = 14,423.$$

Отримане значення критерію Хі-квадрат порівнюємо з табличним при числі ступенів свободи $df = k - 1 = 4 - 1 = 3$ (див. Додаток Д):

$$\chi^2_{\alpha=0,05} = 7,81;$$
$$\chi^2_{\alpha=0,01} = 11,34;$$
$$\chi^2_{\alpha=0,001} = 16,27.$$

Як бачимо, рівень значущості розрахованої нами величини знаходиться в межах $0,001 < p < 0,01$. (Точне значення рівня значущості критерію Хі-квадрат у цьому випадку становить: $p = 0,0024$.)

Таким чином, нульова гіпотеза повинна бути відхилена і прийнята альтернативна гіпотеза про нерівність вибіркових варіанс.

Критерій Левене

Усі розглянуті вище критерії для порівняння вибіркових варіанс можуть бути використані лише у тих випадках, коли вибіркові розподіли слабко відхиляються від нормального (див. § 16). У випадках, коли є інформація про недотримання цієї вимоги, більш коректним способом перевірки гіпотези щодо гомоскедастичності (тобто, рівності вибіркових варіанс) є непараметричні

тести, найбільш розповсюдженим з яких є критерій Левене (W_0) (чи його модифікація – критерій Брауна-Форсайта).

При використанні непараметричних критеріїв перевірки гомоскедастичності для декількох вибірок одночасно, всі вихідні дані попередньо замінюються на модулі їхнього відхилення від оцінки центра розподілу в кожній вибірці:

$$z_{ij} = \left| x_{ij} - M_i \right|,$$

При цьому, якщо в якості оцінки центра розподілу використовується вибіркове середнє арифметичне значення (тобто, $M_i = \bar{x}_i$), ми маємо справу із критерієм Левене. Якщо ж у якості оцінки центра розподілу використовується оцінка вибіркової медіани ($M_i = Me_i$) – із критерієм Брауна-Форсайта.

Далі модифіковані значення використовуються для проведення розрахунків за алгоритмом однофакторного дисперсійного аналізу Р. Фішера (див. § 18). У цьому випадку значення критерію Левене розраховується за формулою:

$$W_0 = \frac{\left(\sum_{i=1}^{k} n_i (\bar{z}_i - \bar{\bar{z}})^2 \right) / (k-1)}{\left(\sum_{ij} (z_{ij} - \bar{z}_i)^2 \right) / \sum_{i=1}^{k} (n_i - 1)}, \qquad (17.13)$$

де k – кількість вибірок;

n_i – обсяг i-тої вибірки;

$\bar{z}_i$ – середнє арифметичне значення i-тої вибірки;

$\bar{\bar{z}}$ – загальне середнє арифметичне значення для всієї сукупності даних.

Оцінка рівня значущості критерію Левене проводиться шляхом порівняння розрахованого значення W_0 із табличним значенням F-критерію Фішера-Снедекора із числом ступенів свободи: $df_1 = k - 1$, $df_2 = N - k$, де N – сумарний обсяг усіх вибірок.

Основна ідея застосування критерію Левене (і його модифікації) полягає в тому, що у випадку низької варіабельності вихідних даних у будь-якій вибірці, модифіковані z-значення будуть мало відхилятися від нуля; у випадку ж високого рівня варіабельності вихідних значень вибірки – значення z, навпаки, будуть високими. Класичний однофакторний дисперсійний аналіз тоді доведе статистичну значущість розходжень між модифікованими значеннями аналізованих вибірок і, відповідно, нерівність рівня мінливості вихідних даних.

Критерій Левене може бути використаний також і у випадку двох вибірок (непараметричний аналог F-критерію Фішера-Снедекора).

Приклад. Необхідно перевірити гіпотезу щодо гомоскедастичності в трьох вибірках корів червоної степової породи щодо вмісту жиру в молоці.

Усі вихідні і проміжні дані приведені в таблиці 17.5.

Таблиця 17.5 – Вихідні дані для перевірки гіпотези щодо гомоскедастичності в трьох вибірках корів щодо вмісту жиру в молоці

j	*i* = 1		*i* = 2		*i* = 3	
	x	*z*	*x*	*z*	*x*	*z*
1	3,75	*0,087*	3,90	*0,045*	3,85	*0,041*
2	4,20	*0,363*	3,89	*0,035*	3,82	*0,011*
3	3,69	*0,147*	3,85	*0,005*	3,88	*0,071*
4	3,56	*0,277*	3,86	*0,005*	3,85	*0,041*
5	3,99	*0,153*	3,85	*0,005*	3,92	*0,111*
6	3,84	*0,003*	3,92	*0,065*	3,72	*0,089*
7	3,87	*0,033*	3,80	*0,055*	3,79	*0,019*
8	3,85	*0,013*	3,90	*0,045*	3,76	*0,049*
9	3,80	*0,037*	3,81	*0,045*	3,76	*0,049*
10	3,82	*0,017*	3,77	*0,085*	3,74	*0,069*
Середні	3,837		3,855		3,809	

Спочатку необхідно розрахувати середні арифметичні значення в межах кожної групи:

$$\bar{x}_1 = 3{,}837\ (\%);$$

$$\bar{x}_2 = 3{,}855\ (\%);$$

$$\bar{x}_3 = 3{,}809\ (\%).$$

Далі, у межах кожної вибірки розраховуємо модифіковані *z*-значення; наприклад, перше значення першої вибірки тоді буде дорівнювати:

$$z_{11} = |3{,}75 - 3{,}837| = 0{,}087,$$

друге значення першої вибірки:

$$z_{12} = |4{,}20 - 3{,}837| = 0{,}363,$$ і т. д.

Модифіковані *z*-значення використовуємо для проведення дисперсійного аналізу Р. Фішера. Не зупиняючись докладно на особливостях розрахунку (вони наведені в § 18), приведемо відразу результати аналізу (табл. 17.6).

Таблиця 17.6 – Результати дисперсійного аналізу

Мінливість	Сума квадратів (*C*)	Число ступенів свободи (*df*)	Варіанса ($\sigma^{2)}$	Дисперсійне відношення (*F*)	Рівень значущості (*p*)
Факторіальна (*X*)	0,03032	2	0,01516	2,704	0,085
Випадкова (*Z*)	0,15140	27	0,00561		
Сумарна (*Y*)	0,18172	29	-		

Таким чином, можна зробити висновок, що наявна деяка тенденція до порушення гомоскедастичності в трьох розглянутих вибірках. Проаналізувавши модифіковані *z*-значення в різних вибірках можна

відзначити, що в першій вибірці зустрічаються дуже високі значення, тоді як у другій і третій – їхні величини досить вирівняні. Це свідчить про більш високий рівень мінливості в першій вибірці, порівняно з іншими.

І дійсно, якщо розглянемо їхні варіанси (0,0293, 0,0024 і 0,0043, відповідно), то можна помітити, що рівень мінливості в першій вибірці майже на порядок вище. (Критерії Кохрена і Бартлетта в цьому випадку дають більш однозначний доказ для відкидання нульової гіпотези).

Рандомізований критерій

У тих випадках, коли обсяги вибірок малі ($n \leq 5$), що не дозволяє використання ні параметричних, ні непараметричних методів порівняння двох вибірок, на допомогу можуть прийти методи ресамплінгу, а саме, **рандомізований критерій** (від англ. random – випадковий).

При залежних вибірках для його розрахунку необхідно провести наступні обчислення:

1) знайти всі парні різниці між значеннями першої та другої порівнюваних вибірок;

2) знайти суму парних різниць із урахуванням їх знаків ($\sum R$);

3) перелічити всі можливі комбінації різниць (N), що призводять до одержання суми рівної чи більшої, ніж отримане значення $\sum R$;

4) розрахувати всі можливі комбінації одержання будь-якої суми на основі величин різниць із урахуванням їх знаків (як додатних, так і від'ємних) шляхом зведення 2 у ступінь n;

5) розрахувати імовірність прийняття нульової гіпотези за формулою:

$$p = \frac{2N}{2^n}, \qquad (17.14)$$

При цьому, якщо розраховане значення p виявляється меншим 0,05 (чи 0,01), то нульова гіпотеза відхиляється на рівні значущості $p < 0,05$ (чи $p < 0,01$).

Приклад. Оцінити вірогідність розходжень між вмістом білка в молоці корів та їхніх дочок. Вихідні дані наведені в таблиці 17.7

Таблиця 17.7 – Вихідні дані для оцінки вірогідності розходжень між вмістом білка в молоці корів та їхніх дочок

Вміст білка в молоці (%)		Різниці
дочок	матерів	
3,1	3,0	+0,1
3,3	3,1	+0,2
3,2	3,2	0
3,0	3,2	-0,2
3,5	3,3	+0,2
Сума		$\sum R$ = +0,3

Сума різниць між значеннями першої і другої вибірок становить $\sum R = +0,3$. Однак, відповідно до нуль-гіпотези, це значення повинне становити нуль. Наскільки вірогідне перевищення дочок над матерями за вмістом білка в молоці?

Перелічимо всі можливі суми, які можна одержати на підставі величин різниці між першою та другою вибірками, при цьому відберемо з них тільки ті, котрі дорівнюють чи перевищують $\sum R = +0,3$. Таких варіантів чотири:

$$0,1 + 0,2 + 0 + 0,2 + 0,2 = 0,7;$$
$$0,1 + 0,2 + 0 + 0,2 - 0,2 = 0,3;$$
$$0,1 - 0,2 + 0 + 0,2 + 0,2 = 0,3;$$
$$-0,1 + 0,2 + 0 + 0,2 + 0,2 = 0,5.$$

Усього ж, використовуючи п’ять отриманих пар різниць, можна розрахувати 32 різних варіанта ($2^5 = 32$), змінюючи знаки перед кожним доданком. Тоді критерій рандомізації буде дорівнювати:

$$p = \frac{2 \cdot 4}{2^5} = \frac{8}{32} = 0,25.$$

Оскільки це значення набагато перевищує 0,05, нуль-гіпотеза не може бути відхилена і, відповідно, статистично не доведено вірогідність розходжень між вмістом білка в молоці дочок і матерів.

Подібний хід розрахунку має критерій рандомізації й у випадку незалежних вибірок. Для його розрахунку необхідно виконати наступне:

1) знайти суми значень, що належать першій і другій вибіркам і далі в аналізі враховується тільки менша сума;

2) використовуючи дані, що належать першій і другій вибіркам, знайти всі можливі комбінації, суми яких дорівнюють чи менші отриманої в пункті 1 величини; при цьому число доданків відповідає чисельності вибірки із меншою сумою;

3) підрахувати число таких комбінацій (N);

4) розрахувати можливе число всіх комбінацій за формулою:

$$C = \frac{(n_x + n_y)!}{n_x! n_y!};$$

5) отримати оцінку критерію рандомізації за формулою:

$$p = \frac{2N}{C}; \qquad (17.15)$$

6) якщо отримана оцінка критерію рандомізації перевищує 0,05, нульова гіпотеза залишається в силі; якщо ж значення критерію виявляється меншим, ніж 0,05 (чи 0,01) вважається вірогідним розходженнями між двома вибірками із рівнем значущості $p < 0,05$ (чи $p < 0,01$).

За певних умов число можливих комбінацій може бути досить великим. Для того, щоб не витрачати час на перебір усіх можливих комбінацій можна знайти їхнє критичне значення – $N_{крит}$:

$$N_{крит} = [0,025 \cdot C] + 1, \qquad (17.16)$$

де вираз $[x]$ означає цілу частину величини x.

Перебір усіх можливих комбінацій можна припинити, якщо вони перевищать $N_{крит}$. У цьому випадку нуль-гіпотеза не відхиляється і вірогідність розходжень вважається не доведеною.

Приклад. Використовуючи критерій рандомізації ще раз проаналізуємо наступні дані (табл. 17.8).

Таблиця 17.8 – Вихідні дані

Експериментальна група (x)	Контрольна група (y)
3,8	3,9
4,0	4,0
4,1	4,3
4,2	4,4
4,5	4,4
$\Sigma x = 20{,}6$	$\Sigma y = 21{,}0$

Менша сума виявляється для експериментальної групи і цю оцінку ми приймаємо як базову. Тепер необхідно розрахувати за формулою 17.16 критичне значення числа комбінацій:

$$N_{крит} = [0{,}025 \times C] + 1 = [0{,}025 \times \left(\frac{(5+5)!}{5! \times 5!}\right)] + 1 = 7.$$

Отже, як тільки число комбінацій значень першої і другої вибірок, що дають у сумі величину менше чи рівну 20,6, досягне 7, можна вважати, що нуль-гіпотеза залишається в силі.

Отже, почнемо перерахування комбінацій. Для цього в групі з меншою сумою знаходимо найбільше значення (у нашому випадку це 4,5) і підставляємо замість нього всі не переважаючі його значення із другої групи:

3,8 + 4,0 + 4,1 + 4,2 + 4,5 = 20,6;
3,8 + 4,0 + 4,1 + 4,2 + 3,9 = 20,0;
3,8 + 4,0 + 4,1 + 4,2 + 4,0 = 20,1;
3,8 + 4,0 + 4,1 + 4,2 + 4,3 = 20,4;
3,8 + 4,0 + 4,1 + 4,2 + 4,4 = 20,5;
3,8 + 4,0 + 4,1 + 4,2 + 4,4 = 20,5.

Далі, знаходимо наступне за величиною значення (у нашому випадку 4,2) і заміняємо не переважаючими його значеннями із другої вибірки:

3,8 + 4,0 + 4,1 + 3,9 + 4,5 = 20,3;
3,8 + 4,0 + 4,1 + 4,0 + 4,5 = 20,4.

Потім повторюємо цю процедуру, заміняючи значення 4,5 значеннями із другої вибірки не переважаючими його:

3,8 + 4,0 + 4,1 + 3,9 + 4,0 = 19,8;
3,8 + 4,0 + 4,1 + 3,9 + 4,3 = 20,1;
3,8 + 4,0 + 4,1 + 3,9 + 4,4 = 20,2;
3,8 + 4,0 + 4,1 + 3,9 + 4,4 = 20,2;
3,8 + 4,0 + 4,1 + 4,0 + 4,3 = 20,2;

3,8 + 4,0 + 4,1 + 4,0 + 4,4 = 20,3;
3,8 + 4,0 + 4,1 + 4,0 + 4,4 = 20,3.

У принципі, можна продовжувати і далі, однак уже на даному етапі ми маємо 15 комбінацій, сума яких не перевищує 20,6, тоді як критичне значення становить 7. Тому не продовжуючи, можна вважати доведеною нуль-гіпотезу про відсутність вірогідних розходжень між двома групами корів за жирномолочністю.

Критерії рандомізації можуть бути використаними для будь-яких обсягів вибірок, однак при їхньому збільшенні обсяг розрахунку настільки збільшується, що використання даного критерію стає вже складним без допомоги ПЕОМ та відповідного програмного забезпечення.

Для визначення вірогідності відмінностей між двома вибірками стосовно центральної тенденції може бути використаний **критерій перестановок** (пермутацій). Його застосування продемонструємо на наступному прикладі.

Приклад. Необхідно визначити, чи є вірогідні відмінності між рівнем жирномолочності в двох групах (контрольній та експериментальній) корів червоної породи під час проведення досліду (табл. 17.9).

Таблиця 17.9 – Вихідні дані для визначення вірогідності відмінності за жирномолочністю між двома групами корів

Контрольна група, %	Експериментальна група, %
3,0	4,4
4,0	4,6
4,1	4,3
4,2	4,4
4,5	4,5
$\overline{X} = 3{,}96$	$\overline{X} = 4{,}46$

Як бачимо, є суттєві відмінності між середніми арифметичними в двох групах, тому нульова гіпотеза повинна бути перевірена. Якщо ми використаємо класичні методи аналізу (в даному випадку, критерій Ст'юдента для незалежних вибірок; див. формулу 17.4), то отримаємо оцінку критерію t = 1,93 при числі ступенів свободи df = 8. Оскільки табличне значення критерію Ст'юдента для такого числа ступенів свободи перевищує отримане ($t_{df=8}$ = 2,31), ми не можемо відхилити нуль-гіпотезу. Але враховуючи, що обидві сукупності мають небагато значень і нам не відомий тип розподілу величин обох вибірок, більш адекватним вибором для перевірки нуль-гіпотези є методи чисельного ресамплінгу, які не базуються на жодних умовах придатності їх використання.

Для даного випадку зробимо наступне. Розрахуємо абсолютну різницю між середніми значення двох вибірок. В нашому випадку вона становить: |3,96 – 4,46| = 0,50. Далі сформуємо із двох наших вибірок один ряд значень,

перші п’ять якого належать контрольній групі, а інші п’ять – експериментальній. Він має наступний вигляд:

3,0	4,0	4,1	4,2	4,5	**4,5**	**4,6**	**4,3**	**4,4**	**4,5**
контрольна група					експериментальна				

Випадковим чином змішаємо вихідні дані, наприклад, наступним чином:

4,2	**4,5**	4,0	**4,6**	**4,5**	4,1	**4,3**	4,5	3,0	**4,4**
контрольна група					експериментальна				

Для цієї штучно створеної псевдовибірки розрахуємо середні арифметичні (4,36 та 4,06, відповідно) та знову оцінимо абсолютну різницю між ними (0,30).

Знову випадковим чином змішаємо вихідні дані і проведемо всі наступні розрахунки.

Зробимо таку процедуру багато разів (наприклад, M). Бажано, щоб ця кількість була в межах 1000-10000.

Далі підрахуємо, скільки разів для наших псевдовибірок абсолютна різниця між середніми була такою ж за величиною або навіть перевищувала оцінку абсолютних різниць, отриману для вихідних даних (наприклад, це відбулося m разів). Тоді рівень значущості критерію перестановок, що використаний для перевірки даної нуль-гіпотези, можна розрахувати за формулою:

$$p = \frac{m}{M+1}. \qquad (17.17)$$

Підґрунтя цього методу досить просте. Чим значніші відмінності між групами (тобто, вище абсолютна різниця між середніми), тим менше шансів отримати таку ж оцінку при випадковому розташуванні значень вихідної вибірки.

Після 500 перестановок для даних із нашого прикладу, лише в 17 випадках абсолютна різниця була такою ж або навіть перевищувала значення 0,50. Таким чином, рівень значущості для даного критерію становить:

$$p = 17 : (500+1) = 0{,}034.$$

Це досить мала величина і, відповідно, нульова гіпотеза щодо рівності середніх у двох групах тварин, що порівнювалися, повинна бути відхилена.

Відмітимо, що класичні методи її перевірки виявилися не в змозі адекватно проаналізувати дану ситуацію.

Контрольні питання:

1. Який критерій використовується для перевірки гіпотези про рівність декількох вибіркових варіанс коли обсяги порівнюваних вибірок не однакові?

2. В яких випадках використовуються критерій Левене (W_0)?

§ 18. Дисперсійний аналіз кількісних ознак

18.1 Поняття про дисперсійний аналіз

Алгоритм *дисперсійного аналізу* був розроблений англійським вченим Р. Фішером у 1925 р. та набув суттєвого розвитку в працях його учня – Йєтса.

Основна мета дисперсійного аналізу – розкладання загальної мінливості ознаки на її складові, що виникають між членами популяції під впливом багатьох різноманітних факторів. У ході аналізу встановлюють частку мінливості, що зумовлена кожним фактором в експерименті, частку мінливості, що зумовлена сумісною дією цих факторів, а також частку мінливості, що є результатом впливу багатьох неорганізованих (випадкових) факторів.

Дисперсійний аналіз широко використовується в зоотехнії та селекції с.-г. тварин:

- при визначенні оцінки коефіцієнтів успадкування (h^2) та повторюваності (w);
- при оцінці плідників за якістю нащадків;
- при оцінці частки генотипової та паратипової мінливості в загальній мінливості тварин, тощо.

Дисперсійний аналіз базується на понятті *дисперсії*, тобто суми квадратів відхилень кожної варіанти вибірки від середнього арифметичного значення:

$$C = \sum (x_i - \bar{x})^2 \ . \qquad (18.1)$$

Р. Фішер показав, що якщо аналізується мінливість у декількох групах, загальна мінливість такого комплексу (Cy) може бути розкладена на дві складові – мінливість ознаки між групами (міжгрупова дисперсія – Cx) та мінливість ознаки всередині груп (внутрішньогрупова дисперсія – Cz). Залежність між цими джерелами варіювання можна виразити рівнянням:

$$\sum (x_{ij} - \bar{x})^2 = n \cdot \sum (\bar{x}_i - \bar{x})^2 + \sum \sum (x_{ij} - \bar{x}_i)^2 , \qquad (18.2)$$

тобто, $Cy = Cx + Cz$.

Алгоритм однофакторного дисперсійного аналізу краще продемонструвати на прикладі.

Приклад. Необхідно визначити, чи впливає вік свиноматок на їх великоплідність. Всі вихідні та проміжні дані наведено в таблиці 18.1.

Дисперсійний комплекс має три градації фактора A, тобто $l = 3$. Всього в аналізі було використано дані щодо 15 свиноматок ($N = 15$).

Для зручності розрахунку окремих компонент загальної дисперсії попередньо необхідно розрахувати три допоміжні величини:

- загальну суму значень ознаки по всіх градаціях фактора: $\Sigma\Sigma\, x_i = 17{,}0$;
- суму відношень квадратів сум значень ознаки по кожній градації до відповідного обсягу груп: $\sum \frac{\left(\sum x_i\right)^2}{n_i} = 19{,}416$;
- загальну суму квадратів значень ознаки по всім градаціям: $\Sigma\Sigma\, x_i^2 = 19{,}56$.

Таблиця 18.1 – Вихідні дані для визначення впливу віку свиноматок на їх великоплідність

	Градації фактору *A* (вік у опоросах)			Суми
	перший	другий	третій	
x_{ij}	0,9	1,1	1,0	
	1,1	1,1	1,3	
	1,0	1,3	1,4	
	1,0	1,2	1,2	
	1,0	1,1	1,3	
n_i	5	5	5	$\Sigma n_i = N = 15$
Σx_i	5,0	5,8	6,2	$\Sigma\Sigma x_i = 17,0$
$\frac{\left(\sum x_i\right)^2}{n_i}$	5,000	6,728	7,688	$\sum \frac{\left(\sum x_i\right)^2}{n_i} = 19,416$
Σx_i^2	5,02	6,76	7,78	$\Sigma\Sigma\, x_i^2 = 19,56$

Для розрахунку міжгрупової та внутрішньогрупової дисперсій необхідно також розрахувати допоміжну величину:

$$H = \frac{\left(\sum\sum x_i\right)^2}{N}. \tag{18.3}$$

Для даних із прикладу значення цієї допоміжної величини становить:

$$H = \frac{17,0^2}{15} = 19,267.$$

Тоді значення загальної, міжгрупової та внутрішньогрупової дисперсії можна розрахувати за формулами:

$$Cy = \sum\sum x_i^{\,2} - H; \tag{18.4}$$

$$Cx = \sum \frac{\left(\sum x_i\right)^2}{n_i} - H; \tag{18.5}$$

$$Cz = Cy - Cx. \tag{18.6}$$

Для даних із прикладу ці значення становлять, відповідно:

Cy = 19,560 – 19,267 = 0,293;

Cx = 19,416 – 19,267 = 0,149;

Cz = 0,293 – 0,149 = 0,144.

Далі необхідно визначити число ступенів свободи для кожної з дисперсій:

$$k_y = N - 1; \tag{18.7}$$

$$k_x = l - 1; \tag{18.8}$$

$$k_z = N - l. \tag{18.9}$$

Таким чином, для даних із нашого прикладу число ступенів свободи для загальної дисперсії становить: $k_y = 15 - 1 = 14$; для міжгрупової дисперсії: $k_x = 3 - 1 = 2$; для внутрішньогрупової: $k_z = 15 - 3 = 12$.

Розраховуємо значення варіанс, тобто середніх квадратів відхилень.

Для цього необхідно оцінки дисперсій віднести до відповідних значень числа ступенів свободи:

$$\sigma_y^2 = \frac{Cy}{k_y}; \tag{18.10}$$

$$\sigma_x^2 = \frac{Cx}{k_x}; \tag{18.11}$$

$$\sigma_z^2 = \frac{Cz}{k_z}. \tag{18.12}$$

Таким чином, загальна варіанса становитиме: $\sigma_y^2 = \frac{0{,}293}{14} = 0{,}0209$; міжгрупова варіанса: $\sigma_x^2 = \frac{0{,}149}{2} = 0{,}0745$; внутрішньогрупова варіанса: $\sigma_z^2 = \frac{0{,}144}{12} = 0{,}0120$.

Частка впливу фактора *A* (в нашому випадку – віку свиноматки) на загальну мінливість ознаки (тобто, великоплідність) та її статистична помилка визначаються за формулами:

$$\eta^2 = \frac{Cx}{Cy}; \tag{18.13}$$

$$SE_{\eta^2} = (1 - \eta^2) \cdot \left(\frac{l - 1}{N - l} \right). \tag{18.14}$$

Таким чином, для даних із прикладу частка мінливості великоплідності, що залежить від віку свиноматки становить:

$$\eta^2 = 0{,}149 : 0{,}293 = 0{,}509$$

зі статистичною помилкою:

$$SE_{\eta^2} = (1 - 0{,}509) \cdot \left(\frac{3 - 1}{15 - 3} \right) = 0{,}082.$$

Тобто, на 50,9% великоплідність залежить від віку свиноматки (номеру опоросу), а на неорганізовані (випадкові) фактори припадає 49,1% загальної мінливості.

Рівень значущості одержаної величини можна оцінити на підставі рівня значущості критерію Р. Фішера:

$$F = \frac{\sigma_x^2}{\sigma_z^2}. \tag{18.15}$$

Для цього розраховане за формулою 18.15 значення дисперсійного відношення необхідно порівняти із табличним значенням критерію Р. Фішера із відповідними числами ступенів свободи, які наведені в Додатку Ж.

У верхній частині таблиці наведено число ступенів свободи для чисельника із формули 18.15, а у лівій – для знаменника.

У нашому випадку дисперсійне відношення дорівнює: $F = \frac{0{,}0745}{0{,}0120} = 6{,}208$.

Це отримане значення необхідно порівняти із табличним при числі ступенів свободи: для чисельника – $df_1 = l - 1 = 2$, для знаменника – $df_2 = N - l = 12$.

Але в Додатку Ж не наведено значення критерію Р. Фішера для таких чисел ступенів свободи ($F_{2,\ 12}$). Для числа ступенів свободи $df_1 = 2$ (верхня строчка таблиці) присутні тільки значення для числа ступенів свободи знаменника $df_2 = 10$ ($F_{2,\ 10} = 4{,}103$) та $df_2 = 15$ ($F_{2,\ 15} = 3{,}682$). Тому необхідно використати лінійну гармонійну інтерполяцію по вертикалі.

У цьому випадку шукане табличне значення критерію Р. Фішера можна розрахувати за формулою:

$$F = F_0 + u \cdot (F_1 - F_0), \qquad (18.16)$$

де $u = \frac{df_1 \cdot (df_0 - df)}{df \cdot (df_0 - df_1)}$;

F_0 – табличне значення критерію Р. Фішера для df_0;

F_1 – табличне значення критерію Р. Фішера для df_1; при цьому $df_0 < df < df_1$.

У нашому випадку $df_0 = 10$, $df = 12$, $df_1 = 15$, тому: $u = \frac{15 \cdot (10 - 12)}{12 \cdot (10 - 15)} = 0{,}5$.

Тоді шукане табличне значення критерію Р. Фішера становитиме:

$$F_{2,\ 12} = 4{,}103 + 0{,}5 \cdot (3{,}682 - 4{,}103) = 3{,}893.$$

Оскільки розраховане значення дисперсійного відношення (F) набагато вище, ніж табличне, вплив фактора A (віку свиноматки) на їх великоплідність вважається вірогідним із рівнем значущості $p < 0{,}05$.

Всі розрахункові дані оформлюються у вигляді таблиці дисперсійного аналізу (табл. 18.2).

Таблиця 18.2 – Результати дисперсійного аналізу

Джерело мінливості	Дисперсія (C)	Число ступенів свободи (df)	Варіанса (σ^2)	Дисперсійне відношення (F)	Сила впливу (η^2)
Фактор A	0,144	2	0,0745	6,208	0,509
Залишкова (Z)	0,149	12	0,0120		0,491
Сумарна (Y)	0,293	14			

Принципової різниці між аналізом багатофакторних комплексів і схемами, що застосовуються при аналізі однофакторних дисперсійних комплексів немає. Багатофакторний аналіз не змінює, а лише дещо ускладнює загальну схему, оскільки поряд з дією кожного фактора окремо, доводиться враховувати і їхню спільну дію на результативну ознаку.

Так, якщо є два регульованих фактора *A* та *B*, то їхній вплив, а також вплив інших факторів на результативну ознаку можна зобразити у виді наступної схеми (рис. 18.1).

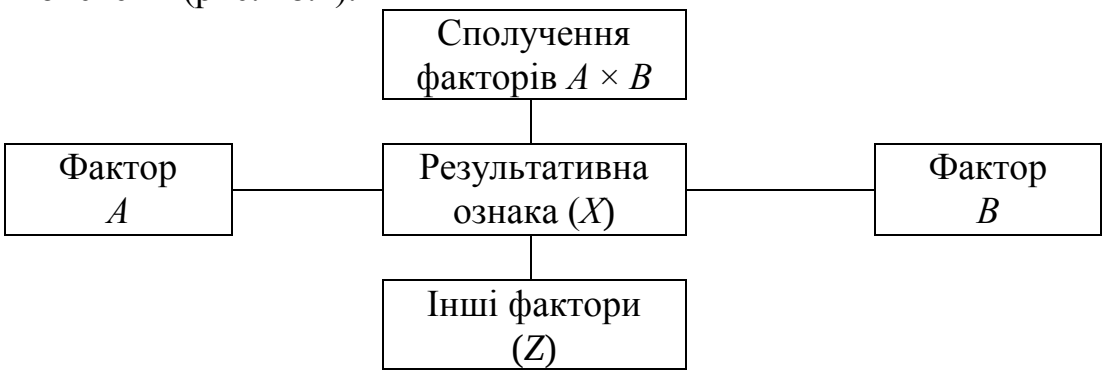

Рисунок 18.1 – Схема впливу факторів на результативну ознаку

У даному випадку загальна дисперсія ($C_У$) містить чотири компоненти варіювання:

$$C_У = C_A + C_B + C_{AB} + C_X.$$

Якщо ж враховуються не два, а три регульованих фактори А, В і С, то, поряд з їхньою індивідуальною дією, можлива ще і дія на ознаку трьох парних сполучень – АВ, АС і ВС, а також їхня спільна дія АВС плюс вплив неорганізованих (випадкових) факторів.

Основні умови організації дво- і багатофакторних дисперсійних комплексів наступні:

а) випадковий вибір вихідних даних;

б) наявність даних по всіх сполученнях усіх градацій усіх факторів, що аналізуються (крім ієрархічного аналізу; див. нижче);

в) повна незалежність факторів, що враховуються;

г) нормальний розподіл вихідних даних.

В якості факторів можуть бути використані:

а) кількісні ознаки, розподілені на градації (температура, вологість, рівень годівлі, хімічний і біологічний вплив і т. п.);

б) якісні (альтернативні) ознаки (генотип, стать, вік, номер лактації або окоту, масть і т. п.).

Важливою вимогою при проведенні дво- і багатофакторного дисперсійного аналізу (ДА) є *рівномірність* дисперсійного комплексу, тобто, однакова кількість варіант по всіх можливих сполученнях факторів.

Аналіз нерівномірних дисперсійних комплексів має свої особливості. Більш докладно з ними можна ознайомитися в книзі Н. А. Плохинского (1969) та ін.

18.2 Алгоритм повного двофакторного дисперсійного аналізу

У популяційній генетиці двофакторний дисперсійний аналіз використовується для більш детального аналізу фенотипової варіації – розкладанні її на гено- і паратипову компоненти. Таким чином, у якості факторів найчастіше використовується генотип особин (фактор *A*) і

особливості утримання, вплив середовища і т. п. (фактор *B*). Розберемо алгоритм повного двофакторного дисперсійного аналізу на наступному прикладі.

Приклад. Вивчався вплив генотипу (порода) і раціону ((комбікорм без преміксу (контроль) і з використанням преміксу) на багатоплідність свиней.

Усі допоміжні величини розраховано і наведено в таблиці 18.3.

Таблиця 18.3 – Допоміжні величини для проведення двофакторного дисперсійного аналізу

	A1		*A2*		Суми
	B1	*B2*	*B1*	*B2*	
	7	8	8	12	
	6	9	10	10	
	8	8	11	9	
	7	10	10	10	
	9	9	9	11	
n	5	5	5	5	$\sum n = N = 20$
$\sum x_i$	37	44	48	52	$\sum\sum x_i = 181$
$\frac{\left(\sum x_i\right)^2}{n}$	273,8	387,2	460,8	540,8	$\sum \frac{\left(\sum x_i\right)^2}{n} = 1662{,}6$
$\sum x_i^2$	279	390	466	546	$\sum\sum x_i^2 = 1681{,}0$
n_A	5 + 5 = 10		5 + 5 = 10		
$\sum x_A$	37 + 44 = 81		48 + 52 = 100		
$\frac{\left(\sum x_A\right)^2}{n_A}$	656,1		1000,0		$\sum \frac{\left(\sum x_A\right)^2}{n_A} = 1656{,}1$
n_B	5 + 5 = 10		5 + 5 = 10		
$\sum x_B$	37 + 48 = 85		44 + 52 = 96		
$\frac{\left(\sum x_B\right)^2}{n_B}$	722,5		921,6		$\sum \frac{\left(\sum x_B\right)^2}{n_B} = 1644{,}1$

У першу чергу необхідно розрахувати величину *H*:

$$H = \frac{\left(\sum\sum x_i\right)^2}{N} = \frac{181^2}{20} = 1638{,}05.$$

Далі, використовуючи цю величину, розраховуємо сумарну дисперсію:

$$C_Y = \sum\sum x_i^2 - H = 1681{,}0 - 1638{,}05 = 42{,}95,$$

і факторіальну дисперсію:

$$C_X = \sum \frac{\left(\sum x_i\right)^2}{n} - H = 1662{,}6 - 1638{,}05 = 24{,}55.$$

Оцінку випадкової (залишкової) дисперсії знаходимо як різницю між сумарною і факторіальною дисперсіями, тобто

$$C_Z = C_Y - C_X = 42{,}95 - 24{,}55 = 18{,}40.$$

Однак факторіальна дисперсія в такому виді являє собою суму дисперсій одночасно для факторів A та B, а також для їхнього сполучення. Тому необхідно знову розкласти факторіальну дисперсію на окремі компоненти:

$$C_A = \sum \frac{\left(\sum x_A\right)^2}{n_A} - H = 1656{,}1 - 1638{,}05 = 18{,}05,$$

$$C_B = \sum \frac{\left(\sum x_B\right)^2}{n_B} - H = 1644{,}1 - 1638{,}05 = 6{,}05.$$

Нарешті, дисперсія спільної дії факторів розраховується за формулою:

$$C_{AB} = C_X - (C_A + C_B) = 24{,}55 - (18{,}05 + 6{,}05) = 0{,}45.$$

Переходимо до розрахунку числа ступенів свободи для кожного джерела варіювання. Для:

- сумарної дисперсії: $k = N - 1 = 20 - 1 = 19$;
- дисперсії по фактору A: $k = a - 1 = 2 - 1 = 1$;
- дисперсії по фактору B: $k = b - 1 = 2 - 1 = 1$;
- дисперсії спільного впливу AB: $k_{AB} = (a - 1)(b - 1) = 1$;
- залишкової дисперсії: $k = N - a \cdot b = 20 - 4 = 16$,

де a – число градацій по фактору A;

b – число градацій по фактору B;

N – загальне число особин, які використані в дисперсійному комплексі.

Розрахунок варіанс (або девіат) проводиться віднесенням значень дисперсій до відповідного числа ступенів свободи:

$$\sigma_A^2 = \frac{C_A}{k_A} = \frac{18{,}05}{1} = 18{,}05;$$

$$\sigma_B^2 = \frac{C_B}{k_B} = \frac{6{,}05}{1} = 6{,}05;$$

$$\sigma_{AB}^2 = \frac{C_{AB}}{k_{AB}} = \frac{0{,}45}{1} = 0{,}45;$$

$$\sigma_Z^2 = \frac{C_Z}{k_Z} = \frac{18{,}4}{16} = 1{,}15.$$

Для того, щоб оцінити вірогідність впливу факторів A, B та їхнього сполучення, необхідно розрахувати дисперсійні відношення:

$$F_A = \frac{\sigma_A^2}{\sigma_Z^2} = \frac{18{,}05}{1{,}15} = 15{,}70;$$

$$F_B = \frac{\sigma_B^2}{\sigma_Z^2} = \frac{6{,}05}{1{,}15} = 5{,}26;$$

$$F_{AB} = \frac{\sigma_Z^2}{\sigma_{AB}^2} = \frac{1{,}15}{0{,}45} = 2{,}56.$$

Необхідно вказати, що при розрахунку дисперсійних відношень ділиться більша варіанса на меншу – найчастіше, факторіальна на випадкову. Однак, у нашому прикладі, при розрахунку дисперсійного відношення для сполучення факторів *AB*, ми повинні ділити випадкову варіансу на факторіальну.

Далі, кожне дисперсійне відношення (*F*) порівнюємо з табличним значенням F-критерію Фішера-Снедекора, причому число ступенів свободи для більшої варіанси знаходиться у верхньому рядку таблиці, а для меншої – у правому стовпчику (див. Додаток Ж).

Усі результати двохфакторного дисперсійного аналізу наведено в таблиці 18.4.

Таблиця 18.4 – Результати двофакторного дисперсійного аналізу

Джерело мінливості	Дисперсія (*C*)	Число ступенів свободи (*df*)	Варіанса (σ^2)	Дисперсійне відношення (*F*)	Сила впливу (η^2)
Фактор *A*	18,05	1	18,05	*15,70*	0,42
Фактор *B*	6,05	1	6,05	*5,26*	0,14
Сполучення *A* × *B*	0,45	1	0,45	2,56	0,01
Залишкова (*Z*)	18,40	16	1,15		0,43
Сумарна (*Y*)	42,95	19			

18.3 Алгоритм двофакторного дисперсійного аналізу без повторюваностей

Ця модифікація дисперсійного аналізу застосовується в тих випадках, коли по кожному сполученню двох розглянутих факторів можна одержати тільки одне значення. Наприклад, для різних бугаїв-плідників можна одержати тільки одну оцінку сумарного обсягу еякуляту за місяць або середній об’єм одного еякуляту; для різних корів – середню жирність молока по місяцях лактації і т. п.

Таким чином, у цьому випадку розглядається вплив як особливостей (генетичних) окремої тварини (фактор *A*), так і особливості варіювання показника в часі (фактор *B*). Принципова відмінність від розглянутого вище алгоритму дисперсійного аналізу полягає в тому, що при відсутності повторюваностей не можна розрахувати частку спільного впливу факторів *A* та *B*.

Алгоритм розрахунку дисперсій, варіанс і дисперсійних відношень продемонструємо на наступному прикладі.

Приклад. Оцінити вплив генотипу корови і порядку місяця лактації на жирномолочність корів червоної степової породи (розглядалися тільки перші повні лактації). Усі вихідні дані наведено в таблиці 18.5.

Для даного дисперсійного комплексу маємо п’ять градацій фактора *A* (тобто, $a = 5$) і десять градацій фактора *B* (тобто, $b = 10$).

Таблиця 18.5 – Вихідні дані для оцінки впливу генотипу корови і місяця лактації на жирномолочність

Номер корови (А)	Номер місяця лактації (*B*)										Σx_a	$(\Sigma x_a)^2$	Σx_a^2
	1	2	3	4	5	6	7	8	9	10			
1	3,3	3,5	3,5	3,5	3,7	4,1	3,6	3,6	3,8	3,7	36,3	1317,69	132,19
2	3,4	3,4	3,5	3,5	3,5	3,4	3,5	3,4	3,6	3,5	34,7	1204,09	120,45
3	3,6	3,3	4,1	3,7	3,4	3,8	3,7	3,5	3,6	3,7	36,4	1324,96	132,94
4	3,4	3,4	3,4	3,6	3,5	3,5	3,8	3,7	3,7	3,5	35,5	1260,25	126,21
5	3,7	3,6	3,7	3,8	3,7	3,6	3,8	3,9	4,0	4,2	38,0	1444,00	144,72
Σx_b	17,4	17,2	18,2	18,1	17,8	18,4	18,4	18,1	18,7	18,6	180,9	6550,99	656,51
$(\Sigma x_b)^2$	302,76	295,84	331,24	327,61	316,84	338,56	338,56	327,61	349,69	345,96	3274,67		

Для того, щоб розрахувати значення дисперсій, спочатку необхідно знайти суми дат по кожному рядку (Σx_a) і квадрати цих величин ($(\Sigma x_a)^2$), а також суми дат по кожному стовпцю (Σx_b) і квадрати цих величин ($(\Sigma x_b)^2$). Далі, необхідно знайти суми цих сум і, крім цього, суму квадратів усіх значень дисперсійного комплексу (Σx^2).

Для даного прикладу ці величини становитимуть: $\Sigma\Sigma x_a = \Sigma\Sigma x_b = 180{,}9$; $\Sigma(\Sigma x_a)^2 = 6550{,}99$; $\Sigma(\Sigma x_b)^2 = 3274{,}67$; $\Sigma\Sigma x^2 = 656{,}51$.

Ці величини використовуються для розрахунку дисперсій (як факторіальних, так і сумарної).

Спочатку необхідно розрахувати величину *H*:

$$H = \frac{\left(\sum\sum x_i\right)^2}{N} = \frac{180{,}9^2}{50} = 654{,}50.$$

Далі, використовуючи цю величину, розраховуємо загальну дисперсію:

$$C_Y = \sum\sum x_i^2 - H = 656{,}51 - 654{,}50 = 2{,}01$$

і факторіальні дисперсії:

$$C_A = \frac{\sum\left(\sum x_a\right)^2}{b} - H = \frac{6550{,}99}{10} - 654{,}50 = 0{,}60;$$

$$C_B = \frac{\sum\left(\sum x_b\right)^2}{a} - H = \frac{3274{,}67}{5} - 654{,}50 = 0{,}43.$$

Оцінку випадкової (залишкової) дисперсії знаходимо як різницю між сумарною і факторіальною дисперсіями, тобто

$$C_Z = C_Y - (C_A + C_B) = 2{,}01 - (0{,}60 + 0{,}43) = 0{,}98.$$

Переходимо до розрахунку числа ступенів свободи для кожного джерела варіювання. Для:

- сумарної дисперсії: $k = N - 1 = 50 - 1 = 49$;
- дисперсії по фактору A: $k = a - 1 = 5 - 1 = 4$;
- дисперсії по фактору B: $k = b - 1 = 10 - 1 = 9$;
- залишкової дисперсії: $k = (a - 1)(b - 1) = 36$;

де a – число градацій по фактору A;

b – число градацій по фактору B;

N – загальне число особин, які використані у дисперсійному комплексі.

Розрахунок варіанс проводиться віднесенням значень дисперсій до відповідного числа ступенів свободи:

$$\sigma_A^2 = \frac{C_A}{k_A} = \frac{0{,}60}{4} = 0{,}150;$$

$$\sigma_B^2 = \frac{C_B}{k_B} = \frac{0{,}43}{9} = 0{,}047;$$

$$\sigma_Z^2 = \frac{C_Z}{k_Z} = \frac{0{,}98}{36} = 0{,}027.$$

Для того, щоб оцінити вірогідність впливу фактора A та B, необхідно розрахувати дисперсійні відношення:

$$F_A = \frac{\sigma_A^2}{\sigma_Z^2} = \frac{0{,}150}{0{,}027} = 5{,}56;$$

$$F_B = \frac{\sigma_B^2}{\sigma_Z^2} = \frac{0{,}047}{0{,}027} = 1{,}74.$$

Усі результати дисперсійного аналізу оформлюються у вигляді таблиці 18.6.

Таблиця 18.6 – Результати дисперсійного аналізу

Джерело мінливості	Дисперсія (C)	Число ступенів свободи (df)	Варіанса (σ^2)	Дисперсійне відношення (F)	Сила впливу (η^2)
Фактор A	0,60	4	0,150	*5,56*	0,29
Фактор B	0,43	9	0,047	1,74	0,21
Залишкова (Z)	0,98	36	0,027		0,50
Сумарна (Y)	2,01	49			

Табличне значення критерію Фішера-Снедекора для числа ступенів свободи 4 і 36 (тобто, по фактору A) становить $F = 2{,}6$, а для числа ступенів свободи 9 і 36 (тобто, по фактору B) – $F = 2{,}2$ (Додаток Ж).

Таким чином, оскільки дисперсійне відношення для фактора A більше, ніж табличне, доведено вплив генотипу корови на жирномолочність. З іншого боку, вплив місяця лактації не доведено, оскільки дисперсійне відношення по цьому фактору виявилося меншим, ніж табличне значення критерію Фішера-Снедекора.

18.4 Алгоритм ієрархічного дисперсійного аналізу

У практиці тваринництва зустрічаються такі дисперсійні комплекси, в яких вільне комбінування факторів один з одним виключене. Ці комплекси, названі ***ієрархічними***, організуються, наприклад, при вивченні впливу батьків на продуктивність їхніх нащадків.

У цих випадках у якості градації фактора *A* використовуються різні батьки, а градації фактора *B* – різні матері тварин. Причому для різних батьків використовуються різні матері, тобто немає повного і вільного комбінування градацій обох факторів.

Для таких комплексів є свої особливості розрахунку дисперсій, варіанс і дисперсійних відношень, що і продемонструємо на прикладі.

Приклад. При вивченні впливу породних властивостей кнурів Барона і Сокола на багатоплідність їхніх дочок, отриманих від шести різних свиноматок, отримано наступні результати (табл. 18.7).

Таблиця 18.7 – Вихідні дані для вивчення впливу породних властивостей кнурів Барона і Сокола на багатоплідність їхніх дочок

	Батьки (*A*)						Суми
	Барон (*A1*)			Сокіл (*A2*)			
	Матері (*B*)						
	B1	*B2*	*B3*	*B4*	*B5*	*B6*	
	7	8	9	10	9	12	
	6	9	7	8	8	10	
	8	8	9	10	7	10	
	7	10	9	11	8	9	
n	4	4	4	4	4	4	$\Sigma n = N = 24$
Σx_i	28	35	34	39	32	41	$\Sigma\Sigma x_i = 209$
$(\Sigma x_i)^2$	784	1225	1156	1521	1024	1681	$\Sigma(\Sigma x_i)^2 = 7391$
Σx_i^2	198	309	292	385	258	425	$\Sigma\Sigma x_i^2 = 1867$
n_A	12			12			
Σx_A	28 + 35 + 34 = 97			39 + 32 + 41 = 112			
$(\Sigma x_A)^2$	9409			12544			$\Sigma(\Sigma x)^2 = 21953$

Насамперед, необхідно розрахувати допоміжні величини, які необхідні для переходу до розрахунку факторіальних, випадкової і сумарної дисперсій.

Першочергово необхідно розрахувати величину *H*:

$$H = \frac{\left(\sum\sum x_i\right)^2}{N} = \frac{209^2}{24} = 1820{,}04.$$

Далі, використовуючи цю величину, розраховуємо сумарну дисперсію:

$$C_Y = \sum\sum x_i^2 - H = 1867 - 1820{,}04 = 49{,}96,$$

і факторіальну дисперсію:

$$C_X = \frac{\sum\left(\sum x_i\right)^2}{n} - H = \frac{7391}{4} - 1820{,}04 = 27{,}71.$$

Оцінку випадкової (залишкової) дисперсії знаходимо як різницю між сумарною і факторіальною дисперсіями, тобто

$$C_Z = C_Y - C_X = 49{,}96 - 27{,}71 = 19{,}25.$$

Однак факторіальна дисперсія в такому вигляді являє собою суму дисперсій одночасно для факторів *A* та *B*, тому необхідно знову розкласти факторіальну дисперсію на окремі компоненти:

$$C_A = \frac{\sum\left(\sum x_A\right)^2}{n_A} - H = \frac{21953}{12} - 1820{,}04 = 9{,}38,$$

$$C_B = \frac{\sum\left(\sum x_i\right)^2}{n} - \frac{\sum\left(\sum x_A\right)^2}{n_A} = \frac{7391}{4} - \frac{21953}{12} = 18{,}33.$$

Особливим чином при ієрархічному дисперсійному аналізі проводиться і розрахунок числа ступенів свободи. Для:

- сумарної дисперсії: $k = N - 1 = 24 - 1 = 23$;
- дисперсії по фактору *A*: $k = a - 1 = 2 - 1 = 1$;
- дисперсії по фактору *B*: $k = b - a = 6 - 2 = 4$;
- залишкової дисперсії: $k = N - b = 24 - 6 = 18$,

де *a* – число градацій по фактору *A*;

b – загальне число градацій по фактору *B*;

N – загальне число особин, які використані у дисперсійному комплексі.

Розрахунок варіанс здійснюється віднесенням значень дисперсій до відповідного числа ступенів свободи:

$$\sigma_A^2 = \frac{C_A}{k_A} = \frac{9{,}38}{1} = 9{,}38;$$

$$\sigma_B^2 = \frac{C_B}{k_B} = \frac{18{,}33}{4} = 4{,}58;$$

$$\sigma_Z^2 = \frac{C_Z}{k_Z} = \frac{19{,}25}{18} = 1{,}07.$$

Для того, щоб оцінити вірогідність впливу фактора *A* та *B*, необхідно розрахувати дисперсійні відношення, що в даному випадку мають особливу форму розрахунку:

$$F_A = \frac{\sigma_A^2}{\sigma_B^2} = \frac{9{,}38}{4{,}58} = 2{,}0;$$

$$F_B = \frac{\sigma_B^2}{\sigma_Z^2} = \frac{4{,}58}{1{,}07} = 4{,}4.$$

У даному випадку значущим виявляється тільки дія фактора *B*, оскільки табличне значення критерію Фішера-Снедекора для числа ступенів свободи 4 і 18 складає $F = 2{,}9$. Отже, доведено лише вплив свиноматок на багатоплідність

їхніх дочок, при тому, що кнури, використані в аналізі, не впливають на рівень багатоплідності дочок.

Результати аналізу заносимо в таблицю 18.8.

Таблиця 18.8 – Результати дисперсійного аналізу

Джерело мінливості	Дисперсія (*C*)	Число ступенів свободи (*df*)	Варіанса (σ^2)	Дисперсійне відношення (*F*)	Сила впливу (η^2)
Фактор *A*	9,38	1	9,38	2,0	-
Фактор *B*	18,33	4	4,58	*4,4*	0,45
Залишкова (*Z*)	19,25	18	1,07		0,55
Сумарна (*Y*)	46,96	23			

Сила впливу фактора розраховується, використовуючи значення виправлених варіанс:

$$\bar{\sigma}_A^2 = \frac{\sigma_A^2 - \sigma_B^2}{nb_0} = \frac{9{,}38 - 4{,}58}{4 \cdot 3} = 0{,}4;$$

$$\bar{\sigma}_B^2 = \frac{\sigma_B^2 - \sigma_Z^2}{n} = \frac{4{,}58 - 1{,}07}{4} = 0{,}88.$$

де b_0 – число градацій фактора *B* у межах кожної градації фактора *A*.

Сумарна дисперсія дорівнює: $\sigma_Y^2 = \bar{\sigma}_A^2 + \bar{\sigma}_B^2 + \sigma_Z^2$. Однак, при її розрахунку включаються тільки ті виправлені факторіальні дисперсії, для яких доведено вплив на ознаку.

У нашому прикладі вона розраховується як $\sigma_Y^2 = \bar{\sigma}_B^2 + \sigma_Z^2 = 0{,}88 + 1{,}07 = 1{,}95,$ оскільки кнури (фактор *A*) не вносять значимого впливу на багатоплідність дочок.

Сила впливу факторів розраховується, як відношення факторіальних виправлених дисперсій до сумарної:

$$h_B^2 = \frac{\bar{\sigma}_B^2}{\sigma_Y^2} = \frac{0{,}88}{1{,}95} = 0{,}45.$$

Дисперсійний аналіз також може бути проведено з використанням методів ресамплінгу. Для цього необхідно використати алгоритм методу перестановок (пермутацій).

Розглянемо особливості використання даного методу на прикладі даних, що було проаналізовано за алгоритмом класичного однофакторного дисперсійного аналізу Р. Фішера (див. табл. 18.1).

У результаті, було встановлено, що має місце вірогідна різниця між середніми арифметичними для груп свиноматок різного віку за великоплідністю. Отримана оцінка дисперсійного відношення становила 6,208 із рівнем значущості $p < 0{,}05$ (табл. 18.2). Але вимоги до використання класичного дисперсійного аналізу для цих даних (нормальність розподілу

значень у межах окремих груп та рівність групових варіанс) важко перевірити, оскільки обсяги вибірок дуже малі (лише по п’ять значень). Більш адекватні висновки дозволить зробити метод перестановок.

Для його проведення зробимо наступне. Випадковим чином змішаємо вихідні значення між трьома групами. Для цієї псевдовибірки, використавши алгоритм класичного дисперсійного аналізу Р. Фішера, розрахуємо дисперсійне відношення (F). Знову змішаємо наші вихідні дані й проведемо всі необхідні розрахунки. Повторимо таку процедуру ще багато разів (наприклад, M).

Після закінчення підрахуємо, скільки разів отримані для псевдовибірок оцінки дисперсійного відношення (F) перевищували або дорівнювали тій оцінці, що була отримана для вихідних даних (для нашого прикладу: F =6,208), наприклад, це становить m разів. Тоді рівень значущості для критерію перестановок, що було використано для дисперсійного аналізу становить:

$$p = \frac{m}{M+1}. \qquad (18.17)$$

Для даних із нашого прикладу, після 500 перестановок лише в 5 випадках отримана оцінка дисперсійного відношення дорівнювала або перевищувала 6,208. Таким чином, нульова гіпотеза щодо рівності всіх середніх може бути відхилена із рівнем значущості:

$$p = 5 : (500+1) = 0{,}010.$$

Аналогічним чином можна визначити рівень значущості для перевірки нуль-гіпотези щодо перевищення нуля оцінкою сили впливу фактора (див. формулу 18.13).

Контрольні питання:

1. Поняття при дисперсійний аналіз, його роль в селекції с.-г. тварин.
2. Основні умови організації дво- і багатофакторних дисперсійних комплексів.
3. У яких випадках застосовується алгоритм двофакторного дисперсійного аналізу без повторюваностей?

§ 19. Кореляційно-регресійний аналіз

Функціональним називається такий зв'язок між двома ознаками, коли кожному значенню одної змінної відповідає строго визначене значення іншої. Але такий тип зв'язку в медико-біологічних дослідженнях майже ніколи не зустрічається.

В зоотехнії найчастіше має місце такий тип зв'язку між ознаками, коли чисельному значенню однієї з них відповідає декілька значень іншої. Такий зв'язок має назву *кореляційного* або *статистичного*.

Кореляційний зв'язок є неповним, він має місце лише при великій кількості спостережень. За допомогою методу кореляційного аналізу можна вирішити два завдання:

- виміряти тісноту зв'язку;
- визначити форму та параметри рівняння зв'язку.

Перше завдання вирішується на підставі оцінки вибіркового значення коефіцієнта кореляції (власне кореляційний аналіз), а друге – за використання методів регресійного аналізу.

Залежно від набору вихідних даних у вибірці, можна визначити наступні типи коефіцієнта кореляції:

- *коефіцієнт парної лінійної кореляції* Пірсона-Браве визначає силу й напрямок лінійного зв'язку між двома ознаками;
- *коефіцієнт множинної кореляції* визначає силу зв'язку між залежною змінною й набором незалежних змінних на підставі рівняння лінійної множинної регресії;
- *коефіцієнт часткової кореляції* визначає силу й напрямок лінійного зв'язку між залежною ознакою й будь-якою з набору незалежних при допущенні, що вплив інших незалежних ознак відсутній.

19.1 Коефіцієнт парної лінійної кореляції Пірсона-Браве

Оцінку вибіркового коефіцієнта парної лінійної кореляції Пірсона-Браве можна отримати за формулою:

$$r = \frac{n \cdot \sum xy - \sum x \cdot \sum y}{\sqrt{[n \cdot \sum y^2 - (\sum y)^2] \cdot [n \cdot \sum x^2 - (\sum x)^2]}}. \qquad (19.1)$$

Для зручності, всі проміжні розрахунки краще проводити у табличному вигляді.

Приклад. Необхідно знайти значення коефіцієнта парної лінійної кореляції між показниками вмісту жиру в молоці корів та їх матерів (n = 15). Всі необхідні попередні розрахунки наведено в таблиці 19.1.

Таблиця 19.1 – Вихідні дані для розрахунку коефіцієнта парної лінійної кореляції між показниками вмісту жиру в молоці корів та їх матерів

№ п/п	Вміст жиру в молоці матерів, % (x)	Вміст жиру в молоці корів, % (y)	x^2	xy	y^2
1	3,77	3,72	14,213	14,024	13,838
2	3,84	4,06	14,746	15,590	16,484
3	4,05	4,12	16,403	16,686	16,974
4	3,84	3,93	14,746	15,091	15,445
5	3,81	3,57	14,516	13,602	12,745
6	3,98	4,19	15,840	16,676	17,556
7	3,70	3,51	13,690	12,987	12,320
8	3,73	3,70	13,913	13,801	13,690
9	3,40	4,04	16,484	16,402	16,322
10	3,80	3,77	14,440	14,326	14,213
11	3,60	3,68	12,960	13,248	13,542
12	3,79	3,52	14,364	13,341	12,390
13	3,72	3,80	13,838	14,136	14,440
14	3,85	3,97	14,823	15,285	15,761
15	3,87	3,8	14,977	14,706	14,440
Суми	56,75	57,38	215,028	217,235	220,161

Таким чином, підставивши всі проміжні значення в формулу 19.1, отримаємо вибіркове значення коефіцієнта кореляції для даних тварин:

$$r = \frac{15 \cdot 217{,}235 - 56{,}75 \cdot 57{,}38}{\sqrt{[15 \cdot 220{,}161 - (57{,}38)^2] \cdot [15 \cdot 215{,}235 - (56{,}75)^2]}} = 0{,}318.$$

Статистична помилка коефіцієнта кореляції розраховується за формулою:

$$SE_r = \sqrt{\frac{1 - r^2}{n - 2}}\ . \qquad (19.2)$$

Вірогідність відхилення отриманого вибіркового значення коефіцієнта кореляції від нуля можна оцінити на підставі критерію Ст'юдента:

$$t = \frac{r}{SE_r}. \qquad (19.3)$$

Статистичну значущість коефіцієнта кореляції можна вважати доведеною, якщо отримане значення критерію Ст'юдента перевищує табличне (див. Додаток И) для числа ступенів свободи: $df = n - 2$.

Для даних із нашого прикладу статистична помилка коефіцієнта кореляції становить:

$$SE_r = \sqrt{\frac{1 - 0{,}318^2}{15 - 2}} = 0{,}263.$$

Тоді оцінка критерію Ст'юдента буде дорівнювати: $t = \frac{0{,}318}{0{,}263} = 1{,}21$, що набагато менше, ніж його табличне значення для 13 ступенів свободи (2,18).

Таким чином, можна зробити висновок, що вибіркове значення коефіцієнта кореляції вірогідно не відхиляється від нуля, а тому зв'язок між вмістом жиру в молоці корів та їх матерів відсутній.

Для зручності в таблиці 19.2 наведено критичні значення коефіцієнта кореляції для різних рівнів ступенів свободи.

Таблиця 19.2 – Критичні значення коефіцієнта кореляції для різних рівнів ступенів свободи

df	$r_{крит}$	*df*	$r_{крит}$	*df*	$r_{крит}$	*df*	$r_{крит}$
10	0,576	19	0,433	28	0,361	80	0,217
11	0,553	20	0,423	29	0,365	90	0,205
12	0,532	21	0,413	30	0,349	100	0,195
13	0,514	22	0,404	35	0,325	125	0,174
14	0,497	23	0,396	40	0,304	150	0,159
15	0,482	24	0,388	45	0,288	200	0,138
16	0,468	25	0,381	50	0,273	300	0,113
17	0,456	26	0,374	60	0,250	400	0,098
18	0,444	27	0,367	70	0,232	500	0,088

Таким чином, для того, щоб зв'язок між вмістом жиру в молоці корів та їх матерів вважати статистично доведеним, значення коефіцієнта кореляції між ними повинно становити якнайменше 0,514.

У загальному вигляді для будь-якого числа ступенів свободи критичне значення коефіцієнта кореляції можна розрахувати за формулою:

$$r_{крит} = \frac{t}{\sqrt{df + t^2}}, \qquad (19.4)$$

де *t* – відповідні значення критерію Ст'юдента.

Наприклад, якщо обсяг вибірки 113 особин, табличне значення критерію Ст'юдента становить, відповідно, 1,98, то критичне значення коефіцієнта кореляції дорівнюватиме:

$$r_{крит} = \frac{1{,}98}{\sqrt{(113-2)+1{,}98^2}} = 0{,}185.$$

Розподіл вибіркових коефіцієнтів кореляції відповідає нормальному лише поблизу нульового значення, тому для побудови довірчого інтервалу вибіркового коефіцієнта кореляції необхідно використовувати його *z*-трансформацію:

$$z = \frac{1}{2}\ln\frac{1+r}{1-r}, \qquad (19.5)$$

статистична помилка якої не залежить від значення коефіцієнта кореляції, а лише від обсягу вибірки:

$$SE_z = \frac{1}{\sqrt{n-3}}. \quad (19.6)$$

Оскільки значення z вже мають нормальний розподіл, 95% довірчий інтервал для цього показника можна побудувати на підставі формули:

$$z - \frac{1{,}96}{\sqrt{n-3}} \le z \le z + \frac{1{,}96}{\sqrt{n-3}}. \quad (19.7)$$

Зворотний перехід від значень z до r проводиться за формулою:

$$r = \frac{e^{2\cdot z} - 1}{e^{2\cdot z} + 1}. \quad (19.8)$$

Для зручності наводимо таблицю переводу значень коефіцієнта кореляції в значення z та навпаки (табл. 19.3).

Таблиця 19.3 – Таблиця переводу значень коефіцієнта кореляції в значення z

r	Значення z для сотих часток r									
	0	1	2	3	4	5	6	7	8	9
0,0	0,000	0,010	0,020	0,030	0,040	0,050	0,060	0,070	0,080	0,090
0,1	0,100	0,110	0,121	0,131	0,141	0,151	0,161	0,172	0,182	0,192
0,2	0,203	0,213	0,224	0,234	0,245	0,255	0,266	0,277	0,288	0,299
0,3	0,310	0,321	0,332	0,343	0,354	0,365	0,377	0,388	0,400	0,412
0,4	0,424	0,436	0,448	0,460	0,472	0,485	0,497	0,510	0,523	0,536
0,5	0,549	0,563	0,576	0,590	0,604	0,618	0,633	0,648	0,662	0,678
0,6	0,693	0,709	0,725	0,741	0,758	0,775	0,793	0,811	0,829	0,848
0,7	0,867	0,887	0,908	0,929	0,950	0,973	0,996	1,020	1,045	1,071
0,8	1,099	1,127	1,157	1,188	1,221	1,256	1,293	1,333	1,376	1,422
0,9	1,472	1,528	1,589	1,658	1,738	1,832	1,946	2,092	2,298	2,647

Наприклад, оцінимо довірчий інтервал для коефіцієнта кореляції, отриманого між вмістом жиру в молоці корів та їх матерів. На цьому прикладі продемонструємо методику використання таблиці 19.3.

Як дано за умовою, коефіцієнт кореляції дорівнює r = 0,318. В таблиці 19.3 таке значення відсутнє. Обираємо такі два найближчі значення r, для яких є табличні значення z, так щоб $r_0 < r < r_1$. Таким чином, r_0 = 0,31 та r_1 = 0,32. Відповідні табличні значення для них складають: z_0 = 0,321 та z_1 = 0,332.

Тоді, використавши формулу лінійної інтерполяції, шукане значення можна отримати за формулою:

$$z = \pm [z_0 + u\cdot(z_1 - z_0)], \quad (19.9)$$

де $u = \frac{r - r_0}{r_1 - r_0}$.

Знак перед значенням z такий же, який має коефіцієнт кореляції.

Для нашого прикладу допоміжна величина дорівнює:

$$u = \frac{0{,}318 - 0{,}31}{0{,}32 - 0{,}31} = 0{,}8.$$

Підставляємо отримане значення в формулу 19.9 і оцінюємо значення:

$$z = 0{,}321 + 0{,}8 \cdot (0{,}332 - 0{,}321) = 0{,}330.$$

Потім розраховуємо довірчий інтервал для отриманої оцінки за формулою 19.7.

Нижня межа 95% довірчого інтервалу дорівнюватиме:

$$z_H = 0{,}330 - \frac{1{,}96}{\sqrt{15-3}} = -0{,}236,$$

а верхня:

$$z_B = 0{,}330 + \frac{1{,}96}{\sqrt{15-3}} = +0{,}896.$$

Для того, щоб від оцінок z повернутися до оцінок r знову використаємо таблицю 19.3.

У цій таблиці немає такого значення r, для якого $z = 0{,}236$ (знак поки не враховуємо). Обираємо два найближчі табличні значення z таким чином, щоб $z_0 < z < z_1$. Тоді, $z_0 = 0{,}234$ та $z_1 = 0{,}245$.

Відповідні значення r для них становлять: $r_0 = 0{,}23$ та $r_1 = 0{,}24$.

Знову використавши лінійну інтерполяцію, шукане значення можна отримати за формулою:

$$r = \pm (r_0 + 0{,}01 \cdot v), \qquad (19.10)$$

де $v = \dfrac{z - z_0}{z_1 - z_0}$.

Знак перед значенням коефіцієнта кореляції ставиться такий, який має значення z.

Підставляємо наші дані та отримуємо значення допоміжної величини:

$$v = \frac{0{,}236 - 0{,}234}{0{,}245 - 0{,}234} = 0{,}182.$$

Тоді шукана оцінка нижньої довірчої межі коефіцієнта кореляції складає:

$$r_H = -(0{,}23 + 0{,}01 \cdot 0{,}182) = -0{,}232.$$

(Оскільки відповідне значення z мало від'ємний знак, від'ємний знак також ставимо й перед значенням r).

Використавши аналогічну процедуру, визначаємо верхню межу 95% довірчого інтервалу. Вона становить $r_B = 0{,}714$.

Таким чином, з імовірністю 95% можна стверджувати, що генеральне значення коефіцієнта кореляції між вмістом жиру в молоці корів та їх матерів знаходиться у межах:

$$-0{,}232 \le r \le 0{,}714.$$

Оскільки нульове значення опинилося в межах довірчого інтервалу, знову підтверджується наш висновок щодо відсутності кореляційного зв'язку між вмістом жиру в молоці корів та їх матерів.

19.2 Лінійна регресія

Регресія – це лінія, вид залежності середньої величини результативної ознаки від факторної. З погляду математики регресія являє собою функцію $y = f(x_1, x_2, ..., x_n)$, яка описує залежність умовного математичного очікування залежної змінної (y) від заданих фіксованих значень незалежної змінної ($x_1, x_2, ..., x_n$).

Рівняння регресії – аналітичне рішення, за допомогою якого відображується зв'язок між досліджуваними ознаками. Розрізняють *прямолінійне* рівняння зв'язку (пряма лінія) і *криволінійне* (парабола, гіпербола, експонента, логарифмічна крива, тощо).

При прямолінійній залежності однієї ознаки від іншої рівняння регресії має вигляд:

$$y = a + b \cdot x, \tag{19.11}$$

де b – коефіцієнт лінійної регресії, тангенс кута нахилу лінії регресії відносно осі OX;

a – точка перетину лінією регресії осі OY.

Коефіцієнт лінійної регресії визначає середній рівень зміни залежної ознаки при зміні незалежної змінної на одиницю.

Визначення коефіцієнтів лінійної регресії відбувається із застосуванням методу найменших квадратів (МНК), розробленого Гаусом ще у XIX сторіччі.

Відповідно до цього методу, коефіцієнти лінійної регресії є рішенням системи лінійних рівнянь:

$$\begin{cases} a \cdot n + b \cdot \sum x = \sum y; \\ a \cdot \sum x + b \cdot \sum x^2 = \sum xy. \end{cases} \tag{19.12}$$

Параметри a та b рівняння лінійної регресії можна визначити за іншими робочими формулами:

$$b = \frac{n \cdot \sum xy - \sum x \cdot \sum y}{n \cdot \sum x^2 - \left(\sum x\right)^2}; \tag{19.13}$$

$$a = \frac{\sum y \cdot \sum x^2 - \sum xy \cdot \sum x}{n \cdot \sum x^2 - \left(\sum x\right)^2}. \tag{19.14}$$

Як бачимо, для розрахунку коефіцієнтів парної лінійної регресії необхідні ті ж самі проміжні розрахунки, що й для розрахунку коефіцієнта парної кореляції. Тому для визначення форми аналітичної залежності жирномолочності дочок від вмісту жиру в молоці їх матерів використовуємо дані, що наведено у таблиці 19.1.

Тоді, коефіцієнти парної лінійної регресії дорівнюватимуть:

$$b = \frac{15 \cdot 217{,}235 - 56{,}75 \cdot 57{,}38}{15 \cdot 215{,}028 - (56{,}75)^2} = 0{,}455$$

та

$$a = \frac{57{,}38 \cdot 215{,}028 - 217{,}235 \cdot 56{,}75}{15 \cdot 215{,}028 - (56{,}75)^2} = 2{,}104.$$

Таким чином, можна зробити висновок, що при зміні вмісту жиру в молоці корів-матерів на 1% у їх дочок рівень жирномолочності підвищується на 0,455%. Аналітична форма зв'язку має наступний вигляд:

$$y = 2{,}104 + 0{,}455 \cdot x.$$

Коефіцієнти парної лінійної регресії, як усі вибіркові показники, мають визначений рівень випадковості, який визначається їх статистичними помилками:

$$SE_b = \sqrt{\frac{\sigma^2_{зал}}{\sum x^2 - \frac{\left(\sum x\right)^2}{n}}}, \tag{19.15}$$

$$SE_a = SE_b \cdot \sqrt{\frac{\sum x^2}{n}}, \tag{19.16}$$

де $\sigma^2_{зал}$ – залишкова варіанса, яка визначається за формулою:

$$\sigma^2_{зал} = \frac{\sum y^2 - a \cdot \sum y - b \cdot \sum xy}{n-2}. \tag{19.17}$$

Перевірка рівня вірогідності відмінності кожного з коефіцієнтів парної лінійної регресії від нуля відбувається із застосуванням критерію Ст'юдента:

$$t_b = \frac{b}{SE_b}, \tag{19.18}$$

$$t_a = \frac{a}{SE_a}. \tag{19.19}$$

Коефіцієнти парної лінійної регресії вважаються значущими, якщо відповідні їм значення критерію Ст'юдента перевищують табличні значення (Додаток И) для числа ступенів свободи $df = n - 2$.

Таким чином, для розрахунку оцінок статистичних помилок коефіцієнтів парної лінійної регресії по-перше необхідно розрахувати значення залишкової варіанси:

$$\sigma^2_{зал} = \frac{220{,}161 - 2{,}104 \cdot 57{,}38 - 0{,}455 \cdot 217{,}235}{15-2} = 0{,}046.$$

Тоді оцінки помилок коефіцієнтів регресії становитимуть, відповідно:

$$SE_b = \sqrt{\frac{0{,}046}{215{,}028 - \frac{(56{,}75)^2}{15}}} = 0{,}377$$

та

$$SE_a = 0{,}377 \cdot \sqrt{\frac{215{,}028}{15}} = 1{,}427 .$$

Таким чином, відповідні значення критерію Ст'юдента дорівнюють:

$$t_b = \frac{0{,}455}{0{,}377} = 1{,}21; \quad t_a = \frac{2{,}104}{1{,}427} = 1{,}47.$$

У Додатку И знаходимо критичне значення критерію Ст'юдента для числа ступенів свободи $df = 15 - 2 = 13$; воно дорівнює 2,16.

Таким чином, ні для коефіцієнта *a*, ні для коефіцієнта *b* рівняння лінійної регресії розрахункові значення критерію Ст'юдента не перевищують табличне значення.

Ще раз підтвердився отриманий вище висновок, про відсутність зв'язку між жирномолочністю корів та їх матерів.

У графічній формі ця залежність має наступний вигляд (рис. 19.1).

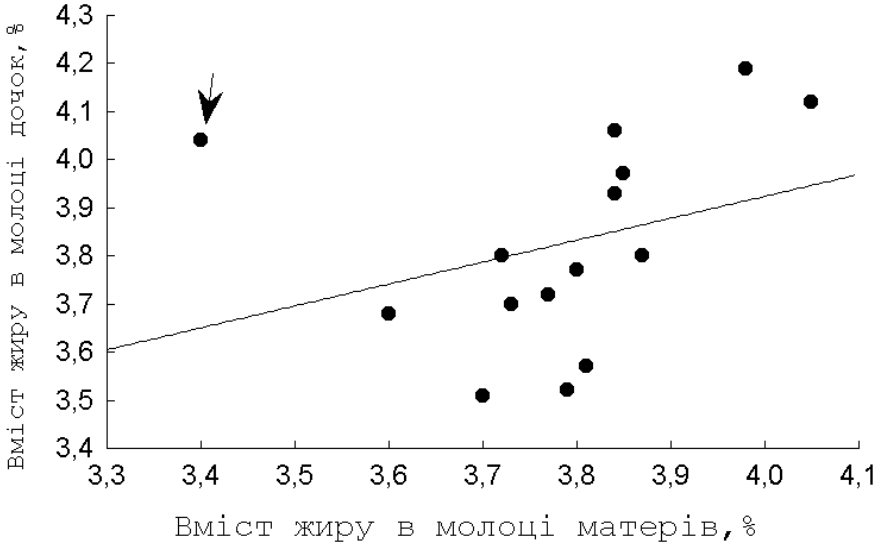

Рисунок 19.1 – Залежність між жирномолочністю корів та їх дочок, побудована на підставі даних таблиці 19.1

Графічний аналіз дозволяє візуалізувати залежність, що аналізується та краще з'ясувати характер цього зв'язку. Це стосується, насамперед, виявлення «викидів», тобто, значень, що значно відхиляються від загальної залежності, але впливають на хід розрахунків.

У нашому випадку точка, яка на графіку позначена стрілкою, значно відхиляється від інших пар «матері-дочки». Якщо її видалити, зв'язок між вмістом жиру в молоці матерів та їх дочок стане вірогідним.

Як і у випадку з іншими статистичними показниками, для оцінок рівня та характеру зв'язку теж можуть бути використані методи ресамплінгу, які більш придатні у випадках, коли мають місце «викиди» (див. рис. 19.1). При цьому для перевірки нуль-гіпотези щодо відсутності зв'язку між двома наборами значень (насамперед, лінійного зв'язку) можуть бути використані як бутстреп-процедура, так і метод перестановок.

Продемонструємо принцип їх використання на прикладі даних, що наведено в таблиці 19.1.

При використанні бутстреп-процедури оцінювання значень коефіцієнта кореляції необхідно зробити наступне. Випадковим чином оберемо із 15 пар значень *X* та *Y* таблиці вихідних даних першу пару. Наприклад, це буде пара (3,84; 4,06). Далі знову із 15 пар вихідних даних оберемо ще одну пару значень *X* та *Y*, при цьому незважаючи на те, чи була вона вже обрана в попередній раз чи ні. В кінцевому рахунку сформуємо першу нашу псевдовибірку, яка буде містити 15 пар значень. При цьому, деякі пари значень можуть бути не обрані в цю псевдовибірку жодного разу, а деякі, навпаки, бути присутні два чи навіть більше разів.

Для цієї псевдовибірки необхідно розрахувати оцінку коефіцієнта кореляції (чи регресії, якщо необхідно).

Далі знову випадковим чином формується друга псевдовибірка із 15 пар значень, які складаються із вихідних даних. І для неї також проводиться оцінка коефіцієнта кореляції.

В кінцевому рахунку, необхідно провести таку процедуру багато разів, наприклад, 500, 1000 або навіть 5000 і отримати відповідну кількість оцінок коефіцієнта кореляції для всіх цих псевдовибірок.

Бутстреп-оцінка коефіцієнта кореляції може бути розрахована за формулою:

$$r_{boot} = \frac{\sum_{i=1}^{M} r_i}{M}, \tag{19.20}$$

а її статистична помилка:

$$SEr_{boot} = \sqrt{\frac{\sum_{i=1}^{M} (r_i - r_{boot})^2}{M-1}}, \tag{19.21}$$

де r_i – оцінка коефіцієнту кореляції, отримана для *i*-тої псевдовибірки.

Нижня та верхня межі 95% довірчого інтервалу коефіцієнта кореляції можуть бути розраховані як 2,5% та 97,5% перцентилі вибірки оцінок коефіцієнта кореляції, отриманих для використаних псевдовибірок.

Для даних, що наведено в таблиці 19.1, ми сформували 50 псевдовибірок. Відповідна бутстреп-оцінка коефіцієнта кореляції, що була оцінена на підставі цих псевдовибірок, дорівнює: $r_{\text{boot}} = 0{,}355 \pm 0{,}362$ із 95% довірчим інтервалом: [-0,326; 0,852].

Як бачимо, і у випадку використання бутстреп-процедури нуль потрапляє у довірчий інтервал, що свідчить про те, що нуль-гіпотеза не може бути відхилена і між показниками жирномолочності корів та їх дочок зв’язок відсутній.

Критерій перестановок, використаний для вирішення цієї ж нульової гіпотези, буде мати наступні етапи розрахунку.

Стовпчик із значеннями ігреків залишаємо без змін, а стовпчик іксів перемішаємо випадковим чином. Для отриманих пар значень розрахуємо оцінку коефіцієнта кореляції. Знову перемішаємо стовпчик іксів і знову розрахуємо для отриманих пар значень оцінку коефіцієнта кореляції. Повторимо таку процедуру ще багато разів (наприклад, *M*).

Після закінчення підрахуємо, скільки разів отримані для псевдовибірок оцінки коефіцієнта кореляції перевищували або дорівнювали тій оцінці, що була отримана для вихідних даних (для нашого прикладу: r = 0,318).

Наприклад, це становить m разів. Тоді рівень значущості для критерію перестановок, що було використано для перевірки нуль-гіпотези щодо рівності нулю оцінки коефіцієнту кореляції, становить:

$$p = \frac{m}{M+1}. \qquad (19.22)$$

У нашому випадку із 500 перестановок в 61 випадку оцінка коефіцієнта кореляції дорівнювала або навіть перевищувала значення, отримане для вихідних даних.

Таким чином, рівень значущості для цієї величини становить:

$$p = 61 : (500+1) = 0{,}122.$$

Це досить високе значення, відповідно, наша нульова гіпотеза не може бути відхилена. Тобто, є висока імовірність, що ми отримали таку високу оцінку коефіцієнта кореляції, хоча насправді зв’язок між цими двома рядами даних відсутній.

Аналогічні розрахунки можуть бути проведені і для перевірки нуль-гіпотези щодо рівності нулю оцінок коефіцієнтів регресії.

19.3 Використання моделей нелінійної регресії в селекції

Математичний аналіз та моделювання використовуються в різних галузях тваринництва вже давно. Найбільше математичні моделі використовуються для опису динаміки живої маси з віком (криві росту живої маси), динаміки продукції молока протягом лактації (лактаційні криві) та динаміки яєчної продукції птиці (криві яєчної продуктивності). Лише у другій половині XX століття було розроблено більше десятка математичних моделей, що дозволяють проводити аналіз та робити прогноз продуктивності на підставі тих чи інших припущень.

Таким чином, головною метою нашої роботи став огляд найбільш поширених математичних моделей кривих росту, лактаційних кривих та кривих яєчної продуктивності та аналіз особливостей їх використання.

Криві росту живої маси

Серед математичних моделей, що використовуються для аналізу росту живої маси, найбільш поширеним є *рівняння Л. фон Берталанффі* (von Bertalanffy, 1938):

$$W_t = W\infty \cdot (1 - \exp(\alpha - \beta \cdot t))^3, \quad (19.23)$$

де *Wt* – жива маса у віці *t*;

W∞ – асимптота живої маси;

α, β – параметри рівняння.

Математико-біологічний сенс цього рівняння наступний: зі збільшенням віку (*t*) ступінь експоненти зростає, але має від'ємний знак, тому вираз у дужках наближається до одиниці. Таким чином, жива маса наближається до свого максимально можливого показника (асимптоти).

У загальному вигляді ступінь виразу, що знаходиться у дужках формули 19.23 також приймається за параметр, оцінку якого необхідно провести на підставі емпіричних даних.

Використання цього рівняння на практиці має свої особливості. Дана модель повинна бути параметризована в два етапи. На першому етапі відбувається розрахунок значення асимптоти (*W*∞), а на другому – оцінювання коефіцієнтів α та β рівняння.

Значення *W*∞ може бути визначено з використанням методу Форда-Волфорда. Цей метод полягає у тому, що значення *W*∞ можна визначити, як точку перетинання системи рівнянь:

$$\begin{cases} W_{t+1} = a + b \cdot W_t; \\ W_{t+1} = W_t. \end{cases} \quad (19.24)$$

При невеликому перетворенні рівняння 19.23 може бути легко перетворене в лінійне рівняння, коефіцієнти якого розраховуються методом найменших квадратів:

$$\ln\left[1 - \sqrt[3]{\frac{W_t}{W_\infty}}\right] = \alpha - \beta \cdot t. \quad (19.25)$$

Рівняння Б. Гомперца (Gompertz, 1825) має вигляд:

$$W_t = W_0 \cdot \exp\left(\frac{A_0(1 - e^{-\alpha \cdot t})}{\alpha}\right), \quad (19.26)$$

де W_0 – жива маса при народженні;

A_0 та α – постійні, що специфічні для виду (чи популяції) й обумовлюють початковий темп росту та швидкість дозрівання, відповідно.

Максимально можлива маса організму (асимптота), тобто W_∞, при використанні рівняння Гомперца становитиме:

$$W_\infty = W_0 \cdot \exp\left(\frac{A_0}{\alpha}\right). \quad (19.27)$$

Рівняння А. Пюттера (Pütter, 1920) має вигляд:

$$W_t = \frac{W\infty}{\exp[\beta(t + \alpha)^p]}, \quad (19.28)$$

де α, β та *p* – специфічні для виду (або популяції) константи.

Ще однією математичною моделлю для аналізу росту живої маси тварин є *рівняння Т. Бріджеса* (Bridges et al., 1986):

$$W_t = W_\infty \cdot (1 - \exp(-\mu \cdot t^\alpha)), \quad (19.29)$$

де W_∞ – середня жива маса, яку досягає тварина при статевому дозріванні;

α – константа кінетичного росту;

μ – константа експоненціального росту.

Це також двопараметрична модель, для якої спочатку необхідно розрахувати оцінку параметра W_∞ (з використанням методу Форда-Волфорда).

Важливий біологічний зміст використаних рівнянь для опису росту живої маси полягає в тому, що після визначеного моменту часу значення залежного параметра (тобто, живої маси) досягає максимально можливого (для даної популяції або даного виду), але ніколи його не перевищує.

Нарешті, ще одна модель може бути запропонована для моделювання процесів росту живої маси – це *логістична модель* (Verhulst, 1838):

$$W_t = \frac{W_\infty}{1 + b \cdot \exp(-c \cdot t)}, \quad (19.30)$$

де *b* та *c* – параметри, які визначають форму кривої.

Лактаційні криві

Серед математичних моделей, що використовуються для опису процесів динаміки молочної продуктивності (насамперед, корів) протягом лактації, найбільш поширеною є *модель П. Вуда* (Wood, 1967):

$$Y_t = a \cdot t^b \cdot \exp(-c \cdot t), \quad (19.31)$$

де Y_t – надій, отриманий протягом t-тої одиниці часу (доби, тижня, декади, місяця);

a, *b* та *c* – параметри моделі Вуда.

Ці параметри мають наступний сенс: *b* – коефіцієнт, що характеризує інтенсивність підйому лактаційної кривої до точки перегину, *c* – коефіцієнт зниження лактаційної кривої після досягнення максимально можливого рівня продуктивності.

На підставі оцінок параметрів моделі Вуда можна розрахувати наступні характеристики лактаційної кривої: сталість лактаційної кривої (Wood, 1970):

$$S = \frac{1}{c^{(b+1)}}, \quad (19.32)$$

дату пікового значення молочної продуктивності протягом лактації (виражається у прийнятих у моделі одиницях часу):

$$t_{peake} = \frac{b}{c}, \quad (19.33)$$

і, нарешті, рівень продуктивності у момент піку:

$$Yt_{peake} = a \cdot \left(\frac{b}{c}\right)^b \cdot \exp(-b). \quad (19.34)$$

Іншою моделлю, що часто застосовується для аналізу лактаційної кривої є *модель Дж. Вілмінка* (Wilmink, 1987):

$$Yt = a + b \cdot \exp(-0{,}01 \cdot t) + c \cdot t \,. \qquad (19.35)$$

Більш складною є *модель Алі та Шаффера* (Ali, Schaeffer, 1987):

$$Yt = a + b \cdot \left(\frac{t}{k}\right) + c \cdot \left(\frac{t}{k}\right)^2 + d \cdot \ln\left(\frac{k}{t}\right) + e \cdot \left(\ln\left(\frac{k}{t}\right)\right)^2 , \qquad (19.36)$$

де k – тривалість лактації у визначених одиницях часу.

Якщо облік молочної продуктивності проводиться щодобово, протягом 305 діб лактації, то k = 305; якщо облік молочної продуктивності проводиться щомісячно, протягом, наприклад, 11 місяців лактації, то приймають, що k = 11.

В моделі Алі та Шаффера, параметр a визначає пікове значення молочної продуктивності, параметри d та e визначають інтенсивність підйому лактаційної кривої до точки перегину, параметри b та c – інтенсивність зниження рівня молочної продуктивності після досягнення нею піку.

Криві динаміки яєчної продуктивності

Серед математичних моделей, що використовують для динаміки яєчної продуктивності у птахівництві найбільш поширеними є дві моделі – МакНеллі та МакМіллана.

Модель Д. МакНеллі (McNally, 1971) є модифікованою моделлю Вуда та має наступний вигляд:

$$Yt = a \cdot t^b \cdot \exp(-c \cdot t + d \cdot \sqrt{t}) \,. \qquad (19.37)$$

Модель І. МакМіллана (McMillan et al., 1970; Yang, Wu, McMillan, 1989) більш складна:

$$Yt = M \cdot (1 - \exp(-c \cdot t)) \cdot \exp(-b \cdot t) \,. \qquad (19.38)$$

де M – параметр шкали;

b – показник, що визначає інтенсивність підйому яєчної продуктивності до моменту піку;

c – показник, що визначає швидкість зниження рівня яєчної продуктивності після піку.

За своєю суттю, модель МакМіллана описує криву яєчної продуктивності, шляхом розкладання її на дві частини. Крива підйому продуктивності визначається першою експоненціальною компонентою моделі 19.38, а крива спаду – другою експоненціальною компонентою моделі. Ця модель також двопараметрична й може бути оцінена лише у два етапи.

На першому етапі проводиться оцінка показника M. Це можна зробити, звернувши увагу на те, що для великих значень t модель 19.38 може бути перетворена (завдяки логарифмуванню) до наступного вигляду:

$$\ln Yt = \ln M - b \cdot t \,. \qquad (19.39)$$

Таким чином, зобразивши показники яєчної продуктивності в напівлогарифмічній шкалі, можна отримати значення ln$M$, як точку перетину рівняння регресії із віссю OY. Але при цьому необхідно взяти до уваги, що у

рівнянні 19.39 можна використовувати лише ті значення продуктивності, які мають місце після досягнення свого максимуму.

Параметри моделі МакМіллана *b* та *c* розраховуються вже після того, як у модель 19.38 підставлено знайдену оцінку *M*.

Інтерполяція лактаційних кривих

Лактаційні криві відображають особливості динаміки формування молочної продуктивності тварин протягом періоду лактаційної діяльності. Їхня форма, насамперед, обумовлюється інтенсивністю наростання рівня продуктивності, часом досягнення максимального рівня продуктивності (асимптоти), величиною асимптоти та швидкістю зниження продуктивності після досягнення асимптоти. Найбільш адекватно лактаційні криві корів можуть бути проаналізовані за допомогою моделі П. Вуда (19.31).

Але при аналізі лактаційної діяльності важливою проблемою є те, що отелення корів дуже рідко відбувається першого числа календарного місяця й, відповідно, показники молочної продуктивності тварин за календарні місяці протягом лактації, що відображені у формі 2-МОЛ, не співпадають із реальними величинами надоїв за фактичні місяці лактації для певної тварини.

Одним із найпростіших методів, здатних нівелювати таке зміщення, є використання лінійної чи нелінійної інтерполяції.

В даній роботі нами запропоновано два методи інтерполяції при аналізі лактаційних кривих корів.

Метод 1. Даний метод ґрунтується на використанні нелінійної інтерполяції на підставі поліному третього ступеня:

$$CumY_t = a + b \cdot t + c \cdot t^2 + d \cdot t^3 , \qquad (19.40)$$

де $CumY_t$ – накопичений надій (у кг) на момент часу від початку лактації *t* (днів).

Наприклад, корова отелилася 27 березня й її надій склав 33 кг молока за березень, 350 – за квітень, 387 – за травень, 371 – за червень, 367 – за липень, 314 – за серпень, 270 – за вересень, 264 – за жовтень, 230 – за листопад, 205 – за грудень і 115 – за січень наступного року.

Таким чином, її накопичений надій становитиме: за п’ять днів березня – 33 кг, за п’ять днів березня та 30 днів квітня (разом 35 днів) – 33 + 350 = 383 кг, за п’ять днів березня, 30 днів квітня та 31 день травня (разом 66 днів) – 33 + 350 + 387 = 770 кг, і т.д. Разом за 311 днів лактації від цієї корови було отримано 2906 кг молока.

Графічно залежність накопиченого надою від моменту часу з початку лактації та апроксимація цієї залежності поліномом третього ступеня наведено на рисунку 19.2.

Рівень адекватності модельної залежності поліномом дуже високий (коефіцієнт детермінації: R^2 = 99,984%), тому можна використовувати отримане рівняння для розрахунку показників надою за певні відрізки часу.

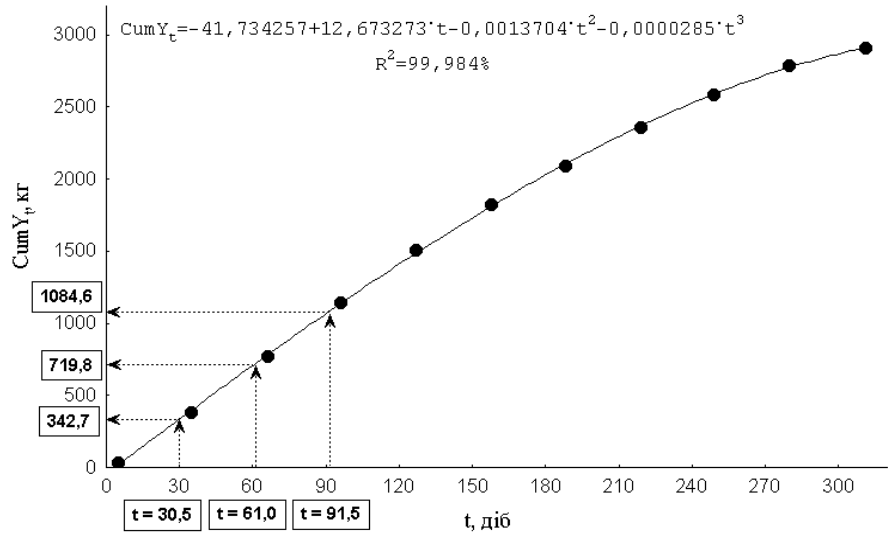

Рисунок 19.2 – Апроксимація накопиченого надою поліномом третього ступеня

Наприклад, за перші 30,5 днів апроксимоване значення накопиченого надою складає 342,7 кг молока, за 61 день від початку лактації – 719,8 кг, за 91,5 днів – 1084,6 кг.

Відрізки часу ми обирали таким чином, щоб 305 днів відповідали 10 місяцям лактації. Таким чином нівелюються відмінності між тривалістю різних календарних місяців протягом року.

Тоді, величина надою за першій місяць лактації для даної тварини становитиме 342,7 кг, за другий – 719,8 – 342,7 = 377,1 кг, за третій – 1084,6 – 719,8 = 364,8 кг, і т. д. А в цілому за 305 днів лактації від цієї тварини отримано 2887,5 кг молока.

Даний метод розрахований на індивідуальний підхід до кожної лактації кожної тварини окремо, тому його результати можуть бути використані при аналізі впливу гено- та паратипових факторів, а також їх сполучення на особливості формування молочної продуктивності корів.

Метод 2. На відміну від попереднього, даний метод розрахований на аналіз даних, отриманих від групи (обов'язково гомогенної) тварин. Гомогенізація повинна проводитися на рівні генотипу, походження, віку, умов утримання і т. п.

Даний метод базується на використанні лінійної інтерполяції, особливості якої можуть не давати зміщення на коротких відрізках лактаційної діяльності, як видно на рисунку 19.2.

Нехай тварина отелилася у k-тий день від початку будь-якого календарного місяця. (Для зручності будемо вважати, що тривалість календарного місяця складає 30 днів.) Тоді, використовуючи лінійну інтерполяцію, можна розрахувати величину надою за перший фактичний місяць її лактації (N_1) за формулою:

$$N_1 = \left(\frac{(30-k)}{30}\right)\cdot HK_1 + \left(\frac{k}{30}\right)\cdot HK_2, \qquad (19.41)$$

де $НК_1$ та $НК_2$ – надої даної тварини за перший та другий календарні місяці.

Оскільки, як ми вказали вище, нас більше цікавить значення не для певної тварини, а очікуване значення для групи тварин, то нам необхідно отримати формулу математичного очікування для виразу, що знаходиться у правій частині формули 19.41. Важливу роль при цьому має середня дата отелення. Оскільки корова може отелитися у будь-який день календарного місяця, можна припустити, що ймовірність цієї події для кожного дня місяця однакова й розподіл отелень формується за рівномірним законом.

Це припущення було перевірено на підставі аналізу розподілу 297 отелень корів української червоної молочної породи племзаводу «Зоря» Херсонської області (рис. 19.3).

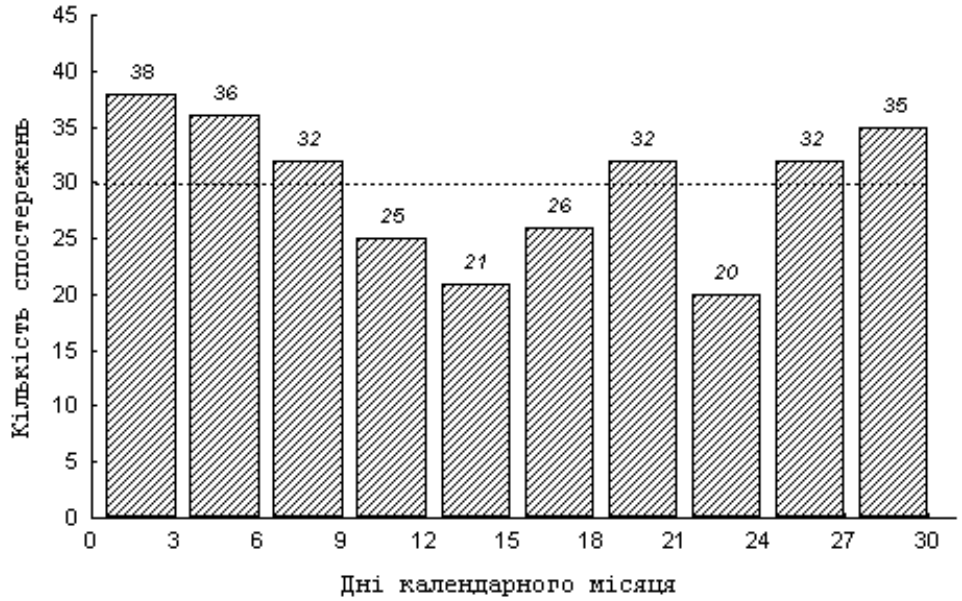

Рисунок 19.3 – Розподіл 297 корів української червоної молочної породи племзаводу «Зоря» Херсонської області за датою отелення

Як видно із рисунка 19.3, емпіричний розподіл отелень досить адекватно характеризується рівномірним розподілом (наведений штриховою лінією), що підтверджується низьким значенням критерію Хі-квадрат К. Пірсона (χ^2 = 12,057; df = 8; p = 0,149).

Як відомо з курсу теорії ймовірності, математичне очікування для змінної, що має рівномірний тип розподілу, дорівнює напівсумі обох межованих значень. Таким чином, середня дата отелення (для групи тварин, чисельність якої прямує до нескінченності) буде дорівнювати: $\bar{k} = \frac{0+30}{2} = 15$ дню від початку календарного місяця.

Якщо це значення підставити у формулу 19.41, то ми отримаємо наступний вираз для математичного очікування надою за перший фактичний місяць лактації:

$$E[N_1] = E\left[\frac{1}{2}HK_1 + \frac{1}{2}HK_2\right]. \qquad (19.42)$$

Згідно властивостей математичного очікування, математичне очікування суми двох випадкових величин дорівнює сумі їх математичних очікувань, при цьому постійне значення можна винести за дужки. Тому формулу (19.42) можна переписати у вигляді:

$$\overline{N}_1 = \frac{1}{2} \cdot \left(\overline{HK}_1 + \overline{HK}_2 \right). \qquad (19.43)$$

де $\overline{HK}_1, \overline{HK}_2$ – оцінки середніх арифметичних показників надою за перший та другий календарні місяці лактації для групи тварин.

В кінцевому рахунку, оцінку середнього арифметичного надою за *i*-тий порядковий місяць лактації для групи тварин можна отримати за наступною формулою:

$$\overline{N}_i = \frac{1}{2} \cdot \left(\overline{HK}_i + \overline{HK}_{i+1} \right). \qquad (19.44)$$

На рисунку 19.4 наведено дві лактаційні криві, одна побудована на підставі оцінок надою за календарні місяці (без урахування інтерполяції), а друга – на підставі оцінок місячних надоїв, отриманих на підставі формули 19.44.

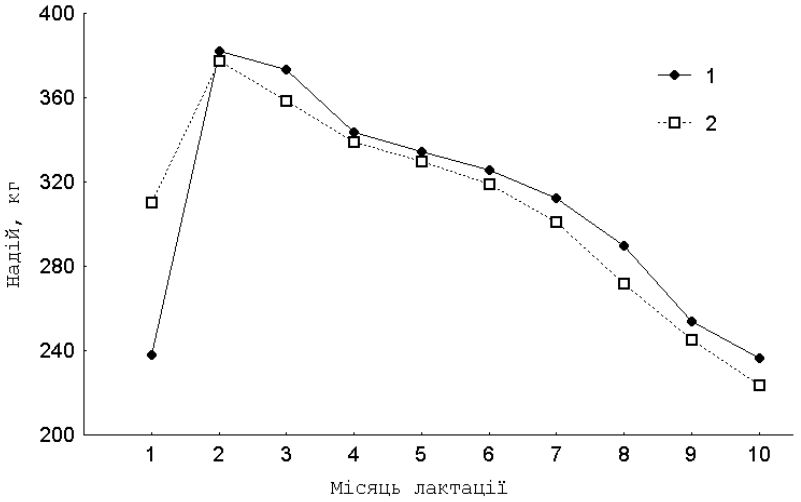

Рисунок 19.4 – Лактаційні криві корів, побудовані на підставі середніх показників надоїв за календарні місяці лактації (1) та з урахуванням лінійної інтерполяції (2)

Як бачимо, відмінності між цими кривими стосуються, насамперед, інтенсивності наростання рівня молочної продуктивності на початку лактації, що, без сумніву, буде призводити до отримання зміщених оцінок показників лактаційної діяльності корів при аналізі їх на підставі математичних моделей (насамперед, моделі П. Вуда).

Контрольні питання:

1. За яких умов доцільно використовувати методи ресамплінгу для оцінок рівня та характеру зв’язку?
2. Які математичні моделі використовуються для опису процесів динаміки молочної та яєчної продуктивності?

§ 20. Ентропійно-інформаційний аналіз кількісних ознак

Останнім часом з'явилося багато публікацій, у яких продемонстровано можливості застосування ентропійно-інформаційного аналізу (ЕІА) у різних сферах біологічної науки, фізіології, медицині й т. ін. В екології, поряд з використанням формули К. Шеннона для оцінки міри біорізноманіття окремих угруповань і біоценозів, ЕІА одержав своє застосування і як метод біоіндикації екосистем стосовно співвідношення мір адаптивності та інадаптивності ознаки чи групи ознак. При цьому приводяться численні приклади застосування ЕІА при вивченні як дискретних (якісних) ознак, що мають поліноміальний розподіл (для яких і було розроблено основні положення теорії інформації), так і кількісних ознак (полігенних), розподіл яких найчастіше відповідає нормальному закону (див. § 15) чи розподілу близькому до нього.

Однак дотепер немає єдиного, теоретично обґрунтованого методу оцінки ентропії для кількісних ознак. Крім того, проаналізувавши приклади застосування ЕІА і методики, що використовуються в цих роботах, ми дійшли висновку, що вони дуже часто мають один істотний недолік, що може вплинути на одержані результати. Це спонукало нас до розробки нового підходу до використання ЕІА для кількісних ознак.

Базисним поняттям теорії інформації є поняття **ентропії,** математично точний зміст якого випливає з робіт К. Шеннона. Ентропія – міра невизначеності деякої ситуації. Її також можна розглядати як міру розсіювання, і в цьому розумінні вона подібна до статистичного поняття «дисперсія». Але якщо дисперсія є адекватною мірою розсіювання тільки для спеціальних розподілів ймовірностей випадкових величин (зокрема, для розподілу Гауса), то ентропія не залежить від типу цього розподілу.

Крім того, ентропія володіє і низкою інших корисних властивостей. По-перше, невизначеність будь-якої системи зростає з ростом числа можливих результатів. А, по-друге, міра невизначеності має властивість адитивності.

Вперше особливості функціонування біологічних систем різного рівня з погляду теорії інформації були розглянуті в роботі І. І. Шмальгаузена. Ним було введено поняття про канали прямого і зворотного зв'язку, по яких передається генетична і фенотипова інформація, розглянуто закономірності кодування і перетворення біологічної інформації.

У. Ешбі вперше запропонував використовувати поняття ентропії для характеристики міри складності системи. Відповідно до його уявлень, складність системи (у тому числі і біологічної) можна охарактеризувати рівнем її розмаїття. Під розмаїттям звичайно розуміється кількість станів, що може набувати система. Крім цього, при оцінці ентропії враховується не тільки абсолютна кількість таких станів, але й імовірність (а у вибіркових дослідженнях імовірність можна замінити частотою), з якою система набуває той чи інший стан. Тоді оцінкою для ентропії може служити наступний вираз:

$$H = -\sum_{i=1}^{k} (p_i \cdot \log_2 p_i), \qquad (20.1)$$

де p_i – імовірність (або частота) того, що система набуде *i*-тий стан із *k* можливих.

Як легко переконатися, максимуму ця величина досягає в тому випадку, коли імовірності прийняття системою кожного із *k* можливих станів однакові. У цьому випадку значення ентропії для такої системи дорівнюватиме:

$$H\max = \log_2 k. \qquad (20.2)$$

У тому випадку, коли система може прийняти лише один стан із частотою рівною 1, ентропія її дорівнює нулю. Таким чином, для будь-якої системи має місце вираз:

$$0 \le H \le H\max. \qquad (20.3)$$

Ентропія, як міра розмаїття й організованості системи, насамперед, характеризує ступінь її невизначеності чи, іншими словами, детермінованості. Система вважається тим більше детермінованою, чим менше значення її ентропії, тобто, чим ближче величина *H* до нуля.

Як ми показали вище, це відбувається в тому випадку, коли один з можливих станів системи має дуже високу імовірність (частоту) прояву. З цих позицій, поняття ентропії можна порівняти з коефіцієнтом успадкування (h^2), уведеним Р. Фішером. Чим вище значення коефіцієнта успадкування ознаки в будь-який групі організмів (популяції), тим менше рівень прояву цієї ознаки залежить від паратипових факторів. Відповідно, тим вищою є детермінованість (у даному випадку фенотипу генотипом) і нижчою ентропія системи.

У своєму застосуванні ЕІА здебільшого розрахований на системи (ознаки), що мають якісне вираження, тобто для яких розходження між окремими станами має дискретний характер. Однак, кількісні (чи безперервні) ознаки також можна аналізувати в термінах і поняттях ЕІА.

У цьому випадку, для всього можливого спектра значень, що може набути ознака, встановлюють деяку міру точності (Δx), у межах якої стани системи виявляються практично нерозрізнені. Тоді розподіл континуальних ознак можна приблизно звести до дискретного. Це рівносильно заміні плавної кривої функції щільності розподілу *f(x)* східчастою ламаною (типу гістограми). У цьому випадку площі прямокутників цієї гістограми зображують імовірність набуття системою того чи іншого стану.

Характерно, що оцінка ентропії для кількісної ознаки виявляється зовсім не залежною від прийнятої точності (Δx). Від точності виміру залежить лише початок відліку, при якому обчислюється ентропія.

Якщо система (ознака) має нормальний розподіл (ідеальний), то в цьому випадку її ентропія, розрахована по гістограмі, буде дорівнювати:

$$H = \log_2 \left[\frac{\sigma\sqrt{2 \cdot \pi \cdot e}}{\Delta x} \right], \qquad (20.4)$$

де σ – середнє квадратичне відхилення ознаки, що аналізується.

Як правило, при порівнянні ознак, що мають різні величини виміру (кг, м, %, градуси і т. ін.), вони попередньо підлягають стандартизації. У цьому випадку кожне вихідне значення у вибірці (x_i) заміняють відповідною *z*-величиною:

$$z = \frac{x_i - \bar{x}}{\sigma}. \qquad (20.5)$$

Характерною рисою цих трансформованих величин є те, що вони мають середнє арифметичне значення рівне 0, та варіансу (і, відповідно, середнє квадратичне відхилення), рівну 1.

Незалежно від того, чому були рівні мінімальне і максимальне значення у вихідній вибірці, для z-трансформованих величин переважна більшість значень розташовуються в межах від –3 до +3 (рис. 20.1).

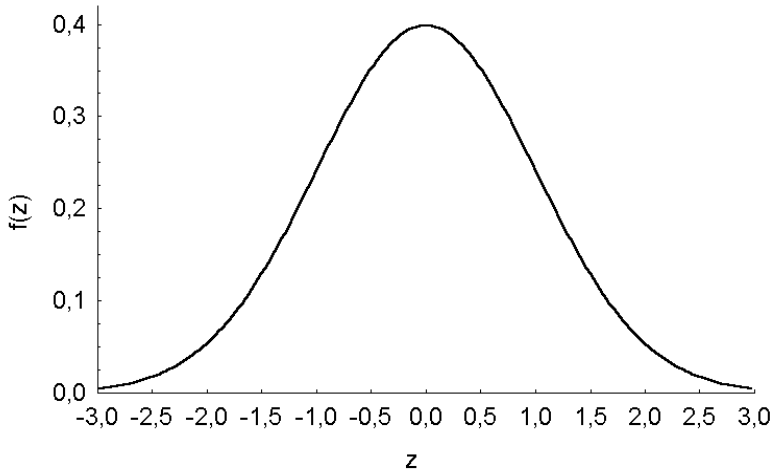

Рисунок 20.1 – Графік щільності розподілу стандартизованих величин

У цьому випадку (при достатньому обсязі вибірки, що містить сотні вимірів), зручніше прийняти величину міри точності однакову для всіх ознак і рівну $\Delta x = 0{,}5$. Це дає нам 12 інтервалів у межах варіаційного ряду. Таким чином, ми приймаємо, що $k = 12$. І замість плавного графіка щільності розподілу *f(z)* ми маємо справу зі східчастою гістограмою (рис. 20.2).

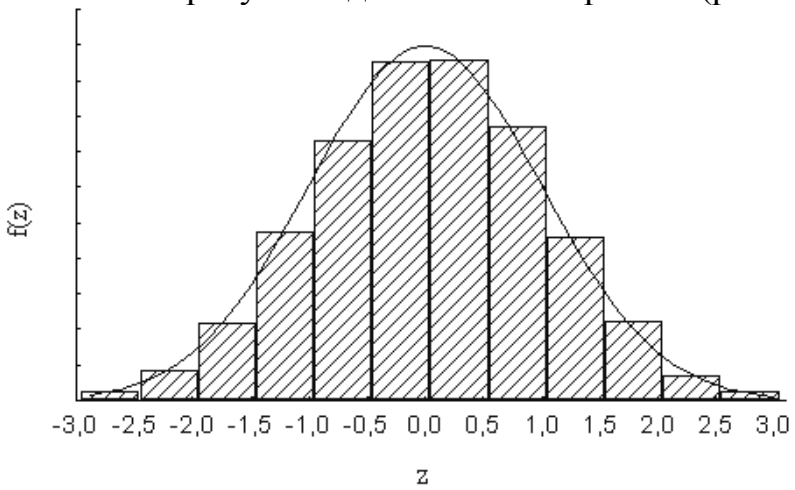

Рисунок 20.2. Гістограма вибіркового розподілу з кривої нормального розподілу

Ентропія для такої системи (кількісної ознаки) може бути розрахована за формулою:

$$H = -\sum_{i=1}^{12} f(z_i) \cdot \log_2 f(z_i), \qquad (20.6)$$

де z_i – середина кожного з 12 інтервалів, а $f(z_i)$ – функція щільності розподілу (відносна частота) для відповідного значення z_i.

Однак у такому випадку ми зіштовхуємося з одним важливим протиріччям.

За визначенням поняття ентропії, максимальне можливе значення ступеня організованості такої системи дорівнює (використовуючи формулу 20.2):

$$H\max = \log_2 12 = 3{,}585 \text{ біт.}$$

Але, з іншого боку, відповідно до формули 20.4, для ідеального нормального розподілу, щільність якого оцінена за гістограмою із 12 інтервалами, ця величина становить:

$$H\max = \log_2 \left[\frac{\sqrt{2 \cdot \pi \cdot e}}{0{,}5} \right] = 3{,}047 \text{ біта.}$$

Більше того, спостерігається інше, ще більш важливе протиріччя. Справа в тому, що будь-яка кількісна ознака має тип розподілу більш-менш близький до нормального (Гауса-Лапласа):

$$f(z) = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{z^2}{2}}. \qquad (20.7)$$

Однак графік щільності нормального розподілу має дзвоноподібну форму (рис. 20.1). Частоти низькі для вкрай малих значень, потім плавно підвищуються для величин, близьких до середнього арифметичного, і знову знижуються в області вкрай великих значень (рис. 20.2).

Таким чином, частоти різних можливих варіантів ознаки (станів системи) при нормальному типі розподілу (чи близькому до нього) розрізняються. Водночас, за визначенням, система досягає максимуму своєї невизначеності (Hmax) тільки в тому випадку, коли ці частоти однакові. Таким чином, у наявності явне протиріччя.

Як відомо, нормальний розподіл – це граничний випадок біноміального розподілу:

$$(p + q)^n, \qquad (20.8)$$

коли $p = q = 0{,}5$ і n має порядок тисяч чи десятків тисяч.

З іншого боку, формула 20.8 описує частоти генотипів для діалельної системи у випадку ко-домінування. І в цьому випадку, величина n відповідає числу пар генів, що одночасно враховуються.

Якщо, наприклад, ми розглядаємо один ген, то частоти, з якими будуть формуватися фенотипи, будуть 1:2:1. Якщо враховується два гени одночасно, частоти відповідних фенотипів будуть 1:3:3:1 і т. д.

Таким чином, ідеальний нормальний розподіл відповідає випадку популяції, що має максимально можливий рівень гетерозиготності. І тільки в

цьому випадку рівень дезорганізованості системи (її ентропія), по визначенню, досягає максимуму.

Для того щоб вирішити ці протиріччя, ми пропонуємо наступний вихід. Пропонується оцінювати ентропію не для величин щільності розподілу z-трансформованих значень вихідної вибірки, а для інтеграла цих оцінок, тобто, використовувати величини:

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \cdot \int_{-\infty}^{+\infty} e^{-\frac{z^2}{2}} dz \,. \qquad (20.9)$$

Графік інтеграла щільності нормального розподілу приведений на рисунку 20.3.

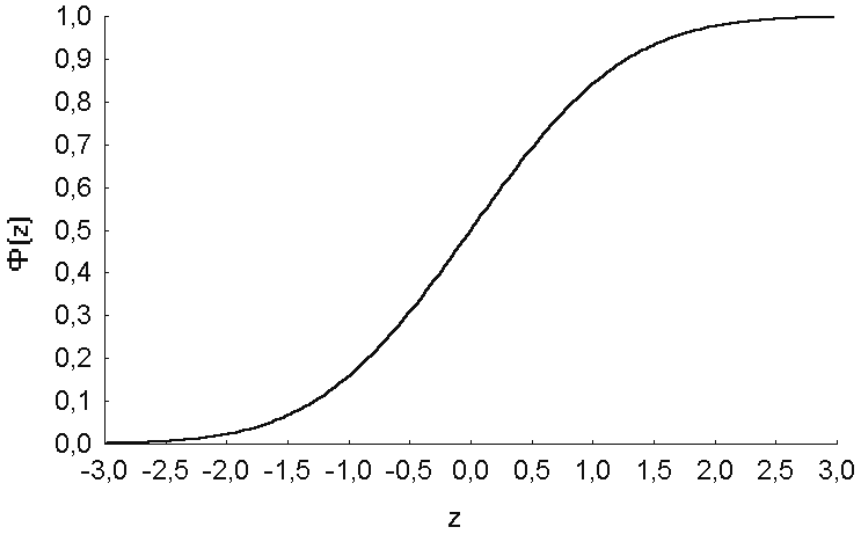

Рисунок 20.3 – Графік інтеграла щільності розподілу стандартизованих величин

Що нам дає цей підхід? По-перше, нові величини – *Ф*(z) для будь-яких ознак (що мають будь-які одиниці виміру) будуть варіювати в межах від 0 до 1. При цьому знімається питання про вибір нижньої границі першого інтервалу гістограми. Ця межа завжди дорівнює нулю.

По-друге, використання інтеграла щільності нормальної кривої приводить до її згладжування. Ця особливість інтеграла щільності нормального розподілу дуже часто використовується в прикладному статистичному аналізі, наприклад, при пробіт-аналізі. Згладжування нормальної кривої дає одну важливу перевагу, а саме – її монотонність, тобто, однакову величину збільшення частоти варіант у вибірці при збільшенні їх абсолютних значень.

Таким чином, гістограма розподілу величин *Ф*(z) буде мати наступний вигляд (рис. 20.4). Отже, значення інтеграла щільності розподілу ознаки будуть мати рівномірний розподіл. І цей розподіл буде тим більше наближатися до рівномірного, чим ближчим є вихідний емпіричний розподіл до нормального.

Як відомо, для рівномірного розподілу, зображеного у вигляді гістограми з числом інтервалів, рівним *k*, ентропія дорівнює значенню, приведеному у формулі 20.2.

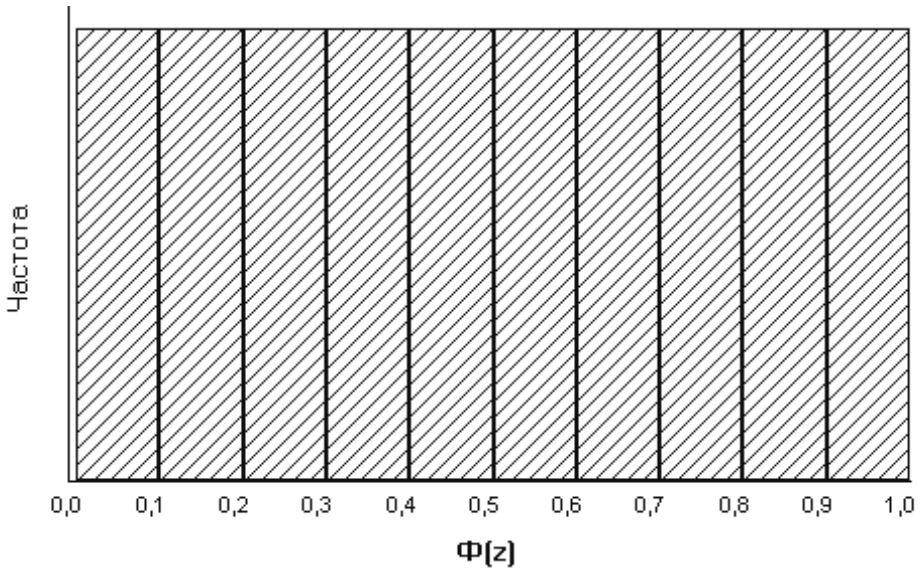

Рисунок 20.4 – Гістограма розподілу величин інтеграла щільності нормального розподілу

А це значить: чим ближче розподіл вихідної ознаки до нормального, тих ближче розподіл інтеграла його щільності до рівномірного і, отже, ентропія такої системи буде прагнути до свого максимуму. І, навпаки, чим сильніше емпіричний розподіл вихідної ознаки відхиляється від нормального, тим сильніше буде відхилятися від рівномірного розподіл інтеграла його щільності і, відповідно, тим нижчим буде значення ентропії цієї системи.

У крайньому випадку, коли усі варіанти у вибірці (чи популяції) будуть рівні, ентропія такої системи, як і випливає по визначенню, буде дорівнювати нулю.

Число інтервалів, на яке можна розбити відрізок [0;1] для інтеграла щільності розподілу (тобто, k), залежить від обсягу вибірки.

Ми можемо запропонувати таке оптимальне число інтервалів, при якому середня частота потрапляння величини в кожний із них не буде меншою 5-10.

Таким чином, для вибірок обсягом 100-200 об’єктів (особин) оптимальним буде 10 інтервалів. У цьому випадку, максимальне значення ентропії такої системи буде дорівнювати $H\max = \log_2 10 = 3{,}322$ біти.

При більшому обсязі наявних вибіркових даних, число інтервалів може бути збільшене. При меншому обсязі, навпаки, зменшене (наприклад, до 5).

Оскільки, оцінка ентропії проводиться на підставі випадкової вибірки, то ця оцінка має свою статистичну помилку, що залежить, насамперед, від обсягу вибірки (n):

$$SE_H = \sqrt{\frac{\sum_{i=1}^{k}\left[p_i \cdot \left(\log_2 p_i\right)^2\right] - H^2}{2 \cdot n}}. \qquad (20.10)$$

Крім безпосередніх оцінок ентропії, можуть бути також використані показники, похідні від них.

Для визначення міри **абсолютної організованості** системи використовується величина:

$$O = Hmax - H. \tag{20.11}$$

Величину **відносної організованості** системи оцінюють за формулою:

$$R = 1 - \frac{H}{H\max}. \tag{20.12}$$

Відповідно до класифікації С. Біра, система для якої $R \le 0{,}1$ є імовірнісною (стохастичною); якщо значення відносної організованості системи $R > 0{,}3$, то така система вважається детермінованою. І, нарешті, система, для якої $0{,}1 < R \le 0{,}3$, є квазідетермінованою (імовірнісно-детермінованою).

З іншого боку, при аналізі численних матеріалів дуже часто виникає завдання оцінити вірогідність різниці отриманих оцінок ентропії у двох чи більше вибірках із наступним обчисленням рівня значущості отриманих різниць. Для перевірки нульової гіпотези пропонується використовувати наступний критерій:

$$t = \frac{\left|H_1 - H_2\right|}{\sqrt{Var(H_1) + Var(H_2)}}, \tag{20.13}$$

де H_1 і H_2 – вибіркові оцінки ентропії в двох порівнюваних сукупностях, а $Var(H_1)$ і $Var(H_2)$ – їх варіанси.

Оцінку варіанси ентропії для відповідної вибірки можна одержати за наступною формулою:

$$Var(H) = \frac{\sum_{i=1}^{s} p_i \cdot \log_2^2 p_i - H^2}{n} - \frac{s-1}{2 \cdot n^2}, \tag{20.14}$$

де n – обсяг вибірки;

s – кількість альтернативних станів чи системи типів елементів у вибірці.

Показано, що критерій (20.13) може бути апроксимований розподілом t-критерію Ст’юдента із числом ступенів свободи:

$$df = \frac{\left[Var(H_1) + Var(H_2)\right]^2}{\frac{Var(H_1)^2}{n_1} + \frac{Var(H_2)^2}{n_2}}, \tag{20.15}$$

де n_1 і n_2 – обсяги порівнюваних вибірок.

Контрольні питання:

1. Поняття про ентропію.
2. Методика визначення абсолютної та відносної організованості системи.
3. Переваги оцінювання ентропії не для величин щільності розподілу z-трансформованих значень вихідної вибірки, а для інтеграла цих оцінок.

ПРЕДМЕТНИЙ ПОКАЖЧИК

СПИСОК ВИКОРИСТАНОЇ ТА РЕКОМЕНДОВАНОЇ ЛІТЕРАТУРИ

- *Алтухов Ю. П.* Генетические процессы в популяциях. – М.: Наука, 1989. – 328 с.
- *Бейли Н.* Статистические методы в биологии. – М.: Изд-во иностранной литературы, 1962. – 260 с.
- *Гиль М. І., Крамаренко С. С.* Генетико-математичне моделювання кількісних ознак в тваринництві: огляд // Збірник наукових праць Сумського НАУ, серія «Тваринництво». – Суми, 2008. – Вип. 6 (2). – С. 49-56.
- *Животовский Л. А.* Показатели популяционной изменчивости по полиморфным признакам // Фенетика популяций. – М.: Наука, 1982. – С. 38-44.
- *Животовский Л. А.* Популяционная биометрия. – М.: Наука, 1991. – 209 с.
- *Компьютерная биометрика* / Под ред. В.Н.Носова. – М.: Изд-во МГУ, 1990. – 232 с.
- *Крамаренко С. С.* Метод использования энтропийно-информационного анализа для количественных признаков // Известия Самарского НЦ РАН. – Самара, 2005. – Т.7, № 1. – С. 242-247.
- *Крамаренко С. С.* Нові методи математичного моделювання лактаційних кривих за допомогою інтерполяції // Новітні технології і скотарстві у XX столітті : матеріали Міжнародної науково-практичної конференції (Миколаїв, 4-6 вересня 2008 р.). – Миколаїв, 2008. – С. 159-164.
- *Лакин Г. Ф.* Биометрия: Издание 2-е. – М.: Высшая школа, 1973. – 343 с.
- *Лакин Г. Ф.* Биометрия: Издание 3-е, перер. и доп. – М.: Высш. школа, 1980. – 293 с.
- *Ли Ч.* Введение в популяционную генетику. – М.: Мир, 1978. – 557 с.
- *Меркурьева Е. К.* Биометрия в селекции и генетике сельскохозяйственных животных. – М.: Колос, 1970. – 424 с.
- *Ней М., Кумар С.* Молекулярная эволюция и филогенетика. – К.: КВІЦ, 2004. – 418 с.
- *Панов Е. Н., Грабовский В. И., Любущенко С. В.* Дивергенция и гибридогенный полиморфизм в комплексе «Черные каменки» *Oenanthe picata* // Зоологический журнал. – 1993. – Т. 72, № 8. – С. 80-96.
- *Патрєва Л. С., Крамаренко С. С.* Ентропійний аналіз кількісних ознак для селекційної оцінки батьківського стада м'ясних курей // Розведення і генетика тварин. – К.: Аграрна наука, 2007. – № 41. – С. 149-153.
- *Плохинский Н. А.* Биометрия. – Новосибирск: Изд-во СО АН СССР, 1970. – 364 с.
- *Плохинский Н. А.* Наследуемость. – Новосибирск: Редакционно-издательский отдел СО АН СССР, 1964. – 196 с.
- *Плохинский Н. А.* Руководство по биометрии для зоотехников. – М.: Колос, 1969. – 256 с.

- *Справочник* по прикладной статистике / Под ред. Э. Ллойда, У. Ледермана: В 2-х т. – М.: Финансы и статистика, 1989, 1990. – 526 с.
- *Терентьев П. В., Ростова Н. С.* Практикум по биометрии. – Л.: Изд-во ЛГУ, 1977. – 152 с.
- *Урбах В. Ю.* Математическая статистика для биологов и медиков. – М.: Изд-во АН СССР, 1964. – 323 с.
- *Урбах В. Ю.* Статистический анализ в биологических и медицинских исследованиях. – М.: Медицина, 1975. – 296 с.
- *Шебаніна О. В., Крамаренко С. С., Ганганов В. М.* Практикум з біометрії: Методи непараметричної статистики. – Миколаїв: МДАУ, 2008. – 166 с.
- *Шеффе Г.* Дисперсионный анализ. – М.: Физматгиз, 1963. – 628 с.
- *Яблоков А. В*. Фенетика: Эволюция, популяция, признак. – М.: Наука, 1980. – 132 с.
- *Яблоков А. В*. Популяционная биология. – М.: Высшая школа, 1987. – 303 с.
- *Яблоков А. В., Ларина Н. И.* Введение в фенетику популяций. – М.: Высшая школа, 1985. – 159 с.
- *Excoffier L., Smouse P. E., Quattro J. M.* Analysis of molecular variance inferred metric distances among DNA haplotypes: Application to human mitochondrial DNA restriction sites // Genetics. – 1992. – V. 131. – P. 479-491.
- *Nei M.* Analysis of gene diversity in subdivided populations // PNAS USA. – 1973. – V. 70. – P. 3321-3323.
- *Nei M.* Estimation of average heterozygosity and genetic distance from a small number individuals // Genetics. – 1978. – V. 89. – P. 583-590.
- *Peakall R., Smouse P. E.* GenAIEx 6: Genetic analysis in Excel. Population genetic software for teaching and research // Molecular Ecology Notes. – 2006. – V. 6. – P. 288-295.
- *Sawada M.* ROOKCASE: An Excel 97/2000 Visual Basic (VB) Add-In for exploring global and local spatial autocorrelation // Bull.Ecol.Soc.Am. – 1999. – V. 80(4). – P. 231-234.
- *Weir B. S., Cockerham C. C.* Estimating F-statistics for the analysis of population structure // Evolution. – 1984. – V. 38. – P. 1358-1370.
- *Wright S.* The genetical structure of populations // Ann. Eugenics. – 1951. – V. 15. – P. 323-354.
- *Yang R.C.* Estimating hierarchical F-statistics // Evolution. – 1998. – V. 52. – P. 950-956.

ДОДАТКИ

ДОДАТОК А

Критичні значення для альтернативи, що має меншу вірогідність зустрітися

n	α		n	α		n	α		n	α	
	0,05	0,01		0,05	0,01		0,05	0,01		0,05	0,01
8	1	1	**31**	10	8	**54**	20	18	**77**	30	27
9	2	1	**32**	10	9	**55**	20	18	**78**	30	28
10	2	1	**33**	11	9	**56**	21	18	**79**	31	28
11	2	1	**34**	11	10	**57**	21	19	**80**	31	29
12	3	2	**35**	12	10	**58**	22	19	**81**	32	29
13	3	2	**36**	12	10	**59**	22	20	**82**	32	29
14	3	2	**37**	13	11	**60**	22	20	**83**	33	30
15	4	3	**38**	13	11	**61**	23	21	**84**	33	30
16	4	3	**39**	13	12	**62**	23	21	**85**	33	31
17	5	3	**40**	14	12	**63**	24	21	**86**	34	31
18	5	4	**41**	14	12	**64**	24	22	**87**	34	32
19	5	4	**42**	15	13	**65**	25	22	**88**	35	32
20	6	4	**43**	15	13	**66**	25	23	**89**	35	32
21	6	5	**44**	16	14	**67**	26	23	**90**	36	33
22	6	5	**45**	16	14	**68**	26	23	**91**	36	33
23	7	5	**46**	16	14	**69**	26	24	**92**	37	34
24	7	6	**47**	17	15	**70**	27	24	**93**	37	34
25	8	6	**48**	17	15	**71**	27	25	**94**	38	35
26	8	7	**49**	18	16	**72**	28	25	**95**	38	35
27	8	7	**50**	18	16	**73**	28	26	**96**	38	35
28	9	7	**51**	19	16	**74**	29	26	**97**	39	36
29	9	8	**52**	19	17	**75**	29	26	**98**	39	36
30	10	8	**53**	19	17	**76**	29	27	**99**	40	37
									100	40	37

ДОДАТОК Б

Біноміальні коефіцієнти

n	Коефіцієнти										
1	1										
2	1	1									
3	1	2	1								
4	1	4	6	4	1						
5	1	5	10	10	5	1					
6	1	6	15	20	15	6	1				
7	1	7	21	35	35	21	7	1			
8	1	8	28	56	70	56	28	8	1		
9	1	9	36	84	126	126	84	36	9	1	
10	1	10	45	120	210	252	210	120	45	10	1

ДОДАТОК В

Довірчі межі для параметра Пуассона

X	$x_{ниж}$	$x_{верх}$	X	$x_{ниж}$	$x_{верх}$	X	$x_{ниж}$	$x_{верх}$	X	$x_{ниж}$	$x_{верх}$
1	*0,03*	*5,57*	**21**	*13,00*	*32,10*	**41**	*29,42*	*55,62*	**61**	*46,65*	*78,38*
2	*0,24*	*7,22*	**22**	*13,79*	*33,31*	**42**	*30,27*	*56,77*	**62**	*47,52*	*79,51*
3	*0,62*	*8,77*	**23**	*14,58*	*34,51*	**43**	*31,12*	*57,92*	**63**	*48,40*	*80,63*
4	*1,09*	*10,24*	**24**	*15,38*	*35,71*	**44**	*31,97*	*59,07*	**64**	*49,27*	*81,76*
5	*1,62*	*11,67*	**25**	*16,18*	*36,90*	**45**	*32,82*	*60,21*	**65**	*50,15*	*82,88*
6	*2,20*	*13,06*	**26**	*16,98*	*38,10*	**46**	*33,68*	*61,36*	**66**	*51,03*	*84,00*
7	*2,81*	*14,42*	**27**	*17,79*	*39,28*	**47**	*34,53*	*62,50*	**67**	*51,91*	*85,12*
8	*3,45*	*15,76*	**28**	*18,61*	*40,47*	**48**	*35,39*	*63,64*	**68**	*52,79*	*86,24*
9	*4,12*	*17,08*	**29**	*19,42*	*41,65*	**49**	*36,25*	*64,78*	**69**	*53,67*	*87,36*
10	*4,80*	*18,39*	**30**	*20,24*	*42,83*	**50**	*37,11*	*65,92*	**70**	*54,53*	*88,47*
11	*5,49*	*19,68*	**31**	*21,06*	*44,00*	**51**	*37,96*	*67,08*	**71**	*55,44*	*89,59*
12	*6,20*	*20,96*	**32**	*21,89*	*45,17*	**52**	*38,83*	*68,21*	**72**	*56,32*	*90,71*
13	*6,92*	*22,23*	**33**	*22,72*	*46,34*	**53**	*39,69*	*69,35*	**73**	*57,20*	*91,82*
14	*7,65*	*23,49*	**34**	*23,55*	*47,51*	**54**	*40,56*	*70,48*	**74**	*58,09*	*92,94*
15	*8,40*	*24,74*	**35**	*24,38*	*48,68*	**55**	*41,42*	*71,61*	**75**	*58,97*	*94,05*
16	*9,15*	*25,98*	**36**	*25,21*	*49,84*	**56**	*42,29*	*72,75*	**76**	*59,86*	*95,16*
17	*9,90*	*27,22*	**37**	*26,05*	*51,00*	**57**	*43,16*	*73,88*	**77**	*60,75*	*96,27*
18	*10,67*	*28,45*	**38**	*26,89*	*52,16*	**58**	*44,03*	*75,00*	**78**	*61,64*	*97,39*
19	*11,44*	*29,67*	**39**	*27,73*	*53,31*	**59**	*44,95*	*76,13*	**79**	*62,53*	*98,50*
20	*12,22*	*30,89*	**40**	*28,58*	*54,47*	**60**	*45,77*	*77,26*	**80**	*63,42*	*99,61*

ДОДАТОК Д

Значення критерію Хі-квадрат К.Пірсона

df	1	2	3	4	5	6	7	8	9	10
χ^2	3,84	5,99	7,82	9,49	11,07	12,59	14,07	15,51	16,92	18,31
df	11	12	13	14	15	16	17	18	19	20
χ^2	19,68	21,03	22,36	23,69	25,00	26,30	27,59	28,87	30,14	31,41
df	21	22	23	24	25	26	27	28	29	30
χ^2	32,67	33,92	35,17	36,42	37,65	38,89	40,11	41,34	42,56	43,77

ДОДАТОК Е

φ-перетворення Р. Фішера

	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
0,00	0,000	0,063	0,089	0,110	0,127	0,142	0,155	0,168	0,179	0,190
0,0	0,000	0,200	0,284	0,348	0,403	0,451	0,495	0,536	0,574	0,609
0,1	0,644	0,676	0,707	0,738	0,767	0,795	0,823	0,850	0,876	0,902
0,2	0,927	0,953	0,976	1,000	1,024	1,047	1,070	1,093	1,115	1,137
0,3	1,159	1,181	1,203	1,224	1,245	1,266	1,287	1,308	1,328	1,349
0,4	1,369	1,390	1,410	1,430	1,451	1,471	1,491	1,511	1,531	1,551
0,5	1,571	1,591	1,611	1,631	1,651	1,671	1,691	1,711	1,731	1,752
0,6	1,772	1,793	1,813	1,834	1,855	1,875	1,897	1,918	1,939	1,961
0,7	1,982	2,004	2,026	2,049	2,071	2,094	2,118	2,141	2,165	2,190
0,8	2,214	2,240	2,265	2,292	2,319	2,346	2,375	2,404	2,434	2,465
0,9	2,498	2,532	2,568	2,606	2,647	2,691	2,739	2,793	2,858	2,941
0,99	2,941	2,952	2,963	2,974	2,987	3,000	3,015	3,032	3,052	3,078

ДОДАТОК Ж

Критичні значення F-критерію Фішера-Снедекора (для $\alpha = 0,05$)

df_2	df_1									
	1	2	3	4	5	6	7	8	9	10
10	4,965	4,103	3,708	3,478	3,326	3,217	3,135	3,072	3,020	2,978
15	4,543	3,682	3,287	3,056	2,901	2,790	2,707	2,641	2,588	2,544
20	4,351	3,493	3,098	2,866	2,711	2,599	2,514	2,447	2,393	2,348
25	4,242	3,385	2,991	2,759	2,603	2,490	2,405	2,337	2,282	2,236
30	4,171	3,316	2,922	2,690	2,534	2,421	2,334	2,266	2,211	2,165
35	4,121	3,267	2,874	2,641	2,485	2,372	2,285	2,217	2,161	2,114
40	4,085	3,232	2,839	2,606	2,449	2,336	2,249	2,180	2,124	2,077
45	4,057	3,204	2,812	2,579	2,422	2,308	2,221	2,152	2,096	2,049
50	4,034	3,183	2,790	2,557	2,400	2,286	2,199	2,130	2,073	2,026
55	4,016	3,165	2,773	2,540	2,383	2,269	2,181	2,112	2,055	2,008
60	4,001	3,150	2,758	2,525	2,368	2,254	2,167	2,097	2,040	1,993
65	3,989	3,138	2,746	2,513	2,356	2,242	2,154	2,084	2,027	1,980
70	3,978	3,128	2,736	2,503	2,346	2,231	2,143	2,074	2,017	1,969
80	3,960	3,111	2,719	2,486	2,329	2,214	2,126	2,056	1,999	1,951
90	3,947	3,098	2,706	2,473	2,316	2,201	2,113	2,043	1,986	1,938
100	3,936	3,087	2,696	2,463	2,305	2,191	2,103	2,032	1,975	1,927
120	3,920	3,072	2,680	2,447	2,290	2,175	2,087	2,016	1,959	1,910
140	3,909	3,061	2,669	2,436	2,279	2,164	2,076	2,005	1,947	1,899
160	3,900	3,053	2,661	2,428	2,271	2,156	2,067	1,997	1,939	1,890
180	3,894	3,046	2,655	2,422	2,264	2,149	2,061	1,990	1,932	1,884
200	3,888	3,041	2,650	2,417	2,259	2,144	2,056	1,985	1,927	1,878
250	3,879	3,032	2,641	2,408	2,250	2,135	2,046	1,976	1,917	1,869
300	3,873	3,026	2,635	2,402	2,244	2,129	2,040	1,969	1,911	1,862
350	3,868	3,022	2,630	2,397	2,240	2,125	2,036	1,965	1,907	1,858
400	3,865	3,018	2,627	2,394	2,237	2,121	2,032	1,962	1,903	1,854
450	3,862	3,016	2,625	2,392	2,234	2,119	2,030	1,959	1,901	1,852
500	3,860	3,014	2,623	2,390	2,232	2,117	2,028	1,957	1,899	1,850
600	3,857	3,011	2,620	2,387	2,229	2,114	2,025	1,954	1,895	1,846
700	3,855	3,009	2,618	2,385	2,227	2,112	2,023	1,952	1,893	1,844
800	3,853	3,007	2,616	2,383	2,225	2,110	2,021	1,950	1,892	1,843
900	3,852	3,006	2,615	2,382	2,224	2,109	2,020	1,949	1,890	1,841
1000	3,851	3,005	2,614	2,381	2,223	2,108	2,019	1,948	1,889	1,840

Продовження додатку Ж

df_2	df_1									
	11	12	13	14	15	16	17	18	19	20
10	2,943	2,913	2,887	2,865	2,845	2,828	2,812	2,798	2,785	2,774
15	2,507	2,475	2,448	2,424	2,403	2,385	2,368	2,353	2,340	2,328
20	2,310	2,278	2,250	2,225	2,203	2,184	2,167	2,151	2,137	2,124
25	2,198	2,165	2,136	2,111	2,089	2,069	2,051	2,035	2,021	2,007
30	2,126	2,092	2,063	2,037	2,015	1,995	1,976	1,960	1,945	1,932
35	2,075	2,041	2,012	1,986	1,963	1,942	1,924	1,907	1,892	1,878
40	2,038	2,003	1,974	1,948	1,924	1,904	1,885	1,868	1,853	1,839
45	2,009	1,974	1,945	1,918	1,895	1,874	1,855	1,838	1,823	1,808
50	1,986	1,952	1,921	1,895	1,871	1,850	1,831	1,814	1,798	1,784
55	1,968	1,933	1,903	1,876	1,852	1,831	1,812	1,795	1,779	1,764
60	1,952	1,917	1,887	1,860	1,836	1,815	1,796	1,778	1,763	1,748
65	1,939	1,904	1,874	1,847	1,823	1,802	1,782	1,765	1,749	1,734
70	1,928	1,893	1,863	1,836	1,812	1,790	1,771	1,753	1,737	1,722
80	1,910	1,875	1,845	1,817	1,793	1,772	1,752	1,734	1,718	1,703
90	1,897	1,861	1,830	1,803	1,779	1,757	1,737	1,720	1,703	1,688
100	1,886	1,850	1,819	1,792	1,768	1,746	1,726	1,708	1,691	1,676
120	1,869	1,834	1,803	1,775	1,750	1,728	1,709	1,690	1,674	1,659
140	1,858	1,822	1,791	1,763	1,738	1,716	1,696	1,678	1,661	1,646
160	1,849	1,813	1,782	1,754	1,729	1,707	1,687	1,669	1,652	1,637
180	1,842	1,806	1,775	1,747	1,722	1,700	1,680	1,661	1,645	1,629
200	1,837	1,801	1,769	1,742	1,717	1,694	1,674	1,656	1,639	1,623
250	1,827	1,791	1,759	1,732	1,707	1,684	1,664	1,645	1,628	1,613
300	1,821	1,785	1,753	1,725	1,700	1,677	1,657	1,638	1,621	1,606
350	1,816	1,780	1,748	1,720	1,695	1,672	1,652	1,633	1,616	1,601
400	1,813	1,776	1,745	1,717	1,691	1,669	1,648	1,630	1,613	1,597
450	1,810	1,774	1,742	1,714	1,689	1,666	1,646	1,627	1,610	1,594
500	1,808	1,772	1,740	1,712	1,686	1,664	1,643	1,625	1,607	1,592
600	1,805	1,768	1,736	1,708	1,683	1,660	1,640	1,621	1,604	1,588
700	1,802	1,766	1,734	1,706	1,681	1,658	1,637	1,619	1,601	1,586
800	1,801	1,764	1,732	1,704	1,679	1,656	1,636	1,617	1,600	1,584
900	1,799	1,763	1,731	1,703	1,678	1,655	1,634	1,615	1,598	1,582
1000	1,798	1,762	1,730	1,702	1,676	1,654	1,633	1,614	1,597	1,581

Продовження додатку Ж

df_2	df_1									
	30	40	50	60	70	80	90	100	150	200
30	1,841	1,792	1,761	1,740	1,724	1,712	1,703	1,695	1,672	1,660
40	1,744	1,693	1,660	1,637	1,621	1,608	1,597	1,589	1,564	1,551
40	1,744	1,693	1,660	1,637	1,621	1,608	1,597	1,589	1,564	1,551
60	1,649	1,594	1,559	1,534	1,516	1,502	1,491	1,481	1,453	1,438
70	1,622	1,566	1,530	1,505	1,486	1,471	1,459	1,450	1,420	1,404
80	1,602	1,545	1,508	1,482	1,463	1,448	1,436	1,426	1,395	1,379
90	1,586	1,528	1,491	1,465	1,445	1,429	1,417	1,407	1,375	1,358
100	1,573	1,515	1,477	1,450	1,430	1,415	1,402	1,392	1,359	1,342
150	1,535	1,475	1,436	1,407	1,386	1,369	1,356	1,345	1,309	1,290
200	1,516	1,455	1,415	1,386	1,364	1,346	1,332	1,321	1,283	1,263

ДОДАТОК И

Критичні точки розподілу Ст'юдента (для $\alpha = 0{,}05$)

df	*t*	*df*	*t*	*df*	*t*
1	12,71	**16**	2,12	**40**	2,02
2	4,30	**17**	2,11	**50**	2,01
3	3,18	**18**	2,10	**60**	2,00
4	2,78	**19**	2,09	**70**	1,99
5	2,57	**20**	2,09	**80**	1,99
6	2,45	**21**	2,08	**90**	1,99
7	2,36	**22**	2,07	**100**	1,98
8	2,31	**23**	2,07	**110**	1,98
9	2,26	**24**	2,06	**120**	1,98
10	2,23	**25**	2,06	**130**	1,98
11	2,20	**26**	2,06	**140**	1,98
12	2,18	**27**	2,05	**150**	1,98
13	2,16	**28**	2,05	**200**	1,97
14	2,14	**29**	2,05	**500**	1,96
15	2,13	**30**	2,04	∞	1,96

ДОДАТОК К

Ординати нормальної кривої

t	Соті частки *t*									
	0	1	2	3	4	5	6	7	8	9
0,0	3989	3989	3989	3988	3986	3984	3982	3980	3977	3973
0,1	3970	3965	3961	3956	3951	3945	3939	3932	3925	3918
0,2	3910	3902	3894	3885	3876	3867	3857	3847	3836	3825
0,3	3814	3802	3790	3778	3765	3752	3739	3726	3712	3697
0,4	3683	3668	3653	3637	3621	3605	3589	3572	3555	3538
0,5	3521	3503	3485	3467	3448	3429	3410	3391	3372	3352
0,6	3332	3312	3292	3271	3251	3230	3209	3187	3166	3144
0,7	3123	3101	3079	3056	3034	3011	2989	2966	2943	2920
0,8	2897	2874	2850	2827	2803	2780	2756	2732	2709	2685
0,9	2661	2637	2613	2589	2565	2541	2516	2492	2468	2444
1,0	2420	2396	2371	2347	2323	2299	2275	2251	2227	2203
1,1	2179	2155	2131	2107	2083	2059	2036	2012	1989	1965
1,2	1942	1919	1895	1872	1849	1826	1804	1781	1758	1736
1,3	1714	1691	1669	1647	1626	1604	1582	1561	1539	1518
1,4	1497	1476	1456	1435	1415	1394	1374	1354	1334	1315
1,5	1295	1276	1257	1238	1219	1200	1182	1163	1145	1127
1,6	1109	1092	1074	1057	1040	1023	1006	0989	0973	0957
1,7	0940	0925	0909	0893	0878	0863	0848	0833	0818	0804
1,8	0790	0775	0761	0748	0734	0721	0707	0694	0681	0669
1,9	0656	0644	0632	0620	0608	0596	0584	0573	0562	0551
2,0	0540	0529	0519	0508	0498	0488	0478	0468	0459	0449
2,1	0440	0431	0422	0413	0404	0396	0387	0379	0371	0363
2,2	0355	0347	0339	0332	0325	0317	0310	0303	0297	0290
2,3	0283	0277	0270	0264	0258	0252	0246	0241	0235	0229
2,4	0224	0219	0213	0208	0203	0198	0194	0189	0184	0180
2,5	0175	0171	0167	0163	0158	0154	0151	0147	0143	0139
2,6	0136	0132	0129	0126	0122	0119	0116	0113	0110	0107
2,7	0104	0101	0099	0096	0093	0091	0088	0086	0084	0081
2,8	0079	0077	0075	0073	0071	0069	0067	0065	0063	0061
2,9	0060	0058	0056	0055	0053	0051	0050	0048	0047	0046
3,0	0044	0043	0042	0041	0039	0038	0037	0036	0035	0034
3,1	0033	0032	0031	0030	0029	0028	0027	0026	0025	0025
3,2	0024	0023	0022	0022	0021	0020	0020	0019	0018	0018
3,3	0017	0017	0016	0016	0015	0015	0014	0014	0013	0013
3,4	0012	0012	0012	0011	0011	0010	0010	0010	0009	0009
3,5	0009	0008	0008	0008	0008	0007	0007	0007	0007	0006

Примітка: В таблиці нуль цілих та кому не вказано

ДОДАТОК Л

Інтеграл імовірності нормальної кривої

t	Соті частки *t*									
	0	1	2	3	4	5	6	7	8	9
0,0	5000	5040	5080	5120	5160	5199	5239	5279	5319	5359
0,1	5398	5438	5478	5517	5557	5596	5636	5675	5714	5753
0,2	5793	5832	5871	5910	5948	5987	6026	6064	6103	6141
0,3	6179	6217	6255	6293	6331	6368	6406	6442	6480	6517
0,4	6554	6591	6628	6664	6700	6736	6772	6808	6844	6879
0,5	6915	6950	6985	7019	7054	7088	7123	7157	7190	7224
0,6	7257	7291	7324	7357	7389	7422	7454	7486	7517	7549
0,7	7580	7611	7642	7673	7703	7734	7764	7794	7823	7852
0,8	7881	7910	7939	7967	7995	8023	8051	8078	8106	8133
0,9	8159	8186	8212	8238	8264	8289	8315	8340	8365	8389
1,0	8413	8438	8461	8485	8508	8531	8554	8577	8599	8621
1,1	8643	8665	8686	8708	8729	8749	8770	8790	8810	8830
1,2	8849	8869	8888	8907	8925	8944	8962	8980	8997	9015
1,3	9032	9049	9066	9082	9099	9115	9131	9147	9162	9177
1,4	9192	9207	9222	9236	9251	9265	9278	9292	9306	9319
1,5	9332	9345	9357	9370	9382	9394	9406	9118	9430	9441
1,6	9452	9463	9474	9484	9495	9505	9515	9525	9535	9545
1,7	9554	9564	9573	9582	9591	9599	9608	9616	9625	9633
1,8	9641	9648	9656	9664	9671	9778	9686	9693	9700	9706
1,9	9713	9719	9726	9732	9738	9744	9750	9756	9762	9767
2,0	9772	9778	9783	9788	9793	9798	9803	9808	9812	9817
2,1	9821	9826	9830	9834	9838	9842	9846	9850	9854	9857
2,2	9861	9865	9868	9871	9875	9878	9881	9884	9887	9890
2,3	9893	9896	9898	9901	9904	9906	9909	9911	9913	9916
2,4	9918	9920	9922	9925	9927	9929	9931	9932	9934	9936
2,5	9938	9940	9941	9943	9945	9946	9948	9949	9951	9952
2,6	9953	9955	9956	9957	9959	9960	9961	9962	9963	9964
2,7	9965	9966	9967	9968	9969	9970	9971	9972	9973	9974
2,8	9974	9975	9976	9977	9977	9978	9979	9979	9980	9981
2,9	9981	9982	9982	9983	9984	9984	9985	9985	9986	9986
3,0	9987	9987	9987	9988	9988	9989	9989	9989	9990	9990
3,1	9990	9991	9991	9991	9992	9992	9992	9992	9993	9993
3,2	9993	9993	9994	9994	9994	9994	9994	9995	9995	9995
3,3	9995	9995	9995	9996	9996	9996	9996	9996	9996	9997
3,4	9997	9997	9997	9997	9997	9997	9997	9997	9997	9998
3,5	9998	9998	9998	9998	9998	9998	9998	9998	9998	9998

Примітка: В таблиці нуль цілих та кому не вказано

ДОДАТОК М

Критичні значення критерію Кохрена (для $\alpha = 0{,}05$)

df	Кількість вибірок (*k*)							
	3	**4**	**5**	**6**	**7**	**8**	**9**	**10**
5	0,707	0,590	0,506	0,445	0,397	0,360	0,329	0,303
10	0,617	0,502	0,424	0,368	0,326	0,293	0,266	0,244
12	0,580	0,466	0,392	0,339	0,298	0,267	0,244	0,224
14	0,560	0,450	0,377	0,325	0,285	0,256	0,232	0,213
16	0,547	0,437	0,364	0,313	0,277	0,246	0,223	0,203
18	0,536	0,425	0,353	0,305	0,267	0,239	0,217	0,199
20	0,526	0,416	0,335	0,298	0,259	0,230	0,210	0,191
25	0,504	0,397	0,329	0,282	0,245	0,219	0,198	0,180
30	0,490	0,383	0,315	0,270	0,233	0,209	0,190	0,173
35	0,479	0,372	0,306	0,262	0,227	0,201	0,183	0,167
40	0,469	0,364	0,298	0,255	0,220	0,195	0,177	0,162
45	0,460	0,357	0,291	0,250	0,215	0,190	0,173	0,157
50	0,455	0,350	0,288	0,246	0,210	0,186	0,170	0,152
60	0,444	0,341	0,279	0,239	0,204	0,179	0,161	0,148
80	0,429	0,329	0,268	0,227	0,195	0,170	0,154	0,140
100	0,417	0,319	0,261	0,220	0,189	0,165	0,150	0,136
150	0,401	0,305	0,250	0,211	0,185	0,160	0,145	0,130
200	0,390	0,299	0,246	0,210	0,180	0,158	0,142	0,130

Навчальне видання

Крамаренко Сергій Сергійович
Луговий Сергій Іванович
Лихач Анна Василівна
Крамаренко Олександр Сергійович

АНАЛІЗ БІОМЕТРИЧНИХ ДАНИХ У РОЗВЕДЕННІ ТА СЕЛЕКЦІЇ ТВАРИН

Навчальний посібник

Технічний редактор: С. С. Крамаренко

Підписано до друку 16.10.2018 р. Формат 60×84/16. Папір офсетн.
Гарнітура Times New Roman.
Друк офс. Умовн. друк. арк. 8,8. Облік видавн. арк. 8,8
Умов. фарбовід. 0,9. Зам. № 537, тир. 100.

Надруковано у видавничому відділі
Миколаївського національного аграрного університету
54020, м. Миколаїв, вул. Георгія Гонгадзе, 9

Свідоцтво суб'єкта видавничої справи ДК № 4490 від 20.02.2013 р.