МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«СЕВЕРО-КАВКАЗСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ»

Пятигорский институт (филиал) СКФУ

Методические указания

по выполнению практических работ
по дисциплине «ДОПОЛНИТЕЛЬНЫЕ ГЛАВЫ МАТЕМАТИКИ»
для студентов направления подготовки 13.03.02 Электроэнергетика и электротехника
Передача и распределение электрической энергии в системах электроснабжения

(ЭЛЕКТРОННЫЙ ДОКУМЕНТ)

Содержание

	Стр.
Практическое занятие 1. Абсолютная и относительная погрешности измерений	5
Практическое занятие 2. Расчет погрешностей непосредственных измерений	7
Практическое занятие 3. Расчет погрешностей косвенных измерений	8
Практическое занятие 4. Вариационные ряды и их графическое изображение	10
Практическое занятие 5. Числовые характеристики выборки, свойства числовых характеристик.	13
Практическое занятие 6. Выборочный метод. Статистическое оценивание.	15
Практическое занятие 7. Понятие интервального оценивания. Доверительная вероятность и предельная ошибка выборки. Оценка характеристик генеральной совокупности по малой выборке	17
Практическое занятие 8-9. Статистическая проверка гипотез. Параметрические критерии	21
Практическое занятие 10-11. Статистическая проверка гипотез. Непараметрические критерии	26
Практическое занятие 12. Проверка гипотез о законах распределения	28
Практическое занятие 13. Корреляционная зависимость. Коэффициент корреляции	31
Практическое занятие 14. Линейная регрессия. Коэффициенты регрессии	33
Практическое занятие 15-16. Однофакторный дисперсионный анализ	35
Практическое занятие 17-18. Двухфакторный дисперсионный анализ	39

1.ЦЕЛЬ И СОДЕРЖАНИЕ

Целью освоения дисциплины «Дополнительные главы математики» является формирование набора универсальных компетенций бакалавра по направлению подготовки 13.03.02 «Электроэнергетика и электротехника».

Задачи освоения дисциплины: формирование представлений о роли и месте математики в современном мире, этапах развития, универсальности ее понятий и представлений; формирование умений конструирования и анализа статистических моделей объектов, систем и процессов при решении задач, связанных со сферой будущей профессиональной деятельности; овладение навыками точного и сжатого выражения математической мысли в устном и письменном изложении, с использованием соответствующей символики.

Целью проведения практических занятий по дисциплине является развитие логического и алгоритмического мышления, формирование знаний по основным разделам дисциплины, необходимым студентам как для освоения базовых и вариативных дисциплин, так и для дальнейшей самостоятельной работы.

В ходе практического занятия студент учится логично, ясно, четко, грамотным математическим языком излагать свои мысли, приводить доводы, формулировать аргументы в защиту своей позиции.

На занятии студент опирается на свои конспекты, сделанные на лекции, собственные выписки из учебников и словарно-справочную литературу.

2. НАИМЕНОВАНИЕ ПРАКТИЧЕСКИХ ЗАНЯТИЙ

№ темы	Наименование тем практических занятий	Обьем часов	Форма проведения
	3 семестр		
2	Абсолютная и относительная погрешности измерений	1,5	Решение разноуровневых задач
2	Расчет погрешностей непосредственных измерений	1,5	Решение разноуровневых задач
2	Расчет погрешностей косвенных измерений	1,5	Решение разноуровневых задач
3	Вариационные ряды и их графическое изображение	1,5	Решение разноуровневых задач
3	Числовые характеристики выборки, свойства числовых характеристик	1,5	Решение разноуровневых задач
4	Выборочный метод. Статистическое оценивание	1,5	
5	Понятие интервального оценивания. Доверительная вероятность и предельная ошибка выборки. Оценка характеристик	1,5	

	генеральной совокупности по малой выборке		
6	Статистическая проверка гипотез. Параметрические критерии	3	Решение разноуровневых задач
6	Статистическая проверка гипотез. Непараметрические критерии	3	
6	Проверка гипотез о законах распределения	1,5	Решение разноуровневых задач
7	Корреляционная зависимость. Коэффициент корреляции	1,5	Решение разноуровневых задач
8	Линейная регрессия. Коэффициенты регрессии	1,5	
9	Однофакторный дисперсионный анализ	3	
10	Двухфакторный дисперсионный анализ	3	
	Итого за 3 семестр	**27**	

3.ВОПРОСЫ И ЗАДАЧИ ПРАКТИЧЕСКИХ ЗАНЯТИЙ

Практическое занятие 1. Абсолютная и относительная погрешности измерений.

Цель: формирование представлений о значении и методах вычисления погрешностей измерений.

Теоретическая часть:

Практическая деятельность человека неразрывно связана с числами, которые можно получать тремя способами: в результате измерений, счета и выполнения математических операций.

Однако:

- любое измерение нельзя выполнить точно: ошибку дает либо прибор, либо наблюдатель;
- счет дает точные результаты, только если количество предметов невелико и если оно постоянно во времени;
- не все математические операции можно выполнить абсолютно точно.

В этих случаях мы имеем дело с приближенными числами. Но при вычислениях важно знать отклонение приближенного значения величины от ее точного значения, для этого вводится понятие абсолютной погрешности приближения.

Абсолютной погрешностью приближения называется модуль разности между точным значением величины и ее приближенным значением.

$\Delta = |a - x|$, где Δ – абсолютная погрешность

a – точное значение величины

x – приближенное значение

$\Delta = |a - x| \Rightarrow a - x = \pm\, \Delta \Rightarrow a = x \pm \Delta$

Пример. Найти абсолютную погрешность приближения 0,44 числа 4/9.

$\Delta = \left|\frac{4}{9} - 0{,}44\right| = \left|\frac{4}{9} - \frac{11}{25}\right| = \left|\frac{100-99}{225}\right| = \frac{1}{225}$

На практике во многих случаях точное значение бывает неизвестно, поэтому абсолютную погрешность найти нельзя. Однако можно дать оценку абсолютной погрешности, если известны приближения с избытком и с недостатком.

Границей абсолютной погрешности Δ приближения называется такое положительное число h больше которого абсолютная погрешность быть не может.

$\Delta = |a - x| \le h$

Пример. $\frac{1}{225} = 0{,}004444... < 0{,}0045$

x - Δ – Нижняя граница (Н.Г.)

x + Δ – Верхняя граница (В.Г.)

Приближенные числа, как и точные записываются как правило при помощи десятичных дробей. Но если в записи точного числа все его цифры верные, то в приближенном некоторые его цифры верные, а другие являются сомнительными.

Цифра называется верной (точно значащей), если абсолютная погрешность числа не превосходит единицы того разряда в котором записана эта цифра. В противном случае она называется сомнительной.

Пример. x = 3,7412 $\pm 0{,}002$

Определить верные и сомнительные цифры.

В.Г. = 3,7412 + 0,002 = 3,7432

Н.Г. = 3,7412 - 0,002 = 3,7392

Верные – 3 и 7, сомнительные 4,1 и 2.

Замечания.

1) В записи приближенного числа сохраняются только верные цифры. x = 3,7

2) Если в десятичной дроби последние верные цифры нули, то они остаются в записи числа.

x = 0,301 $\pm 0{,}001$

В.Г. = 0,302 Н.Г. = 0, 300 $\Rightarrow$ x = 0,30

3) В десятичной записи числа значащими цифрами называются все его верные цифры, начиная с первой слева отличной от нуля.

0, 583; 38,57; 38,507; 29,830

Но абсолютной погрешности не достаточно для полной характеристики приближения.

Если измерять расстояние между двумя городами, которое равно 100 км, с точность до 1 м, то это будет точное измерение, а если с точность до 1м измерена длина участка земли, которая равна 10м, то это грубое измерение.

Относительной погрешностью называется отношение абсолютной погрешности к приближенному значению измеряемой величины. Обычно выражается в процентах.

$\omega = \frac{\omega}{x}$; $\omega\% = \frac{\omega}{x} \cdot 100\%$

Т.о. для более полной оценки точности измерений необходимо определить, какую часть, или сколько процентов, составляет абсолютная погрешность от значения данной величины.

Пример. Сравнить точность двух измерений .

d = 4 $\pm$0,3;   H = 600 $\pm$0,3

ω(d) = $\frac{0,3}{4}=\frac{3}{10}\div 4=\frac{3}{10}\cdot\frac{1}{4}=\frac{3}{40}=0,075=7,5\%$

ω(H) = $\frac{0,3}{600}=\frac{3}{10}\div 600=\frac{3}{10}\cdot\frac{1}{600}=\frac{3}{6000}=\frac{1}{2000}=0,5\cdot 0,001=0,0005=0,05\%$

Второе измерение более точное.

Задачи и упражнения:

1.Вольтметром, имеющим значения γ_m= 1,0 %, $U_н$ = $G_{норм}$, G_k = 450 В, измеряют напряжение U_u, равное 10 В. Оценить погрешности измерений.

2. Вычислить абсолютную погрешность вольтметра В7-26 при измерениях напряжения в цепи постоянного тока. Класс точности вольтметра задан максимально приведенной погрешностью γ_m=±2,5 %. Используемый в работе предел шкалы вольтметра $U_{норм}$=30 В.

3. Найти предельную абсолютную погрешность частного 2,81 : 0,571.

4. Найти сумму приближенных замеров прибора. Найти количество верных знаков: 0,0909 + 0,0833 + 0,0769 + 0.0714 + 0,0667 + 0.0625 + 0,0588+ 0,0556 + 0,0526.

5. Если при трех и более повторных измерениях данным прибором получены одинаковые значения физической величины, то чему равны абсолютные случайная и систематическая погрешности? Относительная погрешность?

6. Определить для вольтметра с пределом измерения 30 В класса точности **0,5** относительную погрешность для точек 5, 10, 15, 20, 25 и 30 В и наибольшую абсолютную погрешность прибора.

Практическое занятие 2. Расчет погрешностей непосредственных измерений.

Цель: формирование представлений о значении и методах вычисления погрешностей измерений.

Теоретическая часть:

Работа химиков, физиков и представителей других профессий часто связана с выполнением количественных измерений различных величин. При этом возникает вопрос анализа достоверности получаемых значений, обработки результатов непосредственных измерений и оценки погрешностей расчетов, в которых используются значения непосредственно измеряемых характеристик (последний процесс также называется обработкой результатов косвенных измерений).

A -измеряемая величина, $\bar{A}$ -среднее значение измеряемой величины, $\Delta\bar{A}$ - абсолютная погрешность среднего значения измеряемой величины, $\varepsilon=\frac{\Delta\bar{A}}{\bar{A}}\cdot 100\%$ - относительная погрешность среднего значения измеряемой величины.

Предположим, что были проведены n измерений одной и той же величины A в одних и тех же условиях. В этом случае можно рассчитать среднее значение этой величины в проведенных измерениях:

$$\overline{A} = \frac{\sum_{i=1}^{n} A_i}{n} \tag{1}$$

Как вычислить погрешность $\overline{A}$? По следующей формуле:

$$\Delta\overline{A} = t_{\gamma, n-1} \frac{\sqrt{\frac{\sum_{i=1}^{n} \left(A_i - \overline{A}\right)^2}{n-1}}}{\sqrt{n}} \tag{2}$$

В этой формуле используется коэффициент Стьюдента $t_{\gamma, n-1}$. Его значения при разных доверительных вероятностях и значениях n приведены в таблице.

Пример. Проводили измерения длины L металлического бруска. Было сделано 10 измерений и получены следующие значения: 10 мм, 11 мм, 12 мм, 13 мм, 10 мм, 10 мм, 11 мм, 10 мм, 10 мм, 11 мм. Требуется найти среднее значение $\overline{L}$ измеряемой величины (длины бруска) и его погрешность $\Delta\overline{L}$.

Решение:

С использованием формулы (1) находим:

$$\overline{L} = \frac{\sum_{i=1}^{10} L_i}{10} = \frac{10+11+12+13+10+10+11+10+10+11}{10} = 10{,}8 \text{ мм}$$

Теперь с использованием формулы (2) найдем абсолютную погрешность $\Delta\overline{L}$ среднего значения $\overline{L}$ при доверительной вероятности $\gamma = 0{,}95$ и числе степеней свободы $f = n - 1 = 10 - 1 = 9$ (используем значение $t_{0.95,9}$ =2,262, взятое из таблицы):

$$\Delta\overline{L} = t_{0.95,9} \cdot \frac{\sqrt{\frac{\sum_{i=1}^{10} \left(L_i - \overline{L}\right)^2}{9}}}{\sqrt{10}} = 2{,}26 \times$$

$$\times \frac{\sqrt{\frac{(10-10{,}8)^2 + (11-10{,}8)^2 + (12-10{,}8)^2 + (13-10{,}8)^2 + (10-10{,}8)^2 + (10-10{,}8)^2 + (11-10{,}8)^2 + (10-10{,}8)^2 + (11-10{,}8)^2}{9}}}{\sqrt{10}} = 0{,}7$$

Запишем результат:

L =10,8±$0{,}7_{0.95}$ мм

Задачи и упражнения:

1. Студент измеряет период колебаний T маятника три раза и получает результаты: 1,6; 1,8; 1,7 с. Чему равны среднее арифметическое значение $\bar{T}$ и среднее квадратическое отклонение $S(\bar{T})$ результата измерений?

2. Определите доверительную случайную погрешность $\varepsilon(x)$, если $S(\bar{x}) = 0{,}03$ мм; $n = 5$; $P = 0{,}95$.

3. Вычислите среднее квадратическое отклонение результата измерения, если доверительная случайная погрешность $\varepsilon = 15$ с, число наблюдений равно 7, доверительная вероятность $P = 0{,}95$.

Практическое занятие 3. Расчет погрешностей косвенных измерений.

Цель: формирование представлений о значении и методах вычисления погрешностей измерений.

Теоретическая часть:

Предположим, что в ходе эксперимента измеряются величины $K_1, K_2, K_3 \ldots K_p$ а затем с использованием полученных значений вычисляется величина D по формуле $D = f(K_1, K_2, K_3)$. При этом погрешности непосредственно измеряемых величин рассчитываются так, как это было описано в предыдущем занятии.

Расчет среднего значения величины D производится по зависимости $D = f(K_1, K_2, K_3)$ с использованием средних значений аргументов K_p.

Погрешность величины D рассчитывается по следующей формуле:

$$\varepsilon_D = \frac{\Delta \bar{D}}{\bar{D}} \cdot 100\% = \sqrt{\sum_{p=1}^{m} \left(\frac{\partial f}{\partial K_p} \cdot \frac{\Delta \bar{K}_p}{\bar{K}_p} \right)^2} \cdot 100\%, \qquad (3)$$

где m - количество аргументов K_p, $\frac{\partial f}{\partial K_p}$ - частные производные функции f по аргументам K_p, $\Delta \bar{K}_p$ - абсолютная погрешность среднего значения аргумента K_p.

Абсолютная погрешность, как и в случае с прямыми измерениями, рассчитывается по формуле $\Delta \bar{D} = \frac{\varepsilon_D \cdot \bar{D}}{100\%}$.

Пример. Было проведено 5 непосредственных измерений величин Q и R. Для величины Q получены значения: 50, 51, 52, 50, 47; для величины R получены значения: 500, 510, 476, 354, 520. Требуется рассчитать значение величины S, определяемой по формуле $S = \ln(Q \cdot R)$ и найти погрешность полученного значения.

Решение: По формуле (1) найдем средние значения величин Q и R:

$$\overline{Q} = \frac{\sum_{i=1}^{5} Q_i}{5} = \frac{50+51+52+50+47}{5} = 50$$

$$\overline{R} = \frac{\sum_{i=1}^{5} R_i}{5} = \frac{500+510+476+354+520}{5} = 472$$

Вычисляем S:

$$\overline{S} = \ln\left(\overline{Q} \cdot \overline{R}\right) = \ln(50 \cdot 472) = 10{,}07$$

Находим в таблице при доверительной вероятности 0,95 и числе степеней свободы $f = n-1 = 5-1 = 4$ значение $t_{0.95,4} = 2{,}776$. По формуле (2) рассчитываем погрешности средних значений величин Q и R:

$$\Delta Q = t_{\gamma,n-1} \frac{\sqrt{\frac{\sum_{i=1}^{n}\left(Q_i - \overline{Q}\right)^2}{n-1}}}{\sqrt{n}} = 2{,}78 \cdot \frac{\sqrt{\frac{\sum_{i=1}^{5} (50-50)^2 + (51-50)^2 + (52-50)^2 + (50-50)^2 + (47-50)^2 +}{4}}}{\sqrt{5}} = 2{,}32$$

$$\Delta R = t_{\gamma,n-1} \frac{\sqrt{\frac{\sum_{i=1}^{n}\left(R_i - \overline{R}\right)^2}{n-1}}}{\sqrt{n}} = 2{,}78 \cdot \frac{\sqrt{\frac{\sum_{i=1}^{5} (500-472)^2 + (510-472)^2 + (476-472)^2 + (354-472)^2 + (520-472)^2 +}{4}}}{\sqrt{5}} = 84{,}5$$

С использованием формулы (3) находим относительную погрешность среднего значения величины S:

$$\varepsilon_S = \frac{\Delta\overline{S}}{\overline{S}} \cdot 100\% = \sqrt{\sum_{p=1}^{m}\left(\frac{\partial f}{\partial K_p} \cdot \frac{\Delta\overline{K}_p}{\overline{K}_p}\right)^2} \cdot 100\% = \sqrt{\left(\overline{R} \cdot \frac{1}{\overline{Q}} \cdot \frac{\Delta\overline{Q}}{\overline{Q}}\right)^2 + \left(\overline{Q} \cdot \frac{1}{\overline{R}} \cdot \frac{\Delta\overline{R}}{\overline{R}}\right)^2} \cdot 100\% =$$

$$= \sqrt{\left(472 \cdot \frac{1}{50} \cdot \frac{2{,}32}{50}\right)^2 + \left(50 \cdot \frac{1}{472} \cdot \frac{84{,}5}{472}\right)^2} \cdot 100\% = 44\%$$

Найдем абсолютную погрешность среднего значения величины S:

$$\Delta\overline{S} = \frac{\varepsilon_S}{100\%} \cdot \overline{S} = \frac{44\%}{100\%} \cdot 10{,}07 = 4{,}43$$

Запишем результат:

$$S = 10{,}07 \pm 4{,}43_{0{,}95}$$

Задачи и упражнения:

1. Студент измеряет четыре длины: $a = (50 \pm 5)$ см; $b = (30 \pm 3)$ см; $c = (40 \pm 1)$ см; $d = (7{,}8 \pm 0{,}3)$ см и вычисляет три суммы: $a + b$, $a + c$, $a + d$. Найдите погрешности этих сумм.

2. Студент получил следующие результаты измерений: $a = (5 \pm 1)$ см;

$b = (18 \pm 2)$ см; $c = (12 \pm 1)$ см; $t = (3{,}0 \pm 0{,}5)$ с; $m = (18 \pm 1)$ г. Вычислите значения величин w_1, w_2, w_3, w_4, w_5, w_6, их абсолютные и относительные погрешности, если рабочие формулы: $w_1 = a + b + c$; $w_2 = a + b - c$; $w_3 = ct$; $w_4 = 4a$; $w_5 = b/2$;

$w_6 = mb/t$. Какую погрешность удобнее рассчитать сначала, абсолютную или относительную для этих рабочих формул ?

3. Тележка скатывается без трения по наклонной плоскости с углом наклона φ. Ускорение тележки $a = g{\cdot}\sin(\varphi)$ (g = 9,8 м/с2 – ускорение свободного падения). Результат измерения угла φ = (5,4 ± 0,1)о. Вычислите ускорение a и его погрешность.

Практическое занятие 4. Вариационные ряды и их графическое изображение.

Цель: формирование основных понятий математической статистики.

Теоретическая часть:

Математическая статистика занимается установлением закономерностей, которым подчинены массовые случайные явления, на основе обработки статистических данных, полученных в результате наблюдений. Двумя основными задачами математической статистики являются:

- определение способов сбора и группировки этих статистических данных;
- разработка методов анализа полученных данных в зависимости от целей исследования, к которым относятся:

а) оценка неизвестной вероятности события; оценка неизвестной функции распределения; оценка параметров распределения, вид которого известен; оценка зависимости от других случайных величин и т.д.;

б) проверка статистических гипотез о виде неизвестного распределения или о значениях параметров известного распределения.

Для решения этих задач необходимо выбрать из большой совокупности однородных объектов ограниченное количество объектов, по результатам изучения которых можно сделать прогноз относительно исследуемого признака этих объектов.

Определим основные понятия математической статистики.

Генеральная совокупность – все множество имеющихся объектов.

Выборка – набор объектов, случайно отобранных из генеральной совокупности.

Объем генеральной совокупности N и *объем выборки* n – число объектов в рассматриваемой совокупности.

Виды выборки:

Повторная – каждый отобранный объект перед выбором следующего возвращается в генеральную совокупность;

Бесповторная – отобранный объект в генеральную совокупность не возвращается.

Для того, чтобы по исследованию выборки можно было сделать выводы о поведении интересующего нас признака генеральной совокупности, нужно, чтобы выборка правильно представляла пропорции генеральной совокупности, то есть была *репрезентативной* (представительной). Учитывая закон больших чисел, можно утверждать, что это условие выполняется, если каждый объект выбран случайно, причем для любого объекта вероятность попасть в выборку одинакова.

Пусть интересующая нас случайная величина X принимает в выборке значение x_1 n_1 раз, x_2 – n_2 раз, …, $x_к$ – $n_к$ раз, причем $\sum_{i=1}^{k} n_k = n$, где n – объем выборки. Тогда наблюдаемые значения случайной величины x_1, x_2,…, $x_к$ называют вариантами, а n_1, n_2,…, $n_к$ – частотами. Если разделить каждую частоту на объем выборки, то получим *относительные частоты* $\omega_i = \frac{n_i}{n}$.

Последовательность вариант, записанных в порядке возрастания, называют вариационным рядом, а перечень вариант и соответствующих им частот или относительных частот – статистическим рядом.

Пример 1:

При проведении 20 серий из 10 бросков игральной кости число выпадений шести очков оказалось равным 1, 1, 4, 0, 1, 2, 1, 2, 2, 0 , 5, 3, 3, 1, 0, 2, 2, 3, 4, 1.

Составим вариационный ряд: 0,1,2,3,4,5. Статистический ряд для абсолютных и относительных частот имеет вид:

x_i	0	1	2	3	4	5
n_i	3	6	5	3	2	1
w_i	0,15	0,3	0,25	0,15	0,1	0,05

Если исследуется некоторый непрерывный признак, то вариационный ряд может состоять из очень большого количества чисел. В этом случае удобнее использовать группированную выборку. Для ее получения интервал, в котором заключены все наблюдаемые значения признака, разбивают на несколько равных частичных интервалов длиной h, а затем находят для каждого частичного интервала n_i – сумму частот вариант, попавших в i-й интервал. Составленная по этим результатам таблица называется *группированным статистическим рядом*.

Для наглядного представления о поведении исследуемой случайной величины в выборке можно строить различные графики. Один из них – полигон частот: ломаная, отрезки которой соединяют точки с координатами (x_1, n_1), (x_2, n_2),…, (x_k, n_k), где x_i откладываются на оси абсцисс, а n_i – на оси ординат. Если на оси ординат откладывать не абсолютные (n_i), а относительные (w_i) частоты, то получим полигон относительных частот.

Этапы первичной обработки выборки:

1.Ранжирование опытных данных (расположение значений признака по убыванию или возрастанию).

2.Частотный анализ (построение статистического ряда, определение относительных частот).

3.Группировка (частотная табуляция выборки).

Задачи и упражнения:

1. Имеются следующие данные об успеваемости 20 студентов группы по статистике в летнюю сессию: 4, 4, 3, 5, 2, 2, 5, 3, 3, 3, 5, 3, 2, 4, 3, 2, 3, 5, 5, 4. Постройте: ряд распределения студентов по баллам оценок; ряд распределения студентов пол уровню успеваемости, выделив в нем две группы студентов: неуспевающие и успевающие. Изобразите каждый из рядов графически.

2.Построить кривую и гистограмму суммы налоговых неуплат, зафиксированных по условным регионам страны, данные по которым приведены в таблице (в млн. руб.):

21	43	72	84
22	54	75	32
26	49	77	45
27	53	78	65
28	54	81	12
32	58	83	34
34	61	84	54
37	65	84	34
39	68	88	41

3.Дана исходная выборка по росту и весу студентов группы:

Рост	160	168	175	169	170	169	162	166	163	160	158	173	162	173	156
Вес	48	58	69	64	69	70	51	60	67	54	48	58	44	50	56

Определите объем выборки. Составьте ранжированный и вариационный ряды по каждому из признаков.

4.Построить гистограмму нагрузки на одного следователя по расследованным уголовным делам по данным таблицы:

Годы	1995	1996	1997	1998	1999	2000	2001	2001	2003
По линии МВД	63,2	49,6	41	31,5	32,3	32,4	37,9	34,2	36,2
По линии прокуратуры	17,2	16,9	15,5	15,1	16	17,7	18,2	18,3	19,7

5.Три варианта исходных данных таблицы ниже – результаты телефонных переговоров в минутах сотрудников трех отдельных служб в течение рабочего дня. Требуется составить интервальный ряд распределения.

№ выборки	1	2	3	4	5	6	7	8	9
Служба 1	1,9	2,6	3,1	3,3	2,1	2,0	4,7	0,9	2,8
Служба 2	6,9	1,1	3,6	1,0	7,0	2,1	8,7	1,0	8,0
Служба 3	4,7	6,8	3,8	3,1	1,8	7,1	9,2	8,1	4,0

6. Дана выборка 7,3,3,6,4,3,5,1,2,1,3. Построить вариационный ряд. Определить размах выборки.

Практическое занятие 5. Числовые характеристики выборки, свойства числовых характеристик.

Цель: формирование основных понятий математической статистики.

Теоретическая часть:

Одной из основных характеристик ряда распределения является средняя арифметическая. Существует две формулы для расчёта средней арифметической: простая и взвешенная. Простую среднюю арифметическую используют, когда данные наблюдений не сведены в вариационный ряд или все частоты равны единице (одинаковы).

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n},$$

где x_i – i-е значение признака; n – объём ряда (число наблюдений).

Если частоты отличны друг от друга, расчёт производится по формуле средней арифметической взвешенной

$$\bar{x} = \frac{\sum_{i=1}^{k} x_i m_i}{\sum_{i=1}^{k} m_i},$$

где x_i – i-е значение признака; m_i – частота i-го значения признака; k – число его значений (вариантов).

При расчёте средней арифметической в качестве весов могут выступать и частости, тогда формула расчёта средней арифметической взвешенной примет следующий вид.

$$\bar{x} = \sum_{i=1}^{k} x_i w_i,$$

где w_i – частость i-го значения признака;

Колеблемость изучаемого признака можно охарактеризовать с помощью различных показателей вариации. К числу основных показателей вариации относятся: математическое ожидание, дисперсия, среднее квадратическое отклонение, коэффициент вариации.

Математическое ожидание – это числовая характеристика случайной величины, со средним арифметическим её наблюдаемых значений, которое является статистической характеристикой вариационного ряда и рассчитывается по формуле:

$$M(X) = \sum_{i=1}^{n} x_i p_i,$$

где p_i – вероятность i-го значения признака.

Дисперсию можно рассчитать по простой и взвешенной формулам имеющим вид

$$D(X) = \frac{\sum_{i=1}^{n}(x_i - \bar{x}_i)^2}{n}; \qquad D(X) = \frac{\sum_{i=1}^{n}(x_i - \bar{x}_i)^2 m_i}{\sum_{i=1}^{k} m_i};$$

Среднее квадратическое отклонение рассчитывается по формуле

$$\sigma(X) = \sqrt{D(X)}.$$

Коэффициент вариации определяется формулой

$$V(X) = \frac{\sigma(X)}{\bar{x}} \cdot 100\%.$$

Задачи и упражнения:

1) При обследовании 50 членов семей рабочих и служащих установлено следующее количество членов семьи: 5; 3; 2; 1; 4; 6; 3; 7; 9; 1; 3; 2; 5; 6; 8; 2; 5; 2; 3; 6; 8; 3; 4; 4; 5; 6; 5; 4; 7; 5; 6; 4; 8; 7; 4; 5; 7; 8; 6; 5; 7; 5; 6; 6; 7; 3; 4; 6; 5; 4. Составьте вариационный ряд распределения частот. Постройте полигон распределения частот, кумуляту. Определите среднее число членов семьи Охарактеризуйте колеблемость размера семьи с помощью показателей вариации

Объясните полученные результаты, сделайте выводы.

2) Имеются данные о еженедельном количестве проданных компьютеров одной из фирм: 398, 412, 560, 474, 544, 690, 587, 600, 613, 457, 504, 477, 530, 641, 359, 566, 452, 633, 474, 499. 580, 606, 344, 455,505, 396, 347, 441, 390, 632, 400, 582. Составьте вариационный ряд. Найдите среднее количество проданных компьютеров. Рассчитайте показатели вариации

3) Администрацию магазина интересует частота покупок калькуляторов. Менеджер в течении января регистрировал данные о покупке МК и собрал следующие данные: 8, 4, 4, 9, 3, 3, 1, 2, 0, 4, 2, 3, 5, 7, 10, 6, 5, 7, 3, 2, 9, 8, 1, 4, 6, 5, 4, 2, 1, 0, 8. Постройте вариационный ряд, определите его числовые характеристики. Какие рекомендации вы дали бы администрации универсама?

4) Число пассажиров одного из рейсов за 30 дней составило: 128, 121, 134, 118, 123, 109, 120, 116, 125, 128, 121, 129, 130, 131, 127, 119, 114, 124, 110, 126, 134, 125, 128, 123, 128, 133, 132, 136, 134, 129. Составьте вариационный ряд. Найдите среднее число пассажиров в рейсе? Рассчитайте показатели вариации. Сделайте анализ полученных результатов.

5) Имеются данные о годовой мощности предприятий в 2003 году

Предприятия с годовой мощностью, тыс.т	Количество предприятий
До 500	27
500 – 1000	11
1000 – 2000	8
2000 – 3000	8
Свыше 3000	2

Постройте гистограмму, кумуляту. Рассчитайте среднюю мощность предприятий. Найдите дисперсию. среднее квадратическое отклонение, коэффициент вариации. Сделайте анализ полученных результатов.

6) По данным выборочного обследования получено следующее распределение по среднедушевому доходу

Среднедушевой доход семьи в месяц, у.е.	до 25	25 – 50	50 – 75	75 – 100	100 – 125	125 – 150	150 и выше
Количество обследованных семей	46	236	250	176	102	78	12

Постройте гистограмму, кумуляту. Рассчитайте среднюю мощность предприятий. Найдите дисперсию. среднее квадратическое отклонение, коэффициент вариации. Сделайте анализ полученных результатов.

Практическое занятие 6. Выборочный метод. Статистическое оценивание.

Цель: формирование основных понятий математической статистики при обработке результатов эксперимента.

Теоретическая часть:

По одному из определений, статистика – это наука, позволяющая распространять выводы, сделанные на основе изучения части совокупности, на всю совокупность. В этом определении заключена сущность выборочного метода и его ведущая роль в статистике.

Все единицы совокупности, обладающие интересующими исследователя признаками, составляют генеральную совокупность.

Часть совокупности, случайным образом отобранная из генеральной совокупности составляют выборочную совокупность – выборку.

Число элементов статистической совокупности называется её объёмом. Объём генеральной совокупности обозначается N, а объём выборки – n.

Случайная выборка из n элементов – это такой отбор, при котором элементы извлекаются по одному из всей генеральной совокупности и каждый из них имеет равный шанс быть отобранным. Такая выборка называется собственно – случайной.

По способу отбора элементов различают два типа случайных выборок: собственно –_случайная бесповторная и собственно –_случайная повторная. Выбор схемы отбора зависит от характера изучаемого объекта.

Пусть из генеральной совокупности извлекается выборка объёмом n, причём значение признака x_1 наблюдаются m_1 раз, x_2 – m_2 раз, …, x_k наблюдается m_k раз,

$\sum_{i=1}^{k} m_i = n$- объём выборки

Статистическим распределением выборки называется перечень возможных значений признака x_i и соответствующих ему частот m_i

Числовые характеристики генеральной совокупности называются параметрами генеральной совокупности. Доля единиц, обладающих тем или иным признаком в генеральной совокупности, называется генеральной долей и обозначается p.

Оценка параметра – это определённая числовая характеристика, полученная из выборки. Когда оценка определяется одним числом её называют точечной оценкой. В качестве точечных оценок параметров генеральной совокупности используются соответствующие выборочные характеристики.

Ошибки выборки.

Так как выборочная совокупность это часть генеральной совокупности, то естественно, что выборочные характеристики не будут точно совпадать с соответствующими генеральными. Ошибка может быть представлена как разность между генеральными и выборочными характеристиками изучаемой совокупности: $\varepsilon = \tilde{\bar{X}} - \bar{X}$, либо $\varepsilon = p - \omega$.

Применительно к выборочному методу из теоремы Чебышева следует, что с вероятностью, сколь угодно близкой к единице, можно утверждать, что при достаточно большом объёме выборки и ограниченной дисперсии генеральной совокупности разность между выборочной средней и генеральной средней будет сколь угодно мала.

$$P(|\tilde{X} - \bar{X}| < \frac{t \cdot \sigma_{\text{ген}}}{\sqrt{n}}) > 1 - \frac{1}{t^2},$$

где $\tilde{X}$ - средняя по совокупности выбранных единиц; $\bar{X}$ - средняя по генеральной совокупности; $\sigma_{\text{ген}}$- среднее квадратическое отклонение в генеральной совокупности.

О величине расхождения между параметром и статистикой

$$\Delta = \frac{t \cdot \sigma}{\sqrt{n}} = t \cdot \mu,$$

можно судить лишь с определённой вероятностью, от которой зависит величина t.

Средняя ошибка выборки $\mu = \frac{\sigma_{\text{ген}}}{\sqrt{n}}$. Согласно центральной предельной теореме Ляпунова, выборочные распределения статистик (при n 30) будут иметь нормальное распределение независимо от того, какое распределение имеет генеральная совокупность. Следовательно,

$$P(|\tilde{X} - \bar{X}| < t\mu) \approx 2\Phi_0(t),$$

где $\Phi_0(t)$- функция Лапласа.

В зависимости от способа отбора средняя ошибка выборки определяется по разному

μ	Собственно случайный отбор	
	повторный	бесповторный
Для средней	$\sqrt{\frac{\sigma^2}{n}}$	$\sqrt{\frac{\sigma^2}{n}(1 - \frac{n}{N})}$
Для доли	$\sqrt{\frac{\omega(1-\omega)}{n}}$	$\sqrt{\frac{\omega(1-\omega)}{n}(1 - \frac{n}{N})}$

Здесь σ^2- выборочная дисперсия значений признака; $\omega(1-\omega)$ - выборочная дисперсия доли значений признака; n – объём выборки; N – объём генеральной совокупности; $\frac{n}{N}$ - доля обследованной совокупности; (1 - $\frac{n}{N}$) – поправка на бесповторность отбора.

Формулы расчёта необходимой численности выборки для собственно случайного отбора определяются в таблице.

μ	Собственно случайный отбор	
	повторный	бесповторный
Для средней	$\frac{t^2\sigma^2}{\Delta^2}$	$\frac{t^2\sigma^2 N}{N\Delta^2 + t^2\sigma^2}$
Для доли	$\frac{t^2\omega(1-\omega)}{\Delta^2}$	$\frac{t^2 N\omega(1-\omega)}{n\Delta^2 + t^2\omega(1-\omega)}$

Задачи и упражнения:

По данным выборки, удовлетворяющей нормальному закону распределения, вычислить:

1) выборочное среднее;

2) исправленное выборочное среднее квадратическое отклонение.

А) 8.0 -1.1 13.5 10.0 2.4 4.1 20.0 12.4 13.4 4.8 7.8 0.0 10.9 13.7 6.6

Б) 31.6 34.9 46.9 42.8 36.0 26.2 28.6 48.5 27.7 45.8 32.0 41.2 39.8 33.1 36.3 53.5 43.9 35.8 32.9 34.4

В) 25.4 31.1 13.2 23.0 19.1 26.5 23.2 29.2 24.8 26.6 29.3 21.4 28.2 38.2 19.9 30.6 24.5 23.2

Г) 10.5 5.5 12.6 27.0 25.0 31.2 15.9 15.3 17.4 32.8 30.3 9.5 17.7 16.4 15.9.

Практическое занятие 7. Понятие интервального оценивания. Доверительная вероятность и предельная ошибка выборки. Оценка характеристик генеральной совокупности по малой выборке.

Цель: формирование основных понятий интервального оценивания и методов оценки характеристик генеральной совокупности по малой выборке.

Теоретическая часть:

При выборке малого объёма точечная оценка может значительно отличаться от оцениваемого параметра, т. е. приводить к грубым ошибкам. По этой причине при небольшом объёме выборки следует пользоваться интервальными оценками. Интервальной называют оценку, которая определяется двумя числами – концами интервала.

Пусть найденная по данным выборки статистическая характеристика Θ^* служит оценкой неизвестного параметра Θ. Будем считать Θ постоянным числом (Θ может быть и случайной величиной). Ясно, что Θ^* тем точнее определяет параметр Θ, чем меньше абсолютная величина разности $\left|\Theta - \Theta^*\right|$.

Другими словами, если $\delta > 0$ и $\left|\Theta - \Theta^*\right| < \delta$, то чем меньше δ, тем оценка точнее. Таким образом, положительное число δ характеризует точность оценки.

Однако статистические методы не позволяют категорически утверждать, что оценка Θ^* удовлетворяет неравенству $\left|\Theta - \Theta^*\right| < \delta$; можно лишь говорить о вероятности γ, с которой это неравенство осуществляется.

Надёжностью (доверительной вероятностью) оценки Θ по Θ^* называют вероятность γ, с которой осуществляется неравенство $\left|\Theta - \Theta^*\right| < \delta$. Обычно надёжность оценки задаётся наперёд, причём в качестве γ берут число, близкое к единице. Наиболее часто задают надёжность, равную 0,95; 0,99 и 0,999.

Пусть вероятность того, что $\left|\Theta - \Theta^*\right| < \delta$, равна γ: $P\left[\left|\Theta - \Theta^*\right| < \delta\right] = \gamma$.

Заменив неравенство $\left|\Theta - \Theta^*\right| < \delta$ равносильным ему двойным неравенством

$-\delta < \Theta - \Theta^* < \delta$, или $\Theta^* - \delta < \Theta < \Theta^* + \delta$, имеем $P\left[\Theta^* - \delta < \Theta < \Theta^* + \delta\right] = \gamma$.

Это соотношение следует понимать так: вероятность того, что интервал $(\Theta^* - \delta, \Theta^* + \delta)$ заключает в себе (покрывает) неизвестный параметр Θ, равна

γ. Доверительным называют интервал $(\Theta^* - \delta, \Theta^* + \delta)$, который покрывает неизвестный параметр с заданной надёжностью γ.

Доверительные интервалы для оценки математического ожидания нормального распределения при известном σ.

Пусть количественный признак X генеральной совокупности распределён нормально, причём среднее квадратическое отклонение σ этого распределения известно. Требуется оценить неизвестное математическое ожидание a по выборочной средней $\bar{x}$. Поставим своей задачей найти доверительные интервалы, покрывающие параметр a с надёжностью γ.

Если случайная величина X распределена нормально, то выборочная средняя $\overline{X}$, найденная по независимым наблюдениям, также распределена нормально. Параметры распределения $\overline{X}$ таковы:

$$M(\overline{X}) = a,\ \ \sigma(\overline{X}) = \frac{\sigma}{\sqrt{n}}.$$

Приняв во внимание, что по условию нам задана вероятность γ, получаем следующую формулу (чтобы получить рабочую формулу, выборочную среднюю вновь обозначим через $\bar{x}$)

$$P\left(\bar{x} - \frac{t\sigma}{\sqrt{n}} < a < \bar{x} + \frac{t\sigma}{\sqrt{n}}\right) = 2\Phi_0(t) = \gamma .$$

Смысл полученного соотношения таков: с надёжностью γ можно утверждать, что доверительный интервал $\left(\bar{x} - \frac{t\sigma}{\sqrt{n}}, \bar{x} + \frac{t\sigma}{\sqrt{n}}\right)$ покрывает неизвестный параметр a; точность оценки $\delta = \frac{t\sigma}{\sqrt{n}}$.

Укажем ещё, что число t определяется из равенства $2\Phi_0(t) = \gamma$, или $\Phi_0(t) = \frac{\gamma}{2}$; по таблице функции Лапласа находят аргумент t, которому соответствует значение функции Лапласа, равное $\frac{\gamma}{2}$.

Надёжность γ=0,95 указывает, что если произведено достаточно большое число выборок, то 95% из них определяет такие доверительные интервалы, в которых параметр действительно заключён; лишь в 5 % случаев он может выйти за границы доверительного интервала.

Доверительные интервалы для оценки математического ожидания нормального распределения при неизвестном σ

Пусть количественный признак X генеральной совокупности распределён нормально, причём среднее квадратическое отклонение σ неизвестно. Требуется оценить неизвестное математическое ожидание a с помощью доверительных интервалов. Разумеется, невозможно воспользоваться результатами предыдущего параграфа, в котором σ предполагалось известным.

Оказывается, что по данным выборки можно построить случайную величину $T = \frac{\overline{X} - a}{S/\sqrt{n}}$, которая имеет распределение Стьюдента с k = n-1 степенями свободы; здесь $\overline{X}$ - выборочная средняя, S – «исправленное» среднее квадратическое отклонение, n – объём выборки.

Пользуясь распределением Стьюдента, находим:

$$P\left(\overline{X}-\frac{t_\gamma S}{\sqrt{n}}<a<\overline{X}+\frac{t_\gamma S}{\sqrt{n}}\right)=\gamma .$$

Значит, доверительный интервал $\left(\bar{x}-\frac{t_\gamma s}{\sqrt{n}},\bar{x}+\frac{t_\gamma s}{\sqrt{n}}\right)$, покрывает неизвестный параметр a с надёжностью γ.

Пример. Случайная величина X – вес полугодовалого поросенка в хозяйстве (то есть в генеральной совокупности) - распределена нормально. По выборке объёма n = 16 найдены выборочная средняя $\bar{x}$=20,2 кг и «исправленное» среднее квадратическое отклонение s=0,8 кг. Оценить неизвестное математическое ожидание при помощи доверительного интервала с надёжностью 0,95.

Решение: Найдём t_γ. Пользуясь таблицей, по γ=0,95 и n = 16 находим t_γ=2,13.

Найдём доверительные границы:

$$\bar{x}-\frac{t_\gamma s}{\sqrt{n}}=20{,}2-\frac{2{,}13\cdot 0{,}8}{\sqrt{16}}=19{,}774кг,$$

$$\bar{x}+\frac{t_\gamma s}{\sqrt{n}}=20{,}2+\frac{2{,}13\cdot 0{,}8}{\sqrt{16}}=20.626кг.$$

Итак, с надёжностью 0,95 неизвестный параметр a заключён в доверительном интервале 19,774<a<20,626 (кг).

Доверительные интервалы для оценки среднего квадратического отклонения σ нормального распределения

Пусть количественный признак X генеральной совокупности распределён нормально. Требуется оценить неизвестное генеральное среднее квадратическое отклонение σ по «исправленному» выборочному среднему квадратическому отклонению s.

Доверительный интервал, покрывающий параметр σ с заданной надёжностью γ находят по следующей формуле:

$$s(1-q)<\sigma<s(1+q).$$

Здесь параметр q определяют, пользуются таблицей приложения 2, а s находят по выборке.

Пример. Случайная величина X – вес полугодовалого поросенка в хозяйстве – (то есть в генеральной совокупности) распределён нормально. По выборке объёма n=25 найдено «исправленное» среднее квадратическое отклонение s=0,8 кг. Найти доверительный интервал, покрывающий генеральное среднее квадратическое отклонение σ с надёжностью 0,95.

Решение: По таблице по данным γ=0,95 и n=25 найдём q=0,32.

Искомый доверительный интервал таков:

0,8 (1 - 0,32)< σ<0,8 1(1+0,32), или

0,544<σ<1,056 (кг).

Замечание. Если q>1, то неравенство примет вид

0<σ<s(1+q).

Задачи и упражнения:

1) С помощью собственно – случайного повторного отбора фирма провела обследование 900 своих служащих. Средний стаж работы в фирме равен 8,7 года, а среднее квадратическое отклонение – 2,7 года. Среди обследованных оказалось 270 женщин. Считая стаж работы служащих распределённым по нормальному закону определите: а) с вероятностью 0,95 доверительный интервал, в котором окажется средний стаж работы всех служащих фирмы; б) с вероятностью 0.9 доверительный интервал, накрывающий неизвестную долю женщин во всём коллективе фирмы.

2) Владелец автостоянки опасается обмана со стороны служащих. В течении года владельцем автостоянки проведено 40 проверок. По данным проверок среднее число автомобилей, оставляемых на ночь на охрану, составило 400 единиц, а стандартное отклонение их числа – 10 автомобилей. Считая отбор собственно случайным, с вероятностью 0,99 оцените с помощью доверительного интервала истинное среднее число автомобилей, оставляемых на ночь. Обоснованы ли опасения владельца стоянки, если по отчётности охранников среднее число автомобилей составляет 395 автомобилей.

3) В 24 из 40 проверок число автомобилей на автостоянке не превышало 400 единиц. С вероятностью 0,98 найдите доверительный интервал для оценки истинной доли дней в течении года, когда число оставляемых на стоянке автомобилей не превышало 400 единиц.

4) Служба контроля Энергосбыта провела выборочную проверку расхода электроэнергии жителями одного из домов. С помощью собственно – случайного отбора выбрано 10 квартир и определён расход электроэнергии в течении месяцев: 125; 78; 102; 140; 90; 45; 50; 125; 115; 112. С вероятностью 0. 95 определите доверительный интервал для оценки среднего расхода электроэнергии на 1 квартиру во всём доме при условии, что в доме 70 квартир, а отбор был: а) повторным; б) бесповторным.

5) С целью изучения размеров выручки киосков была произведена 10% - ая случайная бесповторная выборка из 1000 киосков города. В результате были получены данные о средней выручке составившие 500 у.е. В каких пределах с доверительной вероятностью 0,95 может находиться средняя дневная выручка, если среднее квадратическое отклонение составило 150 у. е.?

6) Фирма торгующая компьютерами собирает информацию о состоянии местного компьютерного рынка. С этой целью из 8 746 лиц в возрасте 18 лет и старше, проживающих в этом городе, отобрано 500 человек. Среди них оказалось 29 человек, планирующих приобрести компьютер в новом году. Оцените долю лиц в генеральной совокупности в возрасте 18 лет и старше, планирующих приобрести компьютер в новом году, если $\alpha = 0,05$

7) Для оценки числа безработных среди рабочих в порядке случайной повторной выборки отобраны 400 человек. 25 из них оказались безработными. Используя 95% доверительный интервал, оцените истинные размеры безработицы.

8) Туристическое агентство утверждает, что для черноморского курорта характерна идеальная погода со среднегодовой температурой 20^0С. Пусть случайно отобраны 35 дней в году. Какова вероятность того, что отклонение средней температуры за отобранные дни от среднегодовой температуры не превысит по абсолютной величине 2^0С, если температура воздуха

распределена по нормальному закону, а стандартное отклонение дневной температуры составляет 4^0С ?

Практическое занятие 8-9. Статистическая проверка гипотез. Параметрические критерии.

Цель: сформировать представление о методах и принципах проверки статистических гипотез.

Теоретическая часть:

Статистической называют гипотезу о виде неизвестного распределения, или о параметрах известных распределений. Наряду с выдвинутой гипотезой рассматривают и противоречащую ей гипотезу. Если выдвинутая гипотеза будет отвергнута, то имеет место противоречащая гипотеза. По этой причине эти гипотезы целесообразно различать.

Нулевой (основной) называют выдвинутую гипотезу H_0.

Конкурирующей (альтернативной) называют гипотезу H_1, которая противоречит нулевой.

Выдвинутая гипотеза может быть правильной или неправильной, поэтому возникает необходимость её проверки. Поскольку проверку производят статистическими методами, её называют статистической. В итоге статистической проверки гипотезы в двух случаях может быть принято неправильное решение, т. е. могут быть допущены ошибки двух родов.

Ошибка первого рода состоит в том, что будет отвергнута правильная гипотеза. Ошибка второго рода состоит в том, что будет принята неправильная гипотеза.

Для проверки нулевой гипотезы используют специально подобранную случайную величину, точное или приближённое распределение которой известно. Обозначим эту величину в целях общности через K.

Статистическим критерием (или просто критерием) называют случайную величину K, которая служит для проверки нулевой гипотезы.

После выбора определённого критерия множество всех его возможных значений разбивают на два непересекающихся подмножества: одно из них содержит значения критерия, при которых нулевая гипотеза отвергается, а другая – при которых она принимается.

Критической областью называют совокупность значений критерия, при которых нулевую гипотезу отвергают.

Областью принятия гипотезы (областью допустимых значений) называют совокупность значений критерия, при которых гипотезу принимают.

Основной принцип проверки статистических гипотез можно сформулировать так: если наблюдаемое значение критерия принадлежит критической области – гипотезу отвергают, если наблюдаемое значение критерия принадлежит области принятия гипотезы – гипотезу принимают.

В области статистики и биометрии в частности применяют два вида статистических критериев: *параметрические,* построенные на основании параметров данной совокупности (например, $\bar{x}$ и s^2_x) и представляющие функции этих параметров, и *непараметрические,* представляющие собой функции, зависящие непосредственно от вариант данной совокупности с их частотами. Первые служат для проверки гипотез о параметрах совокупностей, распределяемых по нормальному закону, вторые — для проверки рабочих гипотез независимо от формы распределения совокупностей, из которых взяты сравниваемые выборки. Применение параметрических критериев связано с необходимостью вычисления выборочных характеристик — средней величины и показателей вариации, тогда как при использовании непараметрических критериев такая необходимость отпадает.

t-критерий Стьюдента (t-распределение). Английский математик В. Госсет (печатавшийся под псевдонимом Стьюдент), в 1908 г. нашел закон распределения величины $t = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$, в которой генеральный параметр σ заменен на его выборочную характеристику s_x, т. е. нашел *закон распределения значений*

$$t = \frac{\bar{x} - \mu}{s / \sqrt{n}}.$$

Открытый Стьюдентом и теоретически обоснованный Р. Фишером *закон t-распределения* служит основой так называемой теории малой выборки, которая характеризует распределение выборочных средних в нормально распределяющейся совокупности в зависимости от объема выборки. t-распределение зависит только от числа степеней свободы $k = n—1$, причем с увеличением объема выборки n t-распределение быстро приближается к нормальному с параметрами $\mu = 0$ и $\sigma = 1$ и уже при n>30 не отличается от него. Это видно из таблицы ниже, в которой приведены табулированные значения t-распределения и нормального распределения для разных значений t.

Распределение	**Нормированное отклонение *t***						
	0,5	**1.0**	**1.5**	**2,0**	**2,5**	**3,0**	**3,5**
Нормальное	0,383	0,683	0,866	0,955	0,988	0,997	0,9995
Стьюдента при n = 3	0,333	0,577	0,728	0,816	0,870	0,905	0,927
n = 20	0,377	0,670	0,850	0,940	0,978	0,993	0,998
n = 30	0,383	0,683	0,866	0,955	0,988	0,997	0,9995

Оценка разности средних. Сравнивая друг с другом две независимые выборки, взятые из нормально распределяющихся совокупностей с параметрами μ_1 и μ_2. Разность μ_1- μ_2 этих параметров обозначим через D, то есть μ_1- μ_2=D, а дисперсию этой разности σ^2_D. Значения генеральных параметров неизвестны, однако по выборкам мы можем найти величины выборочных средних и разность между ними $\bar{x}_1 - \bar{x}_2$, которую обозначим d, то есть $\bar{x}_1 - \bar{x}_2 = d.$

Нулевая гипотеза сводится к предположению, что $\mu_1=\mu_2$, то есть $D=0$. Критерием для проверки H_0-гипотезы служит отношение

$$t = \frac{d - (\mu_1 - \mu_2)}{s_d}$$

где t — переменная величина, следующая t-распределению Стьюдента с числом степеней свободы $k = n_1 + n_2 — 2$, а s_D - ошибка указанной разности, а n_1 и n_2 - объемы первой и второй выборок соответственно.

Так как, согласно H_0-гипотезе, $\mu_1 = \mu_2$, то t-критерий выражается в виде отношения разности выборочных средних к своей ошибке, т. е.

$$t = \frac{d}{s_d}.$$

H_0-гипотезу отвергают, если фактически установленная величина t-критерия (обозначаемая $t_ф$) превзойдет или окажется равной критическому значению t_{kp} этой величины для принятого уровня значимости α и числа степеней свободы $k = n_1 + n_2 — 2$, т. е. при условии $t_ф \geq t_{kp}$.

Ошибку разности средних s_D определяют по формуле:

$$s_D = \sqrt{\frac{\sum(x_i - \overline{x_1})^2 + \sum(x_j - \overline{x_2})^2}{n_1 + n_2 - 2} \cdot \left(\frac{n_1 + n_2}{n_1 n_2}\right)}$$

Пример. Изучали влияние кобальта на массу тела кроликов. Опыт проводили на двух группах животных: опытной и контрольной. Были исследованы кролики в возрасте от полутора до двух месяцев, массой тела 500—600 г. Опыт продолжался полтора месяца. Животных обеих групп содержали на одном и том же кормовом рационе. Однако опытные кролики в отличие от контрольных ежедневно получали добавку к рациону в виде водного раствора по 0,06 г хлористого кобальта на 1 кг живой массы тела. За время опыта животные дали следующие прибавки живой массы тела:

Привесы, г		Отклонения от средней		Квадраты отклонений	
опыт	контроль	опыт $(x_i - \overline{x_1})$	контроль $(x_j - \overline{x_2})$	опыт $(x_i - \overline{x_1})^2$	контроль $(x_j - \overline{x_2})^2$
580	504	58	22	3364	484
692	560	54	34	2916	1 156
700	420	62	106	3844	11236
621	600	17	74	289	5 476
640	580	2	54	4	2916
561	530	77	4	5929	16
680	490	42	36	1764	1 296
630	580	8	54	64	2916
	470		56	ч	3136
Σ = 5104	Σ = 4734	—	—	Σ =18 174	Σ = 28 632
$\overline{x_1}$ =638	$\overline{x_2}$ =526	—	—	Σ =46 806	

Средние арифметические привесов:

в опыте $\overline{x_1}$ = 5104/8= 638 г,

в контроле $\overline{x_2}$=4734/9 = 526 г. Разница $\overline{x_1} - \overline{x_2} = d$ = 112 г. Чтобы установить, достоверна или случайна эта разница, нужно определить ошибку разности средних:

$$s_D = \sqrt{\frac{46806}{8+9-2} \cdot \left(\frac{8+9}{8 \cdot 9}\right)} = \sqrt{736{,}8} = 27{,}14.$$

Отсюда $t_ф$ = 112/27,14= 4,1.

По таблице для уровня значимости α = 0,01 и числа степеней свободы k = 9+8—2= 15 находим t_{kp} = 2,95. Так как $t_ф > t_{kp}$, нулевая гипотеза опровергается на высоком уровне значимости (P<0,01). Разница между средними величинами опыта и контроля оказалась в высшей степени достоверной.

Задачи и упражнения:

1) 7 студентов из 10 сдавали практические работы так, как будто они были списаны друг у друга. На уровне значимости 0,05 определите, случайно ли это, или студенты действительно списывали.

2) На уровне значимости α = 0,025 проверить гипотезу о нормальном распределении генеральной совокупности , если известны эмпирические и теоретические частоты

$m_{(эмп)i}$	5	10	20	25	14	3
$m_{(теор)i}$	6	14	28	18	8	3

3) Техническая норма предусматривает в среднем 40 с на выполнение определённой технологической операции на конвейере. От работающих на этой операции поступили жалобы, что они в действительности затрачивают на неё больше времени. Для проверки жалобы произведены измерения времени выполнения этой операции у 16 работниц, занятых на ней и получено среднее время 42 с . Можно ли по имеющимся хронометрическим данным на уровне значимости α = 0,01 отклонить гипотезу о том, что среднее время выполнения этой операции соответствует норме, если: а) исправленное выборочное отклонение s – 3,5с; б) выборочное среднее отклонение – 3,5 с?

4) Экономический анализ производительности труда предприятий позволил выдвинуть гипотезу о наличии 2 типов предприятий с различной средней величиной показателя производительности труда. Выборочное обследование 42 предприятий 1-й группы дало следующие результаты: средняя производительность труда 119 деталей. Выборочное обследование 35 предприятий 2-й группы показало, что средняя производительность труда составляет 107 деталей. Генеральные дисперсии соответственно равны 126,91 (дет.^{2}) и 136,1 (дет.^{2}). Считая, что выборки извлечены из нормально распределённых генеральных совокупностей X и Y ,на уровне значимости 0.05, проверьте, случайно ли полученное различие средних показателей производительности труда или же имеются 2 типа предприятий с различной средней величиной производительности труда.

5) Предполагается, что применение новых компьютерных технологий сократит время решения задач. Хронометраж времени решения 9 задач без компьютерных технологий дал следующие результаты: среднее время решения 57 минут, исправленная выборочная дисперсия s_x^2 = 186,2 (мин^2). Среднее время решения 15 задач с применением компьютерных технологий 52 минуты, а исправленная выборочная дисперсия s_y^2 = 166,4 (мин^2). На уровне значимости α = 0,01 ответьте позволило ли применение компьютерных технологий сократить время решения задач.

6) Партия изделий принимается в том случае, если вероятность того, что изделие окажется соответствующим стандарту, составляет не менее 0,97. Среди случайно отобранных 200 изделий проверяемой партии оказалось 193 соответствующих стандарту. Можно ли на уровне значимости $\alpha = 0{,}02$ принять партию

7) Для завода изготавливают однотипные детали. Для оценки их качества сделаны выборки из продукции этих заводов и получены следующие результаты

Выборки	Завод №1	Завод №2
Объём выборки	n_1	n_2
Число бракованных деталей	m_1	m_2

На уровне значимости $\alpha = 0{,}025$ определите, имеется ли существенное различие в качестве изготавливаемых деталей?

8) Компания, производящая средства для похудения, утверждает, что приём таблеток в сочетании со специальной диетой позволяет сбросить в среднем в неделю 400гр веса. Случайным образом отобраны 25 человек, использующих эту терапию, и обнаружено, что в среднем еженедельная потеря в весе составила 430гр со средним квадратическим отклонением 110гр. Проверьте гипотезу о том, что средняя потеря в весе составляет 400гр. Уровень значимости $\alpha = 0{,}05$

9) Компания утверждает, что новый вид зубной пасты лучше предохраняет зубы, чем зубные пасты других фирм. Для проверки в случайном порядке выбраны 400 детей, пользовавшихся новой пастой и 300 детей, которые пользовались зубными пастами других фирм. После окончания эксперимента было выяснено, что у 30 детей, использующих новую пасту, и 25 детей из другой группы появились новые признаки кариеса. Имеются ли у компании достаточные основания для утверждения о том, что новый сорт зубной пасты эффективнее. Принять уровень значимости $\alpha = 0{,}05$

10) Инженер по контролю качества проверяет среднее время горения нового вида электроламп. Для проверки в порядке случайной выборки было отобрано 100 ламп, среднее время горения которых составило 1075 часов. Среднее квадратическое отклонение времени горения составляет 100 часов. Используя уровень значимости $\alpha = 0.05$, проверьте гипотезу о том, что среднее время горения ламп – более 1000 часов.

11) Компания, выпускающая в продажу новый сорт кофе, провела проверку вкусов покупателей по случайной выборке из 400 человек и выяснила, что 220 из них предпочли новый сорт всем остальным. Проверьте на уровне значимости $\alpha = 0{,}01$ гипотезу о том, что по крайней мере 52% потребителей предпочтут новый сорт кофе.

Практическое занятие 10-11. Статистическая проверка гипотез. Непараметрические критерии.

Цель: сформировать представление о методах и принципах проверки статистических гипотез.

Теоретическая часть:

Правильное применение параметрических критериев для проверки статистических гипотез основано на предположении о нормальном распределении совокупностей, из которых взяты сравниваемые выборки. Однако это не всегда имеет место, так как не все биологические признаки распределяются нормально. Немаловажным является и то обстоятельство, что исследователю приходится иметь дело не только с количественными, но и с качественными признаками, многие из которых выражаются порядковыми номерами, индексами и другими условными знаками. В таких случаях необходимо использовать *непараметрические критерии.*

Известен целый ряд непараметрических критериев, среди которых видное место занимают так называемые *ранговые критерии,* применение которых основано на ранжировании членов сравниваемых групп. При этом сравниваются не сами по себе члены ранжированных рядов, а их порядковые номера, или ранги. Далее мы рассмотрим некоторые непараметрические критерии, применяемые для проверки нулевой гипотезы при сравнении как независимых, так и зависимых выборочных групп.

U **- критерий Уилкоксона (Манна—Уитни).** Гипотезу о принадлежности сравниваемых независимых выборок к одной и той же генеральной совокупности или к совокупностям с одинаковыми параметрами, т. е. H_0-гипотезу, можно проверить с помощью *рангового критерия Уилкоксона (Манна—Уитни).*

Для расчета U -критерия необходимо:

1. Расположить числовые значения сравниваемых выборок в возрастающем порядке в один общий ряд и пронумеровать члены общего ряда от 1 до $N = n_1 + n_2$. (Эти номера и будут «рангами» членов ряда.)

2. Отдельно для каждой выборки найти суммы рангов и определить величины,

$$U_1 = R_1 - \frac{n_1(n_1+1)}{2}$$

и

$$U_2 = R_2 - \frac{n_2(n_2+1)}{2}$$

которые отображают связь между суммами рангов первой и второй выборки.

3. В качестве U -критерия использовать меньшую величину $U_ф$, которую сравнить с табличным значением U_{kp} Условием для сохранения принятой H_0-гипотезы служит неравенство $U_ф > U_{kp}$. Критические точки U -критерия U_{kp} для n_1, n_2 и принимаемого уровня значимости α содержатся в таблице 3 Приложения.

Пример. На двух группах лабораторных мышей – опытной (n_1 = *9)* и контрольной (n_2 = 11) – изучали воздействие на организм нового препарата. Испытание продолжалось один месяц. После этого масса тела животных, выраженная в граммах, варьировала следующим образом:

В опытной группе *80, 76, 75, 64, 70, 68, 72, 79, 83*.

В контрольной группе 70,78, 60, 80, 62, 68, 73, 60, 71, 66, 69.

Вычислим по выборкам: $\overline{x_1}$ =74,1 и $\overline{x_2}$ =68,8.

Проверим с помощью U -критерия, является ли разность в массе тела между опытной и контрольной группами мышей статистически достоверной.

Суммируя ранги отдельно для каждой группы, находим:
R_1= 4+6+9+12+14+15+17+19+20=112;
R_2= 1+2+3+5+7+8+10+11+13+16+18=94.
Подставляем эти данные в формулы:

$$U_1 = 112 - \frac{9 \cdot 10}{2} = 67; \qquad U_2 = 94 - \frac{11 \cdot 12}{2} = 22$$

Меньшую величину U_2 = 22 сравниваем с табличным U_{kp} значением для $n_1=9$, $n_2=11$ и уровня значимости $\alpha=0,01$, которое равно $U_{kp}=19$.

Поскольку $U_ф > U_{kp}$, отвергнуть проверяемую H_0-гипотезу нельзя. Следовательно, различия, наблюдаемые между этими выборками, статистически недостоверны. Выборки не имеют значимых отличий.

Критерий знаков *z*. В тех случаях, когда результаты наблюдений выражаются не числами, а качественными признаками, принимающими два различных значения (помечаем их знаками плюс (+) и минус (—)), различия между попарно связанными членами сравниваемых выборок оценивают с помощью *критерия знаков z*. Конструкция этого критерия базируется на весьма простых соображениях: если попарно сравниваемые значения двух зависимых выборок существенно не отличаются друг от друга, то число плюсовых и минусовых разностей окажется совершенно одинаковым; если же заметно преобладают плюсы или минусы, это будет указывать на положительное или отрицательное действие изучаемого фактора на результативный признак. Большее число однозначных разностей служит в качестве фактически найденной величины *z*-критерия знаков. При этом нулевые разности, т. е. случаи, не давшие ни положительного, ни отрицательного результата, обозначаемые цифрой 0, в расчет не принимают и число парных наблюдений соответственно уменьшается.

Как и всякий другой выборочный показатель, *z*-критерий знаков является величиной случайной; он служит для проверки H_0-гипотезы, т. е. предположения о том, что совокупности, из которых взяты сравниваемые выборки, имеют одинаковые функции распределения. H_0-гипотеза отвергается, если $z_ф \geq z_{kp}$ для принятого уровня значимости α и числа парных наблюдений *n*, взятых без нулевых разностей. Критические точки z_{kp} для двух уровней значимость и числа парных наблюдений содержатся в таблице.

Задачи и упражнения:

Результаты тестирования по 30-бальной шкале для группы X и группы Y представлены в таблице. Сравнить эффективность двух методов обучения студентов в двух группах для уровня статистической значимости β = 5%.

X	18	10	7	15	14	11	13			
Y	15	20	10	8	16	10	19	7	15	14	29

Практическое занятие 12. Проверка гипотез о законах распределения.

Цель: формирование представления о принципах проверки статистических гипотез.

Теоретическая часть:

Если закон распределения неизвестен, но есть основания предположить, что он имеет определённый вид (назовём его А), то проверяют нулевую гипотезу: генеральная совокупность распределена по закону А.

Проверка гипотезы о предполагаемом законе неизвестного распределения производится при помощи специально подобранной случайной величины – критерия согласия.

Критерием согласия называют критерий проверки гипотезы о предполагаемом законе неизвестного распределения.

Критерий согласия Пирсона. Имеется несколько критериев согласия: χ^2(«хи-квадрат») К. Пирсона, Колмогорова, Смирнова и др. Ограничимся описанием применения критерия Пирсона к проверке гипотезы о нормальном распределении генеральной совокупности (критерий аналогично применяется и для других распределений, в этом состоит его достоинство). С этой целью будем сравнивать эмпирические (наблюдаемые) и теоретические (вычисленные в предположении нормального распределения) частоты.

Обычно эмпирические и теоретические частоты различаются. Например:

эмп. частоты…….6 13 38 74 106 85 30 10 4

теорет. частоты…3 14 42 82 99 76 37 11 2

Случайно ли расхождение частот? Возможно, что расхождение случайно (незначимо) и объясняется либо малым числом наблюдений, либо способом их группировки, либо другими причинами. Возможно, что расхождение частот неслучайно (значимо) и объясняется тем, что теоретические частоты вычислены, исходя из неверной гипотезы о нормальном распределении генеральной совокупности.

Критерий Пирсона отвечает на поставленный выше вопрос. Правда, как и любой критерий, он не доказывает справедливость гипотезы, а лишь устанавливает на принятом уровне значимости её согласие или несогласие с данными наблюдений.

Итак, пусть по выборке объёма n получено эмпирическое распределение:

варианты…… x_i x_1 x_2 x_s

эмп. частоты... n_i n_1 n_2 n_s

Допустим, что в предположении нормального распределения генеральной совокупности вычислены теоретические частоты n_i' (например, так, как в следующем параграфе). При уровне значимости α требуется проверить нулевую гипотезу: генеральная совокупность распределена нормально.

В качестве критерия проверки нулевой гипотезы примем случайную величину

$$\chi^2 = \frac{\sum (n_i - n_i')^2}{n_i'}. \quad (*)$$

Эта величина случайная, т.к. в различных опытах она принимает различные, заранее неизвестные значения. Ясно, что чем меньше различаются эмпирические и теоретические частоты, тем меньше величина критерия (*), и, следовательно, он в известной степени характеризует близость эмпирического и теоретического распределений.

Заметим, что возведением в квадрат разностей частот устраняют возможность взаимного погашения положительных и отрицательных разностей. Делением на n_i' достигают уменьшения каждого из слагаемых; в противном случае сумма была бы настолько велика, что приводила бы к отклонению нулевой гипотезы даже и тогда, когда она справедлива. Разумеется, приведённые соображения не являются обоснованием выбранного критерия, а лишь пояснением.

Доказано, что при $n \to \infty$ закон распределения случайной величины (*) независимо от того, какому закону распределения подчинена генеральная совокупность, стремится к закону распределения χ^2 с k степенями свободы. Поэтому случайная величина (*) обозначена через χ^2, а сам критерий называют критерием согласия «хи квадрат».

Число степеней свободы находят по равенству $k = s - 1 - r$, где s – число групп (частичных интервалов) выборки; r - число параметров предполагаемого распределения, которые оценены по данным выборки.

В частности, если предполагаемое распределение – нормальное, то оценивают два параметра (математическое ожидание и среднее квадратическое отклонение), поэтому r=2 и число степеней свободы $k = s - 1 - r = s - 1 - 2 = s - 3$.

Если, например, предполагают, что генеральная совокупность распределена по закону Пуассона, то оценивают один параметр λ, поэтому r =1 и

$k = s - 2$.

Поскольку односторонний критерий более «жёстко» отвергает нулевую гипотезу, чем двусторонний, построим правостороннюю критическую область, исходя из требования, чтобы вероятность попадания критерия в эту область в предположении справедливости нулевой гипотезы была равна принятому уровню значимости α:

$$P\left[\chi^2 > \chi^2_{кр}(\alpha; k)\right] = \alpha.$$

Таким образом, правосторонняя критическая область определяется неравенством $\chi^2 > \chi^2_{кр}(\alpha; k)$, а область принятия нулевой гипотезы – неравенством $\chi^2 < \chi^2_{кр}(\alpha; k)$.

Обозначим значение критерия, вычисленное по данным наблюдений, через $\chi^2_{набл}$ и сформулируем правило проверки нулевой гипотезы.

Правило. Для того, чтобы при заданном уровне значимости проверить нулевую гипотезу H_0: генеральная совокупность распределена нормально, надо сначала вычислить теоретические частоты, а затем наблюдаемое значение критерия:

$$\chi^2_{набл} = \frac{\sum (n_i - n'_i)^2}{n_i} \quad (**)$$

и по таблице критических точек распределения χ^2, по заданному уровню значимости α и числу степеней свободы $k = s - 3$ найти критическую точку $\chi^2_{кр}(\alpha; k)$.

Если $\chi^2_{набл} < \chi^2_{кр}$ - нет оснований отвергнуть нулевую гипотезу.

Если $\chi^2_{набл} > \chi^2_{кр}$ - нулевая гипотеза отвергается.

Замечание 1. Объём выборки должен быть достаточно велик, во всяком случае, не менее 50. Каждая группа должна содержать не менее 5-8 вариант; малочисленные группы следует объединять в одну, суммируя частоты.

Пример. При уровне значимости 0,05 проверить гипотезу о нормальном распределении генеральной совокупности, если известны эмпирические и теоретические частоты:

эмп. частоты……6 13 38 75 106 85 30 14

теорет. частоты…3 14 42 82 99 76 37 13

Вычислим $\chi^2_{набл}$, для чего составим расчётную таблицу.

1	2	3	4	5	6	7	8
i	n_i	n'_i	$n_i - n'_i$	$(n_i - n'_i)^2$	$\frac{(n_i - n'_i)^2}{n'_i}$	n_i^2	$\frac{n_i^2}{n'_2}$
1	6	3	3	9	3	36	12
2	13	14	-1	1	0,07	169	12,07
3	38	42	-4	16	0,38	1444	34,38
4	74	82	-8	64	0,78	5476	66,78
5	106	99	7	49	0,49	11236	113,49
6	85	76	9	81	1,07	7225	95,07
7	30	37	-7	49	1,32	900	24,32
8	14	13	1	1	0,08	196	15,08
$\sum$	366	366			$\chi^2_{набл} =$ $= 7{,}19$		373,19

Контроль: $\chi^2_{набл}$=7,19:

$\left[\sum \frac{n_i^2}{n'_i}\right] - n = 373{,}19 - 366 = 7{,}19.$ Вычисления произведены правильно.

Найдём число степеней свободы, учитывая, что число групп выборки (число различных вариант) $s = 8;\ k = 8 - 3 = 5.$

По таблице критических точек распределения χ^2, по уровню значимости α=0,05 и числу степеней свободы $k = 5$ находим $\chi^2_{кр}(0{,}05; 5) = 11{,}1.$

Так как $\chi^2_{набл} < \chi^2_{кр}$ - нет оснований отвергнуть нулевую гипотезу. Другими словами, расхождение эмпирических и теоретических частот незначимое. Следовательно, данные наблюдений согласуются с гипотезой о нормальном распределении генеральной совокупности.

Задачи и упражнения:

Используя критерий Пирсона, при уровне значимости 0,05 проверить, согласуется ли гипотеза о нормальном распределении генеральной совокупности X по результатам выборки:

X 0,3 0,5 0,7 0,9 1,1 1,3 1,5 1,7 1,9 2,1 2,3

N 7 9 28 27 30 26 21 25 22 9 5

Практическое занятие 13. Корреляционная зависимость. Коэффициент корреляции.

Цель: формирование представления об элементах корреляционного анализа.

Теоретическая часть:

Зависимость между значениями одной случайной величины и условным математическим ожиданием другой случайной величины носит название статистической.

Чтобы изучить статистическую зависимость, нужно знать условное математическое ожидание случайной величины. Для его оценки необходимо знать аналитический вид двумерного распределения (X,Y). Однако, суждение об аналитическом виде двумерного распределения, сделанного по отдельной ограниченной по объёму выборке, может привести к серьёзным ошибкам. Поэтому идут на упрощение и переходят от условного математического ожидания случайной величины к условному среднему значению, т.е. принимают, что $M(Y)_x = \overline{Y}_x$

Зависимость между значениями одной случайной величины и условным средним значением другой случайной величины носит название **корреляционной** (от англ. correlation - согласование, связь, взаимосвязь, соотношение, взаимозависимость); термин впервые введен Гальтоном в 1888г.

Парный коэффициент корреляции Пирсона (1896 г.) изменяется в пределах от -1 до +1. Значение 0,00 интерпретируется как отсутствие корреляции. Корреляция определяет степень, с которой значения двух переменных пропорциональны друг другу.

Определяют две черты зависимости между переменными: величину зависимости и надежность зависимости.

Надежность зависимости – менее наглядные понятия, чем величина зависимости, однако чрезвычайно важна. Оно непосредственно связано с репрезентативностью той определенной выборки, на основе которой строятся выводы. Другими словами надежность говорит, насколько вероятно, что зависимость подобная найденной, будет вновь обнаружена (подтвердится) на данных другой выборки, извлеченной из той же самой популяции. Если исследование удовлетворяет некоторым специальным критериям, то надежность найденных зависимостей между переменными выборки можно количественно оценить и представить с помощью стандартной статистической меры (называемой p-уровень, или статистический уровень значимости).

Статистическая значимость результата представляет собой оцененную меру уверенности в его правильности. Уровень значимости или p-уровень, - это показатель, находящийся в убывающей зависимости от надежности результата. Более высокий p-уровень соответствует более низкому уровню доверия к найденной в выборке зависимости между переменными. Именно p-уровень представляет собой вероятность ошибки, связанной с распространением наблюдаемого результата на всю популяцию. Чем слабее зависимость между переменными, тем большего объема требуется выборка, чтобы значимо ее обнаружить. Другими словами, если зависимость между переменными почти отсутствует, объем выборки, необходимый для ее значимого обнаружения, почти равен объему всей популяции, которой предполагается бесконечным.

Рабочие формулы коэффициентов корреляции применяют с учетом того, с какой выборкой (большей или малой) мы имеем дело.

Например, <u>для малых выборок</u> удобнее всего пользоваться следующей формулой:

$$r = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sqrt{q_x \cdot q_y}} \quad (1)$$, где

$q_x = \sum x^2 - \frac{\left(\sum x\right)^2}{n}$ и $q_y = \sum y^2 - \frac{\left(\sum y\right)^2}{n}$, где x-варианты первого признака;

y-варианты второго признака; n-число наблюдений в выборке.

Задачи и упражнения:

Определить зависимость содержания витамина С в продукте в зависимости от длительности его обработки при определенной температуре по результатам n=6 измерений (экспериментов). Какое количество витамина С (в долях к первоначальному) будет содержаться в продукте при обработке его в течение t^*_1 и t^*_2 часов. Построить график полученной зависимости и изобразить на координатной плоскости результаты измерений.

1.

x_i	0,11	0,21	0,31	0,41	0,51	0,61	t^*_1	t^*_2
y_i	0,97	0,94	0,9	0,86	0,83	0,78	0,25	0,74

2.

x_i	0,21	0,22	0,23	0,24	0,25	0,26	t^*_1	t^*_2
y_i	0,97	0,94	0,9	0,86	0,83	0,78	0,25	0,74

3.

x_i	0,14	0,24	0,34	0,44	0,54	0,64	t^*_1	t^*_2
y_i	0,97	0,94	0,9	0,86	0,83	0,78	0,25	0,74

Практическое занятие 14. Линейная регрессия. Коэффициенты регрессии.

Цель: формирование представления об элементах и принципах регрессионного анализа.

Теоретическая часть:

Для изучения корреляционных связей большое значение имеет коэффициент регрессии ρ, который показывает, насколько в среднем изменяется признак (Х), если коррелирующий с ним признак (У) изменяется на определенную величину.

Коэффициент регрессии в конкретной выборке имеет два значения, а именно: ρ_{xy} и ρ_{yx}, т.е. прямое и обратное влияние признаков друг на друга. Формула для расчета коэффициента имеет вид:

$$\rho_{xy} = r\frac{S_x}{S_y}; \ \rho_{yx} = r\frac{S_y}{S_x}$$

Статистическую зависимость Y от X описывают с помощью уравнения вида

$$M(Y)_x = f(x)$$

где $M(Y)_x$ - условное математическое ожидание величины Y, соответствующее данному значению x; x – отдельные значения величины X; $f(x)$ - некоторая функция. Это уравнение называется уравнением регрессии Y на X.

Обратную статистическую зависимость можно описать уравнением регрессии X на Y:

$$M(X)_y = \varphi(y)$$

где $M(X)_y$ - условное математическое ожидание величины X, соответствующее данному значению y случайной величины Y; $\varphi(y)$ - некоторая функция.

Функции $f(x)$ и $\varphi(y)$ называют соответственно регрессиями Y на X и X на Y, а их графики – линиями регрессии Y на X и X на Y. Уравнения регрессии выражают математическое ожидание случайной величины Y (или X) для случая, когда другая переменная принимает определенное число.

В зависимости от вида уравнений регрессии и формы соответствующих линий регрессии говорят о различной форме статистической зависимости между изучаемыми величинами – линейной, квадратичной, показательной и т.д.

Если функции $f(x), \varphi(y)$ линейные, т.е. уравнения регрессии можно представить в виде:

$$M(Y)_x = Ax + B$$

$$M(X)_y = Cy + D,$$

где A,B,C,D – некоторые параметры, то описываемые этими уравнениями зависимости Y от X и X от Y называются линейными; линии регрессии при этом – прямые. Если линия регрессии не является прямой, то такую зависимость называют нелинейной.

Для характеристики формы связи между двумя случайными величинами, полученными в результате выборочных наблюдений, используют корреляционную зависимость $\overline{Y_x} = f(x)$ (или $\overline{X_y} = \varphi(y)$). Уравнения, описываемые подобной зависимостью, называют выборочными уравнениями регрессии.

Если функции $f(x)$, $\varphi(y)$ линейные, то выборочные уравнения линейной регрессии Y на Xи X на Y можно представить в виде:

$$\overline{Y_x} = \rho_{yx} x + b$$

$$\overline{X_y} = \rho_{xy} y + d,$$

где $\overline{Y_x}$ и $\overline{X_y}$ - условные средние значения величин Y и X, параметры b и d - оценки B и D, ρ_{yx} и ρ_{xy} - выборочные оценки коэффициентов A и C.

Угловые коэффициенты ρ_{yx} и ρ_{xy} линий регрессии носят названия выборочных коэффициентов регрессии Y на X и X на Y соответственно. Они определяются как:

$$\rho_{yx} = \frac{\overline{XY} - \overline{X}\,\overline{Y}}{S_x^2}; \quad \rho_{xy} = \frac{\overline{XY} - \overline{X}\,\overline{Y}}{S_y^2},$$

где $S_x^2 = \overline{X^2} - (\overline{X})^2$, $S_y^2 = \overline{Y^2} - (\overline{Y})^2$.

Задачи и упражнения:

Найти выборочное уравнение прямой $\bar{y}_x - \bar{y} = r_B \frac{\sigma_y}{\sigma_x}(x - \bar{x})$ регрессии У на Х по данной корреляционной таблице.

1.

Y	X						n_y
	10	15	20	25	30	35	
6	4	2	–	–	–	–	6
12	–	6	2	–	–	–	8
18	–	–	5	40	5	–	50
24	–	–	2	8	7	–	17
30	–	–	–	4	7	8	19
n_x	4	8	9	52	19	8	n=100

2.

Y	X						n_y
	3	13	23	33	43	53	
8	4	2	–	–	–	–	6
10	–	6	2	–	–	–	8
12	–	–	5	40	5	–	50
14	–	–	2	8	7	–	17
16	–	–	–	4	7	8	19
n_x	4	8	9	52	19	8	n=100

3.

Y	X						n_y
	11	15	19	23	27	31	
10	4	2	–	–	–	–	6
20	–	6	2	–	–	–	8
30	–	–	5	40	5	–	50
40	–	–	2	8	7	–	17
50	–	–	–	4	7	8	19
n_x	4	8	9	52	19	8	n=100

4.

Y	X						n_y
	20	23	26	29	32	35	
16	4	2	–	–	–	–	6
19	–	6	2	–	–	–	8
22	–	–	5	40	5	–	50
25	–	–	2	8	7	–	17
28	–	–	–	4	7	8	19
n_x	4	8	9	52	19	8	n=100

5.

Y	X						n_y
	14	18	22	26	30	34	
5	4	2	–	–	–	–	6
10	–	6	2	–	–	–	8
15	–	–	5	40	5	–	50
20	–	–	2	8	7	–	17

25	–	–	–	4	7	8	19
n_x	4	8	9	52	19	8	n=100

Практическое занятие 15-16. Однофакторный дисперсионный анализ.

Цель: формирование представления об элементах и принципах дисперсионного анализа.

Теоретическая часть:

Предположим, что имеется K выборок с объемами n_1, $n_2,...,n_k$, $N=n_1+n_2+...+n_k$, и наблюдения можно представить в виде $x_{ij}=a_j+\varepsilon_{ij}$, где i - номер наблюдения в выборке; j - номер выборки; a_j - групповые математические ожидания; ε_{ij} - случайные ошибки с $M\varepsilon_{ij}=0$, о которых предполагается, что они независимы и одинаково расположены.

Подобная ситуация возникает, когда существует некий фактор, принимающий K различных значений (называемых уровнями), и каждая группа объектов, чьи признаки мы примеряем, подвергается воздействию определенного уровня этого фактора. Методы математической статистики, изучающие воздействие одного фактора на объекты и их признаки, называют в совокупности <u>однофакторным анализом.</u>

Предполагается, что ошибки нормально распределены: $\varepsilon_{ij}\in N(0,\sigma^2)$. Тогда можно изучать влияние фактора, вычисляя дисперсии некоторых величин. Совокупность этих методов называют <u>однофакторным дисперсионным анализом.</u>

Основной гипотезой, нуждающейся в проверке, является гипотеза о равенстве групповых средних $H_0: a_1=a_2=...=a_k$. Иными словами, проверяют гипотезу о том, что фактор вообще не влияет на наблюдения. В случае нормальных ошибок ее можно проверить, вычислив две разные оценки дисперсии.

Рассмотрим группу экспериментальных животных, подвергнутых ультрафиолетовому облучению. В процессе эксперимента измерялась температура тела животных. Результаты измерений были занесены в таблицу:

№ испытания	Уровень фактора А (мощность ультрафиолетового облучения)		
	А1	А2	А3
1	37,4	37,8	38,0
2	37,3	37,9	37,9
3	37,0	37,5	38,4
4	36,9	37,4	38,3
$\bar{x}_j$	37,15	37,65	38,15

Физический фактор А (ультрафиолетовое излучение) имеет $m=3$ постоянных уровней (3 различных мощности облучения). На всех уровнях распределения случайной величины Х (температуры тела животного) предполагается нормальным, а дисперсии одинаковыми, хотя и неизвестными.

В данном эксперименте число проведенных наблюдений при действии каждого из уровней фактора одинаково.

Все значения величины X, наблюдаемые при каждом фиксированном уровне фактора A_j, составляют группу, и в последней строке таблицы представлены соответствующие выборочные групповые средние, вычисленные по формуле

$$\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}.$$

Здесь n – число испытаний, j – номер столбца, i - номер строки, в которой расположено данное значение случайной величины. Общая средняя арифметическая всех nm наблюдений находится как

$$\bar{x} = \frac{1}{m}\sum_{j=1}^{m} \bar{x}_j.$$

Введем следующие понятия:

Факторная сумма квадратов отклонений групповых средних от общей средней $\bar{x}$, которая характеризует рассеивание «между группами» (т.е. рассеивание за счет исследуемого фактора):

$$S_{факт} = n\sum_{j=1}^{m} (\bar{x}_j - \bar{x})^2,$$

Остаточная сумма квадратов отклонений наблюдаемых значений группы от своей групповой средней $\bar{x}_j$, которая характеризует рассеивание «внутри групп» (за счет случайных причин):

$$S_{ост} = \sum_{i=1}^{n} (x_{i1} - \bar{x}_1)^2 + \sum_{i=1}^{n} (x_{i2} - \bar{x}_2)^2 + \ldots + \sum_{i=1}^{n} (x_{im} - \bar{x}_m)^2.$$

Общая сумма квадратов отклонений наблюдаемых значений от общей средней $\bar{x}$:

$$S_{общ} = \sum_{j=1}^{m}\sum_{i=1}^{n} (x_{ij} - \bar{x})^2,$$

Можно доказать следующее равенство:

$$S_{общ} = S_{факт} + S_{ост}.$$

С помощью $S_{общ}$, $S_{факт}$, $S_{ост}$ производится оценка общей, факторной и остаточной дисперсий:

$$s^2_{общ} = \frac{1}{mn-1} S_{общ},$$

$$s^2_{факт} = \frac{1}{m-1} S_{факт},$$

$$s^2_{ост} = \frac{1}{m(n-1)} S_{ост}.$$

В основе однофакторного дисперсионного анализа лежит тесная связь между различием в групповых средних $\bar{x}_j$ и соотношением между двумя видами дисперсий – факторной, которая характеризует влияние фактора А на величину X, и остаточной, которая характеризует влияние случайных причин. Сравнивая факторную дисперсию с остаточной по величине их отношения судят, насколько сильно проявляется влияние фактора.

Для сравнения двух дисперсий используют **показатель критерия Фишера** $F_{эксп} = s^2_{факт} / s^2_{ост}.$

При этом при заданном уровне значимости проверяют <u>нулевую гипотезу о равенстве факторной и остаточной дисперсии</u> (изучаемый фактор не вызывает изменчивости признака) при конкурирующей гипотезе об их неравенстве (изучаемый фактор вызывает изменчивость признака).

По таблице критических значений распределения Фишера – Снедекора при уровне значимости, равном половине заданного уровня α, находят критическое значение $F_{kp}(\alpha/2; k_1; k_2)$. Здесь $k_1 = m-1;\ k_2 = m(n-1)$. Если $F_{эксп} < F_{kp}$, нулевую гипотезу считают согласующейся с результатами наблюдений. Если $F_{эксп} > F_{kp}$, то эту гипотезу отвергают в пользу конкурирующей.

<u>Замечание.</u> Если окажется, что $S^2_{факт} < S^2_{ост}$, следует сделать вывод об отсутствии влияния фактора А на Х.

Если проверка покажет значимость различий между $S^2_{факт}$ и $S^2_{ост}$,следует сделать вывод о существенном влиянии фактора А на Х.

Обычно для упрощенная расчетов фактурную и остаточную дисперсии рассчитывают не по экспериментальным значениям x_{ij} величины Х, а по значениям $y_{ij} = x_{ij} - C$, где постоянная С представляет собой произвольное число, близкое к среднему значению $\bar{x}$ всех результатов наблюдений.

Вернемся к нашему примеру. Вычтем из всех значений x_{ij} постоянное число С=37,5 близкое к общему среднему $\bar{x}$ =37,51 и составим таблицу:

№ испытания	Уровень фактора А (мощность ультрафиолетового облучения)		
	А1	А2	А3
1	-0,1	0,3	0,5
2	-0,2	0,4	0,4
3	-0,5	0	0,9
4	-0,6	-0,1	0,8
	-0,35	0,15	0,65

Общая средняя будет равна $\bar{x} = \sum_{j=1}^{m} \frac{1}{3}$ [illegible]

Определим значения $S_{факт}$, $S_{ост}$

$S_{факт} = \sum_{j=1}^{m}$ [illegible]

$S_{ост} = \sum_{i=1}^{4}(\ldots)^2 + \sum_{i=1}^{4}(\ldots)^2 + \sum_{i=1}^{4}(\ldots)^2 =$ [illegible]

Определим значения факторной и остаточной дисперсий:

$S^2_{факт} = \frac{1}{m-1} S_{факт} = \frac{1}{3-1} \cdot 2$ [illegible],

$S^2_{ост} = \frac{1}{m(n-1)} S_{ост} = \frac{1}{3(4-1)}$ [illegible].

Так как $S^2_{ост} < S^2_{фак}$, следует проверить значимость их различия. Найдем экспериментальное значение критерия:

$$F_{эксп} = \frac{S^2_{факт}}{S^2_{ост}} \approx 1,63.$$

Сравним его с критическим значением распределения Фишера – Снедекора для уровня значимости 0,05:

$$F_{kp}(m-1; n-m) = F_{kp}(0,05; \text{[illegible]}) = \text{[illegible]}.$$

Поскольку $F_{эксп} > F_{kp}$ можно утверждать, что при уровне значимости $\alpha = 0,05$ рассматриваемый физический фактор оказывает влияние на температуру тела животного.

Критерий Фишера указывает на влияние изучаемого фактора (если $F_{эксп} > F_{kp}$) на изменчивость признака. Однако он не указывает на силу влияния этого фактора. В качестве показателя силы влияния фактора на изменчивость признака используют величину η:

$$\eta = \frac{S^2_{факт} - S^2_{ост}}{S^2_{факт} + (n-1)S^2_{ост}}.$$

Оценим силу влияния ультрафиолетового облучения на повышение температуры тела животных:

$$\eta = \frac{10,057 - 1}{10,057 + (4-1)} = 0,8$$ или 80,5%

Таким образом, влияние ультрафиолетового облучения на повышение температуры тела животных составляет 80,5%, а 19,5% обусловлены случайными причинами.

Задачи и упражнения:

1. Проверить, существенны ли различия содержания загрязняющего вещества на трех уровнях (глубинах взятия проб):

№	Уровни замеров		
	1	2	3
1	1,17	2,28	1,80
2	1,52	2,46	2,38
3	1,90	0,88	2,62
4	1,76	2,03	2,91
5	1,54	1,22	1,60
6	0,63	2,29	2,83
7	2,30	1,80	2,13
8	1,32	1,79	2,06
9	0,94	1,61	2,23
10	1,15	2,30	3,06
11	0,75	2,60	1,86
12	2,49	1,76	1,92
13	2,14	2,14	2,16
14	1,62	2,73	2,27
15	1,40		

Практическое занятие 17-18. Двухфакторный дисперсионный анализ.

Цель: формирование представления об элементах и принципах дисперсионного анализа.

Теоретическая часть:

Двухфакторные комплексы по своей структуре более сложны, чем однофакторные.

Объединение в один статистический комплекс допускается только таких факторов, которые независимы друг от друга (например, тип кормления и доза облучения, возраст и пол и т.д.).

Чтобы построить двухфакторную дисперсионную модель все имеющиеся данные представим в виде таблицы, в которой по строкам - уровни A_i фактора А, по столбцам - уровни B_i фактора В, а в соответствующих клетках, или ячейках, таблицы находятся значения признака x_{ijk} (i=1,2…, m; j=1,2…, l; k=1,2…, n):

B A	B$_1$	B$_2$	…	B$_j$	…	B$_l$
A$_1$	$x_{111},...,x_{11k}$	$x_{121},...,x_{12k}$	…	$x_{1j1},...,x_{1jk}$	…	$x_{1l1},...,x_{1lk}$
A$_2$	$x_{211},...,x_{21k}$	$x_{221},...,x_{22k}$	…	$x_{2j1},...,x_{2jk}$	…	$x_{2l1},...,x_{2lk}$
. . .	…	…	…	…	…	…
A$_j$	$x_{i11},...,x_{i1k}$	$x_{i21},...,x_{i2k}$	…	$x_{ij1},...,x_{ijk}$	…	$x_{il1},...,x_{ilk}$
. . .	…	…	…	…	…	…
A$_m$	$x_{m11},...,x_{m1k}$	$x_{m21},...,x_{m2k}$	…	$x_{mj1},...,x_{mjk}$	…	$x_{ml1},...,x_{mlk}$

Двухфакторная дисперсионная модель имеет вид:

$$x_{ijk} = \mu + F_i + G_j + I_{ij} + \varepsilon_{ijk} \quad (14.1)$$

где x_{ijk} - значение наблюдения в ячейке ij с номером k;

μ - общая средняя;

F_i - эффект, обусловленный влиянием i-го уровня фактора А;

G_j - эффект, обусловленный влиянием j-го уровня фактора В;

I_{ij} - эффект, обусловленный взаимодействием двух факторов, т.е. отклонение от средней по наблюдениям в ячейке ij от суммы первых трех слагаемых в модели (14.1);

ε_{ijk} - возмущение, обусловленное вариацией переменной внутри отдельной ячейки.

Полагаем, что ε_{ijk} имеет нормальный закон распределения $N(0;\sigma^2)$, а все математические ожидания F_*, G_*, I_{i*}, I_{*j} равны нулю.

Групповые средние находятся по формулам:

в ячейке -

$$\bar{x}_{ij*} = \frac{\sum_{k=1}^{n} x_{ijk}}{n} \quad (14.2)$$

по строке -

$$\bar{x}_{i**}=\frac{\sum_{j=1}^{l}\bar{x}_{ij*}}{l}, \qquad (14.3)$$

по столбцу –

$$\bar{x}_{*j*}=\frac{\sum_{i=1}^{m}\bar{x}_{ij*}}{m} \qquad (14.4)$$

Общая средняя

$$\bar{x}_{***}=\frac{\sum_{i=1}^{m}\sum_{j=1}^{l}\bar{x}_{ij*}}{ml} \qquad (14.5)$$

Таблица дисперсионного анализа имеет вид:

Компоненты дисперсии	Сумма квадратов	Число степеней свободы	Средние квадраты
Межгрупповая (фактор А)	$Q_1=ln\sum_{i=1}^{m}(\bar{x}_{i**}-\bar{x}_{***})^2$	*m-1*	$s_1^2=\frac{Q_1}{m-1}$
Межгрупповая (фактор В)	$Q_2=mn\sum_{j=1}^{l}(\bar{x}_{*j*}-\bar{x}_{***})^2$	*l-1*	$s_2^2=\frac{Q_2}{l-1}$
Взаимодействие (АВ)	$Q_3=n\sum_{i=1}^{m}\sum_{j=1}^{l}(\bar{x}_{ij*}-\bar{x}_{i**}-\bar{x}_{*j*}+\bar{x}_{***})^2$	*(m-1)(l-1)*	$s_3^2=\frac{Q_3}{(m-1)(l-1)}$
Остаточная	$Q_4=\sum_{i=1}^{m}\sum_{j=1}^{l}\sum_{k=1}^{n}(x_{ijk}-\bar{x}_{ij*})^2$	*mln-ml*	$s_4^2=\frac{Q_4}{mln-ml}$
Общая	$Q=\sum_{i=1}^{m}\sum_{j=1}^{l}\sum_{k=1}^{n}(x_{ijk}-\bar{x}_{***})^2$	*mln-1*	

Можно показать, что проверка нулевых гипотез H_A, H_B, H_{AB} об отсутствии влияния на рассматриваемую переменную факторов А, В и их взаимодействия АВ осуществляется сравнением отношений $s_1^2/s_4^2, s_2^2/s_4^2, s_3^2/s_4^2$.

Если n=1, т.е. при одном наблюдении в ячейке, то не все нулевые гипотезы могут быть проверены, так как выпадает компонента Q_3 из общей суммы квадратов отклонений, а с ней и средний квадрат s_3^2, ибо в этом случае не может быть речи о взаимодействии факторов.

Пример. В таблице приведены суточные привесы (г) отобранных для исследования 18 поросят в зависимости от метода содержания поросят (фактора А) и качества их кормления (фактор В).

Количество голов в группе (фактор А)	Содержание протеина в корме, г (фактор В)	
	B_1=80	B_2=100
A_1=30	530,540,550	600,620,580
A_2=100	490,510,520	550,540,560
A_3=300	430,420,450	470,460,430

Необходимо на уровне значимости $\alpha = 0,05$ оценить существенность (достоверность) влияния каждого фактора и их взаимодействия на суточный привес поросят.

Решение: Имеем *m*=3, *l*=2, *n*=3. Определим (в г) средние значения привеса:

в ячейках – по (формуле 14.2)

$\bar{x}_{11*} = \frac{530+540+550}{3} = 540$ и аналогично $\bar{x}_{12*} = 600;$

[illegible];

по строкам – по (14.3):

$\bar{x}_{1**} = \frac{540+600}{2} = 570$ и аналогично $\bar{x}_{2**} = 528,4;\ \bar{x}_{3**} = 443,2;$

по столбцам – (14.4):

$\bar{x}_{*1*} = \frac{540+506,7+433,3}{3} = 493,3$ и аналогично $\bar{x}_{*2*} = 534,4;$

Общий средний привес – по (14.5):

$\bar{x} = \frac{[illegible]}{6} = [illegible].$

Все средние значения привеса (г) поместим в таблице

Количество голов в группе (фактор А)	Содержание протеина в корме, г (фактор В)		
	B_1=80	B_2=100	$\bar{x}_{i**}$
A_1=30	$\bar{x}_{11*} = 540,0$	$\bar{x}_{12*} = 600,0$	$\bar{x}_{1**} = 570,0$
A_2=100	$\bar{x}_{21*} = 506,7$	$\bar{x}_{22*} = 550,0$	$\bar{x}_{2**} = 528,4$
A_3=300	$\bar{x}_{31*} = 433,3$	$\bar{x}_{32*} = 453,3$	$\bar{x}_{3**} = 443,3$
$\bar{x}_{*j*}$	$\bar{x}_{*1*} = 493,3$	$\bar{x}_{*2*} = 534,4$	$\bar{x}_{***} = 513,9$

Из таблицы следует, что с увеличением количества голов в группе средний суточный привес поросят в среднем уменьшается, а при увеличении содержания протеина в корме - в среднем увеличивается. Но является ли эта тенденция достоверной или объясняется случайными причинами? Для ответа на этот вопрос вычислим необходимые суммы квадратов отклонений:

[illegible];

[illegible];

[illegible]

Средние квадраты находим делением полученных сумм на соответствующие им число степеней свободы *m-1=2, l-1=1; (m-1)(l-1)=2; mln-ml=18-6=12; mln-1=18-1=17.*

Результаты расчета сведем в таблицу.

Очевидно, данные факторы имеют фиксированные уровни, т.е. мы находимся в рамках модели I. Поэтому для проверки существенности влияния факторов А, В и их взаимодействия АВ необходимо найти отношения:

$$F_A=\frac{s_1^2}{s_4^2}=\frac{25005,5}{250,0}=100,0, \quad F_B=\frac{s_2^2}{s_4^2}=\frac{7605,6}{250,0}=30,4;$$

$F_{AB}=\frac{s_3^2}{s_4^2}=\frac{605,5}{250,0}=2,4$ и сравнить их с табличными значениями (см. приложение 6) соответственно $F_{0,05;2;12}=3,88$, $F_{0,05;1;12}=4,75$, $F_{0,05;2;12}=3,88$. Так как $F_A>F_{0,05;2;12}$ и $F_B>F_{0,05;1;12}$, то влияние метода содержания поросят (фактор А) и качества их кормления (фактор В) является существенным. В силу того что $F_{AB}<F_{0,05;2;12}$, взаимодействие указанных факторов незначимо (на 5%-ном уровне).

Компонента дисперсии	Суммы квадратов	Число степеней свободы	Средние квадраты
Межгрупповая (фактор А)	$Q_1=50011,1$	2	$s_1^2=25005,5$
Межгрупповая (фактор В)	$Q_2=7605,6$	1	$s_2^2=7605,6$
Взаимодействие (АВ)	$Q_3=1211,1$	2	$s_3^2=605,6$
Остаточная	$Q_4=3000,0$	12	$s_4^2=250,0$
Общая	$Q=61827,8$	17	

Задачи и упражнения:

В химической лаборатории проверяется влияние температуры (фактор *A*) и катализатора (фактор *B*) на выход продукта химического синтеза. Полученные результаты приведены в таблице. Проведите двухфакторный дисперсионный анализ. При уровне значимости *a* = 0,05 проверьте гипотезу о влиянии факторов *A* и *B* и их комбинации на указанный признак. Предварительно проверьте по критерию Кочрена равенство дисперсий в группах.

	B_1	B_2	B_3
A_1	16; 19; 17; 16	18; 16; 17; 14	16; 16; 18; 13
A_2	22; 22; 19; 23	18; 19; 23; 24	18; 16; 19; 20
A_3	20; 16; 18; 19	18; 17; 19; 19	20; 20; 16; 16
A_4	23; 20; 22; 23	19; 18; 19; 22	20; 19; 20; 22

4.СПИСОК ЛИТЕРАТУРЫ

Основная литература:

Маталыцкий, М. А. Теория вероятностей, математическая статистика и случайные процессы [Электронный ресурс] : учебное пособие / М. А. Маталыцкий, Хацкевич А. Г.. — Электрон. текстовые данные. — Минск : Вышэйшая школа, 2012. — 720 с. — 978-985-06-2105-4. — Режим доступа: http://www.iprbookshop.ru/20289.html

Дополнительная литература:

1.Математическая статистика. Примеры и задачи [Электронный ресурс] : учебное пособие / М. Ю. Васильчик, А. П. Ковалевский, И. М. Пупышев [и др.]. — Электрон. текстовые данные. — Новосибирск : Новосибирский государственный технический университет, 2011. — 84 с. — 978-5-7782-1721-8. — Режим доступа: http://www.iprbookshop.ru/45382.html.

2. Иванов, В. П. Математическая статистика в инженерных задачах [Электронный ресурс] : курс лекций / В. П. Иванов, А. Ю. Лемин. — Электрон. текстовые данные. — М. : Московский государственный строительный университет, ЭБС АСВ, 2016. — 56 с. — 978-5-7264-1362-4. — Режим доступа: http://www.iprbookshop.ru/62622. html.