Министерство образования и науки Российской Федерации

Амурский государственный университет

А.П. Филимонова, Т.А. Юрьева

МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

Практикум

Благовещенск

2016

ББК 22.172 я73
Ф 53

Рекомендовано
учебно-методическим советом университета

Рецензенты:
О.П. Митрохина, зав. каф. высшей математики ДальГАУ, канд. техн.. наук, доцент;
И.А. Голубева, доцент каф. физики АмГУ, канд. физ.-мат. наук

Филимонова А.П., Юрьева Т.А.
Ф 53 Математическая статистика. Практикум / А.П. Филимонова, Т.А. Юрьева. – Благовещенск: Амурский гос. ун-т, 2016.

Практикум содержит материалы для выполнения лабораторных работ по математической статистике, включает краткие теоретические сведения, примеры, задания. Предназначен для студентов всех направлений подготовки, в учебный план которых входит дисциплина «Теория вероятностей и математическая статистика».

ББК 22.172 я73

В авторской редакции.

ВВЕДЕНИЕ

Математическая статистика – это раздел математики, посвященный методам сбора, анализа и обработки статистических данных для научных и практических целей. Статистические данные представляют собой данные, полученные в результате обследования большого числа объектов и явлений. Обработка эмпирических данных, их систематизация, наглядное представление в форме графиков и таблиц, количественное описание посредством основных статистических показателей, формулировка выводов, имеющих прикладное значение для конкретной области человеческой деятельности – все это относится к методам математической статистики.

Практикум предназначен для проведения лабораторных работ по математической статистике.

Каждая тема, представленная в практикуме, содержит основные понятия, примеры выполнения лабораторных работ и восемь вариантов заданий для самостоятельного выполнения.

Задание для лабораторной работы содержит несколько задач и таблицы с дынными по вариантам. Обучающийся должен решить задачи по варианту, номер которого укажет преподаватель. При выполнении лабораторных работ надо придерживаться следующих правил: работу следует выполнять в тетради чернилами черного или синего цвета, оставляя поля, на обложке тетради должны быть ясно написаны фамилия обучающегося, его инициалы, название дисциплины, номер группы; в заголовке работы должны быть указаны номер лабораторной работы и номер варианта.

Перед решением каждой задачи надо выписать полностью ее условие. Решение задач следует излагать подробно и аккуратно, объясняя и мотивируя все действия по ходу решения и делая необходимые рисунки.

Тема 1. ВЫБОРОЧНЫЙ МЕТОД

Основные понятия

Допустим, что исследуем совокупность однородных объектов относительно некоторого количества признака X, X – случайная величина. Например, в качестве такой совокупности можно рассмотреть партию однотипных деталей, количественный признак X этой совокупности – размер детали. Случайный выбор объекта можно рассматривать как независимое наблюдение над величиной X, в результате которого эта величина принимает конкретные числовые значения x.

При исследовании совокупности однородных объектов можно проводить сплошное обследование. В этом случае обследуют каждый объект совокупности. Но на практике к сплошному обследованию прибегают редко, так как совокупность может содержать слишком большое количество объектов или же изучение совокупности объектов связано с материальными затратами.

Обычно из всей совокупности объектов выбирают ее часть. Выборочный метод – это метод статистического обследования, при котором из совокупности выбирают ограниченное число объектов и подвергают их изучению.

Выборочной совокупностью (выборкой) называют совокупность случайно отобранных объектов.

Генеральной совокупностью называют совокупность объектов, из которых производится выборка.

Объемом совокупности (выборочной или генеральной) называют число объектов этой совокупности.

Для того чтобы по данным выборки можно было судить об интересующем нас признаке генеральной совокупности, необходимо, чтобы объекты выборки правильно его представляли. Иначе, выборка должна правильно представлять пропорции генеральной совокупности, то есть выборка должна быть репрезентативной (представительной).

В силу закона больших чисел можно утверждать, что выборка будет репрезентативной, если ее осуществлять случайно: каждый объект выборки отбирается из генеральной совокупности случайным образом, причем все объекты имеют одинаковую вероятность попасть в выборку.

Пусть из генеральной совокупности извлечена выборка, причем значение x_1 наблюдалось n_1 раз, x_2 – n_2 раз,…, x_k наблюдалось n_k раз. Пусть, далее, $n=\sum_{i=1}^{k} n_i$ – объем выборки. Наблюдаемые значения x_i называют вариантами, а последовательность вариант, записанных в возрастающем порядке, – вариационным рядом. Числа n_i называют частотами, а их отношение к объему выборки $w_i=\frac{n_i}{n}$ – относительными частотами. Величина $R=x_{\max}-x_{\min}$ называется размахом варьирования.

Статистическим распределением выборки называется таблица, задающая перечень вариант и соответствующих им частот (таблица 1) или относительных частот (таблица 2). Такую таблицу называют еще дискретным статистическим рядом.

Таблица 1 – Статистическое распределение частот

x_i	x_1	x_2	x_k
n_i	n_1	n_2	n_k

Здесь $\sum_{i=1}^{k} n_i = n$.

Таблица 2 – Статистическое распределение относительных частот

x_i	x_1	x_2	x_k
w_i	w_1	w_2	w_k

Здесь $\sum_{i=1}^{k} w_i = 1$.

При большом числе наблюдений дискретный статистический ряд не дает полезной информации. В этом случае прибегают к группировке данных. Группировка состоит в том, что область на оси Ox, куда попали значения $x_1, x_2,…, x_k$, разбивают на интервалы и подсчитывают частоту попадания значений величины в каждый интервал.

Рекомендуемое число интервалов находится по формуле Серджеса: $k = 1 + 3{,}322 \lg n$, а величину интервала – по формуле: $h = \dfrac{x_{\max} - x_{\min}}{1 + 3.322 \lg n}$. За начало интервала рекомендуется брать величину $x_{нач} = x_{\min} - \dfrac{1}{2} h$.

Вариационный ряд, представленный таблицей, построенной с помощью процедуры группировки, называется интервальным (или сгруппированным) статистическим рядом.

Для наглядного представления статистического распределения пользуются графическим изображением статистических рядов – полигоном и гистограммой.

Полигоном частот называют ломаную линию, отрезки которой соединяют точки (x_1, n_1), (x_2, n_2),…, (x_k, n_k). Например, рисунок 1. Полигоном относительных частот называют ломаную линию, отрезки которой соединяют точки (x_1, w_1), (x_2, w_2),…, (x_k, w_k).

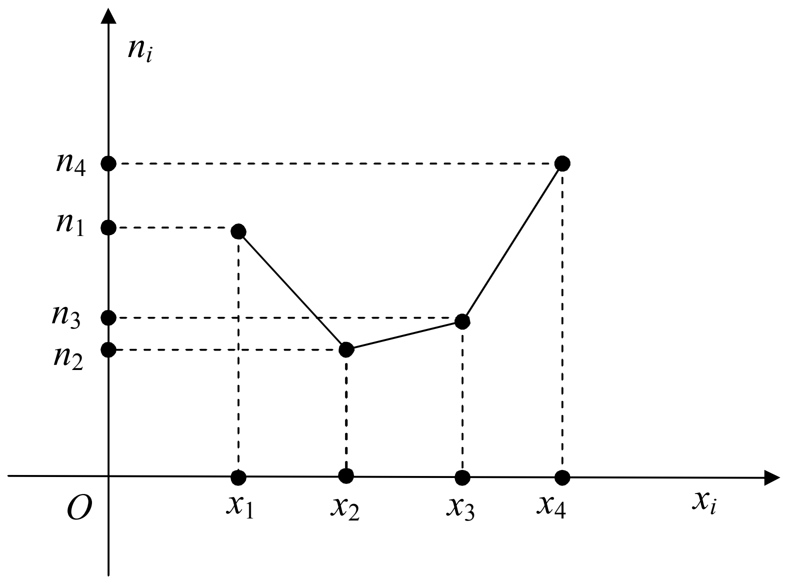


Рис. 1 – Полигон частот

В случае интервального распределения вместо значений x_k берут середины интервалов: $x_k^* = \dfrac{x_{k+1} + x_k}{2}$.

Гистограммой частот называется ступенчатая фигура, состоящая из прямоугольников, основания которых – частичные интервалы длины h, а высоты –

отношения $\frac{n_i}{h}$ (плотность частоты). Например, рисунок 2. Площадь i-го частичного прямоугольника равна $\frac{h \cdot n_i}{h} = n_i$ – сумма частот вариант i-го интервала. Площадь гистограммы частот равна объему выборки.

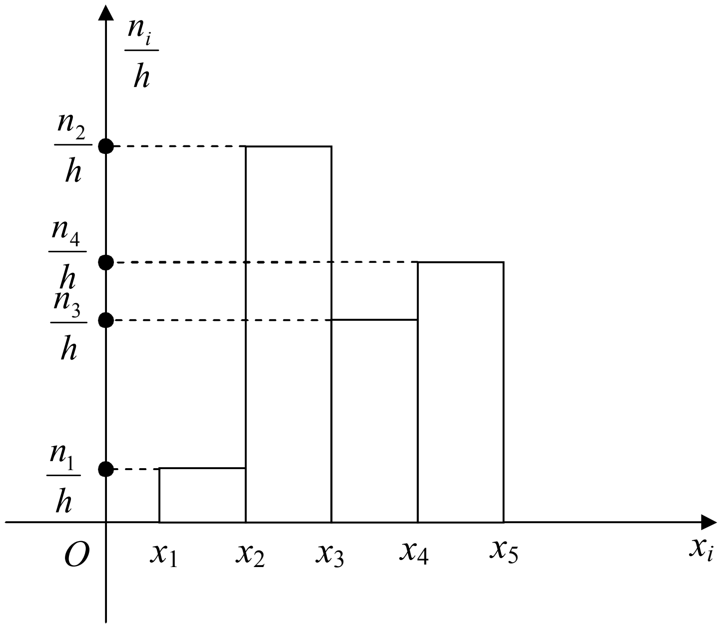


Рис. 2 – Гистограмма частот

Гистограммой относительных частот называется фигура, состоящая из прямоугольников, основаниями которых служат частичные интервалы длины h, а высоты равны отношениям $\frac{w_i}{h}$ (плотность относительной частоты). Площадь гистограммы относительных частот равна сумме всех относительных частот, то есть единице.

Гистограмма строится только для сгруппированного статистического ряда.

Для дискретного вариационного ряда легко находится варианта x_m, для которой частота n_i имеет наибольшее значение. Эта варианта, эмпирическая вероятность которой максимальна, называется модой (M_0).

Для интервального ряда легко находится интервал, у которого частота максимальна: мода находится внутри него. Для ее вычисления используют формулу линейной интерполяции.

Медиана (M_e) – это середина распределения, то есть такая точка, что половина принимаемых значений лежит слева от нее, а половина – справа. Для сгруппированного статистического ряда медиана – это точка, в которой площадь гистограммы делится пополам.

Выборочная средняя (средняя статистического распределения) для сгруппированного ряда находится по формуле: $\overline{x}_B = \frac{1}{n}\sum_{i=1}^{k} x_i^* n_i$.

Центральным моментом порядка k называется число, равное $m_k = \frac{1}{n}\sum_{i=1}^{k}(x_i^* - \overline{x}_B)^k \cdot n_i$ (для сгруппированного ряда). В частности, для k=2 получим выборочную дисперсию $D_B = \frac{1}{n}\sum_{i=1}^{k}(x_i^* - \overline{x}_B)^2 \cdot n_i$. Выборочное среднее квадратическое отклонение находится по формуле: $\sigma_B = \sqrt{D_B}$.

Коэффициентом асимметрии эмпирического распределения называется отношение центрального момента третьего порядка к кубу выборочного среднего квадратического отклонения: $As = \frac{m_3}{\sigma_B^3}$. Если As=0, то распределение имеет симметричную форму, то есть варианты, равноудаленные от $\overline{x}_B$, имеют одинаковую частоту. При As>0 говорят о положительной (правосторонней) асимметрии, при As<0 – об отрицательной (левосторонней) асимметрии. Для нормального распределения As=0.

Эксцесс эмпирического распределения определяется равенством: $E = \frac{m_4}{\sigma_B^4} - 3$ и характеризует «крутость» кривой эмпирического распределения по сравнению с нормальной кривой. Для нормального распределения E=0. Если эксцесс положительный, то кривая эмпирического распределения имеет более высокую и «острую» вершину по сравнению с нормальной кривой; если эксцесс отрицательный, то кривая эмпирического распределения имеет более низкую и «плоскую» вершину, чем нормальная кривая.

Коэффициент асимметрии и эксцесс используют для оценки отклонения эмпирического распределения от нормального. Эти величины свидетельствуют о достоверном отличии эмпирического распределения от нормального в том случае, если они превышают по абсолютной величине свою ошибку репрезентативности в три и более раз: $t_{As} = \frac{|As|}{m_a} \geq 3$, $t_E = \frac{|E|}{m_e} \geq 3$, где $m_a = \sqrt{\frac{6}{n}}$, $m_e = 2\sqrt{\frac{6}{n}}$ – ошибки репрезентативности для коэффициента асимметрии и эксцесса соответственно.

Коэффициентом вариации называется выраженное в процентах отношение выборочного среднего квадратического отклонения к выборочной средней: $V = \frac{\sigma_B}{\overline{x}_B} \cdot 100\%$. Если коэффициент вариации величины, имеющей только положительные значения, высок (например, более 50%), то, как правило, это свидетельствует о неоднородности значений признака.

Допустим, что имеется основание предположить, что изучаемая величина распределена по некоторому определенному закону (предположение о виде закона распределения можно сделать по виду полигона или гистограммы). Чтобы проверить, согласуется ли это предположение с данными наблюдений, вычисляют частоты n'_i наблюдавшихся значений в предположении, что величина X распределена по предполагаемому закону. Иными словами, находят выравнивающие (теоретические) частоты n'_i с помощью равенства: $n'_i = n \cdot p_i$, где n – число испытаний (объем выборки), p_i – вероятность наблюдавшегося значения. Если имеются основания предполагать, что случайная величина X распределена нормально, то выравнивающие частоты могут быть найдены по формуле: $n'_i = \frac{nh}{\sigma_B} \cdot \varphi(u_i)$, где n – число испытаний (объем выборки), h – длина частичного интервала, $u_i = \frac{x_i^* - \overline{x}_B}{\sigma_B}$, $\varphi(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}}$. Значения функции $\varphi(x)$ приведены в приложении 1.

Если эмпирические и теоретические частоты отличаются мало, то можно считать, что теоретическая кривая удовлетворительно отражает данные наблюдений.

Пример выполнения лабораторной работы №1

Задание. Даны наблюдавшиеся значения некоторой случайной величины:

185	151	187	211	155	208	178	193	149	175	193	169
163	166	131	200	173	145	166	216	216	156	174	219

174	161	225	178	188	157	177	183	206	187	209	189
157	180	163	189	196	204	199	242	192	160	123	134
181	172	183	120	164	197	134	204	148	157	133	157

Требуется:

1) построить сгруппированный статистический ряд;

2) построить гистограмму и полигон относительных частот;

3) Найти выборочные точечные характеристики: среднюю, дисперсию, среднее квадратическое отклонение.

Решение. Найдем число интервалов. Так как объем выборки n=60, то $k = 1 + 3{,}322\lg n \approx 6{,}9$. Поэтому следует взять 6-7 интервалов. Так как $x_{\min}$=120, $x_{\max}$=242, то длина интервала примерно равна $h = \dfrac{x_{\max} - x_{\min}}{1 + 3.322\lg n} \approx 17{,}7$. Чтобы не получилось больше 7 интервалов, примем h=20. Тогда $x_{нач} = 120 - \dfrac{1}{2}\cdot 20 = 110$. Договоримся не включать в интервал значения, равные правому концу интервала. Получим сгруппированный статистический ряд (табл. 3).

Таблица 3 – Сгруппированный ряд

Границы интервалов x_i ; x_{i+1}	110-130	130-150	150-170	170-190	190-210	210-230	230-250
Частоты n_i	2	7	15	18	12	5	1

Дальнейшие расчеты оформим в таблицу 4.

Таблица 4 – Расчет данных для построения гистограммы и полигона

1	2	3	4	5	6	7	8
Границы интервалов x_i ; x_{i+1}	110-130	130-150	150-170	170-190	190-210	210-230	230-250
Частоты n_i	2	7	15	18	12	5	1
Относительные частоты $w_i = \dfrac{n_i}{60}$	0,033	0,117	0,250	0,300	0,200	0,083	0,017

Плотность относительной частоты $\frac{w_i}{h}$	0,002	0,006	0,013	0,015	0,010	0,004	0,001
Середины интервалов	120	140	160	180	200	220	240

Построим гистограмму и полигон относительных частот (рис. 3), откладывая по Oy значения плотности относительной частоты $\frac{n_i}{nh}$, а по Ox отмечая интервалы $(x_i\ ;\ x_{i+1})$.

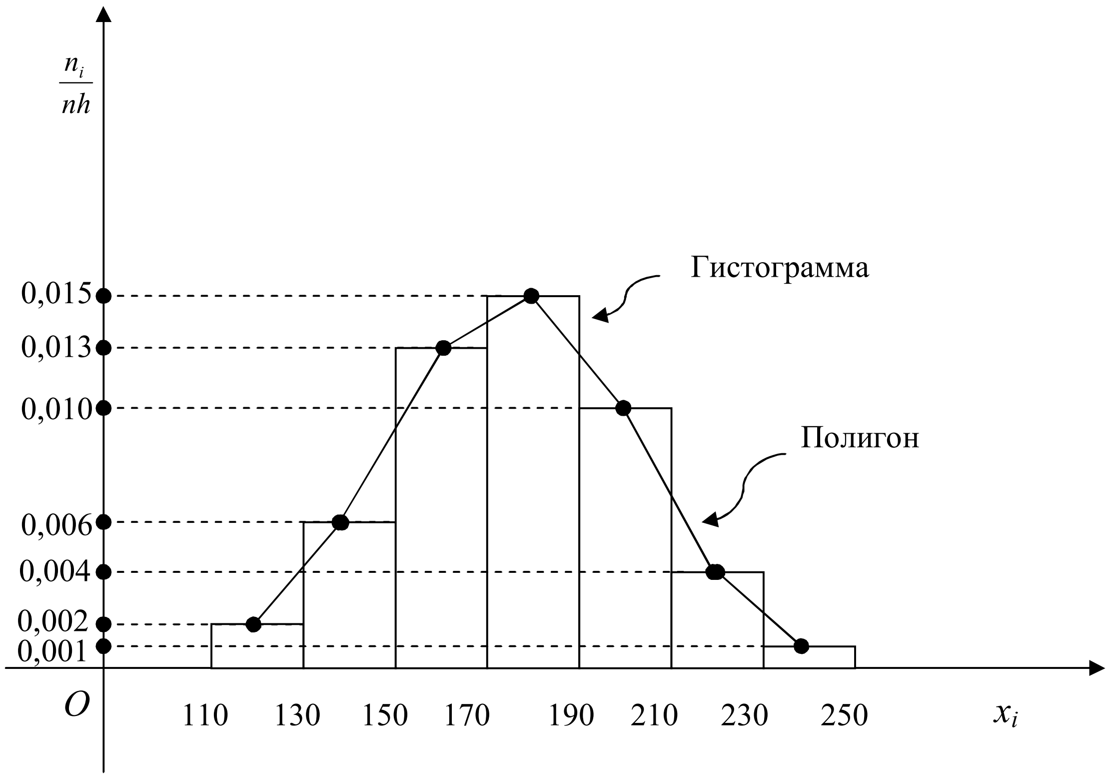


Рис. 3 – Гистограмма и полигон относительных частот

Полигон строим, соединяя отрезками середины верхних сторон прямоугольников гистограммы.

Найдем выборочные точечные характеристики.

Выборочная средняя:

$$\bar{x}_B = \frac{1}{n}\sum_{i=1}^{k} x_i^* n_i = \frac{1}{60}\cdot\left(120\cdot 2 + 140\cdot 7 + 160\cdot 15 + 180\cdot 18 + 200\cdot 12 + 220\cdot 5 + 240\cdot 1\right) \approx 177\,;$$

Выборочная дисперсия:

$$D_B = \frac{1}{n}\sum_{i=1}^{k}(x_i^* - \overline{x}_B)^2 \cdot n_i = \frac{1}{60}((120-177)^2 \cdot 2 + (140-177)^2 \cdot 7 + (160-177)^2 \cdot 15 +$$
$$+ (180-177)^2 \cdot 18 + (200-177)^2 \cdot 12 + (220-177)^2 \cdot 5 + (240-177)^2 \cdot 1) = 669;$$

Выборочное среднее квадратическое отклонение: $\sigma_B = \sqrt{D_B} = \sqrt{669} \approx 25{,}87$.

Пример выполнения лабораторной работы №2

Задание. Даны наблюдаемые значения некоторой случайной величины. Требуется:

1) Найти выборочные точечные характеристики: асимметрию, эксцесс, моду, коэффициент вариации;

2) Проверить гипотезу относительно близости распределения к нормальному.

Решение. В лабораторной работе №1 найдены x_i^*, $\overline{x}_B$, D_B, σ_B, n.

Для того чтобы найти коэффициент асимметрии As и эксцесс E, найдем центральные эмпирические моменты третьего и четвертого порядков:

$$m_3 = \frac{1}{n}\sum_{i=1}^{k}(x_i^* - \overline{x}_B)^3 \cdot n_i = \frac{1}{60}((120-177)^3 \cdot 2 + (140-177)^3 \cdot 7 + (160-177)^3 \cdot 15 +$$
$$+ (180-177)^3 \cdot 18 + (200-177)^3 \cdot 12 + (220-177)^3 \cdot 5 + (240-177)^3 \cdot 1) = -4580;$$

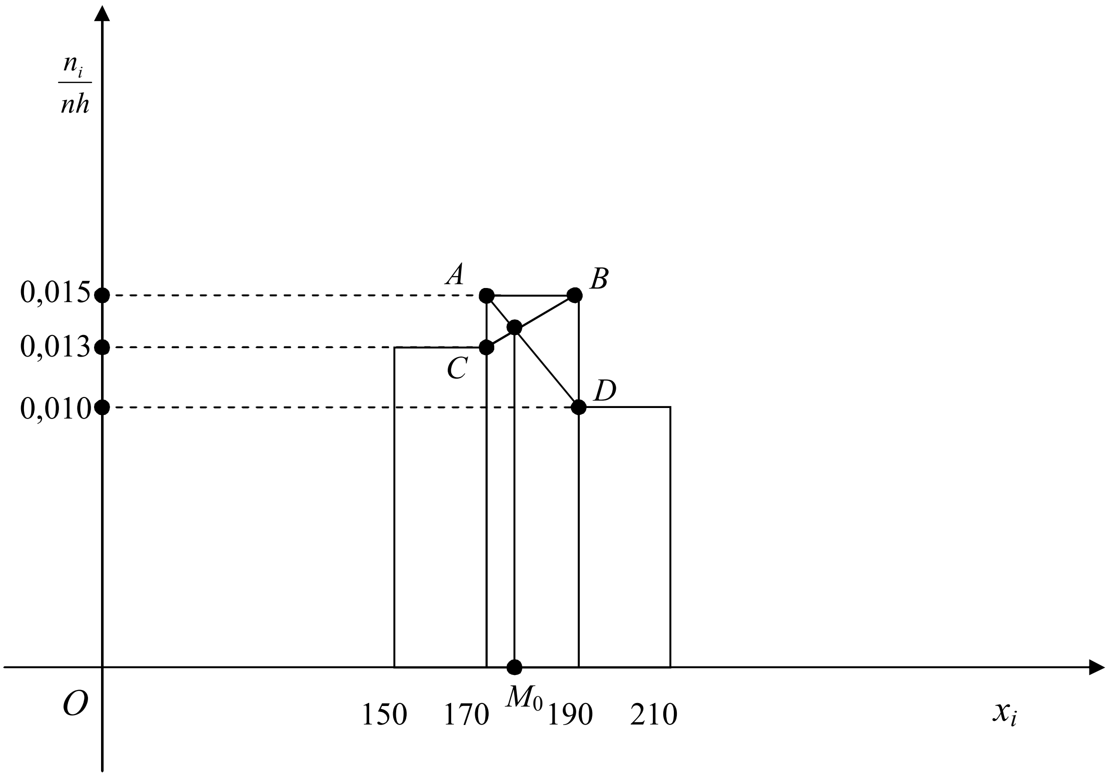
$$m_4 = \frac{1}{n}\sum_{i=1}^{k}(x_i^* - \overline{x}_B)^4 \cdot n_i = \frac{1}{60}((120-177)^4 \cdot 2 + (140-177)^4 \cdot 7 + (160-177)^4 \cdot 15 +$$
$$+ (180-177)^4 \cdot 18 + (200-177)^4 \cdot 12 + (220-177)^4 \cdot 5 + (240-177)^4 \cdot 1) = 1194841;$$

Отсюда коэффициент асимметрии $As = \frac{m_3}{\sigma_B^3} = -\frac{4580}{(25{,}87)^3} = -0{,}26$, а эксцесс $E = \frac{m_4}{\sigma_B^4} - 3 = \frac{1194841}{(25{,}87)^4} - 3 = 2{,}67 - 3 = -0{,}33$.

Таким образом, кривая эмпирического распределения имеет незначительную левостороннюю асимметрию, то есть «длинная часть» кривой распределения расположена слева от математического ожидания. Так как эксцесс отрицательный, то кривая эмпирического распределения имеет более низкую и «плоскую» вершину, чем нормальная кривая.

Коэффициент вариации $V = \frac{\sigma_B}{\overline{x}_B} \cdot 100\% = \frac{25,87}{177} \cdot 100\% = 14,6\%$. По значению коэффициента вариации можно сделать вывод об однородности выборочной совокупности.

Чтобы найти моду M_0, рассмотрим фрагмент гистограммы (лабораторная работа№1). Изобразим прямоугольник с наибольшей высотой и прилегающие к



нему слева и справа прямоугольники (рис. 4).

Рис. 4 – Мода сгруппированного ряда

Отметим отрезки *AD* и *CB*, концы которых имеют координаты: *A*(170; 0,015), *B*(190; 0,015), *C*(170; 0,013), *D*(190; 0,010). Составим уравнения прямых *AD* и *CB* как уравнения прямых, заданных двумя различными точками на плоскости. *AD*: $\frac{x-170}{190-170} = \frac{y-0,015}{0,01-0,015}$. Отсюда *AD*: $y = -0,00025x + 0,0575$. Аналогично *CB*: $\frac{x-170}{190-170} = \frac{y-0,013}{0,015-0,013}$, *CB*: $y = 0,0001x - 0,004$. Решая систему $\begin{cases} y = -0,00025x + 0,05; \\ y = 0,0001x - 0,004, \end{cases}$ найдем абсциссу точки пересечения *AD* и *CB*: $x \approx 176$. Данное значение и есть значение моды M_0. Итак, M_0=176. Эмпирическая вероятность данного значения признака максимальна.

Проверим предположение о близости эмпирического распределения к нормальному.

Найдем выравнивающие частоты и составим таблицу 5:

$$n'_1 = \frac{60 \cdot 20}{25,87} \cdot \varphi\left(\frac{120-177}{25,87}\right) = \frac{1200}{25,87} \cdot \varphi(-2,20) = 1,65,$$

$$n'_2 = \frac{60 \cdot 20}{25,87} \cdot \varphi\left(\frac{140-177}{25,87}\right) = \frac{1200}{25,87} \cdot \varphi(-1,43) = 6,66,$$

$$n'_3 = \frac{60 \cdot 20}{25,87} \cdot \varphi\left(\frac{160-177}{25,87}\right) = \frac{1200}{25,87} \cdot \varphi(-0,66) = 14,89,$$

$$n'_4 = \frac{60 \cdot 20}{25,87} \cdot \varphi\left(\frac{180-177}{25,87}\right) = \frac{1200}{25,87} \cdot \varphi(0,12) = 18,37,$$

$$n'_5 = \frac{60 \cdot 20}{25,87} \cdot \varphi\left(\frac{200-177}{25,87}\right) = \frac{1200}{25,87} \cdot \varphi(0,89) = 12,45,$$

$$n'_6 = \frac{60 \cdot 20}{25,87} \cdot \varphi\left(\frac{220-177}{25,87}\right) = \frac{1200}{25,87} \cdot \varphi(1,66) = 4,67,$$

$$n'_7 = \frac{60 \cdot 20}{25,87} \cdot \varphi\left(\frac{240-177}{25,87}\right) = \frac{1200}{25,87} \cdot \varphi(2,44) = 0,94.$$

Таблица 5 – Эмпирические и теоретические частоты

x_i^*	120	140	160	180	200	220	240
Эмпирические частоты n_i	2	7	15	18	12	5	1
Теоретические частоты n'_i	1,7	6,7	14,9	18,4	12,5	4,7	0,9

Так как эмпирические и теоретические частоты отличаются мало, то можно предположить, что нормальная кривая удовлетворительно отражает данные наблюдений.

Вычислим отклонения коэффициента асимметрии и эксцесса от их ошибок репрезентативности: $m_a = \sqrt{\frac{6}{60}} = 0,316$, $m_e = 2\sqrt{\frac{6}{60}} = 0,632$, $t_{As} = \frac{|As|}{m_a} = \frac{|-0,26|}{0,316} = 0,82$,

$t_E = \frac{|E|}{m_e} = \frac{|-0,33|}{0,632} = 0,52$. Так как $t_{As} < 3$ и $t_E < 3$, то можно сделать вывод, что эмпирическое распределение признака не отличается от нормального.

Варианты лабораторных работ №1 и №2

Лабораторная № 1

Задание. Даны наблюдавшиеся значения некоторой случайной величины:

Требуется:

1) построить сгруппированный статистический ряд;

2) построить гистограмму и полигон относительных частот;

3) найти выборочные точечные характеристики: среднюю, дисперсию, среднее квадратическое отклонение.

Лабораторная № 2

Задание. Даны наблюдаемые значения некоторой случайной величины. Требуется:

1) найти выборочные точечные характеристики: асимметрию, эксцесс, моду, коэффициент вариации;

2) проверить гипотезу относительно близости распределения к нормальному.

Вариант 1

34	25	29	34	12	28	13	28	28	17
23	31	32	23	16	22	34	22	25	28
24	24	25	28	26	19	29	21	30	18
30	30	21	30	19	20	30	34	20	36
28	36	27	17	27	26	26	19	29	24
37	28	31	25	23	33	35	31	22	30
25	26	22							

Вариант 2

37	49	43	38	47	40	43	43	41	44
45	51	47	41	44	43	44	39	43	39
46	41	44	44	46	39	45	42	43	45
40	48	41	43	44	41	43	48	43	46
48	39	44	45	42	43	42	46	47	40
44	46	42							

Вариант 3

95	93	100	94	95	94	101	90	95	103
93	98	95	93	107	100	98	101	90	95

93	95	93	96	97	102	97	106	96	94
99	104	98	109	98	104	95	100	102	96
103	95	94	98	97	94	97	91	96	108
99	91	94	95	94	89	98	102	98	95
103	98	99	94	94	89	106	93	97	99
100	95	92	93	96	97	100	91	93	99

Вариант 4

10	42	34	33	44	29	32	32	59	31
48	26	40	47	33	48	25	37	57	37
43	29	41	36	39	45	37	49	15	18
35	43	50	52	25	22	58	44	27	32
39	12	46	37	34	28	23	38	55	58
33	45	23	44	22	18	48	51	21	41
38	53								

Вариант 5

86	72	67	84	75	51	77	74	55	79
63	58	76	72	53	90	71	52	87	74
60	80	63	59	79	62	74	70	81	91
77	61	56	46	46	71	68	52	68	53
82	99	69	64	49	68	76	57	67	62
48	66	83	96	70	65				

Вариант 6

26	39	16	30	16	34	25	31	34	28
28	27	22	23	31	31	24	17	27	22
25	12	33	25	16	17	28	13	25	24
36	31	27	28	26	26	26	28	26	27
26	34	29	25	30	20	23	37	12	24
34	23	30	23	28	31	32	28	28	24
19	36								

Вариант 7

61	41	27	42	50	58	47	43	39	50
54	26	45	48	37	48	53	60	31	47
51	57	50	40	46	51	42	64	74	39
30	61	45	28	41	49	33	46	37	48
52	45	69	41	47	32	51	59	48	40
49	54	34	44	38	49	43	65	34	46
52	35	66	36	49	46	56	44	36	47
50	55	43	31	44	50	46	29	53	42

Вариант 8

231	213	239	243	264	248	211	296	257	281
297	246	302	232	259	283	240	249	213	255
225	246	269	254	214	258	200	236	278	266
260	293	276	285	264	179	262	288	275	288
284	209	253	237	218	241	299	217	255	220
129	293	206	237	278	239	290	245	267	

Тема 2. ИНТЕРВАЛЬНЫЕ ОЦЕНКИ ПАРАМЕТРОВ РАСПРЕДЕЛЕНИЯ

Основные понятия

Пусть требуется изучить некоторый количественный признак генеральной совокупности. На практике редко встречается такое положение, когда изучаемый закон распределения неизвестен полностью. Чаще всего вид закона распределения известен заранее (из каких-либо теоретических соображений), но требуется найти его параметры. Например, известно, что закон распределения величины X – нормальный. Тогда задача сводится к оценке параметров a и σ.

В подобных случаях можно обойтись выборкой небольшого объема – порядка одного или нескольких десятков. Допустим, что закон распределения величины X содержит параметр Θ, численное значение которого неизвестно. Возникает задача оценки параметра Θ. Для этого производится выборка объема n, в результате которой получают значения количественного признака: $x_1, x_2, \ldots, x_n$. Через эти данные и выражают оцениваемый параметр.

Обозначим Θ^* оценку параметра Θ. Θ^* представляет собой некоторое выражение, зависящее от значений. Таким образом, Θ^* является случайной величиной, принимающей свои значения в результате n опытов. Рассматривая $x_1, x_2, \ldots, x_n$ как независимые случайные величины $X_1, X_2, \ldots, X_n$, можно сказать, что найти статистическую оценку неизвестного параметра распределения – это значит найти функцию от наблюдаемых случайных величин, которая и дает приближенное значение оцениваемого параметра.

Итак, статистической оценкой неизвестного параметра распределения называют функцию от наблюдаемых случайных величин.

Для того, чтобы оценка была доброкачественной, к ней предъявляют ряд требований, основное из которых – несмещенность оценки.

Желательно, чтобы пользуясь величиной Θ^* вместо Θ, не делалось систематически ошибок ни в сторону занижения, ни в сторону завышения, то есть чтобы выполнялось равенство $M(\Theta^*)=\Theta$. Несмещенной называется оценка, если математическое ожидание Θ^* равно оцениваемому параметру Θ.

Пусть некоторая случайная величина X имеет неизвестное математическое ожидание a. Проведем n опытов, в результате каждого из которых получим значения случайной величины X, которые, в свою очередь сами являются случайными величинами: $X_1, X_2, \ldots, X_n$. Поскольку случайные величины $X_1, X_2, \ldots, X_n$ возникают в результате измерения одной и той же величины X, то они имеют тот же закон распределения, что и сама величина, а значит, и тот же параметр $M(X)$: $M(X_i)= M(X)=a$.

В качестве оценки неизвестного математического ожидания a берется среднее арифметическое значений X_k: $a^* = \frac{\sum_{i=1}^{n} X_i}{n} = \overline{X}$.

Имеем: $M(a^*) = M(\overline{X}) = M\left(\frac{\sum_{i=1}^{n} X_i}{n}\right) = \frac{1}{n}\sum_{i=1}^{n} M(X_i) = \frac{1}{n} \cdot na = a$. Отсюда следует, что величина a^* является несмещенной оценкой заменим случайные величины X_i на значения x_i и оценку a^* на выборочную среднюю $\overline{x}_B$. Таким образом, точечной оценкой математического ожидания является выборочная средняя $\overline{x}_B = \frac{\sum_{i=1}^{n} x_i}{n}$ (каждое значение $x_1, x_2, \ldots, x_n$ наблюдалось один раз).

Оценку дисперсии $D(X)$ будем находить по формуле: $D^* = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x}_B)^2$ (каждое значение $x_1, x_2, \ldots, x_n$ наблюдалось один раз). Известно, что $M(D^*) = \frac{n-1}{n}D$; оценка D^* является смещенной, так как $M(D^*) \neq D$. Поэтому рассматривают еще так называемую исправленную дисперсию $S^2 = \frac{n}{n-1}D^*$.Для S^2 имеем $M(S^2) = M\left(\frac{n}{n-1}D^*\right) = D$. S^2 является несмещенной оценкой дисперсии. Явное выражение для S^2: $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$. Естественно в качестве приближения неизвестного параметра брать несмещенную оценку для того, чтобы не делать систематической ошибки в сторону завышения или занижения.

Для оценки генерального среднего квадратического отклонения используют исправленное среднее квадратическое отклонение: $S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2}$. Заменив X_i и $\overline{X}$ на значения x_i и $\overline{x}_B$, получим формулу для нахождения среднего квадратического отклонения: $S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x}_B)^2}$. Можно использовать другой вариант формулы: $S = \sqrt{\frac{n}{n-1}\left(\frac{1}{n}\sum_{i=1}^{n}(x_i)^2 - (\overline{x}_B)^2\right)}$.

При статистической обработке результатов малого числа наблюдений оценка Θ^* неизвестного параметра может сильно отличаться от самого параметра Θ. Поэтому при небольшом объеме выборки принято пользоваться интервальными оценками. Интервальной называется оценка, которая определяется двумя числами – концами интервала. Интервальные оценки позволяют установить точность и надежность оценок.

Оценка Θ^* тем точнее, чем меньше абсолютная величина разности Θ и Θ^*: $|\Theta^* - \Theta|$. Однако, статистические методы не позволяют утверждать, что оценка Θ^* удовлетворяет неравенству $|\Theta^* - \Theta| < \varepsilon$; можно лишь говорить о вероятности этого события: $P(|\Theta^* - \Theta| < \varepsilon) = \gamma$. Положительное число ε характеризует точность оценки. Оценка тем точнее, чем меньше ε. Вероятность γ, с которой осуществляется неравенство $|\Theta^* - \Theta| < \varepsilon$, называется надежностью оценки (доверительной вероятностью).

Пусть $P(|\Theta^* - \Theta| < \varepsilon) = \gamma$. Отсюда следует: $P(\Theta^* - \varepsilon < \Theta < \Theta^* + \varepsilon) = \gamma$. Это соотношение следует понимать так: вероятность того, что интервал $(\Theta^* - \varepsilon < \Theta < \Theta^* + \varepsilon)$ заключает в себе (покрывает) неизвестный параметр Θ равна γ. Интервал $(\Theta^* - \varepsilon, \Theta^* + \varepsilon)$ называется доверительным. Длина доверительного интервала, характеризующего точность оценки, зависит от γ и n. Обычно надежность оценки задается наперед, причем в качестве γ берут число, близкое к единице: 0,95, 0,99, 0,999.

Доверительный интервал для математического ожидания нормально распределенной случайной величины с заданной надежностью γ и известным средним квадратическом отклонением σ имеет вид: $\left(\bar{x}_B - \frac{t_\gamma \cdot \sigma}{\sqrt{n}}; \bar{x}_B + \frac{t_\gamma \cdot \sigma}{\sqrt{n}}\right)$ или $\bar{x}_B - \frac{t_\gamma \cdot \sigma}{\sqrt{n}} < a < \bar{x}_B + \frac{t_\gamma \cdot \sigma}{\sqrt{n}}$, n – объем выборки, t_γ находится из соотношения: $2\Phi(t)=\gamma$. Функция $\Phi(t)$ табулирована (приложение 2). Из данной формулы можно сделать выводы:

1) при возрастании n число ε убывает, следовательно, точность оценки увеличивается;

2) увеличение $\gamma=2\Phi(t)$ приводит к увеличению t ($\Phi(t)$ – возрастающая функция), а значит и к увеличению ε. Таким образом, увеличение надежности оценки влечет за собой уменьшение ее точности;

3) объем выборки можно найти по формуле: $n = \frac{t^2\sigma^2}{\varepsilon^2}$.

Доверительный интервал для оценки математического ожидания нормального распределения при неизвестном σ имеет следующий вид: $\left(\bar{x}_B - \frac{t_\gamma \cdot S}{\sqrt{n}}; \bar{x}_B + \frac{t_\gamma \cdot S}{\sqrt{n}}\right)$. Он покрывает неизвестный параметр a с надежностью γ. Здесь $\bar{x}_B$ – выборочная средняя, S – исправленное среднее квадратическое отклонение, n – объем выборки. Значения t_γ находят по таблице значений $t_\gamma=t(\gamma, n)$ (приложение 4) по заданным γ и n. Выражение $\varepsilon = \frac{t_\gamma \cdot S}{\sqrt{n}}$ определяет точность оценки математического ожидания нормального распределения при неизвестном σ.

Доверительный интервал для оценки среднего квадратического отклонения нормального распределения имеет следующий вид $(S - Sq; S + Sq)$. Он покрывает неизвестный параметр σ с надежностью γ. Здесь S – исправленное среднее квадратическое отклонение. Значения q определяют по таблице значений $q = q(\gamma,$

n) (приложение 5) по заданным γ и n. Выражение $\varepsilon = Sq$ определяет точность оценки среднего квадратического отклонения нормального распределения. Среднее квадратическое отклонение не может принимать отрицательные значения. Поэтому доверительный интервал $(S - Sq; S + Sq)$ строят для случая q<1. Если q>1, то доверительный интервал примет вид: $(0; S + Sq)$.

Пример выполнения лабораторной работы №3

Задание. Имеются результаты измерения некоторой величины: 2,2; 4,5; 3,8; 5,6; 4,6; 3,0; 2,3; 3,5; 4,4; 3,2. Будем рассматривать эти результаты как n реализаций случайной величины X. Далее, будем предполагать, что X имеет нормальное распределение.

Требуется:

1) найти точечные несмещенные оценки математического ожидания a и среднего квадратического отклонения σ;

2) найти доверительный интервал, покрывающий математическое ожидание с заданной доверительной вероятностью γ=0,95;

3) найти погрешность, с которой среднее арифметическое ожидание a случайной величины X, если доверительная вероятность γ=0,99; γ=0,999;

4) найти минимальный объем выборки, чтобы с доверительной вероятностью γ=0,95 можно было утверждать, что, принимая среднее арифметическое за математическое ожидание случайной величины X, допускаем погрешность $\varepsilon = \frac{1}{2}\sigma$.

Решение. Найдем несмещенные оценки для математического ожидания и среднего квадратического отклонения: $\bar{x}_B = \frac{\sum_{i=1}^{n} x_i}{n}$, $S = \sqrt{\frac{1}{n-1} \cdot \sum_{i=1}^{n} \left(x_i - \bar{x}_B\right)^2}$. Для этого заполним таблицу 6.

Заполнив второй столбец таблицы 6, найдем $\bar{x}_B = \frac{37,1}{10} = 3,71$. по последнему столбцу имеем $S = \sqrt{\frac{1}{9} \cdot 10,549} = \sqrt{1,172} \approx 1,083$.

Таблица 6 – Вспомогательные вычисления для $\overline{x}_B$ и S

i	x_i	$x_i - \overline{x}_B$	$\left(x_i - \overline{x}_B\right)^2$
1	2,2	-1,51	2,2801
2	4,5	0,79	0,6241
3	3,8	0,09	0,0081
4	5,6	1,89	3,5721
5	4,6	0,89	0,7921
6	3	-0,71	0,5041
7	2,3	-1,41	1,9881
8	3,5	-0,21	0,0441
9	4,4	0,69	0,4761
10	3,2	-0,51	0,2601
Сумма	37,1	0	10,549

Таким образом, получили точечные несмещенные оценки математического ожидания и среднего квадратического отклонения: $\overline{x}_B = 3{,}71$, $S = 1{,}083$.

Найдем доверительный интервал, покрывающий математическое ожидание с надежностью γ=0,95. Поскольку σ неизвестно, то доверительный интервал имеет вид: $\left(\overline{x}_B - \frac{t_\gamma \cdot S}{\sqrt{n}}; \overline{x}_B + \frac{t_\gamma \cdot S}{\sqrt{n}}\right)$. t_γ найдем по таблице (приложение 4) по γ=0,95, n=10: t_γ=2,26. Найдем точность оценки: $\varepsilon = \frac{t_\gamma \cdot S}{\sqrt{n}} = \frac{2{,}26 \cdot 1{,}083}{\sqrt{10}} = 0{,}77$. Тогда доверительный интервал математического ожидания имеет вид: (3,71–0,77; 3,71+0,77) или (2,94; 4,48), то есть 2,94<a<4,48.

Смысл полученного результата в том, что если будет проведено достаточно большое число серий, состоящих из 10 измерений данной величины, то в 95% случаев из них найденный доверительный интервал накроет математическое ожидание.

Выполняем п. 3 задания. Если γ=0,99, n=10, то t_γ=3,25. Точность оценки при этом будет: $\varepsilon = \frac{t_\gamma \cdot S}{\sqrt{n}} = \frac{3{,}25 \cdot 1{,}083}{\sqrt{10}} = 1{,}11$. Если γ=0,999, n=10, то t_γ=4,78. Точность оценки при этом будет: $\varepsilon = \frac{t_\gamma \cdot S}{\sqrt{n}} = \frac{4{,}78 \cdot 1{,}083}{\sqrt{10}} = 1{,}64$. При увеличении надежности точность оценки снижается.

Минимальный объем выборки найдем по формуле: $n = \frac{t^2\sigma^2}{\varepsilon^2}$. Значение t находим из соотношения $2\Phi(t)= \gamma$=0,95, $\Phi(t)$=0,475. По таблице значений функции $\Phi(t)$ (приложение 2) получаем t=1,96. Тогда $n = \frac{1{,}96^2\sigma^2}{\left(\frac{1}{2}\sigma\right)^2} = 4\cdot 1{,}96^2 \approx 15$.

Таким образом, минимальный объем выборки, позволяющий с доверительной вероятностью γ=0,95 утверждать, что, принимая среднее арифметическое за математическое ожидание случайной величины X, допуская погрешность $\varepsilon = \frac{1}{2}\sigma$.

Пример выполнения лабораторной работы №4

Задание. Признак X распределен в генеральной совокупности нормально с известным средним квадратическим отклонением σ=3. Данные выборочного наблюдения представлены вариационным рядом:

x_i ; x_{i+1}	11-13	13-15	15-17	17-19	19-21	21-23	23-25
n_i	2	7	15	18	12	5	1

По данным выборки требуется:

1) найти доверительный интервал для математического ожидания генеральной совокупности с надежностью 0,99;

2) составить функцию плотности распределения вероятности для предполагаемого нормального закона распределения;

3) определить надежность интервальной оценки (17,5; 17,9) математического ожидания;

4) вычислите несмещенную точечную оценку среднего квадратического отклонения двумя различными способами;

5) считая среднее квадратическое отклонение генеральной совокупности неизвестным, найти для него доверительный интервал с надежностью 0,95;

6) определить каким был объем выборки, если доверительный интервал (0, $3S$) для среднего квадратического отклонения генеральной совокупности ожидается с надежностью 0,999.

Решение. Вычислим середины интервалов x_i^* и, по ним, среднюю арифметическую $\overline{x}_B = \frac{1}{n}\sum_{i=1}^{k} x_i^* n_i = \frac{1}{60}\cdot\left(12\cdot 2 + 14\cdot 7 + 16\cdot 15 + 18\cdot 18 + 20\cdot 12 + 22\cdot 5 + 24\cdot 1\right) \approx 17{,}7$.

Формула для функции плотности вероятности $f(x) = \frac{1}{\sigma\sqrt{2\pi}}\cdot e^{-\frac{(x-a)^2}{2\sigma^2}}$ в нашем случае примет вид: $f(x) = \frac{1}{3\sqrt{2\pi}}\cdot e^{-\frac{(x-17,7)^2}{18}}$.

Так как σ известно, доверительный интервал для математического ожидания нормально найдем по формуле: $\left(\overline{x}_B - \frac{t_\gamma\cdot\sigma}{\sqrt{n}}; \overline{x}_B + \frac{t_\gamma\cdot\sigma}{\sqrt{n}}\right)$. t_γ найдем из соотношения: $2\Phi(t)=\gamma=0{,}99$. $\Phi(t)=0{,}495$. По таблице приложения 2: $t_\gamma=2{,}58$. Искомый доверительный интервал: $\left(17{,}7 - \frac{2{,}58\cdot 3}{\sqrt{60}}; 17{,}7 + \frac{2{,}58\cdot 3}{\sqrt{60}}\right)$ или (16,7;18,7).

Найдем надежность интервальной оценки (17,5; 17,9) математического ожидания генеральной совокупности. Так как средняя арифметическая является центром доверительного интервала, точность оценки ε составит $\varepsilon = \frac{17{,}9 - 17{,}5}{2} = 0{,}2$. С другой стороны $\varepsilon = \frac{t_\gamma\cdot\sigma}{\sqrt{n}}$. Получим уравнение: $0{,}2 = \frac{t_\gamma\cdot 3}{\sqrt{60}}$. Отсюда $t_\gamma=0{,}52$.

По таблице приложения 2: $\Phi(0{,}52)=0{,}1985$. Тогда $\gamma=2\Phi(t)=2\Phi(0{,}52)=$ $=0{,}397$. То есть с вероятностью 0,397 можно ожидать, что математическое ожидание генеральной совокупности попадет в интервал (17,5; 17,9).

Несмещенной точечной оценкой среднего квадратического отклонения является исправленное среднее квадратическое отклонение. Вычислим исправленное среднее квадратическое отклонение двумя способами:

$$S = \sqrt{\frac{1}{n-1}\cdot\sum_{i=1}^{n}\left(x_i^* - \overline{x}_B\right)^2} = \sqrt{\frac{1}{59}\cdot 401{,}3} = 2{,}61;$$

$$S = \sqrt{\frac{n}{n-1}\left(\frac{1}{n}\sum_{i=1}^{n} n_i (x_i^*)^2 - (\overline{x}_B)^2\right)} =$$

$$= \sqrt{\frac{60}{59}\cdot\left(\frac{1}{60}(2\cdot 144 + 7\cdot 196 + 15\cdot 256 + 18\cdot 324 + 12\cdot 400 + 5\cdot 484 + 1\cdot 576)\right) - 17{,}7^2} = 2{,}61.$$

Пусть среднее квадратическое отклонение генеральной совокупности неизвестно. По таблице приложения 5: $q=q(\gamma,\ n)=q(0{,}95,\ 60)=0{,}188<1$. Поэтому доверительный интервал для неизвестного среднего квадратического отклонения будем искать по формуле $(S-Sq;S+Sq)$: $(2{,}61-2{,}61\cdot 0{,}188;2{,}61+2{,}61\cdot 0{,}188)$ или (2,11; 3,10).

Определим, каким был объем выборки, если $(0,\ 3S)$ – доверительный интервал, покрывающий среднее квадратическое отклонение генеральной совокупности с доверительной вероятностью 0,999.Правая граница доверительного интервала $S+Sq=3S$, $q=2$. По таблице приложения 5 для $q=2{,}06\approx 2$ и $\gamma=0{,}999$ находим объем выборки: $n=9$.

Варианты лабораторных работ №3 и №4

Лабораторная работа № 3

Задание. Приводятся результаты измерения некоторой величины, которую будем рассматривать как n реализаций случайной величины X. В предположении, что X имеет нормальное распределение, требуется:

1) найти точечные несмещенные оценки математического ожидания a и среднего квадратического отклонения σ;

2) найти доверительный интервал, покрывающий математическое ожидание с заданной доверительной вероятностью $\gamma=0{,}95$;

3) найти погрешность, с которой среднее арифметическое оценивает математическое ожидание a случайной величины X, если доверительная вероятность $\gamma=0{,}99$; $\gamma=0{,}999$;

4) найти минимальный объем выборки, чтобы с доверительной вероятностью $\gamma=0{,}95$ можно было утверждать, что, принимая среднее арифметическое за математическое ожидание случайной величины X, допускаем погрешность $\varepsilon=\frac{1}{3}\sigma$.

Вариант 1

20,58	22,80	25,4	26,08	23,25	21,42
23,1	24,09	24,02	26,12	22,84	25,10

Вариант 2

43,4	43,6	44,3	44,5	44,6	44,5

44,7	44,7	44,6	44,8	45,1	45,2
45,3					

Вариант 3

31,85	31,36	30,32	30,90	31,7	32,40
31,60	31,12	30,98	31,02	31,05	31,00

Вариант 4

4,8	6,2	6,0	5,9	4,9	6,0
5,6	6,1	5,9	5,8	5,7	5,1
5,5	6,2	5,4			

Вариант 5

20,49	22,67	23,40	25,05	22,15	20,42
24,35	21,20	23,05	24,02		

Вариант 6

30,8	29,9	30,3	29,9	30,4	30,2
27,5	31,1	29,2	29,7	30,8	30,3
31,8	30,7				

Вариант 7

50,52	52,24	51,82	50,46	51,88	52,01
53,00	51,12	51,08	51,92	52,08	51,78

Вариант 8

26,2	27,5	26,0	32,2	35,0	30,2
29,7	26,8	30,0	30,8	32,0	28,8
35,0	30,8				

Лабораторная работа № 4

Задание. Признак X распределен в генеральной совокупности нормально с известным средним квадратическим отклонением σ. Данные выборочного наблюдения представлены вариационным рядом. По данным выборки требуется:

1) найти доверительный интервал для математического ожидания генеральной совокупности с надежностью 0,99;

2) составить функцию плотности распределения вероятности для предполагаемого нормального закона распределения;

3) определить надежность интервальной оценки (α; β) математического ожидания;

4) вычислите точечную оценку среднего квадратического отклонения двумя различными способами;

5) считая среднее квадратическое отклонение генеральной совокупности неизвестным, найти для него доверительный интервал с надежностью 0,95;

6) определить каким был объем выборки, если доверительный интервал (0, 2,33S) для среднего квадратического отклонения генеральной совокупности ожидается с надежностью 0,999.

Вариант 1

Левая граница:	14,5	19,5	24,5	29,5	34,5	39,5	44,5	49,5	54,5
Правая граница:	19,5	24,5	29,5	34,5	39,5	44,5	49,5	54,5	59,5
Частота:	2	4	11	12	18	11	10	8	4

σ=10, α=29; β=47

Вариант 2

Левая граница:	45,5	50,5	55,5	60,5	65,5	70,5	75,5	80,5	85,5
Правая граница:	50,5	55,5	60,5	65,5	70,5	75,5	80,5	85,5	90,5
Частота:	1	4	10	12	18	11	10	8	4

σ=32, α=60; β=78

Вариант 3

Левая граница:	8	15	21	27	33	39	45	51	57
Правая граница:	14	21	27	33	39	45	51	57	63
Частота:	3	4	10	12	18	11	10	7	4

σ=12, α=21; β=52

Вариант 4

Левая граница:	88	94	100	106	112	118	124	130	136
Правая граница:	94	100	106	112	118	124	130	136	142
Частота:	4	4	10	12	18	11	10	7	4

σ=78, α=100; β=130

Вариант 5

Левая граница:	81	87	93	99	105	101	107	113	119
Правая граница:	87	93	99	105	111	107	113	119	125
Частота:	3	4	10	13	17	12	10	7	4

σ=67, α=94; β=114

Вариант 6

Левая граница:	17,5	23,5	29,5	35,5	41,5	47,5	53,5	59,5	65,5
Правая граница:	23,5	29,5	35,5	41,5	47,5	53,5	59,5	65,5	71,5
Частота:	2	5	10	13	16	12	11	7	4

σ=14, α=40; β=52

Вариант 7

Левая граница:	19	22	25	28	31	34	37	40	43
Правая граница:	22	25	28	31	34	37	40	43	46
Частота:	2	5	6	9	12	10	8	5	1

σ=9, α=30; β=35

Вариант 8

Левая граница:	52	55	58	61	64	67	70	73	76
Правая граница:	55	58	61	64	67	70	73	76	79
Частота:	3	5	6	9	12	10	8	5	1

σ=30, α=60; β=70

Тема 3. ПРОВЕРКА СТАТИСТИЧЕСКИХ ГИПОТЕЗ

Основные понятия

При изучении различных совокупностей перед исследователем могут возникать различные задачи. Если необходимо знать закон распределения генеральной совокупности и есть основание предполагать, что он имеет определенный вид (А), то выдвигают гипотезу: генеральная совокупность распределена по закону А. Если закон распределения известен, но неизвестен параметр распределения, однако есть основания предполагать, что неизвестный параметр Θ равен a, то выдвигают гипотезу: $\Theta=a$.

В первом случае речь идет о виде предполагаемого распределения, во втором – о предполагаемой величине параметра известного распределения. Возможны и другие гипотезы, например, о равенстве параметров двух или более распределений.

Статистической называют гипотезу о виде неизвестного распределения или о параметрах известных распределений.

Нулевой (основной) называют выдвинутую гипотезу H_0. Конкурирующей (альтернативной) называют гипотезу H_1, которая противоречит нулевой.

В итоге статистической проверки гипотезы в двух случаях могут быть приняты неправильные решения, то есть 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. Последствия этих ошибок могут быть различными. Вероятность совершить ошибку первого рода принято обозначать через α; ее называют уровнем значимости. Наиболее часто уровень значимости принимают равным 0,05 или 0,01. Если принять уровень значимости α=0,05, то это означает, что в пяти случаях из 100 имеется риск допустить ошибку первого рода (отвергнуть правильную гипотезу).

Для проверки нулевой гипотезы используют специально подобранную случайную величину, точное и приближенное распределение которой известно.

Статистическим критерием (критерием) называют случайную величину K, которая служит для проверки нулевой гипотезы. Например, если проверяют гипотезу о равенстве дисперсий двух нормальных распределений, то в качестве критерия K принимают отношение исправленных выборочных дисперсий: $F=\frac{S_1^2}{S_2^2}$. Эта величина случайная, так как в различных опытах дисперсии принимают различные значения, и распределена по закону Фишера-Снедекора.

Для проверки гипотезы по данным выборок вычисляют частные значения входящих в критерий величин и таким образом получают частное (наблюдаемое) значение критерия.

Наблюдаемым значением критерия $K_{\text{набл}}$ называется значение критерия, вычисленное по выборкам.

После выбора определенного критерия множество всех его возможных значений разбивают на два непересекающихся подмножества: одно из них содержит значения, при которых нулевая гипотеза отвергается, а другое – 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: если наблюдаемое значение критерия принадлежит критической области, то гипотезу отвергают. Если наблюдаемое значение критерия принадлежит области принятия гипотезы – гипотезу принимают.

Критическими точками (границами) $k_{\text{кр}}$ называются точки, отделяющие критическую область от области принятия гипотезы. Правосторонней называют критическую область, определяемую неравенством $K>k_{\text{кр}}$, где $k_{\text{кр}}>0$. Левосторонней называют критическую область, определенную неравенством $K<k_{\text{кр}}$, где $k_{\text{кр}}<0$. Двусторонней называют критическую область, определяемую неравенствами $K<k_1$, $K>k_2$, где $k_2>k_1$.

Правосторонняя критическая область определяется исходя из требования, что при справедливости нулевой гипотезы выполняется соотношение $P(K>k_{кр})=\alpha$. Так как α – малая вероятность, то событие «$K>k_{кр}$» при справедливости нулевой гипотезы в силу принципа практической невозможности маловероятных событий в единичном испытании не должно произойти. Если же оно произошло, то это можно объяснить тем, что нулевая гипотеза ложна.

Левосторонняя критическая область определяется соотношением $P(K<k_{кр})=\alpha$. Двусторонняя критическая область определяется соотношением $P(K<k_1)+P(K<k_{кр})=\alpha$.

Если нулевая гипотеза принята, то это еще не значит, что она доказана, так как один пример этого не доказывает. Более правильно говорить, что данные наблюдений согласуются с нулевой гипотезой, и, следовательно, не дают оснований ее отвергнуть. Отвергают гипотезу более категорично, чем принимают, поскольку достаточно привести один пример, противоречащий некоторому общему утверждению, чтобы это утверждение отвергнуть.

Проверка гипотезы о нормальном распределении генеральной совокупности. Критерий согласия Пирсона

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

Критерием согласия называется критерий проверки гипотезы о предполагаемом законе неизвестного распределения. Рассмотрим критерий Пирсона. Будем рассматривать эмпирические (наблюдаемые) и теоретические (вычисленные в предположении нормального распределения) частоты. Обычно эмпирические и теоретические частоты различаются. Критерий Пирсона отвечает на вопрос: является ли распределение частот случайным (незначимым) или же расхождение вызвано тем, что теоретические частоты вычислены исходя из неверного предположения о нормальном распределении.

Пусть по выборке объема n получено эмпирическое распределение:

Варианты x_i	x_1	x_2	…	x_s
Эмпирические частоты n_i	n_1	n_2	…	n_s

В предположении нормального распределения генеральной совокупности вычислим теоретические частоты n'_i.

При уровне значимости α требуется проверить нулевую гипотезу: генеральная совокупность распределена нормально. В качестве критерия проверки нулевой гипотезы примем случайную величину $\chi^2 = \sum_{i=1}^{s} \frac{(n_i - n'_i)^2}{n'_i}$. Эта величина случайная, так как в различных опытах она принимает различные, заранее неизвестные значения.

Доказано, что при $n \to \infty$ закон распределения данной случайной величины независимо от того, какому закону распределения подчинена генеральная совокупность, стремится к закону распределения χ^2 с k степенями свободы. Число степеней свободы находится по формуле: k=s-1-r, где s – число групп выборки, r – число параметров предполагаемого распределения. В случае нормального распределения – 2 параметра (математическое ожидание и среднее квадратическое отклонение), поэтому r=2 и число степеней свободы k=s-3.

Так как односторонний критерий более жестко отвергает нулевую гипотезу, чем двусторонний, строят правостороннюю критическую область (значения χ^2 положительны): $P(\chi^2 > \chi^2_{кр}(\alpha;k)) = \alpha$.

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

Правило проверки нулевой гипотезы. Для того чтобы при заданном уровне значимости проверить нулевую гипотезу H_0: генеральная совокупность распределена нормально, надо:

1) вычислить теоретические частоты;

2) найти наблюдаемое значение критерия по формуле $\chi^2_{набл} = \sum_{i=1}^{s} \frac{(n_i - n'_i)^2}{n'_i}$;

3) по таблице критических точек распределения χ^2 (приложение 6) по указанному уровню значимости α и числу степеней свободы k=s-3 найти критические точки $\chi^2_{кр}(\alpha;k)$;

4) если $\chi^2_{набл} < \chi^2_{кр}$, то нет оснований отвергнуть нулевую гипотезу; если $\chi^2_{набл} > \chi^2_{кр}$ – нулевую гипотезу отвергают.

Вычисление теоретических частот. Пусть эмпирическое распределение задано в виде вариационного ряда:

Варианты x_i	x_1	x_2	…	x_s
Эмпирические частоты n_i	n_1	n_2	…	n_s

Тогда теоретические частоты вычисляются по формуле: $n'_i = \frac{nh}{\sigma_B} \cdot \varphi(u_i)$, где n – объем выборки, $h = x_{i+1} - x_i$, $u_i = \frac{x_i^* - \overline{x}_B}{\sigma_B}$, $\overline{x}_B = \frac{1}{n}\sum_{i=1}^{k} x_i n_i$, $\sigma_B = \sqrt{\frac{1}{n}\sum_{i=1}^{k}(x_i - \overline{x}_B)^2 \cdot n_i}$, $\varphi(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}}$; $\varphi(u)$ – табулирована (приложение 1). σ_B можно найти и по формуле: $\sigma_B = \sqrt{\frac{1}{n}\sum_{i=1}^{k} x_i^2 n_i - (\overline{x}_B)^2}$.

Проверка статистических гипотез о параметрах распределения

Пусть генеральные совокупности X и $У$ распределены нормально. Из совокупности X извлечена выборка объемом n_x, а из совокупности $У$ – объемом n_y. Пусть для этих выборок вычислены исправленные дисперсии S_x^2 и S_y^2 соответственно. Требуется проверить при уровне значимости α нулевую гипотезу о равенстве дисперсий генеральных совокупностей: H_0: $D_x = D_y$.

Для проверки гипотезу о равенстве дисперсий двух нормальных распределений используют случайную величину, распределенную по закону Фишера-Снедекора $F = \frac{S_x^2}{S_y^2}$, где в числителе большая из вычисленных исправленных выборочных дисперсий.

Критическое значение $F_{крит}$ следует находить с помощью таблицы распределения Фишера-Снедекора (приложение 7) по уровню значимости α и числу степеней свободы $k_1 = n_x - 1$ и $k_2 = n_y - 1$,

где k_1 — число степеней свободы большей (по величине) дисперсии;

k_2 — число степеней свободы меньшей (по величине) дисперсии;

n_x — объем выборки большей (по величине) дисперсии;

n_y — объем выборки меньшей (по величине) дисперсии.

При этом, если в качестве альтернативной гипотезы выступает H_1: $D_x > D_y$, то $F_{крит} = F(\alpha, k_1, k_2)$, если H_1: $D_x \neq D_y$, то $F_{крит} = F\left(\frac{\alpha}{2}, k_1, k_2\right)$. Если вычисленное значение $F < F_{крит}$, то нет оснований отвергать нулевую гипотезу.

Пусть требуется проверить нулевую гипотезу о равенстве математических ожиданий нормально распределенных генеральных совокупностей H_0: $M_x = M_y$.

Здесь возможны случаи:

1) дисперсии генеральных совокупностей известны;

2) дисперсии генеральных совокупностей неизвестны и выборки имеют объем больший 30;

3) дисперсии генеральных совокупностей неизвестны и выборки имеют объем меньший 30.

В первом случае используется нормированная нормально распределенная случайная величина $Z = \frac{\bar{x}_B - \bar{y}_B}{\sqrt{\frac{D_x}{n_x} + \frac{D_y}{n_y}}}$. Если альтернативная гипотеза H_1: $M_x > M_y$ или H_1: $M_x < M_y$,то $Z_{крит}$ находят по таблице функции Лапласа (приложение 2) из соотношения $\Phi(Z_{крит}) = \frac{1-2\alpha}{2}$. Если H_1: $M_x \neq M_y$, то по той же таблице $Z_{крит}$ находят из соотношения $\Phi(Z_{крит}) = \frac{1-\alpha}{2}$. При $|Z| < Z_{крит}$ нет оснований отвергать нулевую гипотезу.

Во втором случае используется приближенно нормально распределенная величина $Z'=\dfrac{\overline{x}_B-\overline{y}_B}{\sqrt{\dfrac{(D_x)_B}{n_x}+\dfrac{(D_y)_B}{n_y}}}$, для вычисления которой находят выборочные дисперсии $(D_x)_B$ и $(D_y)_B$. Вычисление критического значения, в зависимости от вида конкурирующей гипотезы, и правило принятия нулевой гипотезы аналогичны первому случаю.

В третьем случае предварительно требуется выполнить проверку гипотез о равенстве дисперсий генеральных совокупностей. Гипотеза о равенстве математических ожиданий генеральных совокупностей (в случае равенства дисперсий) проверяется с помощью случайной величины имеющей распределение t-Стьюдента: $T=\dfrac{\overline{x}_B-\overline{y}_B}{\sqrt{(n_x-1)S_x^2+(n_y-1)S_y^2}}\sqrt{\dfrac{n_x n_y(n_x+n_y-2)}{n_x+n_y}}$. Критическое значение $t_{крит}$ следует находить с помощью таблицы распределения Стьюдента (приложение 3) по уровню значимости α и числу степеней свободы $k = n_x+n_y-2$. Если альтернативная гипотеза H$_1$: $M_x>M_y$ или H$_1$: $M_x<M_y$, то α выбирают по нижней строке таблицы приложения 3. Если H$_1$: $M_x\neq M_y$, то α выбирают по верхней строке таблицы приложения 3. При $|T|<t_{крит}$ нет оснований отвергать нулевую гипотезу.

Пример выполнения лабораторной работы №5

Задание. Используя критерий Пирсона, при уровне значимости α=0,05 проверить, согласуется ли гипотеза о нормальном распределении генеральной совокупности с эмпирическим распределением выборки.

Дана выборка объема n=60:

x_i	0,3	0,5	0,7	0,9	1,1	1,3	1,5	1,7	1,9
n_i	4	6	7	8	10	9	7	5	4

Решение. Составим таблицу 7.

Таблица 7 – Расчет теоретических частот

1	2	3	4	5	6	7	8
x_i	n_i	$x_i n_i$	$x_i-\overline{x}_B$	$x_i^2 n_i$	$u_i=\dfrac{x_i-\overline{x}_B}{\sigma_B}$	$\varphi(u_i)$	$n'_i=\dfrac{nh}{\sigma_B}\varphi(u_i)$
0,3	4	1,2	-0,8	0,36	-2	0,054	2

1	2	3	4	5	6	7	8
0,5	6	3	-0,6	1,5	-1,5	0,1295	4
0,7	7	4,9	-0,4	3,43	-1	0,242	7
0,9	8	7,2	-0,2	6,48	-0,5	0,3521	11
1,1	10	11	0	12,1	0	0,3989	12
1,3	9	11,7	0,2	15,21	0,5	0,3521	11
1,5	7	10,5	0,4	15,75	1	0,242	7
1,7	5	8,5	0,6	14,45	1,5	0,1295	4
1,9	4	7,6	0,8	14,44	2	0,054	2
$\sum_{i=1}^{s}$	60	65,6		83,72			60

По результату колонки 3 находим $\overline{x}_B = \frac{\sum_{i=1}^{9} x_i n_i}{60} = \frac{1}{60} \cdot 65{,}6 = 1{,}1$.

По результату колонки 5 находим $\sigma_B = \sqrt{\frac{\sum_{i=1}^{9} x_i^2 n_i}{60} - \left(\overline{x}_B\right)^2} = \sqrt{\frac{1}{60} \cdot 83{,}72 - (1{,}1)^2} \approx 0{,}4$. Так как $h = x_{i+1} - x_i = 0{,}2$, то теоретические частоты будем вычислять по формуле: $n'_i = \frac{60 \cdot 0{,}2}{0{,}4} \cdot \varphi(u_i) = 30 \cdot \varphi(u_i)$.

Для того чтобы вычислить значение $\chi^2_{набл} = \sum_{i=1}^{s} \frac{\left(n_i - n'_i\right)^2}{n'_i}$, составим таблицу 8.

Таблица 8 – Расчет значения $\chi^2_{набл}$

n_i	n'_i	$n_i - n'_i$	$\left(n_i - n'_i\right)^2$	$\frac{\left(n_i - n'_i\right)^2}{n'_i}$	$\left(n_i\right)^2$	$\frac{\left(n_i\right)^2}{n'_i}$
4	2	2	4	2	16	8
6	4	2	4	1	36	9
7	7	0	0	0	49	7
8	11	-3	9	0,82	64	5,82
10	12	-2	4	0,33	100	8,33
9	11	-2	4	0,36	81	7,36
7	7	0	0	0	49	7
5	4	1	1	0,25	25	6,25
4	2	2	4	2	16	8
$\sum = 60$				$\chi^2_{набл} = 6{,}76$		$\sum = 66{,}76$

Для контроля вычислений формулу $\chi^2_{набл} = \sum_{i=1}^{s} \frac{\left(n_i - n'_i\right)^2}{n'_i}$ преобразуем к виду $\chi^2_{набл} = \sum_{i=1}^{s} \frac{\left(n_i\right)^2}{n'_i} - n$ (контроль $\chi^2_{набл} = 6{,}76$). Так как $\chi^2_{набл} = \sum_{i=1}^{s} \frac{\left(n_i\right)^2}{n'_i} - n = 66{,}76 - 60 = 6{,}76$, то вычисления произведены правильны.

Найдем число степеней свободы, учитывая, что число групп выборки (число различных вариант) равно s=9: k=9-3=6. По таблице критических точек распределения χ^2 (приложение 6) по уровню значимости α=0,05 и числу степеней свободы k=6, находим $\chi^2_{кр}(0{,}05;6) = 12{,}6$.

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

Пример выполнения лабораторной работы №6

<u>Задание.</u> 1. Из двух нормально распределенных генеральных совокупностей извлечены выборки:

x_i:	41	50	35	45	53	30	57	20	44	36	48	55	28		
y_i:	40	55	52	38	20	25	47	53	48	54	51	39	49	46	45

Требуется при уровне значимости α=0,05 проверить гипотезу о равенстве математических ожиданий генеральных совокупностей, если:

а) дисперсии генеральных совокупностей D_x=130 и D_y=115;

б) дисперсии генеральных совокупностей неизвестны.

В качестве конкурирующей гипотезы принять H_1: $M_x \neq M_y$.

2. По двум независимым выборкам объемов n_x=100, n_y=120 вычислены выборочные средние $\bar{x}_B$=32,4 и $\bar{y}_B$=30,1 и выборочные дисперсии $(D_x)_B$=15, $(D_y)_B$=25.

Проверить при уровне значимости α=0,05 гипотезу о равенстве математических ожиданий генеральных совокупностей при конкурирующей гипотезе H_1: $M_x<M_y$.

<u>Решение.</u> 1. а) Вычислим выборочные средние:

$$\bar{x}_B = \frac{1}{13} \cdot 542 = 41{,}69;$$

$$\bar{y}_B = \frac{1}{15} \cdot 662 = 44{,}13.$$

Так как дисперсии генеральных совокупностей известны, наблюдаемое значение критерия найдем по формуле: $Z = \dfrac{\overline{x}_B - \overline{y}_B}{\sqrt{\dfrac{D_x}{n_x} + \dfrac{D_y}{n_y}}} = \dfrac{41{,}69 - 44{,}13}{\sqrt{\dfrac{130}{13} + \dfrac{115}{15}}} = -0{,}58$.

Критическое значение найдем из соотношения $\Phi(Z_{крит}) = \dfrac{1 - 0{,}05}{2} = 0{,}495$. По таблице (приложение 2) $Z_{крит}$=2,58. Так как $|Z| < Z_{крит}$ нет оснований отвергать нулевую гипотезу.

б) Дисперсии генеральных совокупностей неизвестны и объемы выборок меньше 30. Поэтому проверим сначала гипотезу о равенстве дисперсий генеральных совокупностей, используя критерий Фишера-Снедекора.

Вычисляем исправленные выборочные дисперсии:

$S_x^2 = \dfrac{1}{n_x - 1}\sum\limits_{i=1}^{n_x}(x_i - \overline{x}_B) = 128{,}06$ – исправленная дисперсия первой выборки;

$S_x^2 = \dfrac{1}{n_y - 1}\sum\limits_{i=1}^{n_y}(y_i - \overline{y}_B) = 105{,}98$ – исправленная дисперсия второй выборки.

Так как первая дисперсия больше, ее помещаем в числитель формулы $F = \dfrac{S_x^2}{S_y^2} = \dfrac{128{,}06}{105{,}98} = 1{,}21$.

Критическое значение $F_{крит}$ найдем с помощью таблицы распределения Фишера-Снедекора (приложение 7) по уровню значимости α=0,05 и числу степеней свободы $k_1=n_x–1=13–1=12$ и $k_2=n_y–1=15–1=14$. $F_{крит} = F(\alpha, k_1, k_2) = F(0{,}05;12;14) = 2{,}53$.

Вычисленное значение критерия меньше критического, следовательно, гипотезы генеральных совокупностей можно считать равными.

Далее сравним средние:

$$T = \frac{41{,}69 - 44{,}13}{\sqrt{(13-1)\cdot 128{,}06 + (15-1)\cdot 105{,}98}} \cdot \sqrt{\frac{13\cdot 15(13+15-2)}{13+15}} = -0{,}597.$$

Критическое значение $t_{крит}$ находим с помощью таблицы приложения 3, выбирая α по верхней строке (H$_1$: $M_x \neq M_y$).

Число степеней свободы $k=n_1+n_2-2=13+15-2=26$. $t_{крит}=t(0,05;26)=2,06$.

Наблюдаемое значение по модулю меньше критического, поэтому нет оснований отвергать нулевую гипотезу.

2. Воспользуемся приближенно нормально распределенной величиной $Z'=\dfrac{\overline{x}_B-\overline{y}_B}{\sqrt{\dfrac{(D_x)_B}{n_x}+\dfrac{(D_y)_B}{n_y}}}=\dfrac{32,4-30,1}{\sqrt{\dfrac{15}{100}+\dfrac{25}{120}}}=3,59$. Так как альтернативная гипотеза H_1: $M_x<M_y$, то $Z_{крит}$ найдем из соотношения $\Phi(Z_{крит})=\dfrac{1-2\alpha}{2}=\dfrac{1-2\cdot 0,05}{2}=0,495$. По таблице функции Лапласа (приложение 2) $Z_{крит}$=2,56. Наблюдаемое значение критерия больше критического, следовательно, нулевую гипотезу отвергаем в пользу конкурирующей. Иными словами, средние генеральных совокупностей различаются значимо.

Варианты лабораторных работ №5 и №6

Лабораторная работа №5

Задание. Используя критерий Пирсона, при уровне значимости α=0,05 проверить, согласуется ли гипотеза о нормальном распределении генеральной совокупности с эмпирическим распределением выборки.

Вариант 1

x_i	8	11	14	17	20	23	26	29	32
n_i	6	14	17	22	25	20	18	10	8

Вариант 2

x_i	3	5	7	9	11	13	15	17	19
n_i	6	10	12	13	15	14	10	6	4

Вариант 3

x_i	0,3	0,5	0,7	0,9	1,1	1,3	1,5	1,7	1,6
n_i	6	9	10	12	13	10	8	7	5

Вариант 4

x_i	1	3	5	7	9	11	13	15	17

n_i	6	7	10	13	14	12	8	6	4

Вариант 5

x_i	4	6	8	10	12	14	16	18	20
n_i	5	9	14	16	18	15	10	9	4

Вариант 6

x_i	5	8	11	14	17	20	23	26	29
n_i	5	9	12	13	14	12	10	6	4

Вариант 7

x_i	2	5	8	11	14	17	20	23	26
n_i	10	11	13	15	16	14	12	11	8

Вариант 8

x_i	1	2	3	4	5	6	7	8	9
n_i	2	6	10	13	22	20	13	10	4

Лабораторная работы №6

Задание.

1. Из двух нормально распределенных генеральных совокупностей извлечены выборки.

Требуется при уровне значимости α=0,05 проверить гипотезу о равенстве математических ожиданий генеральных совокупностей, если:

а) дисперсии генеральных совокупностей D_x и D_y;

б) дисперсии генеральных совокупностей неизвестны.

В качестве конкурирующей гипотезы принять H_1: $M_x \neq M_y$.

2. По двум независимым выборкам объемов n_x, n_y вычислить выборочные средние и выборочные дисперсии.

Проверить при уровне значимости α=0,05 гипотезу о равенстве математических ожиданий генеральных совокупностей при конкурирующей гипотезе H_1: $M_x < M_y$.

1. Вариант 1

x_i: 17 18 19 20 20 D_x=2

y_i:	18	18	19	20	20	21	$D_y=1$

Вариант 2

x_i:	0,52	0,53	0,54	0,55	0,55		$D_x=0{,}00020$
y_i:	0,53	0,53	0,54	0,55	0,55	0,56	$D_y=0{,}00015$

Вариант 3

x_i:	0,12	0,13	0,34	0,55	0,71		$D_x=0{,}07$
y_i:	0,1	0,25	0,54	0,56	0,56	0,69	$D_y=0{,}05$

Вариант 4

x_i:	23	35	41	52	63		$D_x=230$
y_i:	32	35	45	56	67	70	$D_y=260$

Вариант 5

x_i:	5,6	5,8	5,9	6,1	7,3		$D_x=0{,}4$
y_i:	5,1	5,3	5,5	5,9	5,9	7	$D_y=0{,}5$

Вариант 6

x_i:	2,3	3,5	4,1	5,2	6,3		$D_x=2{,}3$
y_i:	3,2	3,5	4,5	5,6	6,7	7,1	$D_y=2{,}5$

Вариант 7

x_i:	12	13	34	55	60		$D_x=500$
y_i:	10	25	54	56	56	73	$D_y=550$

Вариант 8

x_i:	1,52	1,53	1,53	1,55	1,55		$D_x=0{,}00020$
y_i:	1,52	1,53	1,54	1,55	1,55	1,55	$D_y=0{,}00015$

2. Вариант 1

x_i:	8	11	14	17	20	23	26	29	32
n_{xi}	5	26	27	27	28	20	14	8	5
y_i:	8	11	14	17	20	23	26	29	32
n_{yi}	6	14	17	22	25	20	18	10	8

Вариант 2

x_i:	3	5	7	9	11	13	15	17	19
n_{xi}	11	15	22	26	28	25	17	10	6
y_i:	3	5	7	9	11	13	15	17	19
n_{yi}	6	10	12	13	15	14	10	6	4

Вариант 3

x_i:	0,3	0,5	0,7	0,9	1,1	1,3	1,5	1,7	1,6
n_{xi}	10	12	16	18	24	20	10	6	4
y_i:	0,3	0,5	0,7	0,9	1,1	1,3	1,5	1,7	1,6
n_{yi}	6	9	10	12	13	10	8	7	5

Вариант 4

x_i:	1	3	5	7	9	11	13	15	17
n_{xi}	10	12	16	18	24	20	10	6	4
y_i:	1	3	5	7	9	11	13	15	17
n_{yi}	6	7	10	13	14	12	8	6	4

Вариант 5

x_i:	4	6	8	10	12	14	16	18	
n_{xi}	6	9	10	12	13	10	8	7	5
y_i:	4	6	8	10	12	14	16	18	20
n_{yi}	5	9	14	16	18	15	10	9	4

Вариант 6

x_i:	5	8	11	14	17	20	23	26	
n_{xi}	6	9	10	12	13	10	8	7	5
y_i:	5	8	11	14	17	20	23	26	29
n_{yi}	5	9	12	13	14	12	10	6	4

Вариант 7

x_i:	2	5	8	11	14	17	20	23	26
n_{xi}	2	6	10	13	22	20	13	10	4

y_i:	2	5	8	11	14	17	20	23	26
n_{yi}	10	11	13	15	16	14	12	11	8

Вариант 8

x_i:	1	2	3	4	5	6	7	8	9
n_{xi}	10	11	13	15	16	14	12	11	8
y_i:	1	2	3	4	5	6	7	8	9
n_{yi}	2	6	10	13	22	20	13	10	4

Тема 4. МЕТОД НАИМЕНЬШИХ КВАДРАТОВ

Основные понятия

При решении некоторых исследовательских задач приходится использовать формулы, составленные на основе наблюдений (эмпирические формулы). Самым распространенным способом нахождения эмпирических формул является метод наименьших квадратов. Рассмотрим идею метода на примере отыскания формулы, описывающей линейную зависимость между двумя величинами.

Пусть изучается система двух количественных признаков (X, $У$). В результате n независимых опытов получены n пар чисел: (x_1, y_1), (x_2, y_2),…, (x_n, y_n), где x_1, x_2, …, x_n – значения случайной величины X, а y_1, y_2,…, y_n – значения случайной величины $У$.

Будем считать, что между рассматриваемыми величинами существует приближенная линейная зависимость: $y = ax + b$. Предположение о наличии линейной зависимости можно сделать, опираясь на изображение точек (x_1, y_1), (x_2, y_2),…, (x_n, y_n) в прямоугольной системе координат xOy. Если точки группируются вдоль прямой линии, то такое предположение естественно.

Пусть требуется найти формулу $y = ax + b$, то есть конкретные значения коэффициентов a и b, наиболее точно описывающую зависимость между величинами $У$ и X.

Подставляя в искомую формулу значения переменной X, получим последовательность значений функции: $ax_1 + b$, $ax_2 + b$, …, $ax_n + b$. Это ординаты точек прямой, абсциссы которых являются наблюдаемыми значениями переменной X. Ординаты наблюдений (y_1, y_2,…, y_n) могут отличаться от ординат точек прямой. Разность между точным значением функции $ax_i + b$ в точке x_i и соответствующим значением y_i, полученным в результате опыта называют отклонением $\varepsilon_i = ax_i + b - y_i$. Очевидно, чем меньше отклонения, тем точнее формула описывает зависимость между $У$ и X. Логичным было бы найти сумму всех отклонений и устремить ее к нулю. Но отклонения ε_i могут быть различных знаков и при суммировании могут взаимно уничтожаться, поэтому рассматривают сумму квадратов отклонений: $F(a,b)=\sum_{i=1}^{n}\varepsilon_i^2=\sum_{i=1}^{n}\left(ax_i+b-y_i\right)^2$.

В методе наименьших квадратов на значения a и b накладывается условие – они должны доставлять минимум сумме квадратов отклонений $F(a,b)$. Требуется найти a и b, удовлетворяющие этому условию. Необходимым условием экстремума функции двух переменных является равенство нулю ее частных производных:

$$\begin{cases}\dfrac{\partial F}{\partial a}=2\sum_{i=1}^{n}(ax_i+b-y_i)\cdot x_i=2a\sum_{i=1}^{n}x_i^2+2b\sum_{i=1}^{n}x_i-2\sum_{i=1}^{n}y_ix_i=0,\\ \dfrac{\partial F}{\partial b}=2\sum_{i=1}^{n}(ax_i+b-y_i)=2a\sum_{i=1}^{n}x_i+2bn-2\sum_{i=1}^{n}y_ix_i=0.\end{cases}$$

Отсюда $\begin{cases}a\sum_{i=1}^{n}x_i^2+b\sum_{i=1}^{n}x_i=\sum_{i=1}^{n}y_ix_i,\\ a\sum_{i=1}^{n}x_i+bn=\sum_{i=1}^{n}y_i.\end{cases}$

Полученная система линейных уравнений с неизвестными a и b называется нормальной системой метода наименьших квадратов.

Вычислив по эмпирическим данным все суммы и подставив их в систему уравнений, можно найти решение системы: $a=a_0$, $b=b_0$. Решение системы (a_0; b_0) является точкой минимума функции $F(a,b)$. Доказательство этого факта мы рассматривать не будем.

Искомая эмпирическая формула линейной зависимости примет вид $y=a_0x+b_0$.

Из системы нормальных уравнений можно вывести формулы для расчета коэффициентов a и b.

$$a=\frac{n\cdot\sum_{i=1}^{n}x_iy_i-\sum_{i=1}^{n}y_i\sum_{i=1}^{n}x_i}{n\sum_{i=1}^{n}x_i^2-\left(\sum_{i=1}^{n}x_i\right)^2},\quad b=\frac{\sum_{i=1}^{n}x_i^2\sum_{i=1}^{n}y_i-\sum_{i=1}^{n}x_i\sum_{i=1}^{n}x_iy_i}{n\sum_{i=1}^{n}x_i^2-\left(\sum_{i=1}^{n}x_i\right)^2}.$$

Важно понимать смысл параметров a и b и уметь их адекватно интерпретировать. Коэффициент a показывает, на сколько единиц в среднем изменится (при a>0 увеличится, при a<0 уменьшится) величина Y при увеличении X на единицу. Параметр b интерпретируется в зависимости от того, какой смысл имеют изучаемые величины.

Будем считать, что между рассматриваемыми величинами существует приближенная квадратичная зависимость: $y = ax^2 + bx + c$.

Отклонение будет определяться по формуле $\varepsilon_i = ax_i^2 + bx_i + c - y_i$, а сумма квадратов отклонений является функцией трех переменных: $F(a,b,c) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \left(ax_i^2 + bx_i + c - y_i\right)^2$.

Можно показать, что для определения коэффициентов a, b и c следует решить систему уравнений:

$$\begin{cases} a\sum_{i=1}^{n} x_i^4 + b\sum_{i=1}^{n} x_i^3 + c\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_i x_i^2, \\ a\sum_{i=1}^{n} x_i^3 + b\sum_{i=1}^{n} x_i^2 + c\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i x_i, \\ a\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i + cn = \sum_{i=1}^{n} y_i. \end{cases}$$

Пример выполнения лабораторной работы №7

Задание. Даны результаты эксперимента

x_i	10	12	13	14	15	16	18	19	20	21
y_i	32	34	35	36	36	37	38	40	39	40

Требуется:

1) в предположении, что между x и y существует линейная зависимость, определить ее эмпирическое уравнение;

2) в предположении, что между x и y существует квадратичная зависимость, определить ее эмпирическое уравнение;

3) найти сумму квадратов отклонений для найденных зависимостей сравнить качество приближений.

Решение. Составим вспомогательную таблицу 9.

Таблица 9 – Расчеты сумм нормальной системы уравнений

№	x_i	y_i	x_i^2	x_i^3	x_i^4	$x_i y_i$	$x_i^2 y_i$
1	10	32	100	1000	10000	320	3200
2	12	34	144	1728	20736	408	4896

№	x_i	y_i	x_i^2	x_i^3	x_i^4	$x_i\,y_i$	$x_i^2\,y_i$
3	13	35	169	2197	28561	455	5915
4	14	36	196	2744	38416	504	7056
5	15	36	225	3375	50625	540	8100
6	16	37	256	4096	65536	592	9472
7	18	38	324	5832	104976	684	12312
8	19	40	361	6859	130321	760	14440
9	20	39	400	8000	160000	780	15600
10	21	40	441	9261	194481	840	17640
$\sum_{i=1}^{10}$	158	367	2616	45092	803652	5883	98631

Найдем уравнение линейной зависимости. Для этого составим систему уравнений $\begin{cases} a\cdot 2616 + b\cdot 158 = 5883, \\ a\cdot 158 + b\cdot 10 = 367. \end{cases}$

Решим систему по формулам Крамера: $\Delta=\begin{vmatrix} 2616 & 158 \\ 158 & 10 \end{vmatrix}=196$, $\Delta a=\begin{vmatrix} 5883 & 158 \\ 367 & 10 \end{vmatrix}=844$, $\Delta b=\begin{vmatrix} 2616 & 5883 \\ 158 & 367 \end{vmatrix}=30558$. Отсюда $a=\dfrac{844}{196}\approx 0{,}71$, $b=\dfrac{30558}{196}\approx 25{,}6$;

$y=0{,}71x+25{,}6$ – искомое эмпирическое уравнение линейной зависимости.

Далее найдем уравнение квадратичной зависимости. Нормальная система уравнений примет вид: $\begin{cases} a\cdot 803625 + b\cdot 45062 + c\cdot 2616 = 98631, \\ a\cdot 45062 + b\cdot 2616 + c\cdot 158 = 5883, \\ a\cdot 2616 + b\cdot 158 + c\cdot 10 = 367. \end{cases}$

Решая систему получим: a=0,31, b=–9,18, c=100,94. $y=0{,}31x^2-9{,}18x+100{,}94$ – искомое эмпирическое уравнение квадратичной зависимости.

Найдем сумму квадратов отклонений для найденных зависимостей. Вычисления оформим в таблицу 10.

Таблица 10 – Вычисление квадратов отклонений

x_i	y_i	$0{,}71x_i+25{,}6$	ε_i	$(\varepsilon_i)^2$	$0{,}31x_i^2-9{,}18x_i+100{,}94$	ε'_i	$(\varepsilon'_i)^2$
10	32	32,7	0,7	0,49	40,14	8,14	66,2596
12	34	34,12	0,12	0,0144	35,42	1,42	2,0164
13	35	34,83	-0,17	0,0289	33,99	-1,01	1,0201
14	36	35,54	-0,46	0,2116	33,18	-2,82	7,9524
15	36	36,25	0,25	0,0625	32,99	-3,01	9,0601
16	37	36,96	-0,04	0,0016	33,42	-3,58	12,8164

x_i	y_i	$0{,}71x_i + 25{,}6$	ε_i	$(\varepsilon_i)^2$	$0{,}31x_i^2 - 9{,}18x_i + 100{,}94$	ε'_i	$(\varepsilon'_i)^2$
18	38	38,38	0,38	0,1444	36,14	-1,86	3,4596
19	40	39,09	-0,91	0,8281	38,43	-1,57	2,4649
20	39	39,8	0,8	0,64	41,34	2,34	5,4756
21	40	40,51	0,51	0,2601	44,87	4,87	23,7169
$\sum_{i=1}^{10}$				2,6816			134,242

Как видим из таблицы, сумма квадратов отклонений для квадратичной зависимости многократно превышает сумму квадратов отклонений для линейной зависимости. Таким образом, линейная зависимость дает более качественное приближение, чем квадратичная.

Варианты лабораторной работы №7

<u>Задание.</u> Даны результаты эксперимента

Требуется:

1) в предположении, что между x и y существует линейная зависимость, определить ее эмпирическое уравнение;

2) в предположении, что между x и y существует квадратичная зависимость, определить ее эмпирическое уравнение;

3) найти сумму квадратов отклонений для найденных зависимостей сравнить качество приближений.

Вариант 1

X	7,5	7	8,3	8,4	6,9	7,7	8,1	7,6	7,9	8,2
Y	26	27	22	21	27	25	21	21	23	22

Вариант 2

X	36	31	32	34	30	33	31	34	35	32
Y	9,3	10,1	9,5	9,7	10,2	9,4	9,6	10,0	9,5	9,2

Вариант 3

X	9,7	10,4	10,3	9,8	10,1	10,2	10	9,9	9,6	9,8
Y	3,5	3,1	3,2	3,4	3	3,3	3,1	3,4	3,5	3,2

Вариант 4

X	10,2	9,5	9,7	9,8	9,4	9,6	10	10,1	9,3	9,6

Y	3,2	3,4	3,2	3,3	3,7	3,8	3	3,1	3,6	3,5

Вариант 5

X	4,5	4,3	4,7	4,7	5,2	4,6	4,9	4,2	4,3	5
Y	6,1	5	6,2	6,7	9	6,5	8,2	4,5	4,6	6,8

Вариант 6

X	8,3	7,2	6,9	9	9	9,5	9,5	9,1	7,5	7
Y	5,6	4,2	1,8	8,4	5,6	10,7	9	6,8	3,1	4,8

Вариант 7

X	42	50	52	48	49	56	53	46	55	45
Y	5	5,9	7,8	6,3	7,3	6,8	6,5	6,2	7	6,2

Вариант 8

X	8,5	9,4	9,2	10,4	10,1	9,8	9,3	8,7	9,9	9,5
Y	5,6	6,3	6	7	6,4	5,9	6,1	4,9	6,8	6,5

Тема 5. ЭЛЕМЕНТЫ ТЕОРИИ КОРРЕЛЯЦИИ

Основные понятия

Во многих задачах требуется установить и оценить зависимость изучаемой случайной величины $У$ от одной или нескольких других величин.

Две случайные величины могут быть связаны либо функциональной зависимостью, либо статистической зависимостью, либо могут быть независимыми.

Строгая функциональная зависимость реализуется редко. Если $У$ зависит от случайных факторов Z_1, Z_2, V_1, V_2, а X зависит от Z_1, Z_2, U_1, U_2, то между $У$ и X имеется статистическая зависимость, так как среди случайных факторов есть общие, а именно Z_1, Z_2.

Статистической называется зависимость, при которой изменение одной из величин влечет изменение распределения другой величины. В частности, статистическая зависимость проявляется в том, что при изменении одной из величин изменяется среднее значение другой величины; в этом случае статистическая зависимость называется корреляционной.

Определить форму связи – значит выявить механизм получения зависимой случайной переменной. При изучении статистических зависимостей форму связи можно характеризовать функцией регрессии (линейной, квадратной, показательной и т.д.)

Функцией регрессии случайной величины $У$ относительно X называется условное математическое ожидание $M(У/X{=}x)$ случайной величины $У$, рассматриваемое как функция от x, то есть $M(У/X{=}x)=\varphi(x)$.

Функция регрессии имеет важное значение при статистическом анализе зависимостей между величинами и может быть использована для прогнозирования одной из случайных переменных, если известно значение другой случайной переменной.

В качестве оценки условного математического ожидания используют условное среднее $\overline{y}_x$, которое так же является функцией от x: $\overline{y}_x = \varphi^*(x)$.

Условным средним $\overline{y}_x$ называют среднее арифметическое наблюдаемых значений *У*, соответствующих значению *X*=*x*, которое приняла величина *X*. Уравнение $\overline{y}_x = \varphi^*(x)$ называется выборочным уравнением регрессии *У* на *X* и используется для характеристики формы связи при изучении зависимости величин *У* и *X*. График $\varphi^*(x)$ называется выборочной линией регрессии, которая показывает, как в среднем зависит *У* от *X*. Аналогично определяется выборочное уравнение регрессии *X* на *У*.

Основными задачами корреляционного анализа являются:

1) установление формы корреляционной зависимости (линейная, квадратичная, и т.д.);

2) оценка тесноты корреляционной связи.

Наиболее часто функции регрессии оказываются линейными. Теснота корреляционной связи оценивается по величине рассеяния значений вокруг условного среднего $\overline{y}_x$. Большое расстояние говорит о слабой зависимости *У* от *X*; малое расстояние указывает на наличие достаточно сильной зависимости.

Вопрос о т ом, что принять за зависимую переменную, а что – за независимую, следует решать применительно к каждому конкретному случаю.

Знание статистической зависимости между случайными переменными имеет большое практическое значение: с ее помощью можно прогнозировать значение зависимой случайной переменной в предположении, что независимая переменная примет определенное значение. Однако, поскольку статистической зависимости относится к осредненным условиям, прогнозы не могут быть безошибочными.

Пусть изучается система количественных признаков (*X*, *У*). В результате n независимых опытов получены n пар чисел: (x_1, y_1), (x_2, y_2),…, (x_n, y_n), где x_1, x_2, …, x_n – значения случайной величины *X*, а y_1, y_2,…, y_n – значения случайной величины *У*.

Данные пары значений можно изобразить точками на плоскости. Если эти точки оказываются расположенными вблизи некоторой прямой, то можно предположить, что между величинами X и $У$ существует линейная связь. Тогда возникает вопрос: как отыскать прямую, которая наилучшим образом прилегает ко всем точкам. Другими словами, задача сводится к тому, чтобы по данным наблюдений найти коэффициенты a и b выборочного уравнения $\overline{y}_x = ax + b$ прямой линии регрессии $У$ на X.

Коэффициент a в этом уравнении называется выборочным коэффициентом регрессии $У$ на X.

Коэффициенты a и b находят методом наименьших квадратов (см. тему «Метод наименьших квадратов»):

$$a = \frac{n \cdot \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}, \; b = \frac{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}.$$

В формуле для нахождения a разделим числитель и знаменатель на n^2:

$$a = \frac{\frac{1}{n} \cdot \sum_{i=1}^{n} x_i y_i - \frac{1}{n} \cdot \sum_{i=1}^{n} y_i \cdot \frac{1}{n} \cdot \sum_{i=1}^{n} x_i}{\frac{1}{n} \cdot \sum_{i=1}^{n} x_i^2 - \left(\frac{1}{n} \cdot \sum_{i=1}^{n} x_i\right)^2}.$$ Учитывая, что $\frac{1}{n} \cdot \sum_{i=1}^{n} x_i y_i = M(XУ)$, $\frac{1}{n} \cdot \sum_{i=1}^{n} x_i = M(X)$, $\frac{1}{n} \cdot \sum_{i=1}^{n} y_i = M(У)$, $\frac{1}{n} \cdot \sum_{i=1}^{n} x_i^2 = M(X^2)$, имеем

$$a = \frac{M(XУ) - M(X)M(У)}{M(X^2) - M^2(X)} = \frac{M(XУ) - M(X)M(У)}{D(X)} = \frac{M(XУ) - M(X)M(У)}{\sqrt{D(X)} \cdot \sqrt{D(X)}} \cdot \frac{\sqrt{D(X)}}{\sqrt{D(X)}}.$$

Введем обозначение: $r_B = \dfrac{M(XУ) - M(X)M(У)}{\sqrt{D(X)} \cdot \sqrt{D(X)}}$. Тогда коэффициент a можно переписать в виде: $a = r_B \cdot \dfrac{\sigma_y}{\sigma_x}$, r_B – выборочный коэффициент корреляции, который измеряет силу (тесноту) линейной связи между величинами X и $У$. Для расчета выборочного коэффициента корреляции используется рабочая формула:

$$r_B = \frac{n \cdot \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i}{\sqrt{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \cdot \sqrt{n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2}}.$$

Свойства коэффициента корреляции:

1. Значения r_B заключены в промежутке от -1 до 1;

2. Если X и $У$ независимы, то r_B=0 (для независимых случайных величин $M(XУ) = M(X)M(У)$). Но если r_B=0, то это не означает, что X и $У$ независимы, так как r_B характеризует тесноту линейной связи, а между X и $У$ может существовать другая связь (нелинейная).

Если r_B>0, то увеличение признака X в среднем приводит к увеличению признака $У$ (связь между величинами X и $У$ прямая). Если r_B<0, то с увеличением признака X в среднем признак $У$ уменьшается (связь между X и $У$ обратная).

Если $|r_B|$=1, то между X и $У$ существует линейная функциональная зависимость, не искажаемая действием случайных факторов.

Для качественной оценки тесноты корреляционной связи между X и $У$ можно воспользоваться таблицей Чеддока (табл. 11).

Таблица 11 – Оценка тесноты корреляционной связи

Диапазон изменения $\|r_B\|$	0,1–0,3	0,3–0,5	0,5–0,7	0,7–0,9	0,9–1
Характер тесноты связи	Слабая	Умеренная	Заметная	Высокая	Весьма высокая

Выборочный коэффициент корреляции r_B является точечной оценкой коэффициента корреляции r. Равенство r_B=0 еще не свидетельствует о том, что r=0, следовательно, о некоррелированности случайных величин X и $У$. Чтобы выяснить, находятся ли случайные величины в корреляционной зависимости, нужно проверить значимость выборочного коэффициента корреляции, то есть установить, достаточна ли его величина для обоснования вывода о наличии корреляционной связи. Для этого проверяют нулевую гипотезу H_0: r=0.

Объем выборки может быть любой. Вычисляют статистику $t = r\sqrt{\dfrac{n-2}{1-r^2}}$, которая имеет распределение Стьюдента с k=n-2 степенями свободы. Для проверки нулевой гипотезы по уровню значимости α и числу степеней свободы k находят по таблице распределения Стьюдента критическое (приложение 3) значение $t_{\alpha,k}$, удовлетворяющее условию: $P(|t| \ge t_{\alpha,k}) = \alpha$. Если $|t| \ge t_{\alpha,k}$, то нулевую гипотезу об отсутствии корреляционной связи между величинами X и $У$ следует отвергнуть. Переменные считаются зависимыми. При $|t| < t_{\alpha,k}$ нет оснований отвергнуть нулевую гипотезу.

Пример выполнения лабораторной работы №8

Задание. Дана выборка

X	51	50	33	40	42	51	52	51	55	36	53
$У$	70	56	31	48	73	72	40	66	76	34	63

По заданной выборке:

1) найти уравнение прямой линии регрессии $У$ на X;

2) оценить тесноту линейной связи, вычислив выборочный коэффициент корреляции;

3) проверить гипотезу о значимости коэффициента корреляции при уровне значимости 0,1.

Решение. Составим вспомогательную таблицу 12.

Таблица 12 – Расчет коэффициента корреляции

№	x_i	y_i	x_i^2	y_i^2	$x_i y_i$
1	51	70	2601	4900	3570
2	50	56	2500	3136	2800
3	33	31	1089	961	1023
4	40	48	1600	2304	1920
5	42	73	1764	5329	3066
6	51	72	2601	5184	3672
7	52	40	2704	1600	2080
8	51	66	2601	4356	3366
9	55	76	3025	5776	4180
10	36	34	1296	1156	1224
11	53	63	2809	3969	3339
$\sum_{i=1}^{11}$	514	629	24590	38671	30240

Отсюда $a = \dfrac{11 \cdot 30240 - 629 \cdot 514}{11 \cdot 24590 - 514^2} = 1{,}4$; $b = \dfrac{24590 \cdot 629 - 514 \cdot 30240}{11 \cdot 24590 - 514^2} = -9{,}6$.

Тогда уравнение регрессии имеет вид: $\overline{y}_x = 1{,}4x - 9{,}6$.

Далее, $r_B = \dfrac{11 \cdot 30240 - 629 \cdot 514}{\sqrt{11 \cdot 24590 - 514^2} \cdot \sqrt{11 \cdot 38671 - 629^2}} \approx 0{,}68$.

По значению коэффициента корреляции $r_B = 0{,}68$ можно сделать вывод о том, что между величинами *X* и *У* существует заметная линейная корреляционная связь.

Выдвинем нулевую гипотезу H_0: $r=0$. Вычислим статистику *t*: $t = 0{,}68\sqrt{\dfrac{9}{1-0{,}68^2}} = 2{,}78.$ По числу степеней свободы *k*=11–2=9 и уровню значимости *α*=0,05 из таблицы приложения 3 находим: *t*(0,05; 9)=2,26. Так как *t* >*t*(0,05; 9), то нулевую гипотезу отвергаем. Переменные *X* и *У* зависимы.

Варианты лабораторной работы № 8

Вариант 1

X	9,7	10,4	10,3	9,8	10,1	10,2	10,0	9,9	9,6	9,8
Y	3,5	3,1	3,2	3,4	3,0	3,3	3,1	3,4	3,5	3,2

Вариант 2

X	36	31	32	34	30	33	31	34	35	32
Y	9,3	10,1	9,5	9,7	10,2	9,4	9,6	10,0	9,5	9,2

Вариант 3

X	7,5	7,0	8,3	8,4	6,9	7,7	8,1	7,6	7,9	8,2
Y	26	27	22	21	27	25	21	21	23	22

Вариант 4

X	8,3	7,2	6,9	9,0	9,0	9,5	9,5	9,1	7,5	7,0
Y	5,6	4,2	1,8	8,4	5,6	10,7	9,0	6,8	3,1	4,8

Вариант 5

X	42	50	52	48	49	56	53	46	55	45
Y	5,0	5,9	7,8	6,3	7,3	6,8	6,5	6,2	7,0	6,2

Вариант 6

X	8,5	9,4	9,2	10,4	10,1	9,8	9,3	8,7	9,9	9,5
Y	5,6	6,3	6,0	7,0	6,4	5,9	6,1	4,9	6,8	6,5

Вариант 7

X	10,2	9,5	9,7	9,8	9,4	9,6	10,0	10,1	9,3	9,6
Y	3,2	3,4	3,2	3,3	3,7	3,8	3,0	3,1	3,6	3,5

Вариант 8

X	4,5	4,3	4,7	4,7	5,2	4,6	4,9	4,2	4,3	5,0
Y	6,1	5,0	6,2	6,7	9,0	6,5	8,2	4,5	4,6	6,8

Корреляционная таблица.

Отыскание прямых регрессий по сгруппированным данным

При большом числе наблюдений одно и тоже значение величины *X* может встречаться n_x раз, значение *У* – n_y раз, а пара (*x*, *y*) – n_{xy} раз. Поэтому данные группируют в виде таблицы, которая называется корреляционной. Примером может служить таблица 13.

Таблица 13 – Корреляционная таблица

У \ X	5	15	25	35	45	n_y
30	7	1				8
32	2	7	1			10
34	1	5	4	1		11
36		1	15	10	8	34
38			3	12	15	30
40				1	6	7
n_x	10	14	23	24	29	100

В первой строке таблицы указаны наблюдаемые значения признака *X*, в первом столбце – наблюдаемые значения признака *У*. На пересечении строк и столбцов находятся частоты n_{xy} наблюдаемых пар значений признаков. В последнем столбце записываются суммы частот строк, а в последней строке – суммы частот столбцов. Очевидно, что $\sum_{j=1}^{k} n_x = \sum_{i=1}^{k} n_y = n$.

Если в корреляционной таблице заполненные клетки вблизи той или другой диагонали, то есть смысл отыскивать линейную функцию регрессии.

Уравнение регрессии *У* на *X*, $\overline{y}_x - \overline{y} = r_B \cdot \frac{\sigma_y}{\sigma_x}(x - \overline{x})$, $r_B = \frac{\sum n_{xy} xy - n\sum \overline{x} \cdot \overline{y}}{n\sigma_x\sigma_y}$ – выборочный коэффициент корреляции, где $\sigma_x^2 = \overline{x}^2 - (\overline{x})^2$, $\sigma_y^2 = \overline{y}^2 - (\overline{y})^2$.

Аналогично можно получить уравнение регрессии *X* на *У*: $\overline{x}_y - \overline{x} = r_B \cdot \frac{\sigma_x}{\sigma_y}(y - \overline{y})$.

Пример выполнения лабораторной работы №9

Задание. По корреляционной таблице 14

1) в прямоугольной системе координат построить эмпирическую линию регрессии *У* на *X*;

2) оценить тесноту линейной связи;

3) провести гипотезу о значимости коэффициента корреляции при уровне значимости 0,05;

4) составить линейное уравнение регрессии *У* на *X* и построить линию регрессии в той же системе координат, где построена эмпирическая линия регрессии;

5) использую полученное уравнение регрессии, оценить ожидаемое среднее значение признака *У* при $x = x_0 = 45$.

Таблица 14 – Корреляционная таблица к примеру лабораторной работы № 9

У \ *X*	20	30	40	50	60	n_y
1	8	2				10
3	12	20	8			40
5			10	1		11
7			9	6	2	17
9			10	4	8	22
n_x	20	22	37	11	10	100

<u>Решение.</u> Для построения ломанной регрессии *У* на *X* найдем условное среднее $\overline{У}_x$.

При *X*=20 признак *У* имеет распределение: $\begin{vmatrix} У & 1 & 3 \\ n_i & 8 & 12 \end{vmatrix}$. То есть условное среднее $\overline{У}_{x=20} = \dfrac{1 \cdot 8 + 3 \cdot 12}{8 + 12} = 2{,}2$. При *X*=30 признак *У* имеет распределение: $\begin{vmatrix} У & 1 & 3 \\ n_i & 2 & 20 \end{vmatrix}$. Тогда $\overline{У}_{x=30} = \dfrac{1 \cdot 2 + 3 \cdot 20}{2 + 20} = 2{,}8$. При *X*=40 признак *У* имеет распределение: $\begin{vmatrix} У & 3 & 5 & 7 & 9 \\ n_i & 8 & 10 & 9 & 10 \end{vmatrix}$. Тогда, $\overline{У}_{x=40} = \dfrac{3 \cdot 8 + 5 \cdot 10 + 7 \cdot 9 + 9 \cdot 10}{8 + 10 + 9 + 10} = 6{,}1$.

Аналогично получаем: $\overline{У}_{x=50} = 7{,}5$; $\overline{У}_{x=60} = 8{,}6$.

Результаты оформим в таблицу 15, выражающую зависимость $\overline{У}$ от *X*.

Таблица 15 – Зависимость $\overline{У}$ от *X*

X	20	30	40	50	60
$\overline{У}$	2,2	2,8	6,1	7,5	8,6

В прямоугольной системе координат отмечаем точки $A_i(x_i, \bar{y}_i)$ и соединяем их отрезками (рис. 5).

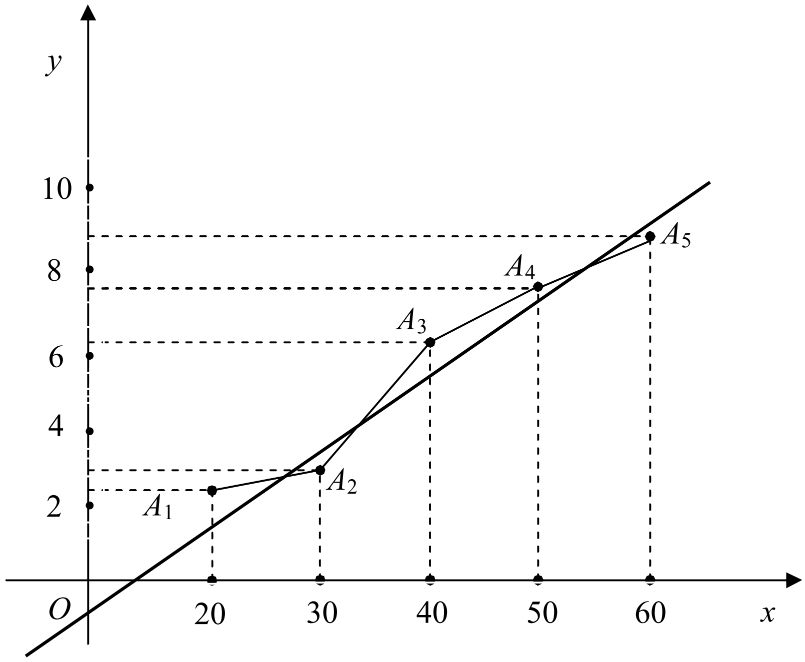


Рис. 5 – Линия регрессии

Построенная эмпирическая линия регрессии свидетельствует о том, что меду признаками X и $У$ существует линейная зависимость.

Оценим тесноту связи, вычислив выборочный коэффициент корреляции:

$$\bar{x} = \frac{1}{100}(20\cdot 20 + 30\cdot 2{,}2 + 40\cdot 37 + 50\cdot 11 + 60\cdot 10) = 36{,}9,$$

$$\bar{y} = \frac{1}{100}(1\cdot 10 + 3\cdot 40 + 5\cdot 11 + 7\cdot 17 + 9\cdot 22) = 5{,}02,$$

$$\bar{x}^2 = \frac{1}{100}(20^2\cdot 20 + 30^2\cdot 2{,}2 + 40^2\cdot 37 + 50^2\cdot 11 + 60^2\cdot 10) = 1505,$$

$$\bar{y}^2 = \frac{1}{100}(1^2\cdot 10 + 3^2\cdot 40 + 5^2\cdot 11 + 7^2\cdot 17 + 9^2\cdot 22) = 52{,}6,$$

$$\Sigma n_{xy}xy = 1\cdot 20\cdot 8 + 1\cdot 30\cdot 2 + 3\cdot 20\cdot 12 + 3\cdot 30\cdot 20 + 3\cdot 40\cdot 8 + 5\cdot 40\cdot 10 + 5\cdot 50\cdot 1 + 7\cdot 40\cdot 9 + 7\cdot 50\cdot 6 +$$
$$+ 7\cdot 60\cdot 2 + 9\cdot 40\cdot 10 + 9\cdot 50\cdot 4 + 9\cdot 60\cdot 80 = 21130,\ \frac{1}{n}\Sigma n_{xy}xy = \frac{1}{100}\cdot 21130 = 211{,}3;$$

$$\sigma_x^2 = 1505 - (36{,}9)^2 = 143{,}39;\ \sigma_y^2 = 32{,}6 - (5{,}02)^2 = 7{,}3996,\ \sigma_y = 2{,}72,\ \sigma_x = 11{,}97;$$

$r_B = \dfrac{21130 - 100\cdot 36{,}9\cdot 5{,}02}{100\cdot 11{,}97\cdot 2\cdot 72} = 0{,}80$. Отсюда следует, что корреляционная зависимость между X и $У$ высокая.

Выдвинем нулевую гипотезу H_0: $r=0$. Вычислим статистику $t = 0{,}8\sqrt{\frac{98}{1-0{,}8^2}} = 13{,}20.$ По числу степеней свободы k=100–2=98 и уровню значимости α=0,05 из таблицы находим: $t_{0,05;\ 98}$=1,98. Так как $|t| > t_{0.05;98}$, то есть основания отвергнуть нулевую гипотезу. Величины X и $У$ коррелированы.

Уравнение выборочной регрессии $У$ на X: $\overline{y}_x - 5{,}02 = 0{,}8 \cdot \frac{2{,}72}{11{,}97}(x - 36{,}9)$. В итоге $\overline{y}_x = 0{,}18x - 1{,}62$. Линию регрессии строим в системе координат по двум точкам (рис.).

Ожидаемое значение $У$ при $x = 45$ найдем, подставив в полученное уравнение значение $x = 45$: $\overline{y}_x = 0{,}18 \cdot 45 - 1{,}62 = 6{,}48$.

Варианты лабораторной работы №9

Задание. По корреляционной таблице:

1) в прямоугольной системе координат построить эмпирическую линию регрессии $У$ на X;

2) оценить тесноту линейной связи;

3) провести гипотезу о значимости коэффициента корреляции при уровне значимости 0,05;

4) составить линейное уравнение регрессии $У$ на X и построить линию регрессии в той же системе координат, где построена эмпирическая линия регрессии;

5) использую полученное уравнение регрессии, оценить ожидаемое среднее значение признака $У$ при $x = x_0$.

Вариант 1

$У$ \ X	8	9	10	11	12	n_y
10	2	3	5			10
11	2	6	20	7		35
12	1	3	10	9	5	28
13	1	2	5	4	7	19
14			2	3	3	8
n_x	6	14	42	23	15	100

$x_0 = 10,5$

Вариант 2

$У$ \ X	5	10	15	20	25	n_y
2	2	4				6
3		4	4			8
4			7	15	8	30
5			2	10	8	20
6					6	6
n_x	2	8	13	25	22	70

$x_0 = 24$

Вариант 3

$У$ \ X	20	35	50	65	80	n_y
2			1	4	1	6
3			7	7	2	16
4		4	12	2		18
5		8	6			14
6	2	4				6
n_x	2	16	26	13	3	60

$x_0 = 70$

Вариант 4

$У$ \ X	15	16	17	18	19	n_y
2				4	2	6
4			6	4	1	11
6		5	10	3		18
8	1	3	7			11
10	3	10	1			14
n_x	4	18	24	11	3	60

$x_0 = 17,5$

Вариант 5

$У$ \ X	10	15	20	25	30	n_y
4				4	4	8
8			2	6	3	11
12		2	10	3		15
16		6	4			10
20	1	3	5			9
24	4	3				7
n_x						60

$x_0 = 17$

Вариант 6

$У \backslash X$	15	25	35	45	55	n_y
2	8	7	2			17
2,5	2	16	8	6	2	34
3		9	12	12	4	37
3,5			2	4	5	11
4					1	1
n_x	10	32	24	22	12	100

$x_0 = 48$

Вариант 7

$У \backslash X$	4-6	6-8	8-10	10-12	12-14	n_y
5-19			2	3	1	6
19-33		4	5	1		10
33-47		8	5	5		18
47-61	3	8	2			13
61-75	2	1				3
n_x	5	21	14	9	1	50

$x_0 = 11$

Вариант 8

$У \backslash X$	8-16	16-24	24-32	32-40	40-48	n_y
3-7				3	2	5
7-11		1	8	4	3	16
11-15	1	8	16	5	1	31
15-19	2	3	12			17
19-23	6	4	1			11
n_x	9	16	37	12	6	80

$x_0 = 45$

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1 Кацман Ю.Я. Теория вероятностей, математическая статистика и случайные процессы: Томск: Томский политехнический университет, 2013. – 131 с.

2 Климов Г.П. Теория вероятностей и математическая статистика. – М.: Московский государственный университет имени М.В. Ломоносова, 2011. – 368 с.

3 Крамер Г. Математические методы статистики. – Москва-Ижевск: Регулярная и хаотическая динамика, 2003. – 648 с.

4 Прохоров Ю.В. Лекции по теории вероятностей и математической статистике. – М.: Московский государственный университет имени М.В. Ломоносова, 2012. – 254 с

5 Теория вероятностей и математическая статистика. Базовый курс с примерами и задачами. – М.: ФИЗМАТЛИТ, 2002. – 224 с

Приложения

Приложение 1

Значения функции $\phi(x)=\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$

	0	1	2	3	4	5	6	7	8	9
0.0	0.3989	0.3989	0.3989	0.3988	0.3986	0.3984	0.3982	0.3980	0.3977	0.3973
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	0.2420	0.2396	0.2371	0.2347	0.2323	0.2299	0.2275	0.2251	0.2227	0.2203
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	0.0540	0.0529	0.0519	0.0508	0.0498	0.0488	0.0478	0.0468	0.0459	0.0449
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	00.96	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	0.0044	0.0043	0.0042	0.0041	0.0039	0.0038	0.0037	0.0036	0.0035	0.0034
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	0008	0008	0008	0008	0006
3.6	0006	0006	0006	0005	0005	0005	0005	0005	0005	0004
3.7	0004	0004	0004	0004	0004	0004	0003	0003	0003	0003
3.8	0003	0003	0003	0003	0003	0002	0002	0002	0002	0002
3.9	0002	0002	0002	0002	0002	0002	0002	0002	0001	0001
4.0	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
4.1	0.0001338									
4.5	0.000016									

Приложение 2

Значения функции $\Phi(x)=\frac{1}{\sqrt{2\pi}}\int_0^x e^{-\frac{z^2}{2}}dz$.

x	Ф(x)	x	Ф(x)	x	Ф(x)	x	Ф(x)
0,00	0,0000	***0,47***	0.1808	***0,94***	0.3264	***1,41***	0.4207
0,01	0,0010	***0,48***	0.1844	***0,95***	0.3289	***1,42***	0.4222
0,02	0,0080	***0,49***	0.1879	***0,96***	0.3315	***1,43***	0.4236
0,03	0,0120	***0,50***	0.1915	***0,97***	0.3340	***1,44***	0,4251
0,04	0,0160	***0,51***	0.1950	***0,98***	0.3365	***1,45***	0,4265
0,05	0,0199	***0,52***	0.1985	***0,99***	0.3389	***1,46***	0,4279
0,06	0,0230	***0,53***	0.2019	***1,00***	0.3443	***1,47***	0,4292
0,07	0,0279	***0,54***	0.2054	***1,01***	0.3438	***1,48***	0,4306
0,08	0,0319	***0,55***	0.2088	***1,02***	0.3461	***1,49***	0,4319
0,09	0,0359	***0,56***	0.2123	***1,03***	0.3485	***1,50***	0,4332
0,10	0,0398	***0,57***	0.2157	***1,04***	0.3508	***1,51***	0,4345
0,11	0,0438	***0,58***	0.2190	***1,05***	0.3531	***1,52***	0,4357
0,12	0,0478	***0,59***	0.2224	***1,06***	0.3554	***1,53***	0,4370
0,13	0,0517	***0,60***	0.2257	***1,07***	0.3577	***1,54***	0,4382
0,14	0,0557	***0,61***	0.2291	***1,08***	0.3599	***1,55***	0,4394
0,15	0,0596	***0,62***	0.2324	***1,09***	0.3621	***1,56***	0,4406
0,16	0,0636	***0,63***	0.2357	***1,10***	0.3643	***1,57***	0,4418
0,17	0,0675	***0,64***	0.2389	***1,11***	0.3665	***1,58***	0,4429
0,18	0,0714	***0,65***	0.2422	***1,12***	0.3686	***1,59***	0,4441
0,19	0,0753	***0,66***	0.2454	***1,13***	0.370.8	***1,60***	0,4452
0,20	0,0793	***0,67***	0.2486	***1,14***	0.3729	***1,61***	0,4463
0,21	0,0832	***0,68***	0.2517	***1,15***	0.3749	***1,62***	0,4474
0,22	0,0871	***0,69***	0.2549	***1,16***	0.3770	***1,63***	0,4484
0,23	0,0910	***0,70***	0.2580	***1,17***	0.3790	***1,64***	0,4495
0,24	0,0948	***0,71***	0.2611	***1,18***	0.3810	***1,65***	0,4505
0,25	0,0987	***0,72***	0.2612	***1,19***	0.3830	***1,66***	0,4515
0,26	0,1026	***0,73***	0.2673	***1,20***	0.3849	***1,67***	0,4525
0,27	0,1064	***0,74***	0.2703	***1,21***	0.3869	***1,68***	0,4535
0,28	0,1103	***0,75***	0.2734	***1,22***	0.3883	***1,69***	0,4545
0,29	0,1141	***0,76***	0.2764	***1,23***	0.3907	***1,70***	0,4554
0,30	0,1179	***0,77***	0.2794	***1,24***	0.3925	***1,71***	0,4564
0,31	0,1217	***0,78***	0.2823	***1,25***	0.3944	***1,72***	0,4573
0,32	0,1255	***0,79***	0.2852	***1,26***	0.3962	***1,73***	0,4582
0,33	0,1293	***0,80***	0.2881	***1,27***	0.3980	***1,74***	0,4591
0,34	0,1331	***0,81***	0.2910	***1,28***	0.3997	***1,75***	0,4599
0,35	0,1368	***0,82***	0.2939	***1,29***	0.4015	***1,76***	0,4608
0,36	0,1406	***0,83***	0.2967	***1,30***	0.4032	***1,77***	0,4616
0,37	0,1443	***0,84***	0.2995	***1,31***	0.4049	***1,78***	0,4625
0,38	0,1480	***0,85***	0.3023	***1,32***	0.4066	***1,79***	0,4633
0,39	0,1517	***0,86***	0.3051	***1,33***	0.4082	***1,80***	0,4641
0,40	0,1554	***0,87***	0.3078	***1,34***	0.4099	***1,81***	0,4649
0,41	0,1591	***0,88***	0.3106	***1,35***	0.4115	***1,82***	0,4656
0,42	0,1628	***0,89***	0.3133	***1,36***	0.4131	***1,83***	0,4664

x	Φ(x)	x	Φ(x)	x	Φ(x)	x	Φ(x)
0,43	0,1664	***1,94***	0,4738	***2,32***	0,4898	***2,76***	0,4971
0,44	0,1700	***1,95***	0,4744	***2,34***	0,4904	***2,78***	0,4973
0,45	0,1736	***1,96***	0,4750	***2,36***	0,4909	***2,80***	0,4974
0,46	0,1772	***1,97***	0,4756	***2,38***	0,4913	***2,82***	0,4976
0,90	0.3159	***1,98***	0,4761	***2,40***	0,4918	***2,84***	0,4977
0,91	0.3186	***1,99***	0,4767	***2,42***	0,4922	***2,86***	0,4979
0,92	0.3212	***2,00***	0,4772	***2,44***	0,4927	***2,88***	0,4980
0,93	0.3238	***2,02***	0,4783	***2,46***	0,1931	***2,90***	0,4981
1,37	0.4147	***2,04***	0,4793	***2,48***	0,4934	***2,92***	0,4982
1,38	0.4162	***2,06***	0,4303	***2,50***	0,4938	***2,94***	0,4984
1,39	0.4177	***2,08***	0,4812	***2,52***	0,4941	***2,96***	0,4985
1,40	0.4192	***2,10***	0,4821	***2,54***	0,4945	***2,98***	0,4986
1,84	0,4671	***2,12***	0,4830	***2,56***	0,4948	***3,00***	0,49865
1,85	0,4678	***2,14***	0,4838	***2,58***	0,4951	***3,20***	0,49931
1,86	0,4686	***2,16***	0,4846	***2,60***	0,4953	***3,40***	0,49966
1,87	0,4693	***2,18***	0,4854	***2,62***	0,4956	***3,60***	0,499841
1,88	0,4699	***2,20***	0,4861	***2,64***	0,4959	***3,80***	0,499928
1,89	0,4706	***2,22***	0,4868	***2,66***	0,4961	***4,00***	0,499968
1,90	0,4713	***2,24***	0,4875	***2,68***	0,4963	***4,50***	0,499997
1,91	0,4719	***2,26***	0,4881	***2,70***	0,4965	***5,00***	0,499997
1,92	0,4726	***2,28***	0,4887	***2,72***	0,4967		
1,93	0,4732	***2,30***	0,4893	***2,74***	0,4969		

Приложение 3

Критические точки распределения Стьюдента

Число степеней свободы *k*	**Уровень значимости *a(двусторонняя критическая область)***					
	0,10	0,05	0,02	0,01	0,002	0,001
1	6,31	12,7	31,82	63,7	318,3	637,0
2	2,92	4,30	6,97	9,92	22,33	31,6
3	2,35	3,18	4,54	5,84	10,22	12,9
4	2,13	2,78	3,75	4,60	7,17	8,61
5	2,01	2,57	3,37	4,03	5,89	6,86
6	1,94	2,45	3,14	3,71	5,21	5,96
7	1,89	2,36	3,00	3,50	4,79	5,40
8	1,86	2,31	2,90	3,36	4,50	5,04
9	1,83	2,26	2,82	3,25	4,30	4,78
10	1,81	2,23	2,76	3,17	4,14	4,59
11	1,80	2,20	2,72	3,11	4,03	4,44
12	1,78	2,18	2,68	3,05	3,93	4,32
13	1,77	2,16	2,65	3,01	3,85	4,22
14	1,76	2,14	2,62	2,98	3,79	4,14
15	1,75	2,13	2,60	2,95	3,73	4,07
16	1,75	2,12	2,58	2,92	3,69	4,01
17	1,74	2,11	2,57	2,90	3,65	3,96
18	1,73	2,10	2,55	2,88	3,61	3,92
19	1,73	2,09	2,54	2,86	3,58	3,88
20	1,73	2,09	2,53	2,85	3,55	3,85
21	1,72	2,08	2,52	2,83	3,53	3,82
22	1,72	2,07	2,51	2,82	3,51	3,79
23	1,71	2,07	2,50	2,81	3,49	3,77
24	1,71	2,06	2,49	2,80	3,47	3,74
25	1,71	2,06	2,49	2,79	3,45	3,72
26	1,71	2,06	2,48	2,78	3,44	3,71
27	1,71	2,05	2,47	2,77	3,42	3,69
28	1,70	2,05	2,46	2,76	3,42	3,66
29	1,70	2,05	2,46	2,76	3,40	3,66
30	1,70	2,04	2,46	2,75	3,39	3,65
40	1,68	2,02	2,42	2,70	3,31	3,55
60	1,67	2,00	2,39	2,66	3,23	3,46
120	1,66	1,98	2,36	2,62	3,17	3,37
∞	1,64	1,96	2,33	2,58	3,09	3,29
	0,05	0,025	0,01	0,005	0,001	0,0005
	Уровень значимости α (односторонняя критическая область)					

Приложение 4

Таблица значений t_γ

n	γ			*n*	γ		
	0,95	0,99	0,999		0,95	0,99	0,999
5	2,78	4,60	8,61	20	2,093	2,861	3,883
6	2,57	4,03	6,86	25	2,064	2,797	3,745
7	2,45	3,71	5,96	30	2,045	2,756	3,659
8	2,36	3,50	5,40	35	2,032	2,720	3,600
9	2,31	3,36	5,04	40	2,023	2,708	3,558
10	2,26	3,25	4,78	45	2,016	2,692	3,527
11	2,23	3,17	4,59	50	2,009	2,679	3,502
12	2,20	3,11	4,44	60	2,001	2,662	3,464
13	2,18	3,05	4,32	70	1,996	2,649	3,439
14	2,16	3,01	4,22	80	1,991	2,640	3,418
15	2,14	2,98	4,14	90	1,987	2,633	3,403
16	2,13	2,95	4,07	100	1,984	2,627	3,392
17	2,12	2,92	4,01	120	1,980	2,617	3,374
18	2,11	2,90	3,96	∞	1,96	2,576	3,291
19	2,10	2,88	3,92				

Приложение 5

Таблица значений q_γ

n	γ			*n*	γ		
	0,95	0,99	0,999		0,95	0,99	0,999
5	1,37	2,67	5,64	20	0,37	0,58	0,88
6	1,09	2,01	3,88	25	0,32	0,49	0,73
7	0,92	1,62	2,98	30	0,28	0,43	0,63
8	0,80	1.38	2,42	35	0,26	0,38	0,56
9	0,71	1,20	2,06	40	0,24	0,35	0,50
10	0,65	1,08	1,80	45	0,22	0,32	0,46
11	0,59	0,98	1,60	50	0,21	0,30	0,43
12	0,55	0,90	1,45	60	0,188	0,269	0,38
13	0,52	0,83	1,33	70	0,174	0,245	0,34
14	0,48	0,78	1,23	80	0,161	0,226	0,31
15	0,46	0,73	1,15	90	0,151	0,211	0,29
16	0,44	0,70	1,07	100	0,143	0,198	0,27
17	0,42	0,66	1,01	150	0,115	0,160	0,221
18	0,40	0,63	0,96	200	0,099	0,136	0,185
19	0,39	0,60	0,92	250	0,089	0,120	0,162

Критические точки распределения χ^2.

Число степеней свободы ***k***	**Уровень значимости** ***a***					
	0,01	0,025	0,05	0,95	0,975	0,99
1	6,6	5,0	3,8	1,0039	0,00098	0,00016
2	9,2	7,4	6,0	0,103	0,051	0,020
3	11,3	9,4	7,8	0,352	0,216	0,115
4	13,3	11,1	9,5	0,711	0,484	0,297
5	15,1	12,8	11,1	1,15	0,831	0,554
6	18,8	14,4	12,6	1,64	1,24	0,872
7	18,5	16,0	14,1	2,17	1,69	1,24
8	20,1	17,5	15,5	2,73	2,18	1,65
9	21,7	19,0	16,9	3,33	2,70	2,09
10	23,2	20,5	18,3	3,94	3,25	2,56
11	24,7	21,9	19,7	4,57	3,82	3,05
12	26,2	23,3	21,0	5,23	4,40	3,57
13	27,7	24,7	22,4	5,89	5,01	4,11
14	29,1	26,1	23,7	6,57	5,63	4,66
15	30,6	27,5	25,0	7,26	6,26	5,23
16	32,0	28,8	26,3	7,96	6,91	5,81
17	33,4	30,2	27,6	8,67	7,56	6,41
18	34,8	31,5	28,9	9,39	8,23	7,01
19	36,2	32,9	30,1	10,1	8,91	7,63
20	37,6	34,2	31,4	10,9	9,59	8,26
21	38,9	35,5	32,7	11,6	10,3	8,90
22	40,3	36,8	33,9	12,3	11,0	9,54
23	41,6	38,1	35,2	13,1	11,7	10,2
24	43,0	39,4	36,4	13,8	12,4	10,9
25	44,3	40,6	37,7	14,6	13,1	11,5
26	45,6	41,9	38,9	15,4	13,8	12,2
27	47,0	43,2	40,1	16,2	14,6	12,9
28	48,3	44,5	41,3	16,9	15,3	13,6
29	49,6	45,7	42,6	17,7	16,0	14,3
30	50,9	47,0	43,8	18,5	16,8	15,0
35	57,3	53,2	49,0	22,5	20,5	17,2
40	63,7	59,3	55,8	26,5	24,4	22,7
45	79,0	65,4	61,6	30,6	28,4	26,0
50	76,1	71,4	67,5	34,8	32,3	29,7
60	88,3	83,3	79,0	43,2	40,5	37,5

Приложение 8

Критические точки распределения Фишера-Снедекора

α=0,01												
	k_1											
k_2	1	2	3	4	5	6	7	8	9	10	11	12
1	4052	4999	5403	5625	5764	5889	5928	5981	6022	6056	6082	6106
2	98,49	99,01	99,17	99,25	99,33	99,30	99,34	99,36	99,36	99,40	99,41	99,42
3	34,12	38,81	29,46	28,71	28,24	27,91	27,67	27,49	27,34	27,23	27,13	27,05
4	21,20	18,00	16,69	15,98	15,52	15,21	14,96	14,80	14,66	14,54	14,45	14,37
5	16,26	13,27	12,06	11,39	10,97	10,67	10,45	10,27	10,15	10,05	9,96	9,89
6	13,74	10,92	9,78	9,15	8,75	8,47	8,26	8,10	7,98	7,87	7,79	7,72
7	12,25	9,55	8,45	7,85	7,46	7,19	7,00	6,84	6,71	6,62	6,54	6,47
8	11,26	8,65	7,59	7,01	6,63	6,37	6,19	6,03	5,91	5,82	5,74	5,67
9	10,56	8,02	6,99	6,42	6,06	5,80	5,62	5,47	5,35	5,26	5,18	5,11
10	10,04	7,56	6,55	5,99	5,64	5,39	5,21	5,06	4,95	4,85	4,78	4,71
11	9,86	7,20	6,22	5,67	5,32	5,07	4,88	4,72	4,63	4,54	4,46	4,40
12	9,33	6,93	5,95	5,41	5,06	4,82	4,65	4,50	4,39	4,30	4,22	4,16
13	9,07	6,70	5,74	5,20	4,86	4,62	4,44	4,30	4,19	4,10	4,02	3,96
14	8,86	6,51	5,56	5,03	4,69	4,46	4,28	4,14	4,03	3,94	3,86	3,80
15	8,68	6,36	5,42	4,89	4,56	4,32	4,14	4,00	3,89	3,80	3,73	3,67
16	8,53	6,23	5,29	4,77	4,44	4,20	4,03	3,89	3,78	3,69	3,61	3,55
17	8,40	6,11	5,18	4,67	4,44	4,10	3,93	3,79	3,68	3,59	3,52	3,45
α=0,05												
	k_1											
k_2	1	2	3	4	5	6	7	8	9	10	11	12
1	161	200	216	225	230	234	237	239	241	242	243	244
2	18,51	19,00	19,16	19,25	19,30	19,33	19,36	19,37	19,38	19,39	19,40	19,41
3	10,13	9,55	9,28	9,12	9,01	8,94	8,88	8,84	8,81	8,78	8,76	8,74
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96	5,93	5,91
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,788	4,74	4,70	4,68
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06	4,03	4,00
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,63	3,60	3,57
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,34	3,31	3,28
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,13	3,10	3,07
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,97	2,94	2,91
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,86	2,82	2,79
12	4,75	3,88	3,49	3,26	3,11	3,00	2,92	2,85	2,80	2,76	2,72	2,69
13	4,67	3,80	3,41	3,18	3,02	2,92	2,84	2,77	2,72	2,67	2,63	2,60
14	4,60	3,74	3,34	3,11	2,96	2,85	2,77	2,70	2,65	2,60	2,56	2,53
15	4,54	3,68	3,29	3,06	2,90	2,79	2,70	2,64	2,59	2,55	2,51	2,48
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49	2,45	2,42
17	4,45	3,59	3,20	2,96	2,81	2,70	2,62	2,55	2,50	2,45	2,41	2,38

СОДЕРЖАНИЕ

Введение	3
Тема 1. Выборочный метод	4
Основные понятия	4
Пример выполнения лабораторной работы № 1	9
Пример выполнения лабораторной работы № 2	12
Варианты лабораторных работ № 1 и № 2	15
Тема 2. Интервальные оценки параметров распределения	17
Основные понятия	17
Пример выполнения лабораторной работы № 3	21
Пример выполнения лабораторной работы № 4	23
Варианты лабораторных работ № 3 и № 4	25
Тема 3. Проверка статистических гипотез	29
Основные понятия	29
Пример выполнения лабораторной работы № 5	35
Пример выполнения лабораторной работы № 6	37
Варианты лабораторных работ № 5 и № 6	39
Тема 4. Метод наименьших квадратов	43
Основные понятия	43
Пример выполнения лабораторной работы № 7	45
Варианты лабораторной работы № 7	47
Тема 5. Элементы теории корреляции	49
Основные понятия	49
Пример выполнения лабораторной работы № 8	53
Варианты лабораторной работы № 8	55
Пример выполнения лабораторной работы № 9	55
Варианты лабораторной работы № 9	58
Библиографический список	61
Приложения	62

Анна Павловна Филимонова,
канд. физ.-мат. наук, доц. кафедры общей математики и информатики АмГУ;

Татьяна Александровна Юрьева,
канд. пед. наук, доц. кафедры общей математики и информатики АмГУ;

Математическая статистика. *Практикум.*

Заказ 744.