А.М. Сазонова

ТЕОРИЯ ВЕРОЯТНОСТИ И МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

Модуль 3
АНАЛИЗ ВАРИАЦИОННЫХ РЯДОВ

Могилев 2010

МИНИСТЕРСТВО ОБРАЗОВАНИЯ РЕСПУБЛИКИ БЕЛАРУСЬ

Учреждение образования
«МОГИЛЕВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
им. А.А. КУЛЕШОВА»

А.М. Сазонова

ТЕОРИЯ ВЕРОЯТНОСТИ И МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

Модуль 3. АНАЛИЗ ВАРИАЦИОННЫХ РЯДОВ

МЕТОДИЧЕСКИЕ РЕКОМЕНДАЦИИ

Могилев 2010

УДК 519.21(075)
ББК 22.3
С12

Печатается по решению редакционно-издательского совета УО «МГУ им. А.А. Кулешова»

Рецензент
кандидат физико-математических наук
доцент МГУ им. А.А. Кулешова
В.Н. Борбат

Сазонова, А.М.

С12 Теория вероятности и математическая статистика. Модуль 3: Анализ вариационных рядов: метод. рекомендации / А.М. Сазонова. – Могилев: УО «МГУ им. А.А. Кулешова», 2010. – 72 с.: ил.

Методические рекомендации отражают практику преподавания предмета в Могилевском государственном университете им. А.А. Кулешова

Предназначаются для студентов вузов, обучающихся по экономическим специальностям, по социологии, преподавателей и слушателей курсов повышения квалификации, применяющих в своей практике вероятностные и статистические методы.

УДК 519.21(075)
ББК 22.3

Часть 2. МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

Модуль 3: Анализ вариационных рядов

Математическая статистика – это наука, которая, основываясь на методах теории вероятностей, изучает (систематизирует, обрабатывает) и использует статистические данные для получения научных и практических выводов.

3.1. Генеральная совокупность и выборка

Для получения статистических данных необходимо провести обследование соответствующих объектов. Если их количество велико, то приходится для обследования отбирать только часть, т.е. проводить выборочное обследование. Обследование объектов всей совокупности иногда на практике не имеет смысла, т.к. в результате обследования они разрушаются. Иногда реально существующую совокупность объектов для обследования можно мысленно дополнить любым количеством таких же однородных объектов, чтобы наблюдаемые значения случайной величины можно было бы мысленно продолжать в неизменных условиях как угодно долго (например, совокупность звонков, поступивших в справочное бюро за неделю отпускного периода можно дополнить гипотетической совокупностью в следующих неделях этого периода). Неизменность условий означает неизменность только тех всех условий, которые можно проконтролировать при проведении наблюдений. Прочие неконтролируемые условия изменяются, что приводит к случайности результатов наблюдений.

Определение. Совокупность всех мыслимо возможных качественно однородных объектов данного вида, над которыми проводятся наблюдения с целью получения конкретных значений определенной случайной величины, или совокупность результатов всех мыслимых наблюдений, проводимых в неизменных условиях над одной из случайных величин, связанных с данным видом объектов, называется ***генеральной совокупностью.*** Число объектов (наблюдений) N генеральной совокупности называют ***объемом генеральной совокупности.***

Заметим, что ***генеральная совокупность*** *объектов данного вида и* ***соответствующая совокупность значений случайной величины*** $X(\omega)$ ***не***

различаются, хотя понятие генеральной совокупности шире понятия случайной величины, т.к. любое значение случайной величины может быть результатом нескольких наблюдений. При этом не следует смешивать понятие генеральной совокупности с реально существующими совокупностями (например, поступившая на склад из цеха продукция является реально существующей совокупностью, которую нельзя назвать генеральной, т.к. выпуск этой продукции можно мысленно продолжить сколь угодно долго).

Генеральная совокупность бывает:

- *конечная и реально существующая, например, такую совокупность образуют жители г. Могилева в фиксированный момент времени;*
- *бесконечная и реально существующая, например, такую совокупность образует множество действительных чисел из интервала* (0; 1);
- *воображаемая (гипотетическая) конечная или бесконечная. Например, совокупность всех мысленно возможных выпущенных, выпускаемых теперь и в будущем на одном оборудовании (в одинаковых условиях) изделий образует бесконечную генеральную совокупность.*

Определение. Часть отобранных объектов из генеральной совокупности или результаты наблюдений над ограниченным числом объектов из этой совокупности называется ***выборочной совокупностью (статистическими данными)*** или ***выборкой***. Число объектов (наблюдений) n выборочной совокупности называют ***объемом выборки.*** Полагают, что N значительно больше n, т.е. $N>>n$.

Различают конкретную и случайную выборки. *Конкретная выборка* $x_1, x_2, ..., x_n$ – это конечная последовательность чисел – реализации случайной величины $X(\omega)$. *Случайная выборка* – это весь мыслимый набор конкретных выборок, или, более строго, – это последовательность $X_1, X_2, ..., X_n$ независимых одинаково распределенных случайных величин, распределение каждой из которых совпадает с распределением генеральной случайной величины $X(\omega)$.

При изучении случайной величины X рассматриваются функции, которые характеризуют эту случайную величину. Такие функции от конкретной выборки называют ***статистиками (выборочными статистиками). В теоретических исследованиях статистику рассматривают как функцию случайной выборки, чтобы статистика стала случайной величиной и ее распределение позволяло бы сделать вывод о распределении самой исследуемой случайной величины.***

Суть выборочного метода в математической статистике и состоит в том, что по выборке судят о свойствах генеральной совокупности в целом. Для этого выборочная совокупность должна быть ***репрезентативной (представительной)***, что обеспечивается объемом

выборки и случайностью отбора из генеральной совокупности ее элементов, каждый из которых имеет одинаковую вероятность попадания в выборку.

Различают пять основных типов выборок:

1. *Собственно случайная*: *повторная* (после выбора элементы возвращаются обратно), *бесповторная* (выбранные элементы не возвращаются).

И для выборки с возвратом, и для выборки без возврата вероятность того, что объект попадет в выборку, не изменяется при переходе от одного испытания к другому, т.е. с вероятностной точки зрения условия испытаний не изменяются. Однако если в выборке с возвратом испытания независимы, то в выборке без возврата испытания зависимы (например, для урновой схемы условная вероятность не совпадает с безусловной). Условие независимости является одним из основных используемых в теоремах теории вероятности, поэтому в дальнейшем будем предполагать, что имеет место случайная выборка с возвратом, и при этом иметь в виду, что выражение «случайная выборка с возвратом» равносильно выражению «испытания независимы и проведены в одинаковых условиях».

2. *Типическая*, когда генеральная совокупность предварительно разбивается на группы типических элементов, и выборка осуществляется из каждой: *равномерные выборки* (при равенстве объемов групп выбирают одинаковое количество элементов из каждой); *пропорциональные* (численность выборки формируется пропорционально численностям или средним квадратическим отклонениям групп генеральной совокупности); Комбинированные (численность выборок пропорциональна и средним квадратическим отклонениям, и численностям групп генеральной совокупности).

3. *Механическая*, когда отбор элементов осуществляется через определенный интервал.

4. *Серийная*, когда отбор производится не по одному элементу, а сериями для проведения сплошного обследования.

5. *Комбинированная*, когда используются различные комбинации вышеуказанных методов.

После получения выборочной совокупности объектов все эти объекты обследуют по отношению к определенной случайной величине (или случайному событию), и в результате этого получают наблюдаемые данные, которые обрабатываются.

3.2. Вариационный ряд

Пусть случайная величина X описывает некоторый признак генеральной совокупности. Из генеральной совокупности осуществлена выборка $\{x_1, x_2, x_3, ..., x_n\}$ объема n. Элементы этой выборки представляют собой значения случайной величины X. Эти значения упорядочивают по возрастанию, что называется ***ранжированием*** выборки. Различные значения x_i называют ***вариантами.*** Число m_i повторов в выборке варианты x_i называют ***частотой*** этого ***варианта***. ***Частостью, относительной частотой*** или ***долей*** варианта называют число $\hat{p}_i = w_i = m_i / n$.

Частоты и частости называются ***весами***.

Зафиксируем некоторое число x. Количество m_x вариант, значения которых меньше x, называют ***накопленной частотой***:

$$m_x = \sum_{x_i < x} m_i$$

Накопленной частостью называют отношение накопленной частоты к объему выборки:

$$w_x = \frac{m_x}{n} = \frac{1}{n}\sum_{x_i < x} m_i$$

Заметим, что предел частости при неограниченном увеличении объема выборки является статистической вероятностью. Естественно считать частость w_i выборочным аналогом (вычисленной по выборочным данным) вероятности p_i появления значения x_i случайной величины X.

Вариационным (***статистическим***) ***рядом*** называют таблицу, состоящую из ранжированных значений вариант и соответствующих им весов. Вариационные ряды бывают дискретными и интервальными.

Дискретным называют вариационный ряд, представляющий выборку значений дискретной случайной величины.

x_i	-2	0	1	3
m_i	8	10	7	5
m_i / n	8/30	10/30	7/30	5/30

Интервальным (непрерывным) называют ряд, представляющий выборку значений непрерывной случайной величины.

Нецелесообразно также построение дискретного ряда для дискретной случайной величины, число возможных значений которой велико. В подобных случаях строят интервальный (вариационный) ряд распределения.

Для построения интервального вариационного ряда множество значений вариант разбивают на полуинтервалы $[a_i, a_{i+1})$, т.е. производят группировку. Количество интервалов k рекомендуется вычислять по формуле Стерджесса:

$$k \approx 1+3.322 \lg n = 1+1.4 \ln n$$

Длина интервала равна $h = \frac{R}{k}$, где число $R = x_{max} - x_{min}$ называют ***размахом варьирования.*** Подсчитывают число значений m_i (частоты), попавших в полуинтервал $[a_i, a_{i+1})$, $i = 1, 2, \ldots, k$. Контроль: $\sum_i m_i = n$.

Если окажется, что h дробное число, то за длину частичного интервала берут либо ближайшее целое число, либо ближайшую простую дробь. За начало первого интервала рекомендуется брать величину $a_1 = x_{min} - 0{,}5h$. *Конец последнего интервала* a_{k+1} *должен удовлетворять условию* $a_{k+1} - h \le x_{max} < a_{k+1}$

Составляют интервальный вариационный ряд:

$[a_i, a_{i+1})$	$[a_1, a_2)$	$[a_2, a_3)$	...	$[a_k, a_{k+1})$
m_i	m_1	m_2	...	m_k

или

$[a_i, a_{i+1})$	$[a_1, a_2)$	$[a_2, a_3)$	...	$[a_k, a_{k+1})$
m_i / n	m_1 / n	m_2 / n	...	m_k / n

Контроль: $\sum_{i=1}^{k} \frac{m_i}{n} = 1$.

Иногда интервальный вариационный ряд для простоты исследований условно заменяют дискретным (и, наоборот, при большом числе наблюдений). В этом случае серединное значение i-го интервала принимают за вариант x_i *с соответствующей частотой* m_i.

3.3. Выборочные аналоги интегральной и дифференциальной функций распределения. Полигон, гистограмма

Из теории вероятностей известно, что закон распределения (или просто распределение) случайной величины X можно задать:

– для случайной дискретной величины – рядом распределения или функцией распределения

$$F(x) = P(X<x);$$

– для случайной непрерывной величины – интегральной функцией $F(x) = P\ (X < x)$ или дифференциальной функцией распределения $p(x) = F'(x)$.

Будем говорить, что если СВ X распределена по некоторому закону $F(x)$ – теоретическая функция распределения , то это и генеральная совокупность распределена по закону $F(x)$. Аналогом (оценкой функции распределения) в математической статистике является ***выборочная (или эмпирическая) функция распределения*** (или функция распределения выборки): $\hat{F}_n(x) = \frac{m_x}{n}$ – относительная частота события $\{X < x\}$.

Выборочная функция распределения при больших n близка к теоретической. Свойства теоретической и эмпирической функций распределения одинаковы. **Сформулируйте их самостоятельно.**

График функции распределения выборки, построенной по дискретному вариационному ряду, также имеет ступенчатый вид, а аналитически может быть задана системой:

$$\hat{F}_n(x) = \begin{cases} 0, & \text{если} \quad x \le x_1, \\ \sum_{i=1}^{i-1} w_x - \text{накопленная частость}, & \text{если} \quad x_{i-1} < x \le x_i, \\ 1, & \text{если} \quad x > x_v, \ \text{где} \quad x_v - \text{наибольшее значение варианта } x. \end{cases}$$

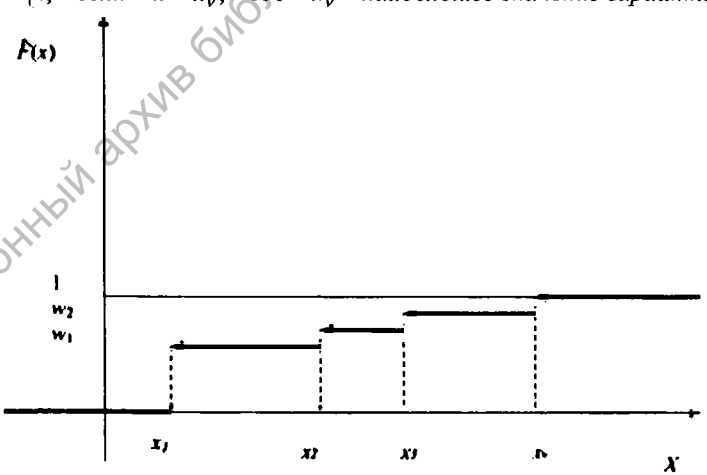
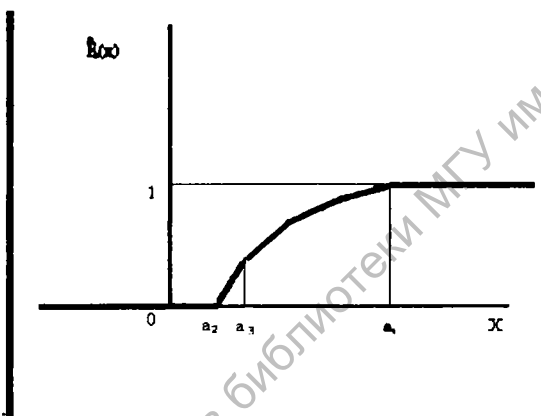


График эмпирической функции распределения (функции распределения выборки), построенной по ***интервальному*** вариационному ряду, представляет собой ломаную, соединяющую точки накопленных частостей, соответствующих правым концам интервалов, растянув при этом крайние интервалы до бесконечности

$$\hat{F}_n(x) = \begin{cases} 0, & \text{если } x \leq a_1, \\ \sum_{i=1}^{i-1} W_x \quad \text{накопленная частость}, & \text{если } a_{i-1} < x \leq a_i,\ i = 1, 2, 3 \ldots k. \\ 1, & \text{если } x > a_{k+1}. \end{cases}$$



Этот график выборочной функции распределения $\hat{F}_*(x)$ дает представление о графике теоретической функции распределения $F(x)$ случайной величины X.

В теории вероятности дифференциальная функция $p(x) = F'(x)$, тогда в приближении $p(x) \approx \frac{F(x+\Delta x) - F(x)}{\Delta x}$. Поэтому естественно выборочным аналогом функции $p(x)$ (оценкой функции плотности) считать функцию $\hat{p}_n(x) = \frac{\hat{F}_*(x+\Delta x) - \hat{F}_*(x)}{\Delta x}$ – плотность частости на промежутке $[x, x+\Delta x)$, т.к. $\hat{F}(x+\Delta x) - \hat{F}(x)$ – частость попадания вариант в промежуток $[x, x+\Delta x)$.

Аналитически выборочную функцию плотности задают соотношением:

$$\dot{p}_n(x)=\begin{cases}0, & \text{если} \quad x \le a_1,\\ m_i/(nh), & \text{если} \quad a_i < x \le a_{i+1}, i=1,2,3,...,k,\\ 0, & \text{если} \quad x > a_{k+1}, \quad \text{где} \quad a_{k+1} - \text{конец последнего } k - \text{го полуинтервала.}\end{cases}$$

Площадь области под графиком этой функции равна единице.

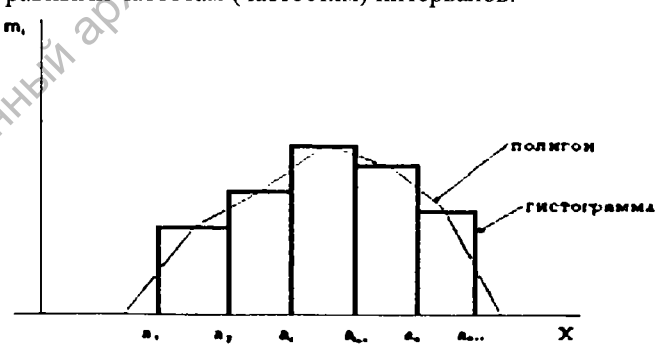
Ломаная, соединяющая точки (x_i, m_{xi}), где m_x – накопленные частоты, а x_i – **значения вариант** для дискретного ряда, или **середины интервалов** для интервального вариационного ряда, называется ***кумулятой (кумулятивной кривой)***.

Вариационные ряды графически изображают и с помощью ***полигона*** или ***гистограммы***.

Полигон частот (частостей) представляет собой ломаную, соединяющую точки плоскости с координатами (x_i, m_i) (или (x_i, m_i/n)) для дискретного статистического ряда, а для интервального вариационного ряда $x_i = (a_{i+1} + a_i)/2$ – середина полуинтервала $[a_i, a_{i+1})$.

Полигон частостей, построенный по дискретному вариационному ряду дискретной случайной величины, называют ***многоугольником распределения частостей*** *– это выборочный аналог многоугольника распределения вероятностей. Полигон для интервального вариационного ряда дает первоначальное представление о дифференциальной функции распределения.*

Гистограмма частот (частостей) изображает только интервальный статистический ряд, имеет вид ступенчатой фигуры из прямоугольников с основаниями, равными длине интервалов h, и высотами, равными частотам (частостям) интервалов.



Контрольные вопросы:

1. Что является аналогом случайной величины в математической статистике?

2. Является ли совокупность всех значений случайной величины генеральной совокупностью?

3. Каким требованиям должна удовлетворять выборка из генеральной совокупности?

4. Каково различие между генеральной и выборочной совокупностями?

5. В чем состоит суть выборочного метода?

6. Какие виды выборочных совокупностей Вы знаете?

7. Что является аналогом варианта в теории вероятностей?

8. Как составляются вариативные ряды?

9. Как можно из частоты получить частость (относительную частоту)?

10. Как определяются значения эмпирической функции распределения?

11. Какими свойствами обладает выборочный аналог функции распределения случайной величины?

12. Какими свойствами обладает выборочный аналог дифференциальной функции распределения?

13. В чем отличие графиков эмпирической функции и кумулянты?

14. Какие вариационные ряды изображаются полигоном, гистограммой?

Пример 1. *Отделом технического контроля завода было проверено 10 партий одинакового числа изделий в каждой партии. Число обнаруженных бракованных изделий в партиях приведено в таблице:*

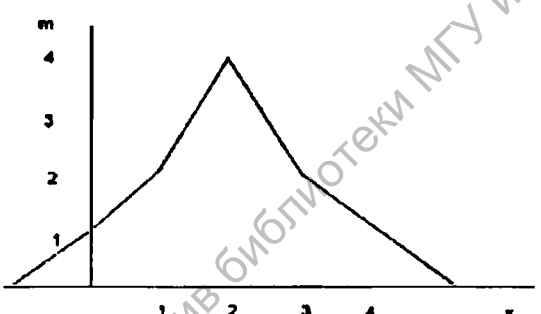
Номер партии	1	2	3	4	5	6	7	8	9	10
Количество бракованных изделий (x_i)	3	2	1	3	0	2	2	1	4	2

Составьте статистический ряд распределения частостей наблюдаемых значений дискретной случайной величины X – числа бракованных изделий в партии. Постройте полигон, кумулянту, график эмпирической функции распределения.

Решение. *Проранжируем исходный ряд, подсчитаем частоту и частость вариант:*
0, 1, 1, 2, 2, 2, 2, 3, 3, 4. *В результате получим дискретный вариационный ряд:*

Количество бракованных изделий (x_i)	Количества партий, m_i	Относительная частота, частость, m_i/n
0	1	0,1
1	2	0,2
2	4	0,4
3	2	0,2
4	1	0,1
$\sum$	10	1

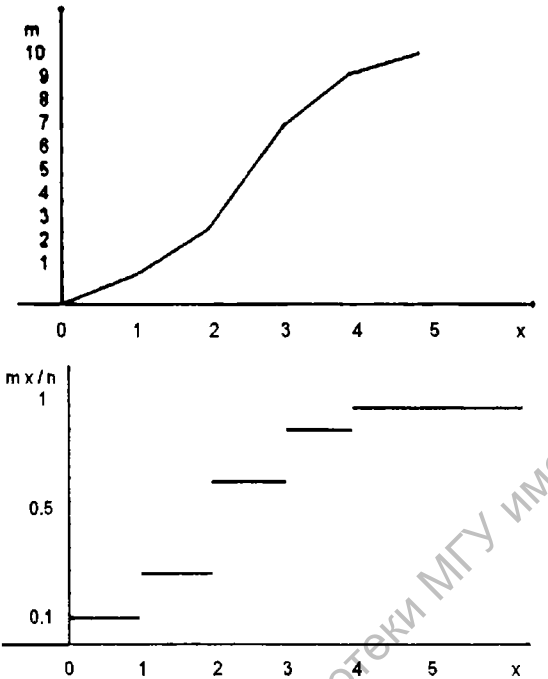
График полигона имеет следующий вид:



По данным дискретного вариационного ряда находим накопленные частоты и частости:

x_i	0	1	2	3	4	5
m_{xi}	0	1	3	7	9	10
m_{xi}/n	0	0.1	0.3	0.7	0.9	1

На следующих рисунках изображены соответственно кумулянта и график эмпирической функции:



Пример 2. *Для определения средней суммы вкладов в банке бесповторной выборкой произведено обследование 60 вкладов (в условных единицах):*

6, 15, 11, 12, 9, 9, 6, 10, 8, 8, 11, 7, 6, 9, 4, 10, 10, 7, 11, 7, 5, 7, 8, 5, 12, 8, 7, 10, 8, 10, 8, 11, 7, 7,9 ,5, 6, 7, 10, 7, 8, 8, 7, 5, 10, 8, 9, 15, 6, 7, 10, 11, 7, 10, 9, 14, 13, 11, 12, 11.

Постройте интервальный вариационный ряд, гистограмму, полигон частот, графики эмпирической функции распределения и эмпирической плотности распределения.

Решение. *По статистическим данным определяем* x_{max}=15; x_{min}=4. *Разобьем множество значений выборки на интервалы. Число интервалов по формуле Стерджесса равно*

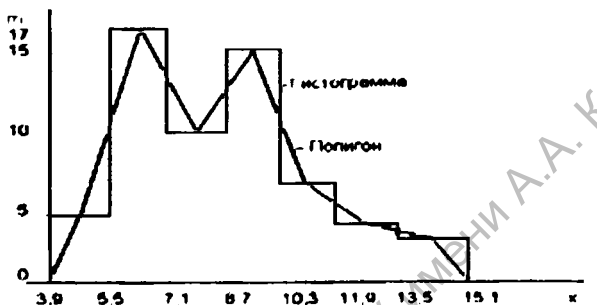
$$k \approx 1 + 1.4 \ln 60 = 6.907;\ \textit{Примем}\ k = 7.$$

Длина частичного интервала $h \approx \frac{x_{max} - x_{min}}{k} = \frac{15-4}{7} \approx 1{,}6$. *Выберем первый интервал так, чтобы в нем содержался вариант* x_{min}, *а последний седьмой интервал содержал* x_{max}. *Например,* a_1=3,9, a_7=13,5.

Подсчитывая число вариант, попадающих в каждый интервал, получим вариационный ряд частот:

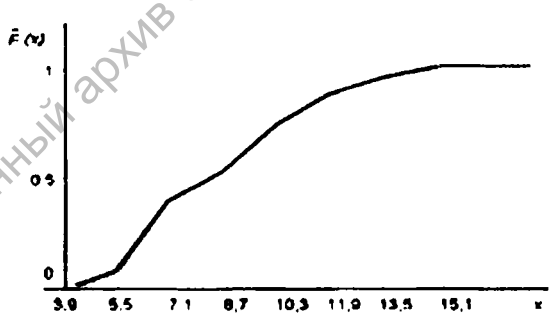
$[a_i, a_{i+1})$	[3,9; 5,5)	[5,5; 7,1)	[7.1. 8,7)	[8,7; 10,3)	[10,3; 11,9)	[11,9; 13,5)	[13,5;15,1)
m_i	5	17	9	15	7	4	3

Контроль: 5+17+9+15+7+4+3=60.



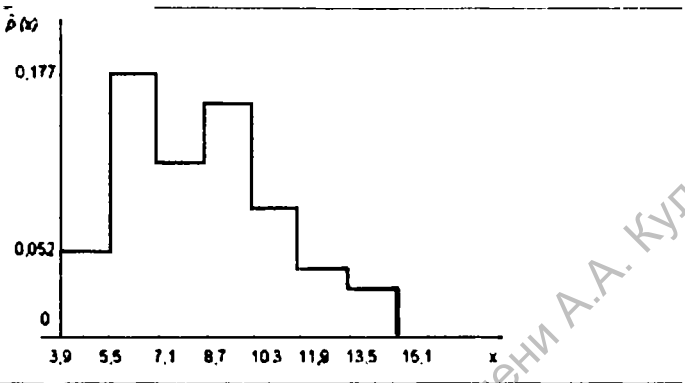
Для построения эмпирической функции распределения вычислим накопленные относительные частоты:

a_i	3.9	5.5	7.1	8.7	10.3	11.9	13.5	15.1
m_{xi}/n	0	5/60	22/60	31/60	46/60	53/60	57/60	1



Для построения графика функции эмпирической плотности вероятности для каждого интервала определим значения: $\frac{W_i}{h} = \frac{m_i}{nh}$

$[a_i, a_{i+1})$	[3,9; 5,5)	[5,5; 7,1)	[7,1; 8,7)	[8,7; 10,3)	[10,3; 11,9)	[11,9; 13,5)	[13,5;15,1)
W_i/h	0,052	0,177	0,094	0,157	0,073	0,042	0,032



3.4. Точечная оценка

Числовые характеристики (*статистические характеристики или оценки*) вариационных рядов являются аналогами числовых характеристик распределения теории вероятностей:

- характеристики положения середины – средняя арифметическая выборки (выборочная средняя), мода, медиана;
- характеристики разброса признака вокруг середины – выборочная дисперсия, выборочное среднее квадратическое отклонение, коэффициент вариации, размах вариации;
- характеристика меры скошенности – выборочный коэффициент асимметрии;
- характеристика островершинности – выборочный эксцесс (или коэффициент крутости);
- характеристики моментов – выборочные моменты и др.

Предполагаемый по изображению вариационного ряда закон распределения генеральной совокупности (СВ X) $F(x,\Theta)$ становится определенным, если известен параметр Θ этого распределения. По имеющейся выборке можно лишь дать оценку $\hat{\Theta}$, приблизительное значение этого параметра, как функцию вариант, т.е. $\hat{\Theta} = \hat{\Theta}_n = u(x_1, x_2, ..., x_n)$. Поскольку значение функции изображается точкой на числовой прямой, то эту оценку называют ***точечной оценкой***.

Точечная оценка является случайной величиной, функцией вариант: $\Theta \approx \Theta_n = u(x_1, x_2, ..., x_n)$, и должна быть «близкой» к истинному значению Θ, т.е. должна быть качественной. Причем, качество оценки определяется не по конкретной выборке, а по всему мысленному набору конкретных выборок, т.е. случайной выборке. Для этого оценка должна удовлетворять требованиям **состоятельности, несмещенности, эффективности, достаточности.**

Оценка Θ называется ***состоятельной*** оценкой генеральной характеристики Θ, если для любого $\varepsilon > 0$ выполняется равенство $\lim_{n \to \infty} P\left(|\Theta_n - \Theta| < \varepsilon\right) = 1$.

Несостоятельные оценки не используются.

Оценка $\dot{\Theta}$ называется ***несмещенной*** оценкой генеральной характеристики Θ, если для любого фиксированного числа n наблюдений выполняется равенство $M(\Theta_n) = \Theta$.

Это свойство желательно, но не обязательно. Если полученная оценка оказалась смещенной, то ее можно поправить, чтобы она стала несмещенной. Оценка может быть смещенной, но асимптотически несмещенной.

Несмещенная оценка $\dot{\Theta}$ называется ***несмещенной эффективной*** оценкой генеральной характеристики Θ, если среди любых других подобных оценок гой же характеристики она имеет наименьшую дисперсию: $D(\Theta_n) \to \min$, т.е. она самая точная.

Оценка называется **достаточной**, если она учитывает все сведения выборки относительно оцениваемого параметра генеральной совокупности (СВ X).

Так **частость** $w = \frac{m}{n}$ является состоятельной, несмещенной и эффективной **оценкой вероятности** p события.

Пусть генеральная совокупность объема N содержит M элементов, обладающих некоторым характеристическим признаком. Для **генеральной доли** этого признака $p = \frac{M}{N}$ несмещенной, состоятельной **оценкой** будет **выборочная доля** $w = \frac{m}{n}$, где m – число элементов выборки объема n, обладающих этим признаком.

Примером оценок наиболее употребительных числовых характеристик являются выборочные моменты как оценки соответствую-

щих теоретических моментов (***метод моментов***, *предложенный английским статистиком Карлом Пирсоном*):

$$\hat{\nu}_i = \overline{x^i} = \frac{1}{n}\sum_{k=1}^{n} x_k^i,$$

$$\hat{\mu}_i = \overline{(x-\bar{x})^i} = \frac{1}{n}\sum_{k=1}^{n}(x_k-\bar{x})^i$$

Выборочный центральный момент *i*-ого порядка

Замечание. Метод моментов отличается простотой, однако оценки, найденные этим методом, как правило, являются смещенными и малоэффективными.

К оценкам генерального среднего для нормально распределенной случайной величины относятся:

• **среднее арифметическое выборки $\bar{x}$ есть оценка математического ожидания $MX=a$ (или генеральной средней $\bar{X}$)**

Эта оценка является состоятельной, несмещенной и эффективной точечной оценкой:

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

или для сгруппированных вариант $\hat{MX} = \bar{x} = \frac{1}{n}\sum_{i=1}^{k} x_i m_i$.

Замечание. Статистические наблюдения над одномерной случайной величиной X могут производиться двумя способами:

а) измерением некоторой характеристики а n раз одного элемента;

б) измерением некоторой характеристики а n качественно однородных элементов, образующих генеральную совокупность.

В результате получают выборку объема n. Если произвели n наблюдений этой некоторой постоянной величины а, без систематических ошибок, т.е. MX = a, то под математическим ожиданием понимается ***истинное значение*** *этой случайной величины.*

Если же наблюдения проводятся над всеми качественно однородными элементами, то математическое ожидание случайной величины X является ***генеральным средним*** $\bar{X}$.

• **Медиана x_{me}** – это среднее, полученное путем выявления «центрального» значения в перечне данных, расположенных в ранжи-

рованном порядке. При наличии n значений медиана соответствует $\frac{n+1}{2}$ порядковому номеру или $\frac{m_{est.}}{n} = 0{,}5$.**(Сравните с определением медианы в теории вероятностей)**

Медиана практически не чувствительна к значительным отклонениям отдельных крайних значений наборов вариант. Это свойство называют устойчивостью, которое страхует от случайных ошибок и отдельных недостоверных данных.

Пример 3. *Определите медиану заработной платы, исходя из данного дискретного вариационного ряда:*

Заработная плата (у.е.) – x_i	150	200	250	300	350
Количество сотрудников – m_i	6	8	5	3	2

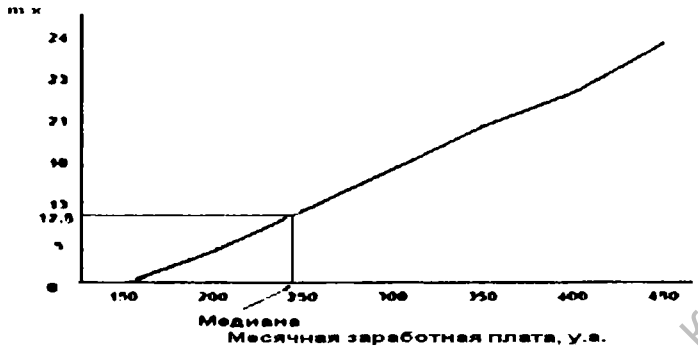
Решение: *Общее количество сотрудников* $n = \sum_{i=1}^{5} m_i = 24$. *Отсюда, медиана есть* $\frac{n+1}{2} = \frac{24+1}{2} = 12{,}5$ *– е значение. Теперь необходимо из данных выбрать 12,5-е значение. Есть 6 сотрудников с зарплатой 150 у.е., есть 8 сотрудников с зарплатой 200 у.е. 6+8=14, т.е. 12,5 – значение это 200 у.е. Итак,* $x_{me} = 200$.

Пример 4. *Определите медиану полугодовой заработной платы по данным интервального статистического ряда:*

Месячная заработная плата (у е.) – x_i	150-	200-	250-	300-	350-	400-
Количество сотрудников – m_i	5	8	5	3	2	1

Решение: *Как и в предыдущем примере медиана соответствует 12,5-му значению, которое находится в интервале 200-250 у.е. Надо определить фактическое значение 12,5 -е значение. Для этого строят кумулянту (или график эмпирической функции распределения) по накопленным частотам (частостям) в соответствии с таблицей*

Месячная заработная плата (у.е.) -x_i	150	200	250	300	350	400	450
Накопленная частота m_{xi}	0	5	13	18	21	23	24



Составим уравнение прямой, проходящей через точки A(200; 5) и B(250; 13):

$\frac{x-200}{250-200}=\frac{y-5}{13-5}$, или $y = 0{,}16x\text{-}27$. *Для* $y = 12{,}5$ *находим* $x_{me} = 246{,}875$.

• **Мода** x_{mo} – это средняя, получаемая путем установления наиболее часто встречающегося значения в наборе данных. (**Сравните с определением моды в теории вероятностей**).

Пример 6. *Определите моду по условию примера 4.*

Решение*: Наиболее часто встречающееся* (8 *раз*) *значение* 200. *Значит,* x_{mo} = 200 у.е.

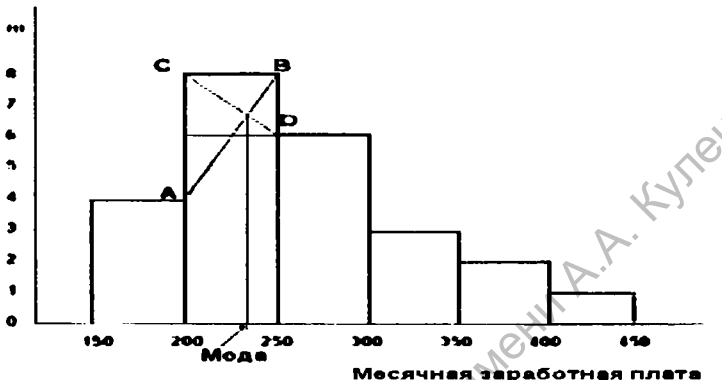
Пример 7. *Определите моду по следующему интервальному вариационному ряду:*

Месячная заработная плата (у.е.) -x_i	150-	200-	250-	300-	350-	400-
Количество сотрудников -m_i	4	8	6	3	2	1

Решение*: Из данной таблицы видно, что наиболее часто повторяется интервал 200-250 у.е.*

Отсюда естественно предположить, что мода находится в пределах этого интервала. Можно определить моду как середину интервала, т.е. 225 у.е. И хотя в этом есть резон, все же лучше определить среднее относительно значений частот по обе стороны наибольшего значения. Мы видим, что значение частот для интервала меньшего 200-250 меньше значения частот для интервала, большего 200-250. Поэтому более вероятно, что мода находится во вто-

рой половине интервала группировки 200-250, т.е. больше 225. Один из принятых методов получения приемлемого значения моды состоит в использовании гистограммы частот.



Найдем пересечение прямых AB *и* СД, *где* A(200;4) , B(250;8), C(200;8) *и* D(250;6). *Уравнение прямой* AB: $\frac{x-200}{250-200}=\frac{y-4}{8-4}$ *или* $y = 0{,}08x-12$.

Уравнение прямой CD: $\frac{x-200}{250-200}=\frac{y-8}{6-8}$ *или* $y = -0{,}04x+16$.

Решив систему уравнений $\begin{cases} y = 0{,}08x - 12, \\ y = -0.04x + 16, \end{cases}$ *получим* $x_{mo} = 233\frac{1}{3}$ у.е.

Данный способ определения моды во многих практических ситуациях позволяет лучше других методов. В том числе метода средней арифметической. Установить «среднее» значение.

Три метода получения «средней», описанные выше, равнозначны, хотя у каждого есть свои достоинства и недостатки:

Метод	Достоинства	Недостатки
Средняя арифметическая	• рассчитывается по формуле • простота понимания	• может быть искажено экстремальными значениями • не всегда репрезентативен с точки зрения данных
Мода	• простота понимания • оптимален с точки зрения выявления "типичного значения из совокупности данных"	• определение на основе графика, хотя можно вывести и аналитический вариант

Метод	Достоинства	Недостатки
		• не подходит для распределения с "особенностями", т.е. включающего в себя два и более максимума
Медиана	• фактическое "центральное" значение • обычно считается наиболее репрезентативным значением	• определение на основе графика или соответствующей математической формуле

К оценкам рассеивания нормально распределенной случайной величины относятся:

• **Выборочная дисперсия** $s^2 = \frac{1}{n}\sum(x_i - \bar{x})^2$ для несгруппированных данных ($s^2 = \frac{1}{n}\sum(x_i - \bar{x})^2 m_i$ для сгруппированных данных) является состоятельной, *смещенной* точечной оценкой $\hat{D}X$ дисперсии DX случайной величины X.

Поэтому, если объем выборки $n < 30$, то используется **несмещенная (исправленная)** оценка дисперсии, которая вычисляется по формуле:

$$\hat{D}X = \hat{\sigma}^2 = s^2_{\text{испр}} = s^2_{\text{несм}} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \text{ (для несгруппированных данных)} =$$

$$= \frac{1}{n-1}\sum_{i=1}^{k}(x_i - \bar{x})^2 m_i \text{ (для сгруппированных данных)}.$$

? Составьте формулу выборочного коэффициента асимметрии и выборочного эксцесса, используя метод моментов, самостоятельно.

• **Выборочное среднее квадратическое отклонение** $s = \sqrt{\hat{D}X}$.

Методы определения точечных оценок

Общим методом определения точечных оценок неизвестных параметров законов распределения, отличных от нормального, является **метод максимального правдоподобия.** Этот метод был предложен в 1912 году английским статистиком Р. Фишером. Данные выборки

рассматриваются как реализация n-мерной случайной величины (X_1, X_2, …, X_n), составляющие которой независимы и имеют плотности распределения $p(X,\Theta_i)$. Строится функция правдоподобия $L(x_1, x_2, ..., x_n) = P((X_1=x_1)(X_2=x_2)...(X_n=x_n))$, которая максимизируется оценкой $\hat{\Theta}_i = u(x_1, x_2, ..., x_n)$ из условия: $\frac{\partial L}{\partial \Theta_i} = 0$ (или $\frac{\partial \ln L}{\partial \Theta_i} = 0$) (исходя из эмпирического правила: «более вероятные события происходят чаще, чем события менее вероятные»).

Метод наибольшего правдоподобия обладает рядом преимуществ по сравнению с методом моментов. Он всегда приводит к асимптотически эффективным и состоятельным оценкам. Но недостатком метода является то, что иногда оценки наибольшего правдоподобия являются смещенными, кроме того, для их нахождения часто приходится осуществлять сложные математические выкладки.

Кроме метода моментов и метода максимального правдоподобия оценки могут находиться и другими методами, например, **методом наименьших квадратов (принцип Лежандра)**. Суть метода состоит в том, что, например, из формул зависимости двух случайных величин X и Y вида $y = f(x)$ наиболее соответствующей опытным данным считается та, для которой сумма квадратов отклонений эмпирических данных от вычисленных является наименьшей. Кроме того, в практике применяется и ряд оценок, предложенных интуитивно без теоретического обоснования. Так для оценки математического ожидания по выборке объема n применяются, например,

- средняя гармоническая $x_{гарм} = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + ... + \frac{1}{x_n}}$;
- средняя геометрическая $x_{геом} = \sqrt[n]{x_1 x_2 .. x_n}$;
- степенные средние $x_{степ} = \sqrt[k]{\frac{x_1^k + x_2^k + ... + x_n^k}{n}}$;
- медиана x_{me};
- мода x_{mo}.

Для оценки дисперсии по выборке иногда применяются:

- коэффициент вариации $\upsilon = \frac{s}{\bar{x}}\%$ показывает как на выборочную среднюю влияет разброс данных.
- размах вариации $R = x_{max} - x_{min}$.

Эти оценки являются оценками «хуже по качеству», чем оценки, полученные методом наибольшего правдоподобия или методом моментов. Поэтому их применяют как дополнительные эмпирические характеристики, описывающие центр распределения или рассеивания измерений случайной величины X.

Контрольные вопросы.

1. Какую совокупность изучают в математической статистике?
2. Является ли совокупность всех значений случайной величины генеральной совокупностью?
3. Как определяется выборочная совокупность?
4. В чем различие между частотой и частостью?
5. Что такое вариационный ряд?
6. Как определяется эмпирическая функция распределения?
7. Как определяются выборочные средние для дискретного и интервального вариационного ряда?
8. Как изменится выборочное среднее, если все варианты ряда умножить на 3?
9. В чем отличие выборочной и исправленной выборочной дисперсиями?

Контрольные задания 3.1 – 3.4

Сводная статистика. Первичная обработка данных. Сравнение средних.

1. Постройте соответствующие графики на основе следующих наборов данных:

- В ходе исследования способа передвижения работников к месту работы были получены следующие результаты:

Способ передвижения	Автомобиль	Поезд	Автобус	Мотоцикл	Пешком	Другой
Количество работников	78	12	22	8	30	10

- Покажите разбивку общих расходов крупной окружной больницы по статьям расходов:

Статья расходов	Персонал	Оборудование	Здания	Услуги
% от общих расходов	40	25	20	15

• Месячный доход среднего магазина электроники за последние 36 месяцев:

Доход(10 тыс. $)	40-	50-	60-	70-	80-
Количество месяцев	3	7	14	8	4

• Сравните данные объема продаж грех компаний за последние четыре года:

	2006	**2007**	**2008**	**2009**
Предприятие А	30	25	26	32
Предприятие Б	18	22	28	33
Предприятие В	24	26	19	14

2. В таблице приведены данные по количеству работников, опоздавших на работу за последние пятьдесят дней. (Фиксировались только опоздания свыше пяти минут.):

15	22	8	26	10	6	1	16	10	17
12	18	7	2	12	15	7	23	13	3
20	9	0	12	16	10	20	11	7	9
11	4	10	19	6	3	8	14	28	14
5	24	9	15	11	13	16	11	8	14

• Составьте на основании этих данных вариационный ряд и вычертите гистограмму и полигон.

• Из данных полученной таблицы рассчитайте значения средней арифметической и среднеквадратического отклонения.

• Сравните полученные значения с данными по второй компании, где за аналогичный период средняя арифметическая равна 18 опоздавшим, а среднеквадратическое отклонение – 3,5 опоздавшим.

3. Найдите среднюю арифметическую, медиану и моду по следующим данным:

• Распределение возрастов выборки из 40 работников:

Возрастной диапазон (лет)	20-	30-	40-	50-	60-
Количество работников	6	15	10	7	2

• Процент брака в 30 выборках, произведенных на линии:

Процент брака	0-	2-	4-	6-	8-	10-
Количество выборок	2	5	9	8	5	1

• Почасовая ставка всех работников (за исключением управленческого персонала) в крупной компании обрабатывающей отрасли промышленности:

Почасовая ставка($):	3.00-	4.00-	5.00-	6.00-	7.00-	8.00-	9.00-
Процент персонала	20	34	30	10	4	1	1

4. Рассчитайте среднюю арифметическую и среднеквадратическое отклонение на основании следующих наборов данных:

• Цена акций при закрытии торгов на фондовой бирже за период в 20 дней:

Максимальная цена за акцию ($)	5.00-	5.20-	5.40-	5.60-	5.80-	6.00-
Количество дней	2	3	7	4	3	1

• Диаметр выборки из 80 шайб, применяемых в мостостроительстве:

Размеры (мм)	20-	22-	24-	26-	28-
Количество изделий	16	26	18	12	8

• Расстояния, зафиксированные группой торговых представителей за одну неделю:

Расстояние (миль)	200-	300-	400-	500-	600-	700-
Количество представителей	3	4	10	3	4	2

5. На основании данных таблицы прокомментируйте различия в ценах акций двух компаний. (Цифры приведены в условных единицах, а цены даны в момент закрытия торгов за последние 60 дней).

	Компания	
	А	Б
Средняя арифметическая	4,00	4,40
Среднеквадратическое отклонение	1,50	0,60

Можно ли сказать, что цены на акции компании А более неустойчивы, чем цены на акции компании Б?

6. Постройте диаграммы на основании следующих наборов данных:

• Недельная заработная плата произвольной выборки работников:

Недельная заработная плата ($)	200-	250-	300-	350-	400-	450-	500-
Количество работников	4	14	20	17	11	7	3

• Количество сверхурочных часов, отработанных группой работников за неделю:

Количество сверхурочных	0-	2-	4-	6-	8-	10-
Количество работников	2	6	13	15	8	5

• Количество работников, опоздавших на работу за период в 65 дней:

Количество опоздавших работников	0	1	2	3	4	5	6	7
Количество дней	25	13	7	9	5	2	3	1

7. Последнее обследование предпочтений телезрителей двух известных сериалов, показанных на телевидении, по возрастным группам аудитории дало следующие результаты (цифры приведены как процент данной возрастной категории от общего количества зрителей):

Возраст (лет)	10-	20-	30-	40-	50-	60-	70-	80-	90-
Программа А	0	2	7	34	23	19	9	5	1
Программа Б	13	40	34	12	1	0	0	0	0

Найдите среднюю арифметическую и среднеквадратическое отклонение возраста зрителей этих двух программ. Прокомментируйте различия в возрасте между двумя группами и, по возможности, объясните их.

3.5 Интервальное оценивание

Найденная из генеральной совокупности с функцией распределения $F(x,\Theta)$ по выборке объема n точечная оценка $\dot{\Theta}$ неизвестного параметра Θ не позволяет непосредственно ответить на вопрос: какую ошибку совершаем при замене точного неизвестного параметра Θ его приближенным значением (оценкой) $\dot{\Theta} = u(x_1, x_2, ..., x_n)$.

По свойству вероятностей непрерывной случайной величины $P(\Theta = \dot{\Theta}) = 0$. *Поэтому с вероятностью равной единице можно утверждать, что совершена ошибка* $\Delta = |\Theta - \dot{\Theta}|$, *принимая за точное значение параметра* Θ *его оценку* $\dot{\Theta}$.

В связи с этим, пользуются **интервальной оценкой,** основанной на определении некоторого интервала (доверительного интервала), который с определенной вероятностью $\gamma = 1 - \alpha$ накрывает неизвестное значение параметра Θ, где γ называют **доверительной вероятностью** или **надежностью**, а α – уровень значимости или риск соверше-

ния **предельной ошибки** Δ. Статистические методы позволяют говорить о выполнении равенства $P(|\Theta - \dot{\Theta}| < \Delta) = \gamma$. Заметим, что доверительный интервал в данном случае является симметричным относительно оценки: $\dot{\Theta} - \Delta < \Theta < \dot{\Theta} + \Delta$.

Существуют несколько подходов к построению доверительных интервалов.

Первый подход основан на подборе такой непрерывной и строго монотонной по Θ функции $\psi(\dot{\Theta}, \Theta)$, называемой статистикой (выборочной статистикой) или выборочной характеристикой (критерием), чтобы ее закон распределения был известен и не зависел от Θ. Тогда при каждом конечном объеме выборки определяется доверительный интервал.

Второй подход использует асимптотические свойства точечных оценок и поэтому пригоден для больших объемов выборки.

Рассмотрим примеры построения доверительных интервалов параметров нормального распределения $N(a, \sigma)$ $a = MX$ и $\sigma^2 = DX$ случайной величины X.

- ***Интервальная оценка MX при известной дисперсии***

Для построения доверительного интервала математического ожидания (генеральной средней) при известной дисперсии составляют статистику, функционально зависимую от наблюдений и связанную с MX, например, для **повторного** отбора $u = \dfrac{\bar{x} - a}{\dfrac{\sigma}{\sqrt{n}}}$, которая распределена по нормальному закону с параметрами 0, 1.т.е. $N(0,1)$, не зависящая от параметра a и как функция параметра a непрерывна и строго монотонна. В силу четности закона распределения $N(0,1)$ имеем $P(-u_{\alpha/2} < u < u_{\alpha/2}) = 1 - \alpha$, откуда $\bar{x} - u_{\alpha/2}\sigma/\sqrt{n} < a < \bar{x} + u_{\alpha/2}\sigma/\sqrt{n}$. При этом $\Delta = u_{\alpha/2}\sigma/\sqrt{n}$ – предельная ошибка. Число $u_{\alpha/2}$ находят по таблицам большой функции Лапласа при условии, что $\Phi(u_{\alpha/2}) = \dfrac{2}{\sqrt{2\pi}} \int_0^t e^{-t^2/2} dt = 1 - \alpha = \gamma.$

При $n \leq 30$ вместо $u_{\alpha/2}$ берут t, распределенное по закону Стьюдента с $\nu = n - 1$ степенями свободы.

Если объем выборки велик, эту оценку можно использовать и при отсутствии нормального распределения генеральной совокупности.

- ***Интервальная оценка MX при неизвестной дисперсии.***

Точечными оценками MX и σ^2 служат средняя выборки $\bar{x}$ и $s^2_{несмещ}$ соответственно. Построение доверительного интервала для $a = MX$ основано на статистике $t(n-1) = \dfrac{\bar{x} - a}{s_{несмещ}/\sqrt{n}}$, которая имеет распределение Стьюдента с $\nu = n - 1$ степенями свободы. Не зависящее от a, и как функция параметра a непрерывна и строго монотонны. Исходя из четности функции распределения Стьюдента, имеем

$P(-t_{\alpha/2} < t(n-1) < t_{\alpha/2}) = 1-\alpha$, откуда $\bar{x} - t_{\alpha/2}s/\sqrt{n} < a < \bar{x} + t_{\alpha/2}s/\sqrt{n}$

и предельная ошибка при повторной выборке $\Delta = t_{\alpha/2}s/\sqrt{n}$.

- ***Интервальная оценка σ нормального распределения***

Требуется выполнение соотношений $P(|\sigma - s| < \Delta) = 1-\alpha = \gamma$ или $P(s - \Delta < \sigma < s + \Delta) = \gamma$. Для того, чтобы можно было пользоваться таблицами, преобразуем двойное неравенство в $s(1 - \Delta/s) < \sigma < s(1 + \Delta/s)$. Положив $\Delta/s = q$, найдем q. Для этого рассмотрим статистику $\chi^2 = \dfrac{S^2(n-1)}{\sigma^2}$, которая распределена по закону «хи-квадрат» с $n - 1$ степенями свободы и функцией плотности распределения $p(\chi, n)$. Вероятность того, что последнее двойное неравенство будет осуществлено, равна $\int_{\sqrt{n-1}(1+q)}^{\sqrt{n-1}/(1-q)} p(\chi,n)d\chi = \gamma$. Из этого уравнения по заданным n и γ по таблице находят q.

Замечание. При $\nu = n-1 > 30$ случайная величина $\chi^2(\nu)$ имеет распределение $N(\sqrt{\nu - 1/\sqrt{2}}, 1/\sqrt{2})$, поэтому с вероятностью, равной 1-α

$\dfrac{2(n-1)s^2}{(\sqrt{2n-3} + u_\alpha)^2} < \sigma^2 < \dfrac{2(n-1)s^2}{(\sqrt{2n-3} - u_\alpha)^2}$, где $\Phi(u_\alpha) = 1 - \alpha = \gamma$.

Сведем сведения в таблицу:

Параметр	Оценка	Предельная ошибка Δ		Примечание
		Повторная выборка	Бесповторная выборка	
a (или $\bar{X}$)	$\bar{x}$	$t\sqrt{\dfrac{s^2}{n}}$	$t\sqrt{\dfrac{s^2}{n}(1-\dfrac{n}{N})}$	$X \in N(a, \sigma)$ Для известной дисперсии $\sigma^2 = s^2$

Параметр	Оценка	Предельная ошибка Δ		Примечание
		Повторная выборка	Бесповторная выборка	
σ	$s_{несмещ.}$	$s(1-q) < \sigma < s(1+q)$		$q = q(\gamma,\nu)$ *находят по таблице при* $\nu = n-1$, *например, [2, с. 464]*
p	$w = \frac{m}{n}$	$t\sqrt{\frac{w(1-w)}{n}}$	$t\sqrt{\frac{w(1-w)}{n}(1-\frac{n}{N})}$	Схема Бернулли
λ	$\bar{x}$	$(\sqrt{\bar{x}} - u_{\alpha/2}\frac{1}{2\sqrt{n}})^2 < \lambda < (\sqrt{\bar{x}} + u_{\alpha/2}\frac{1}{2\sqrt{n}})^2$		Распределение Пуассона $P(X=x) = \frac{\lambda^x}{x!}e^{-\lambda}, x = 0,1,..$

Параметр	Оценка	Необходимый объем выборки n		Примечание
		Повторная выборка	Бесповторная выборка	
a (или $\bar{X}$)	$\bar{x}$	$\frac{t^2 s^2}{\Delta^2}$	$\frac{t^2 s^2 N}{t^2 s^2 + \Delta^2 N}$	$X \in N(a, \sigma)$ Для известной дисперсии $\sigma^2 = s^2$
p	$w = \frac{m}{n}$	$\frac{t^2 w(1-w)}{\Delta^2}$	$\frac{t^2 N w(1-w)}{t^2 w(1-w) + \Delta^2 N}$	Схема Бернулли

Где

- *t – квантиль распределения, соответствующий уровню значимости α:*

при $n>30$ $t=u_{\alpha/2}$ *– квантиль нормального распределения и* $\Phi(u_{\alpha/2}) = \frac{2}{\sqrt{2\pi}}\int_0^x e^{-t^2/2}dt = 1-\alpha = \gamma$;

при $n \le 30$ *t – квантиль распределения Стьюдента с* $\nu = n-1$ *степенями свободы для двусторонней области;*

- *s^2 – выборочная дисперсия; при* $n<30$ *вместо* s^2 *берут* $s^2_{несмещ.}$ *(помня это, обозначения в дальнейшем не будем менять);*
- *N – объем генеральной совокупности;*
- *n – объем выборки.*

Пример 8. Для отрасли, включающей 1500 фирм, составлена случайная выборка из 21 фирмы. По выборке оказалось, что в фирме в среднем работают 78,5 человека при выборочном среднем квадратическом отклонении s = 27 человек. Исходя из 95%-ного доверительного интервала, оцените среднее число работающих в фирме по всей отрасли. Предполагается, что количество работников фирмы подчиняется нормальному закону распределения. Пользуясь 95%-ым доверительным интервалом, оцените вариацию работающих в фирме по всей отрасли.

Решение:

При $\nu=n-1=20$ и $\alpha \approx 1-0{,}95=0{,}05$; $s_{несмещ}=\sqrt{\frac{n}{n-1}}s=27{,}7$ найдем по таблице распределения Стьюдента $t=2{,}09$. Доверительный интервал $\bar{x}-t\sqrt{\frac{s^2_{несм}}{n}(1-\frac{n}{N})}<a<\bar{x}+t\sqrt{\frac{s^2_{несм}}{n}(1-\frac{n}{N})}$ примет вид (66; 91). С вероятностью 95% можно утверждать, что этот интервал накроет среднее число работающих в фирме по всей отрасли.

По условию $n=21$, $s=27$; $\gamma=1-\alpha=0{,}95$. $s_{несмещ}=\sqrt{\frac{n}{n-1}}s=27{,}7$. Определим по таблице $q_{\gamma;n}=q_{0,95;21}=0{,}370$. Тогда $\Delta=q_{\alpha}s_{несм}=10{,}249$ и $17{,}45<\sigma<37{,}9$.

Пример 9. Учреждение коммунального хозяйства желает на основе выборки оценить среднюю квартплату за квартиры определенного типа с надежностью не менее 99% и погрешностью 5 у.е. Предполагая, что квартплата имеет нормальное распределение со средним квадратическим отклонением, не превышающим 25 у.е., найдите минимальный объем выборки.

Решение:

По условию требуется найти такой объем выборки n, при котором $P(|\bar{x}-\bar{X}|<5)\geq 0.99$, где $\bar{x}$ и $\bar{X}$ – выборочная и генеральная средние соответственно. По таблице определим число $t=u_{\alpha/2}$ для $\alpha=1-0{,}99=0{,}01$ и $\Phi(u_{\alpha/2})=0{,}99$. $t=2{,}58$. При $\Delta=5$ и $\sigma=25$ из формулы необходимого объема получим $n=\frac{t^2\sigma^2}{\Delta^2}=166{,}4$. С ростом $1-\alpha$ и уменьшением Δ происходит рост n, поэтому $n\geq 166.4$, т.е. $n_{min}=167$ (конечно, при уменьшении верхней границы для σ будет уменьшаться и n_{min}).

Пример 10. В случайной выборке из 36 аспирантов, специализирующихся по экономике управления на предприятии столичных университетов, 21 человек оказались прибывшими из региональных университетов. Оцените долю аспирантов в обследованных университетах, которые прибыли из регионов, и число таких аспирантов среди 1000 аспирантов при 95%-ном доверительном интервале.

Решение: Определим доверительный интервал с надежностью 0,95 для генеральной доли p: $\frac{m}{n} - t\sqrt{\frac{w(1-w)}{n}(1-\frac{n}{N})} < p < \frac{m}{n} + t\sqrt{\frac{w(1-w)}{n}(1-\frac{n}{N})}$, где $m = 21$, $n = 36$, $N = 1000$,

$t = u_{\alpha/2} = 1{,}96$. Доверительный интервал (0,425; 0,741). Доверительный интервал для числа таких аспирантов среди 1000 человек будет (425; 741).

Контрольные вопросы.

1. Что вкладывается в смысл «хорошая оценка»?
2. Какие Вы знаете методы оценивания параметров?
3. Является ли выборочная средняя хорошей оценкой математического ожидания (генеральной средней)?
4. Какие точечные оценки дисперсии Вы знаете?
5. Что является точечной оценкой вероятности?
6. Почему в статистике используется интервальная оценка? Как определяется интервальная оценка?
7. Проанализируйте таблицы предельной ошибки и необходимого объема выборки.
8. Как изменится необходимый объем выборки, если предельную ошибку уменьшить в два раза?

Контрольные задания 3.5

1. Для исследования востребованности услуг некоторой фирмы осуществлен опрос случайных прохожих, среди которых надо найти пять пользующихся услугами. Оказалось, что 9, 15, 19, 22, 31-ый прохожие пользуются услугами этой фирмы. Оцените вероятность того, что случайный прохожий пользуется услугами фирмы. Для этой вероятности найдите 90%-ный доверительный интервал.

2. Из 300 работников учреждения случайным образом отобрано 20 человек, средняя зарплата которых составила 350 у.е., а среднее квадратическое отклонение 70 у.е. Предположив, что зарплата рас-

пределена по нормальному закону, определите с 98%-ной доверительной вероятностью среднюю зарплату в учреждении и суммарные затраты учреждения на зарплату в месяц

3. Для отрасли, включающей 1300 фирм, была составлена случайная выборка из 48 фирм. По выборочным данным оказалось, что в фирме работают в среднем 73,5 человек при среднем квадратическом отклонении 18 человек;

а) пользуясь 97%-ым интервалом, оцените среднее число работающих в фирме по всей отрасли и общее число работающих в отрасли;

б) пользуясь 94%-ым доверительным интервалом, оцените вариацию работающих в фирме по всей отрасли.

4. С целью установления известности продукции фирма опросила в каждом из пяти районов города по 40 человек. Количество знающих продукцию фирмы оказалось следующим; 25; 20; 30; 15; 20.

а) методом моментов оцените степень известности продукции фирмы;

б) постройте 90%-ый и 95%-ый доверительные интервалы для степени известности продукции. Какой из интервалов шире и почему?

в) пользуясь 95%-ым доверительным интервалом, оцените число жителей среди 1000, знакомых с продукцией фирмы.

5. Распределение 200 станков, потребовавших остановки для регулировки при зарегистрированных разладках за 10 лет представлено в таблице:

Время остановки (часы)	0-2	2-4	4-6	6-8	8-10	10-12
Число станков	129	48	16	4	2	1

Предположив, что время остановки станков имеет показательный закон распределения, найдите точечную и 95%-ую интервальные оценки для среднего числа станков, потребовавших остановки за 10 лет.

6. Имеется 200 дачных участков населения. В результате выборочного обследования 80 дачных участков оказалось, что средняя выборочная урожайность овощей составила 240 ц с гектара при среднем квадратическом отклонении 58 ц с га. Известно, что 36% общей площади посевов овощей занимали огурцы. С надежностью 0,95 определите границы, в которых находится средняя урожайность овощей на всех дачных участках и удельный вес посевов огурцов. Сколько необходимо обследовать дачных участков, чтобы предельная ошибка выборки по признакам уменьшилась в 1,5 раза?

7. Взято 18 проб молока, поступившего на реализацию. Средняя жирность молока составила 3,5%, при среднем квадратическом отклонении 0,4%. Какова вероятность того, что средняя жирность молока всех партий не выйдет за границы от 3,4% до 3,6%?

8. Для того, чтобы определить отношение избирателей к предложению администрации города относительно выпуска облигаций, проводится опрос 200 человек в каждом из двух районов. В одном районе это предложение поддержали 95 человек, в другом – 120. Пользуясь 95%-ными доверительными интервалами, оцените разность процентных долей лиц, поддерживающих предложение администрации города в обследованных районах.

3.6. Проверка статистических гипотез

Статистической гипотезой называют некоторое предположение относительно генеральной совокупности, проверяемое по выборочным данным. Статистические гипотезы делятся на:

- ***параметрические*** – это гипотезы о значении параметра распределения известного вида;
- ***непараметрические*** – это гипотезы о виде (модели закона) предполагаемого распределения.

Параметрическая гипотеза является **простой**, если в ней содержится предположение об одном значении параметра, в противном случае параметрическую гипотезу называют **сложной.** Как правило, выделяют некоторую основную (нулевую) гипотезу H_0 и конкурирующую (альтернативную, противоположную) гипотезу H_1, являющуюся логическим отрицанием H_0. Например, H_0: $a = 5$. H_1: $a \neq 5$ или H_1: $a=7$, или H_1: $a>5$, H_1:$a<5$ и т.д.

Статистическим критерием *K (выборочная статистика)* называют случайную величину, с помощью которой принимают решение о принятии или отклонении нулевой гипотезы. Критерий *K*, применяемый для проверки параметрической гипотезы, называют **критерием значимости**, а для проверки непараметрической гипотезы – **критерием согласия.**

В общем случае схема построения критерия такова: все выборочное пространство делится на две взаимодополняющие области – область принятия нулевой гипотезы H_0 и область отклонения основной гипотезы H_0 *(область принятия альтернативной гипотезы H_1)*, которая называется ***критической***; если выборочная статистика попа-

ла в критическую область, то нулевая гипотеза отклоняется в пользу альтернативной; если же выборочная статистика попала в область принятия нулевой гипотезы, то принимается H_0 и отклоняется H_1

Выбор между гипотезами H_0 и H_1 может сопровождаться ошибками двух родов:

• будет принята гипотеза H_1, тогда как на самом деле верной является H_0 – это ***ошибка первого рода***, ее вероятность обозначают α, которую называют ***уровнем значимости*** или риском:

$P(H_1|H_0) = \alpha$, которая обычно принимает стандартные значения: 0,1; 0,05; 0,01; 0,005; 0,001;

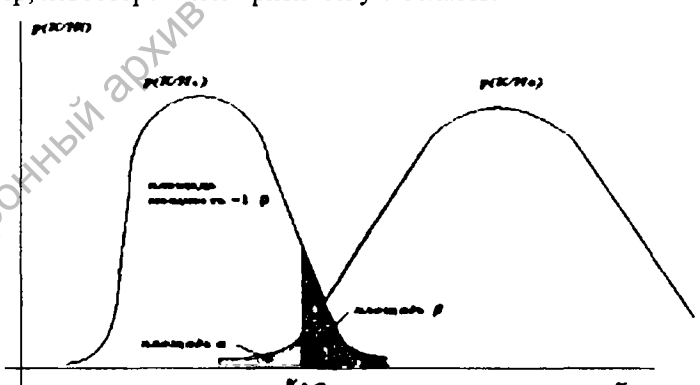
• будет принята гипотеза H_0, тогда как на самом деле верной является H_1 – это **ошибка второго рода**, ее вероятность обозначают β: $P(H_0|H_1) = \beta$.

Правильное решение также может быть двух родов:

• будет принята гипотеза H_0, тогда как и на самом деле верна в генеральной совокупности гипотеза H_0; вероятность такого решения $P(H_0|H_0) = 1 - \alpha$;

• будет принята гипотеза H_1, тогда как и на самом деле верна в генеральной совокупности; вероятность такого решения $P(H_1|H_1) = 1 - \beta$ называют ***мощностью критерия К***.

Дадим геометрическую интерпретацию вероятностей ошибок первого рода, второго рода и мощности критерия K, имеющего, например, левостороннюю критическую область.



Снижение ошибки первого рода α увеличивает вероятность ошибки второго рода. Вероятность совершения ошибки второго рода является функцией объема выборки n, уровня значимости α, характе-

ра альтернативной гипотезы H_1, применяемого критерия K. При этом выполняются предельные условия: $\lim_{n\to\infty}\alpha = 0, \lim_{n\to\infty}\beta = 0, \lim_{n\to 0}\beta = 1$. Откуда следует, что доказательства истинности нулевой и альтернативной гипотез становятся достоверными только при бесконечно большом объеме выборки. При этом же условии имеем единственный способ одновременного уменьшения вероятностей ошибок первого и второго рода. Выбор уровня значимости зависит от тяжести последствий ошибок первого и второго рода. Конечно, нельзя положить $\alpha = 0$, т.к. тогда $\beta = 1$ и будут приниматься все нулевые гипотезы, в том числе и неправильные. Поэтому **будем рассматривать такие статистические критерии – статистические критерии значимости – когда уровень значимости фиксируется заранее**, что позволяет с фиксированным риском принять только одно решение: отклонить нулевую гипотезу или принять ее, а вероятность ошибки второго рода (вероятность принятия ложной нулевой гипотезы) остается неизвестной.

Алгоритм проверки статистических гипотез

- по выборочным данным формируют нулевую и альтернативную гипотезы;
- задают уровень значимости α (0,1; 0,01; 0,005);
- по наблюдениям находят значение критерия K – выборочной статистики как функции наблюдения;
- определяют критическую область, квантили (критические значения) критерия K по таблицам в соответствии с видом альтернативной гипотезы (при верной нулевой гипотезе вероятность попадания значения критерия в критическую область равна α);
- сравнивают расчетное значение критерия с квантилем:

если расчетное значение критерия попало в область принятия нулевой гипотезы, то нулевую гипотезу принимают,

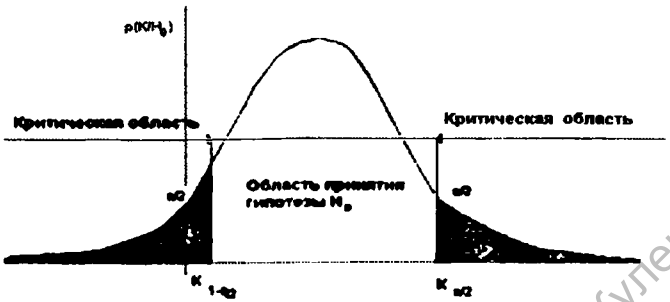
если расчетное значение критерия попало в критическую область, то нулевая гипотеза отклоняется в пользу альтернативной.

Односторонние и двусторонние критические области критерия значимости.

- ***формулировка двустороннего критерия:*** H_0: $\mu = \mu_0$; H_1: $\mu \neq \mu_0$;
- ***формулировка одностороннего критерия:***

H_0: $\mu = \mu_0$; H_1: $\mu > \mu_0$ – **правостороннего;**

H_0: $\mu = \mu_0$; H_1: $\mu < \mu_0$ – **левостороннего.**



Двусторонняя критическая область
$P(K_{1-\alpha/2} < K < K_{\alpha/2}) = 1-\alpha$

Замечание: если распределение статистического критерия является четной функцией (например, распределение Стьюдента, нормальное стандартное распределение), то $K_{1-\alpha} = -K_{\alpha}$

Сведем применение конкретных статистических критериев значимости для проверки параметрических гипотез в следующую таблицу:

ПАРАМЕТРИЧЕСКИЕ КРИТЕРИИ ЗНАЧИМОСТИ				
Гипотеза	**Критерий K (выборочная статистика)**	**Распределение**	**Критическая область**	**Примечание**
H_0: $a = a_0$ H_1: $a \neq a_0$	$u = \frac{x - a_0}{\sigma}\sqrt{n}$	$u \in N(0;1)$	$\lvert u \rvert \geq u_{\alpha/2}$	$X \in N(a;\sigma)$ *) *σ известно*
H_0: $a = a_0$ H_1: $a \neq a_0$	$t = \frac{\bar{x} - a_0}{s}\sqrt{n-1}$	Стьюдента с $\nu = n-1$ степенями свободы	$\lvert t \rvert \geq t_{\alpha/2,\nu}$	$X \in N(a;\sigma)$ *σ неизвестно*
H_0: $a_1 = a_2$ H_1 $a_1 \neq a_2$	$u = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$	$u \in N(0;1)$	$\lvert u \rvert \geq u_{\alpha/2}$	$X_1 \in N(a_1;\sigma_1)$ 1) $X_2 \in N(a_2;\sigma_2)$ $n_1 > 30$; $n_2 > 30$ (дисперсии известны, выборки независимые) 2) Для больших независимых выборок с

ПАРАМЕТРИЧЕСКИЕ КРИТЕРИИ ЗНАЧИМОСТИ				
				нормально распределенными генеральными совокупностями и неизвестными дисперсиями пользуются исправленными выборочными дисперсиями 3)Для больших объемов независимых выборок произвольно распределенных генеральных совокупностей с неизвестными дисперсиями
H_0: $a_1 = a_2$ H_1: $a_1 \neq a_2$	$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$	Стьюдента с $\nu = n_1 + n_2 - 2$ степенями свободы	$\lvert t \rvert \geq t_{\alpha/2,\nu}$	$X_1 \in N(a_1;\sigma_1)$ $X_2 \in N(a_2\sigma_2)$ $\sigma_1^2 = \sigma_2^2$ (малые независимые выборки)
H_0: $\sigma^2 = \sigma_0^2$ H_1: $\sigma^2 > \sigma_0^2$	$\chi^2 = \frac{ns^2}{\sigma_0^2} = \frac{1}{\sigma_0^2}\sum (x_i - \bar{x})^2$	χ^2 распределение с $\nu=n-1$ степенями свободы **)	$\chi^2 \geq \chi^2_{\alpha,\nu}$	$X \in N(a;\sigma)$
H_0: $\sigma_1^2 = \sigma_2^2$ H_1: $\sigma_1^2 > \sigma_2^2$	$F = \frac{s_{1\ несмещ}^2}{s_{2\ несмещ}^2}$	F распределение (Фишера) с $\nu_1 = n_1 - 1$, $\nu_2 = n_2 - 2$ степенями свободы	$F \geq F_{\alpha,n_1,n_2}$	$X_1 \in N(a_1;\sigma_1)$ $X_2 \in N(a_2\sigma_2)$

ПАРАМЕТРИЧЕСКИЕ КРИТЕРИИ ЗНАЧИМОСТИ				
H_0: $p = p_0$ H_1: $p \neq p_0$	$u = \dfrac{\dfrac{m}{n} - p_0}{\sqrt{\dfrac{p_0 q_0}{n}}}$, *n порядка нескольких десятков, $np_0 > 5$*	$u \in N(0;1)$	$\lvert u \rvert \geq u_{\alpha/2}$	$P_n(x) = C_n^x p^x q^{n-x}$ $\dfrac{m}{n} \in N(p; \dfrac{\sqrt{pq}}{n})$
H_0: $p_1 = p_2$ H_1 $p_1 \neq p_2$	$u = \dfrac{\dfrac{m_1}{n_1} - \dfrac{m_2}{n_2}}{\sqrt{\dfrac{pq}{n}}}$ $\bar{p} = \dfrac{m_1 + m_2}{n_1 + n_2}; n = \dfrac{n_1 n_2}{n_1 + n_2}$	$u \in N(0;1)$	$\lvert u \rvert \geq u_{\alpha/2}$	$P_{n_i}(x) = C_{n_i}^x p^x q^{n_i - x}$ $\dfrac{m_i}{n_i} \in N(p_i; \dfrac{\sqrt{p_i q_i}}{n_i})$
H_0: $\sigma_1^2 = \sigma_2^2 = \ldots$ $= \sigma_k^2$ H_1: $\sigma_1^2 \neq \sigma_2^2 \neq$ $\neq \sigma_k^2$	$\chi^2 =$ $\dfrac{\sum_{i=1}^{k} (n_i - 1) \ln(s^2 / s_i^2)}{1 + \dfrac{1}{3(k-1)} \cdot c}$ *где* $\bar{s}^2 = \dfrac{\sum_{i=1}^{k} (n_i - 1) s_i^2}{\sum_{i=1}^{k} (n_i - 1)}$, $c = \sum_{i=1}^{k} \dfrac{1}{n_i - 1} - \dfrac{1}{\sum_{i=1}^{k} (n_i - 1)}$	χ^2 распределение с $\nu = k-1$ степенями свободы	$\chi^2 \geq \chi^2_{\alpha,\nu}$	$X_i \in N(a_i; \sigma_i)$ Критерий Бартлетта $n = \sum_{i=1}^{k} n_i$
H_0: $a_1 = a_2 = \ldots$ $= a_k$ H_1 $a_1 \neq a_2 \neq \ldots$ $\neq a_k$	$F = \dfrac{s_1^2}{s_2^2}$, $s_1^2 = \dfrac{s_{межгр}^2}{s_2^2}$, $s_1^2 = \dfrac{1}{k-1} \sum_{i=1}^{k} (\bar{x}_i - \bar{x})^2 n_i$, $s_2^2 = \dfrac{1}{n-k} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2$	F распределение (Фишера) с $\nu_1 = k-1$, $\nu_2 = n-k$ степенями свободы	$F \geq F_{\alpha,\nu_1,\nu_2}$	Однофакторный дисперсионный анализ $n = \sum_{i=1}^{k} n_i$
H_0: $r = r_0 = 0$ H_1: $r \neq r_0$	$t = \dfrac{\hat{r}\sqrt{n-2}}{\sqrt{1-\hat{r}^2}}$ $\hat{r} = \dfrac{\overline{xy} - \bar{x}\,\bar{y}}{s_x s_y}$ $\hat{r} = \dfrac{\sum_{i=1}^{n} (x_i - x)(y_i - \bar{y})}{n s_x s_y}$	t-распределение Стьюдента с $\nu = n-2$ степенями свободы	$\lvert t(n-2) \rvert < t_\alpha$	*Выборочный коэффициент корреляции $\hat{r}$ называют **статистически значимым** при заданном уровне значимости α, если*

				уровне значимости α, если при этом α нулевую гипотезу отклоняют в пользу альтернативной.

**) Риск ошибки второго рода, меньший β, обеспечивается объемом выборки*

$$n \geq \frac{(u_{2\alpha}+u_{2\beta})^2\sigma^2}{(a_1-a_0)^2},\ \text{где}\ a_1 \neq a_0,\quad \Phi(u_{2\alpha}) = \frac{2}{\sqrt{2\pi}}\int_0^x e^{-t^2/2}dt = 1-2\alpha.$$

***) Если число степеней свободы $\nu > 30$, то критическую точку $\chi^2_{\alpha,\nu}$ можно найти из равенства Уилсона-Гилферти:*

$$\chi^2_{\alpha,\nu} = \nu\left(1-\frac{2}{9\nu}+u_\alpha\sqrt{\frac{2}{9\nu}}\right)^3,\quad \text{где значение функции Лапласа}$$

$$\Phi(u_\alpha) = \frac{2}{\sqrt{2\pi}}\int_0^{u_\alpha} e^{-\frac{t^2}{2}}dt = 1-2\alpha$$

Пример 11. *Предприниматель утверждает, что он получает заказы в среднем по крайней мере от 35% предполагаемых клиентов. Можно ли при 5%-ном уровне значимости считать это утверждение неверным, если предприниматель получил заказы от 30 из 100 отобранных потенциальных клиентов?*

Решение. *Сформулируем нулевую гипотезу:* H_0: $p = p_0 = 0{,}35$. *Альтернативная гипотеза имеет вид:* H_1: $p < 0{,}35$; $n = 100$, $np_0 = 100 \cdot 0{,}35 = 35$; $m = 30$.

По виду альтернативной гипотезы критическая область левосторонняя. Используем критерий (статистику) $u = \dfrac{\frac{m}{n}-p_0}{\sqrt{\frac{p_0q_0}{n}}}$, *численное значение которого равно* $u = -1.05$. *Так как уровень значимости* $\alpha = 0{,}05$, *а табличное значение левостороннего критерия (критическая точка)* $u_{\alpha/2} = -1{,}96 < 0{,}35$, *то нулевую гипотезу отвергаем в*

пользу альтернативной, т.е. с утверждением предпринимателя не согласимся.

***Пример 12**. Торговая компания желает открыть в областном центре филиал. Известно, что фирма будет работать с прибылью, если еженедельный средний доход жителей города превышает 500 у.е. Известно также, что дисперсия дохода $\sigma^2 = 490$ у.е.*

а) Можно ли утверждать при объеме выборки $n = 100$ и уровне значимости $\alpha=0,05$ фирма будет работать прибыльно?

б) Определите вероятность того. Что при применении правила принятия решения, полученного в предыдущем пункте, будет совершена ошибка второго рода. Если в действительности средний доход за неделю достигает 510 у.е.

в) Полагая альтернативное значение генерального среднего 520 у.е., определите объем выборки. При котором риск ошибки первого рода не превысит 0,025, а риск ошибки второго рода не превысит 0,05.

***Решение**. а) Сформулируем нулевую и альтернативную гипотезы. $H_0 : a = a_0 = 500$,*

$H_1: a = a_1 > 500$.

Значение дисперсии σ^2 известно. В этом случае используем критерий $u = \frac{\bar{x} - a_0}{\sigma}\sqrt{n}$ и правостороннюю критическую область. Если $u_{расчетное} \geq u_\alpha$, то принимается альтернативная гипотеза H_1, то есть при $u_{расчетное} = \frac{\bar{x} - 500}{\sqrt{490}}\sqrt{100} \geq u_\alpha = 1,65$. Откуда $\bar{x} \geq 503,6$. Значит, принимается альтернативная гипотеза об открытии филиала.

б) Альтернативное значение еженедельного среднего дохода равно 510 у.е., при этом выполняется гипотеза H_1: $a_1 = 510 > 500$. В этом случае вероятность ошибки второго рода

$$\beta = P(H_0|H_1) = P((\bar{X} < C \Big| \bar{X} \in N(a_1, \frac{\sigma}{\sqrt{n}})) = \frac{1}{2} + \frac{1}{2}\Phi(\frac{C - a_1}{\sigma}\sqrt{n}) = \frac{1}{2} + \frac{1}{2}\Phi(\frac{a_0 + u_\alpha\sigma/\sqrt{n} - a_1}{\sigma\sqrt{n}}) =$$

$$= \frac{1}{2} - \frac{1}{2}\Phi((a_1 - a_0)\sqrt{n}/\sigma - u_\alpha) = 0,5 - \frac{1}{2}\Phi(10 \cdot 10/\sqrt{490} - 1,65) = 0,002$$

в) При заданном α=0,025 риск ошибки второго рода, меньший β=0,05, обеспечивается объемом выборки $n \geq \frac{(u_{2\alpha} + u_{2\beta})^2 \cdot \sigma^2}{(a_1 - a_0)^2} =$

$$= \frac{(u_{0,05} + u_{0,1})^2 \cdot 490}{(520 - 500)^2} = \frac{(1,96 + 1.65)^2 490}{20^2} = 15,\ddot{9}6.$$

Пример 13. *При изготовлении изделий как по новой, так и по старой технологиям расход сырья как случайная величина имеет нормальное распределение. Влияет ли технология на средний расход сырья на одно изделие при уровне значимости α=0,05, если имеются следующие результаты наблюдений:*

	Старая технология				*Новая технология*		
x_i – *расход сырья*	101	104	105	100	101	103	105
m_i – *число изделий*	2	5	4	1	4	7	1

Решение. *Определим выборочные средние, соответственно по старой технологии* $\bar{x}_c = \frac{1}{11}(101\cdot 2 + 104\cdot 5 + 105\ 4) = 103{,}82$ *и по новой технологии* $\bar{x}_н = \frac{1}{13}(100\cdot 1 + 101\cdot 4 + 103\cdot 7 + 105\cdot 1) = 102{,}3$, *а также несмещенные выборочные дисперсии*

$$s_c^2 = \frac{1}{10}((101-103{,}82)^2\cdot 2 + (104-103{,}82)^2\cdot 5 + (105-103{,}82)^2\cdot 4) = 2{,}16;$$

$$s_н^2 = \frac{1}{12}((100-102{,}3)^2\cdot 1 + (101-102{,}3)^2\cdot 4 + (103-102{,}3)^2\cdot 7 + (105-102{,}3)^2\ 1) = 1{,}89$$

Генеральные дисперсии по условию не известны и неизвестно, равны ли они. Прежде чем сравнивать генеральные средние, проверим гипотезу H_0: $\sigma_1^2 = \sigma_2^2$ *при альтернативной гипотезе* H_1: $\sigma_1^2 > \sigma_2^2$. *Согласно критерию* $F = \frac{s^2_{c\ несмещ}}{s^2_{н\ несмещ}}$ *вычислим* $F_{расчет} = \frac{2{,}16}{1{,}89} = 1{,}14$. *По таблицам определим* $F_{0.05;10;12}$=2,76>1,14, *поэтому принимаем нулевую гипотезу о равенстве генеральных дисперсий.*

Теперь проверим гипотезу о равенстве генеральных средних H_0: $a_1 = a_2$ *при альтернативной гипотезе* H_1: $a_1 > a_2$. *Согласно критерия* $t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}(\frac{1}{n_1} + \frac{1}{n_2})}}$ *вычислим* $t_{расчет.} = \frac{103{,}82 - 102{,}3}{\sqrt{\frac{11\cdot 2{,}16 + 13\cdot 1{,}89}{11+13-2}(\frac{1}{11} + \frac{1}{13})}} = 2{,}5036.$

Табличное значение $t_{0.05;22}$=1,717 < 2,5036, *поэтому принимаем альтернативную гипотезу, то есть применение новой технологии снижает средние затраты сырья на одно изделие.*

ПРОВЕРКА ГИПОТЕЗЫ О МОДЕЛИ ЗАКОНА РАСПРЕДЕЛЕНИЯ ГЕНЕРАЛЬНОЙ СОВОКУПНОСТИ

О виде закона распределения генеральной совокупности ($F(x)$ *или* $p(x)$) можно судить по графику выборочной плотности распределения, гистограмме, полигону. Параметры Θ_i закона распределения

могут быть известными, а обычно неизвестными и их заменяют на выборочные значения, оценки этих параметров $\hat{\Theta}_i$. Полной же уверенности в том, что в результате получится истинный закон распределения, к которому принадлежит данная выборка, не существует. Поэтому ставится вопрос лишь о том, что на определенном уровне доверия выбранный закон ***согласуется*** с данными выборки, что и определило название **критерия согласия.**

В качестве примера рассмотрим критерий согласия Пирсона (критерий χ^2).

Сформулируем нулевую гипотезу H_0: выборка извлечена из совокупности, имеющей распределение с функцией $F(x) = F_0(x, \Theta_i)$, где значения параметров $\Theta_i (i=1, 2, ...l)$ известны (или заменены их оценками в случае, когда они не известны).

H_1: $F(x) \neq F_0(x_i, \Theta_i)$.

Алгоритм проверки гипотезы:

- весь диапазон выборочных значений (вариант x_i, объем выборки $n \geq 50$) разбивают на k полуинтервалов $\Delta_i = [a_i, a_{i+1}), i = 1,2,...k$, длиной h, где число m_i вариант в i-ом интервале и $\sum_{i=1}^{k} m_i = n$.

Рекомендация: $a_1 = x_{min} - 0,5\ h$, $a_k \leq x_{max} < a_{k+1}$ (см. п. 3.2)

- Для каждого интервала вычисляют теоретические вероятности p_i попадания случайной величины X, растянув крайние интервалы вариационного ряда до бесконечности.

$p_i = P(X \in \Delta_i) = F(a_{i+1}) - F(a_i)$, $\sum_{i=1}^{k} p_i = 1$.

- Подсчитывают соответствующие теоретические частоты np_i, причем если для некоторых интервалов они меньше 5, то их объединяют с соседними, чтобы $np_i \geq 5$. Новое число интервалов обозначим k^*.

- При $n \to \infty$ статистика $\sum_{i=1}^{k^*} \frac{(m_i - np_i)^2}{np_i}$ имеет распределение χ^2 с $\nu = k^* - l - 1$ степенями свободы l – число неизвестных параметров предполагаемой функции распределения $F(x)$, оцениваемых по результатам наблюдений (если все параметры предполагаемого закона известны точно, то $l = 0$). Величина $\chi^2 = \sum_{i=1}^{k^*} \frac{(m_i - np_i)^2}{np_i}$ называется **критерием согласия** χ^2 или **критерием согласия Пирсона**. Чем ближе к нулю наблюдаемое значение критерия, тем вероятнее, что нулевая гипотеза справедлива. Поэтому для проверки нулевой гипо-

тезы используется **правосторонняя критическая область.** Для заданного уровня значимости α определяем ее $[\chi^2_{\alpha,\nu}, +\infty)$.

- Если расчетное значение критерия согласия $\chi^2_{расчет} < \chi^2_{\alpha,\nu}$, то принимают нулевую гипотезу. В противном случае нулевая гипотеза отклоняется в пользу альтернативной.

Пример 14. *Для работников отрасли исследуется нормальный закон распределения средней заработной платы при уровне значимости* $\alpha = 0{,}05$. *Для этого было обследовано 100 человек. Результаты представлены в таблице:*

Зарплата в у.е.	200-202	202-204	204-206	206-208	208-210	210-212	212-214	214-216	216-218
Количество человек	1	4	9	22	28	19	11	5	1

Нормальное распределение имеет два параметра – математическое ожидание $MX = a$ *и дисперсия* $DX = \sigma^2$. *В задаче они не известны, поэтому найдем их оценки*

$$\hat{a} = \bar{x} = \frac{1}{100}(201\cdot 1 + 203\cdot 4 + 205\cdot 9 + 207\cdot 22 + 209\cdot 28 + 211\cdot 19 + 213\cdot 11 + 215\cdot 5 + 217\cdot 1) = 209{,}08$$

$$\hat{\sigma}^2 = s^2_{испр} = \frac{1}{99}((201-209{,}08)^2\cdot 1 + (203-209{,}08)^2\cdot 4 + (205-209{,}08)^2\cdot 9 + (207-209{,}08)^2\cdot 22 +$$

$$+ (209-209{,}08)^2\cdot 28 + (211-209{,}08)^2\cdot 19 + (213-209{,}08)^2\cdot 11 + (215-209{,}08)^2\cdot 5 +$$

$$+ (217-209{,}08)^2\cdot 1) = 9{,}448080 \quad \hat{\sigma} = 3{,}0738.$$

Для определения расчетного значения критерия $\chi^2_{расчет}$ *составим таблицу в соответствии с алгоритмом:*

№ i-го интервала	$a_i - a_{i+1}$	m_i		$F(a_i)=1/2(1+\Phi(\frac{a-\hat{a}}{\hat{\sigma}}))$	$F(a_{i+1})=1/2(1+\Phi(\frac{a_{i+1}-\hat{a}}{\hat{\sigma}}))$	$p_i=F(a_{i+1})-F(a_i)$	np_i		$\chi^2=\frac{(m_i-np_i)^2}{np_i}$
1	200-202	1	14	0,0000	0,0081	0,0081	0,81	15,795	0,20399
2	202-204	4		0,0081	0,0490	0,0409	4,09		
3	204-206	9		0,0490	0,15795	0,10895	10,895		
4	206-208	22		0,15795	0,3624	0,20445	20,445		0,11827
5	208-210	28		0,3624	0, 6175	0,2551	25,51		0,01574
6	210-212	19		0, 6175	0,8292	0,2117	21,17		0,22243
7	212-214	11	17	0,8292	0,9454	0,1162	11,62	17,08	0,000375
8	214-216	5		0,9454	0,9879	0,0425	4,25		
9	216-218	1		0,9879	1	0,0121	1,21		
Всего		100				1,0000	100		$\chi^2_{расчет}$=0,560805

По таблице распределения Пирсона для уровня значимости $\alpha = 0,05$ и $\nu = k-l-1 = 5-2-1=2$ степенями свободы найдем $\chi^2_{0,005;2}$=5,99. так как $\chi^2_{расчет}$=0,560805<5,99, то гипотезу о нормальном распределении средней заработной платы принимаем.

КРИТЕРИЙ ОДНОРОДНОСТИ

Пусть имеются l независимых выборок, каждая объемом n_i соответственно, где i =1, 2, …, l; $\sum_{i=1}^{l} n_i = n$. Проверяется гипотеза об однородности выборок. H_0: выборки извлечены из одной и той же совокупности.

Алгоритм проверки гипотезы:

- Для каждой выборки составляется вариационный интервальный ряд с одинаковым числом интервалов (групп) k, где m_{ij} – частота наблюдений из i-ой выборки в j-ом интервале, т.е. $\sum_{i=1}^{l}\sum_{j=1}^{k} m_{ij} = n$, $\sum_{j=1}^{k} m_{ij} = n_{i\bullet} = n_i$, $\sum_{i=1}^{l} m_{ij} = n_{\bullet j}$

• Предполагая справедливой гипотезу H_0, подсчитывают частости p_j принадлежности наблюдений к каждому из интервалов (групп): $p_j = \frac{n_{\bullet j}}{n}$, а затем ожидаемые частоты $\hat{m}_{ij} = n_{i\bullet} p_j = n_{i\bullet} n_{\bullet j} / n$.

• Вычисляют расчетное значение критерия Пирсона $\chi^2 = \sum_{i=1}^{l} \sum_{j=1}^{k} \frac{(m_{ij} - \hat{m}_{ij})^2}{\hat{m}_{ij}}$, которое при справедливости нулевой гипотезы и при $\hat{m}_{ij} > 5$ имеет χ^2 – распределение со степенями свободы $\nu = (l-1)(k-1)$. Критическая область правосторонняя.

Пример 15. *Фирма предполагает, что применение новой технологии улучшит долю качественной продукции. Выборочные совокупности контроля двух партий продукции, изготовленных по старой и новой технологиям, представлены в таблице.*

Технологии / Изделия	Новая	Старая	Всего
качественные	$m_{11} = 180$	$m_{21} = 205$	$n_{\bullet 1} = 385$
бракованные	$m_{12} = 15$	$m_{22} = 20$	$n_{\bullet 2} = 35$
Всего	$n_1 = n_{1\bullet} = 195$	$n_2 = n_{2\bullet} = 225$	$n = 420$

При уровне значимости $\alpha = 0{,}005$ проверьте справедливость предположения.

Решение. *Имеется две выборки $l = 2$, каждая из которых разбита на две группы $k = 2$. То есть требуется проверить нулевую гипотезу о равенстве вероятностей (p_{11} и p_{21}) изготовления качественных изделий соответственно при новой и старой технологиях:*

H_0: $p_{11} = p_{21}$, где $p_{12} = 1 - p_{11}$ и $p_{22} = 1 - p_{21}$ – вероятности изготовления бракованных изделий соответственно при новой и старой технологиях.

Если нулевая гипотеза верна, то ожидаемые частоты будут такими:

$$\hat{m}_{11} = n_{1\bullet} n_{\bullet 1} / n = 195 \quad 385/420 = 178{,}75;$$
$$\hat{m}_{21} = n_{2\bullet} n_{\bullet 1} / n = 225 \quad 385/420 = 206{,}25;$$
$$\hat{m}_{12} = n_{1\bullet} n_{\bullet 2} / n = 195 \quad 35/420 = 16{,}25;$$
$$\hat{m}_{22} = n_{2\bullet} n_{\bullet 2} / n = 225 \quad 35/420 = 18{,}75.$$

Все ожидаемые частоты $\dot{m}_y > 5$, *поэтому вычислим наблюдаемой значение статистического критерия*

$$\chi^2_{расчет} = \frac{(180-178{,}75)^2}{178{,}75} + \frac{(205-206{,}25)^2}{206{,}25} + \frac{(15-16{,}25)^2}{16{,}25} + \frac{(20-18{,}75)^2}{18{,}75} = 0{,}196$$

При уровне значимости $\alpha = 0{,}005$ *и* $\nu = (l\text{-}1)(k\text{-}1) = (2\text{-}1)(2\text{-}1) = 1$ *степеней свободы по таблице определяем значение правосторонней критической точки* $\chi^2_{0{,}005;1} = 3{,}84$. *Так как расчетное значение критерия меньше значения критической точки, то нулевую гипотезу об изменении выхода качественной продукции при новой технологии не принимаем.*

3.7. Введение в дисперсионный анализ

Метод дисперсионного анализа позволяет проверить, оказывает ли влияние на математические ожидания случайных величин определенные факторы, которые можно изменять в ходе эксперимента, выбрать важные факторы и оценить степень их влияния. Если на математические ожидания (генеральные средние) оказывает влияние только один фактор, то соответствующий критерий значимости называется однофакторным дисперсионным анализом, если несколько – многофакторным дисперсионным анализом.

Суть метода состоит в разложении общей вариации (общей дисперсии) изучаемого показателя на части, соответствующие совместному и раздельному влиянию факторов, и статистическом изучении этих частей с целью выяснения приемлемости гипотез о существовании этих влияний.

ОДНОФАКТОРНЫЙ ДИСПЕРСИОННЫЙ АНАЛИЗ

В процессе применения однофакторного дисперсионного анализа варианты разбиваются, в зависимости от степени действия (уровня) фактора, на группы. Суть однофакторного дисперсионного анализа заключается в разбиении общей дисперсии случайной величины на два независимых слагаемых – факторную (межгрупповую), порождаемую воздействием исследуемого фактора, и остаточную (внутригрупповую), обусловленную различными другими неучтенными и случайными факторами, т.е. $s^2_{общ} = s^2_{факт} + s^2_{ост.}$

Алгоритм однофакторного дисперсионного анализа:

Пусть результаты независимых наблюдений x_{ij} влияния фактора на некоторый результативный признак разбили на k групп по n_i (i=1, 2, …, k; j=1, 2, …, n_i) как выборки из нормальных генеральных совокупностей $X_i \in N(a_i;\sigma_i)$. Параметры a_i, σ_i хотя и неизвестны, но предполагается, что

$\sigma_1^2 = \sigma_2^2 = ... = \sigma_k^2$. Проверка этого условия осуществляется с помощью критерия Бартлетта (см. таблицу проверки параметрических гипотез).

- Формулируется нулевая гипотеза H_0: $a_1 = a_2 = ... = a_k$ против альтернативной H_1- не все математические ожидания равны между собой;
- Подсчитывают выборочные средние каждой из групп $\bar{x}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij}$, $i = 1,2,...,k$;
- Посчитывают выборочную среднюю всех наблюдений $\bar{x} = \sum_{i=1}^{k}\sum_{j=1}^{n_i} x_{ij}$, $\sum_{i=1}^{k} n_i = n$;
- Вычисляют расчетное значение выборочной статистики $F_{расчет} = \frac{\frac{1}{k-1}\sum(\bar{x}_i - \bar{x})^2 n_i}{\frac{1}{n-k}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)^2}$. При справедливости нулевой гипотезы эта выборочная статистика имеет F-распределение с $\nu_1 = k - 1$, и $\nu_2 = n - k$ степенями свободы. Критическая область правосторонняя.

Если $F_{расчет.} \geq F_{a,\nu_1,\nu_2}$, то нулевая гипотеза отвергается в пользу альтернативной (фактор влияет значимо).

Если $F_{расчет.} < F_{a,\nu_1,\nu_2}$, то нулевая гипотеза принимается (фактор влияет не значимо).

Для удобства все вычисления можно расположить в таблице дисперсионного анализа:

Источник вариации (изменчивости)	Суммы квадратов отклонений	Число степеней свободы	Дисперсия	Наблюдаемое (расчетное) значение критерия
Фактор (между группами)	$S^2_{факт} = \sum_{i=1}^{k}(\bar{x}_i - \bar{x})^2 n_i$ **факторная вариация**	$k-1$	$s^2_{факт} = s_1^2 = \frac{1}{k-1} S^2_{факт}$	$F_{расчет} = \frac{s_1^2}{s_2^2}$
Остаточный (внутри групп)	$S^2_{ост} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)^2$ **остаточная вариация**	$n-k$	$s^2_{ост} = s_2^2 = \frac{1}{n-k} S^2_{ост}$	$F_{расчет} = \frac{\frac{1}{k-1}\sum(\bar{x}_i - \bar{x})^2 n_i}{\frac{1}{n-k}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)^2}$
Общая (полная) вариация	$S^2_{общ} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \bar{x})^2$ **общая вариация**	$n-1$	$s^2_{общ} = s^2 = \frac{1}{n-1} S^2_{общ}$	

Пример 16. *Группа социологов провела исследование влияния четырех способов рекламирования товара на объем продажи. Для этого в каждом из районных центров, где использовались различные способы рекламы, были собраны сведения об объемах продажи товара в денежном выражении в четырех случайно отобранных магазинах и вычислены соответствующие выборочные характеристики:*

Способ рекламы (k)	1	2	3	4
Объем продаж	142	149	147	152
	140	150	148	151
	144	150	150	155
	143	152	149	154
n_i	4	4	4	4
$\bar{x}_i$	142,25	150,25	148,5	153,00
$s^2_{i(испр)} = \hat{\sigma}_i^2$	2,92	1,58	1,67	3,3

Можно ли считать влияние доказанным при уровне значимости 5%?

Решение. *Фактором является способ рекламы. Даны четыре его уровня. Поставлена задача о различии по своему влиянию эти уровни. Допустим, что независимость наблюдений гарантируется*

организацией эксперимента, а объем продаж при каждом способе рекламы имеет нормальный закон распределения со своими математическими ожиданиями и одинаковыми дисперсиями. Используя критерий Бартлетта, убедимся предварительно в справедливости гипотезы $H_0: \sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2$. *Вычислим*

$$\chi^2_{расчет} = \frac{\sum_{i=1}^{k}(n_i - 1)\ln(s^2/s_i^2)}{1 + \frac{1}{3(k-1)} \cdot c} = \frac{3\sum_{i=1}^{4}\ln(2{,}3675/s_i^2)}{1 + \frac{1}{3 \cdot 3} \cdot \frac{15}{12}} = \frac{1{,}9042 \cdot 36}{41} = 1{,}67,$$

где $\bar{s}^2 = \frac{\sum_{i=1}^{k}(n_i - 1)s_i^2}{\sum_{i=1}^{k}(n_i - 1)} = \frac{(4-1)\ 2{,}92 + (4-1)\cdot 1{,}58 + (4-1)\cdot 1{,}67 + (4-1)\cdot 3{,}3}{12} = 2{,}3675,$

$$c = \sum_{i=1}^{4}\frac{1}{n_i - 1} - \frac{1}{\sum_{i=1}^{4}(n_i - 1)} = \frac{4}{3} - \frac{1}{12} = \frac{15}{12}.$$

По таблице определяем критическое значение выборочного критерия $\chi^2_{0,005;3} = 7{,}815$, *которое больше расчетного, поэтому принимаем предварительную нулевую гипотезу о равенстве дисперсий.*

Проверим теперь основную нулевую гипотезу дисперсионного анализа

H_0: $a_1 = a_2 = a_3 = a_4$. *Для этого вычислим выборочную среднюю* $\bar{x} = \frac{1}{4}(142{,}25 + 150{,}25 + 148{,}5 + 153{,}00) = 148{,}5$ *и суммы квадратов отклонений от средних (факторную и остаточную):*

$S^2_{факт} = (142{,}25 - 148{,}5)^2 \cdot 4 + (150{,}25 - 148{,}5)^2 \cdot 4 + (148{,}5 - 148{,}5)^2 \cdot 4 + (153{,}00 - 148{,}5)^2 \cdot 4 = 249{,}5;$

$S^2_{ост} = (142 - 142{,}25)^2 + (140 - 142{,}25)^2 + (144 - 142{,}25)^2 + (143 - 142{,}25)^2 + (149 - 150{,}25)^2 +$

$+ (150 - 150{,}25)^2 \cdot 2 + (152 - 150{,}25)^2 + (147 - 148{,}5)^2 + (148 - 148{,}5)^2 + (150 - 148{,}5)^2 +$

$(149 - 148{,}5)^2 + (152 - 153)^2 + (151 - 153)^2 + (155 - 153)^2 + (154 - 153)^2 = 28{,}5.$

Общая сумма квадратов отклонений

$S^2_{общ} = (142 - 148{,}5)^2 + (140 - 148{,}5)^2 + (144 - 148{,}5)^2 + (143 - 148{,}5)^2 + (149 - 148{,}5)^2 +$

$+ (150 - 148{,}5)^2 \cdot 3 + (152 - 148{,}5)^2 + (147 - 148{,}5)^2 + (148 - 148{,}5)^2 + (149 - 148{,}5)^2 +$

$(152 - 148{,}5)^2 + (151 - 148{,}5)^2 + (155 - 148{,}5)^2 + (154 - 148{,}5)^2 = 278.$

Убеждаемся, что $S^2_{общ} = 278 = 249{,}5 + 28{,}5 = S^2_{факт} + S^2_{ост}$

Вычисляем расчетное значение критерия

$$F_{расчет} = \frac{\frac{1}{k-1}\sum(\bar{x}_i - \bar{x})^2 n_i}{\frac{1}{n-k}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)^2} = \frac{\frac{1}{4-1} \cdot 249{,}5}{\frac{1}{16-4} \cdot 28{,}5} = 35{,}02.$$

По распределению Фишера-Снедекора с $\nu_1 = k - 1 = 4-1 = 3$ и $\nu_2 = n - k = 16-4 = 12$ *степенями свободы для уровня значимости* $\alpha = 0{,}05$ *определим критическое значение*

$F_{0,05;3;12} = 3{,}49$. *Расчетное значение больше критического, поэтому нулевую гипотезу отклоняем, а принимаем альтернативную гипотезу о значимом влиянии фактора, что и требовалось доказать.*

Для измерения степени влияния фактора на результативный признак используют ***выборочный коэффициент детерминации*** $d = \frac{\hat{\sigma}_{факт}}{\hat{\sigma}_{общ}}$, который показывает, какую долю выборочной дисперсии составляет дисперсия групповых средних (или какую долю общей дисперсии можно объяснить зависимостью результативного признака X от фактора)..

Контрольные вопросы:

1. Приведите примеры статистических параметрических гипотез (основной и альтернативной) из области экономической деятельности.
2. Какова связь ошибки первого рода и уровня доверия?
3.
4. Какого вида бывают критические области?
5. Основную гипотезу отклоняют, если значение статистики критерия попадает в критическую область. Почему?
6. Что такое мощность критерия?
7. Как называются критерии при проверке параметрических статистических гипотез?
8. Какой вид имеет статистика критерия для проверки гипотезы о значении генерального среднего при известной и неизвестной дисперсиях?
9. Какой вид имеет статистика критерия для проверки гипотезы о значении дисперсии?
10. Какой вид имеет статистика для проверки гипотезы о величине доли признака?
11. Приведите примеры статистических гипотез (основной и альтернативной) из области коммерческой или биржевой деятельности.
12. Партия изделий принимается, если дисперсия контролируемого размера не превышает 0,2. По выборке $n = 40$ изделий вычислена $s^2 = 0{,}25$. можно ли принять партию при уровне значимости $\alpha = 0{,}05$?

13. Какое распределение используется при применении критерия Пирсона?

14. В чем суть однофакторного дисперсионного анализа?

Контрольные задания 3.6-3.7

1. Прогноз задолженности квартплаты по ЖЭУ таков: средняя задолженность равна 110 у.е. Среднее квадратическое отклонение задолженности $\sigma = 18$ у.е. Выборочные подсчеты по девяти ЖЭУ дали среднюю задолженность 123 у.е. Принимается ли прогноз или отвергается при $\alpha = 0{,}05$? Какова вероятность, что вывод будет ошибочным?

2. Статистика по страховой компании утверждает, что только 3 из каждых 10 визитов страхового агента заканчиваются заключением договора о страховании. Однако агент Разумов в результате 100 визитов заключил 38 договоров. Если Вы – начальник агента Разумова, что Вы решите: случайны его результаты или они свидетельствуют о его высокой квалификации?

3. Сторонники строительства атомной электростанции утверждают, что по крайней мере 50% населения поддерживают данное строительство. Из 300 опрошенных 35% высказались за данный проект, а 33% – за поиск новых, других энергоносителей. При 5%-ном уровне значимости проверьте справедливость утверждения сторонников атомной ЭС; сравните количество сторонников АЭС и других видов энергоносителей.

4. На протяжении месяца в произвольно отобранных районах *A* и *B* фирма проводила новые мероприятия по расширению торговых услуг, а в двух других произвольно отобранных районах *C* и *D* торговля велась традиционными методами. Объемы продаж за текущий и за предыдущий месяцы по четырем районам таковы:

Район	*A*	*B*	*C*	*D*
Объем продаж за предыдущий месяц	75	45	30	150
Объем продаж за текущий месяц	115	75	40	170

В районах, где фирма проводила специальные мероприятия, она, по-видимому,

добилась больших успехов, чем в остальных районах.

• Определите ожидаемый объем продаж по каждому району за текущий месяц (для вычисления доли суммарного объема продаж, ожидаемой для каждого района в текущем месяце, используйте данные о продажах в предыдущем месяце);

• Существенно ли при 5%-ном уровне значимости различие между распределениями наблюдаемых и ожидаемых частот для текущего месяца?

5. Проверьте гипотезу о законе распределения, который определяется двумя параметрами, причем значение одного из них известно по следующим данным:

m_i	2	5	15	14	15	20
np_i	1	3	11	15	18	23

6. При 1%-ном уровне значимости сравните четыре фирмы по качественному составу годной продукции, классифицированной по сортам, по следующим выборочным данным:

Сорт	Фирма			
	1	2	3	4
1	149	179	122	279
2	107	112	188	128
3	38	55	31	43

7. С целью установления известности продукции фирма опросила в каждом из 100 населенных пунктов по 20 человек. Распределение числа незнакомых с продукцией фирмы представлено в таблице:

x_i	0	1	2	3	4	5	6
Количество населенных пунктов	60	23	10	4	3	1	1

При 5%-ном уровне значимости можно ли считать, что число незнакомых с продукцией распределено по закону Пуассона?

8. Определите дисперсию каждой из следующих групп:

I	7	8	10	11	13	15
II	21	23	24	26	27	29

Сравните дисперсию 12 значений, полученных путем объединения этих групп, с групповыми дисперсиями.

9. Из совокупности экспертов извлечена случайная выборка в 10 человек. Каждый эксперт оценивал случайно выбранные 20 студенческих работ по пятибалльной шкале. Общая вариация оценок $S^2_{общ}$=1832,4, вариация оценок, обусловленная различиями экспертов $S^2_{факт}$=94,33. Кто в среднем отличается больше: эксперты или студенческие работы, которые оценивает один и тот же эксперт?

10. В течение пяти лет использовались три различные технологии по изготовлению продукции.

Необходимо установить влияние различных технологий на продуктивность по данным таблицы:

Год	Технология (фактор A)		
	A_1	A_2	A_3
1	1,2	0,9	1,7
2	1,4	0,8	1,8
3	1,5	1,2	1,0
4	1,1	0,9	1,3
5	1,3	1,2	1,2

3.8. Корреляционный и регрессионный анализ

Обозначим через X независимую случайную величину, а через Y – зависимую случайную величину.

Если каждому значению СВ X соответствует единственное значение СВ Y, то зависимость величины Y от X называют ***функциональной***.

Если каждому значению СВ X соответствует целое распределение значений СВ Y, то зависимость величины Y от X называют ***стохастической или вероятностной зависимостью***. *Например, со стохастической зависимостью встречались в дисперсионном анализе. В роли переменной X выступает фактор A, уровни фактора A – это значения переменной X, каждому такому значению соответствует не одно, а множество непредсказуемых значений Y.*

Пусть случайные величины X и Y находятся в стохастической зависимости. Для каждого фиксированного $X = x$ определим математическое ожидание СВ Y: $M(Y/X=x)$ – ***условное математическое ожидание***. Если при изменении x условные математические ожидания $M(Y/X=x)$ изменяются, то говорят, что имеет место ***корреляционная зависимость величины Y от X***; если же условные математические ожидания остаются неизменными, то говорят, что корреляционная зависимость величины Y от X отсутствует.

Примером корреляционой связи является статистическая зависимость между частями человеческого тела.

Основной задачей корреляционного анализа является определение тесноты связи между переменными X и Y и количественная оценка тесноты этой связи. Установление же формы зависимости, оценка

функции регрессии и ее параметров являются задачами регрессионного анализа.

Корреляционная зависимость двух случайных величин (**парная корреляция**) задается моделью $X = X(Y,Z), Y = Y(X,Z)$, где Z – набор внешних случайных факторов. Парная корреляция изучает характеристики взаимосвязи двух случайных величин. Основой получения этих характеристик служит совместное распределение случайных величин $F(x,y) = P(X < x, Y < y)$, а также ***генеральный коэффициент корреляции*** $r(X,Y) = r = \dfrac{M(X - MX)(Y - MY)}{\sigma_X \sigma_Y}$, средние квадратические отклонения $\sigma_X = \sqrt{DX}, \sigma_y = \sqrt{DY}$.

Основной оценкой коэффициента корреляции (выборочный аналог генерального коэффициента корреляции), тесноты связи между переменными X и Y, является ***выборочный коэффициент корреляции***:

$$\hat{r} = r_в = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\hat{\sigma}_x \cdot \hat{\sigma}_y} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n \hat{\sigma}_x \hat{\sigma}_y}.$$

Пусть значения случайных величин X и Y – пары действительных чисел (x_i, y_j), где i=1, 2, …, l; j = 1, 2, …, m. Тогда

$$\hat{r} = \frac{n \sum_{i=1}^{l} \sum_{j=1}^{m} x_i y_j m_{ij} - \left(\sum_{i=1}^{l} x_i m_{i\bullet} \right) \left(\sum_{j=1}^{m} y_j m_{\bullet j} \right)}{\sqrt{n \sum_{i=1}^{l} x_i^2 m_{i\bullet} - \left(\sum_{i=1}^{l} x_i m_{i\bullet} \right)^2} \sqrt{n \sum_{j=1}^{m} y_j^2 m_{\bullet j} - \left(\sum_{j=1}^{m} y_j m_{\bullet j} \right)^2}}.$$

Свойства выборочного коэффициента корреляции аналогичны свойствам генерального коэффициента корреляции:

- $-1 \le \hat{r} \le 1$. Чем ближе $|\hat{r}|$ к 1, тем теснее связь между переменными X и Y.
- При умножении переменных X и Y на одно и то же число, коэффициент корреляции не изменяется.
- Условие $\hat{r} = \pm 1$ является необходимым и достаточным для существования между X и Y линейной функциональной зависимости.

В отличие от генерального коэффициента корреляции r выборочный коэффициент $\hat{r}$ является случайной величиной, т.к. он вычисляется по выборочным данным. Если $\hat{r} \neq 0$, то появляется вопрос, объясняется ли это линейной зависимостью между X и Y или это вызвано случайными факторами. Поэтому проверяется гипотеза об от-

сутствии корреляционной связи между переменными X и Y (или о значимости выборочного коэффициента корреляции) в виде: $H_0 . r = 0$. При справедливости этой гипотезы выборочная статистика $t = \frac{\hat{r}\sqrt{n-2}}{\sqrt{1-\hat{r}^2}}$ имеет распределение Стьюдента с $\nu = n - 2$ степенями свободы. Критическая область двусторонняя.

Основной задачей корреляционного анализа является определение тесноты связи между переменными X и Y и количественная оценка тесноты этой связи.

Установление же формы зависимости, оценка функции регрессии и ее параметров являются задачами регрессионного анализа.

Функция $\varphi(x) = M(Y / X = x)$, описывающая изменение условного математического ожидания СВ Y при изменении значений x переменной X, называется ***функцией регрессии Y на X***. Аналогично определяется ***функция регрессии X на Y***: $\psi(y) = M(X / Y = y)$. Их графики называются соответственно линиями регрессии y по x и x по y и проходят через точку с координатами $(\bar{X}, \bar{Y})$.

Линии условных дисперсий характеризуют, насколько точно линии регрессии передают изменение одной случайной величины при изменении другой:

$$\sigma^2_{Y/X} = D(Y / X = x) = M((Y - \varphi(x))^2 / X = x),$$
$$\sigma^2_{X/Y} = D(X / Y = y) = M((X - \psi(y))^2 / Y = y).$$

Средние из условных дисперсий характеризуют точность прогноза одной случайной величины с помощью другой на всем диапазоне изменения последней: $\overline{\sigma^2_{Y/X}} = M((Y \quad \varphi(x))^2), \overline{\sigma^2_{X/Y}} = M((X \quad \psi(y))^2)$.

Прямолинейные регрессии $\bar{Y}_X = \varphi(x) - a_1 + b_1 x, \quad \bar{X}_Y = \psi(y) - a_2 + b_2 y$ задаются следующими коэффициентами:

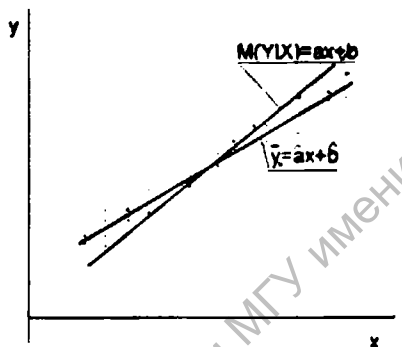
$$b_1 = \frac{\mu_{11}}{\sigma_X^2} = r\frac{\sigma_Y}{\sigma_X}, \quad b_2 = \frac{\mu_{11}}{\sigma_Y^2} = r\frac{\sigma_X}{\sigma_Y}, \quad a_1 = MY - b_1 MX, \quad a_2 = MX - b_2 MY.$$

Выборочными аналогами линейной регрессии служат уравнения $\bar{y}_x = \hat{a}_1 + \hat{b}_1 x, \quad \bar{x}_y = \hat{a}_2 + \hat{b}_2 y$, где коэффициенты определяются из следующих систем нормальных уравнений Гаусса: $\begin{cases} n\hat{a}_1 + \hat{b}_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i, \\ \hat{a}_1 \sum_{i=1}^{n} x_i + \hat{b}_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i \end{cases}$ *и* $\begin{cases} n\hat{a}_2 + \hat{b}_2 \sum_{i=1}^{n} y_i = \sum_{i=1}^{n} x_i, \\ \hat{a}_2 \sum_{i=1}^{n} y_i + \hat{b}_2 \sum_{i=1}^{n} y_i^2 = \sum_{i=1}^{n} x_i y_i \end{cases}$ соответственно.

Если X и Y – система нормально распределенных случайных величин, то линейные уравнения регрессии Y на X и X на Y можно записать:

$$\bar{y}_x - \bar{y} = \hat{r}\frac{\hat{\sigma}_y}{\hat{\sigma}_x}(x - \bar{x}),$$

$$\bar{x}_y - \bar{x} = \hat{r}\frac{\hat{\sigma}_x}{\hat{\sigma}_y}(y - \bar{y}).$$



В случае нелинейных регрессий степень концентрации распределения вблизи линии регрессии Y по X показывает ***генеральное корреляционное отношение*** $\rho_{Y/X} = \sqrt{\dfrac{M(M(Y/X) - MY)^2}{M(Y - MY)^2}} = +\sqrt{1 - \dfrac{\sigma^2_{Y/X}}{\sigma^2_Y}}$. По выборочной совокупности определяют ***эмпирическое корреляционное отношение*** $\hat{\rho}_{Y/X} = \sqrt{\dfrac{(\bar{\bar{Y}}_X - \bar{\bar{Y}})^2}{(Y - \bar{\bar{Y}})^2}} = +\sqrt{1 - \dfrac{\hat{\sigma}^2_{Y/X}}{\hat{\sigma}^2_Y}}$.

Квадрат генерального корреляционного отношения называется ***генеральным коэффициентом детерминации***; он показывает, какую долю дисперсии величины Y составляет дисперсия условных математических ожиданий, или, иначе говоря, какая доля дисперсии DY объясняется корреляционной зависимостью Y от X.

Очевидно, что квадрат корреляционного отношения меняется в пределах от 0 до 1. Он равен 1 тогда и только тогда, когда $\sigma^2_{Y/X} = 0$, то есть все распределение сосредоточено на линии регрессии (имеет место функциональная зависимость). Корреляционное отношение равно нулю тогда и только тогда, когда линия регрессии Y по X представляет собой горизонтальную прямую линию, проходящую через центр

распределения – точку с координатами $((MX, MY)$ *или* $(\bar{X}, \bar{Y}))$, то есть если Y и X некоррелируемы. Во всех случаях доказано, что $r^2 \le \rho^2_{Y/X}$, $r^2 \le \rho^2_{X/Y}$.

Примером регрессионной зависимости является зависимость между урожайностью определенной сельскохозяйственной культуры и влияниями на нее природными и экономическими факторами.

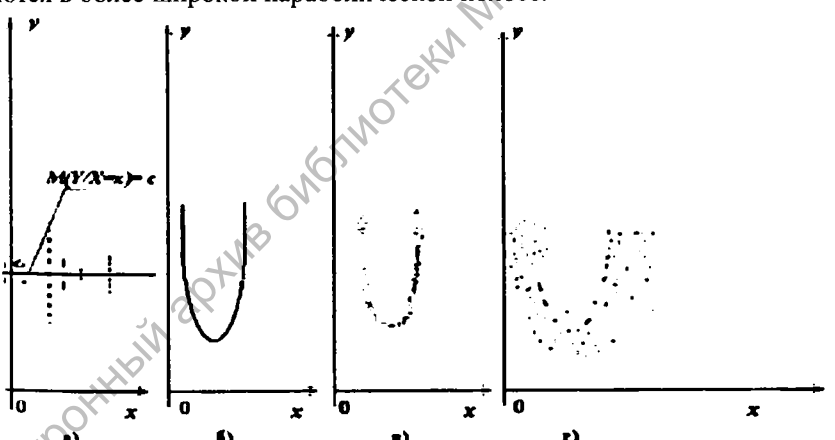
Проиллюстрируем указанные свойства диаграммами рассеивания или ***корреляционным полем*** (графическое изображение точек с координатами (x_i, y_j) соответствующих наблюдаемым значениям переменных X и Y):

а) корреляционная зависимость отсутствует: условные математические ожидания не изменяются, $\rho_{Y/X} = 0$;

б) зависимость Y от X функциональная: $\rho_{Y/X} = 1$;

в) стохастическая зависимость довольно сильная: точки сконцентрированы в относительно узкой параболической полосе;

г) стохастическая зависимость не сильная: точки концентрируются в более широкой параболической полосе.



Экспериментальные данные задаются ***корреляционной таблицей***, где отражаются значения пар (x_i, y_j), где $i = 1, 2, \ldots, l$; $j = 1, 2, \ldots, m$, случайных величин X и Y и соответствующие им частоты m_{ij} появления.

Для интервального задания случайных величин X и Y в качестве x_i и y_j берутся соответственно середины интервалов.

Пример 17. *Даны распределения 69 предприятий по стоимости основных промышленно-производственных фондов (X) и объема выпуска продукции (Y):*

Объем выпуска продукции (Y) / Стоимость основных фондов (X)	Середина интервала y_j	0-2	2-4	4-6	6-8	8-10	$m_{\cdot j}$
	Середина интервала x_i	1	3	5	7	9	
0-0,2	0,1	2	2				4
0,2-0,4	0,3	2	8	12			22
0,4-0,6	0,5		2	18	8		28
0,6-0,8	0,7			4	3		7
0,8-1,0	0,9				3	2	5
1,0-1,2	1,1					3	3
$m_{i\cdot}$		4	12	34	14	5	$\sum_{i=1}^{5}\sum_{j=1}^{6} m_{ij} = n = 69$

Проверить значимость коэффициента корреляции на уровне значимости 5%.

Исследовать зависимость объема выпуска продукции от стоимости основных промышленно-производственных фондов.

Вычислим выборочный коэффициент корреляции, составив для удобства таблицу:

i, j	x_i	$m_{i\cdot}$	y_i	$m_{\cdot i}$	$x_i m_{i\cdot}$	$y_i m_{\cdot i}$	x_i^2	$x_i^2 m_{i\cdot}$	y_i^2	$y_i^2 m_{\cdot i}$
1	1	4	0,1	4	4	0,4	1	4	0,01	0,04
2	3	12	0,3	22	36	6,6	9	108	0,09	1,98
3	5	34	0,5	28	170	14	25	850	0,25	7
4	7	14	0,7	7	98	4,9	49	686	0,49	3,43
5	9	5	0,9	5	45	4,5	81	405	0,81	4,05
6			1,1	3		3,3			1,21	3,63
Σ		69		69	353	33,7		2053		20,13

$$\bar{x} = \frac{1}{69}(1\cdot 4 + 3\cdot 12 + 5\cdot 34 + 7\cdot 14 + 9\cdot 5) = 5{,}12;$$

$$\bar{y} = \frac{1}{69}(0{,}1\cdot 4 + 0{,}3\cdot 22 + 0{,}5\cdot 28 + 0{,}7\cdot 7 + 0{,}9\cdot 5 + 1{,}1\cdot 3) = 0{,}49;$$

$$s_X^2 = \overline{x^2} - (\bar{x})^2 = \frac{1}{69}(1\cdot 4 + 9\cdot 12 + 25\cdot 34 + 49\cdot 14 + 81\cdot 5) - 5{,}12^2 = 3{,}44; \quad s_X = 1{,}88.$$

$s_y^2 = \overline{y^2} - (\overline{y})^2 = \frac{1}{69}(0{,}1^2 \cdot 4 + 0{,}3^2 \cdot 22 + 0{,}5^2 \cdot 28 + 0{,}7^2 \cdot 7 + 0{,}9^2 \cdot 5 + 1{,}1^2 \cdot 3) - 0.49^2 = 0{,}0516;\ s_Y = 0{,}23.$

$\sum_{i=1}^{5}\sum_{j=1}^{6} x_i y_j m_{ij} = 1 \cdot 0{,}1\ \ 2 + 1\ \ 0{,}3 \cdot 2 + 3\ \ 0{,}1 \cdot 2 + 3\ \ 0{,}3 \cdot 8 + 3\ \ 0{,}5 \cdot 2 + 5\ \ 0{,}3 \cdot 12 + 5\ \ 0{,}5\ \ 18 + 5 \cdot 0{,}7 \cdot 4 +$

$+ 7 \cdot 0{,}5 \cdot 8 + 7 \cdot 0{,}7 \cdot 3 + 7 \cdot 0{,}9 \cdot 3 + 9 \cdot 0{,}9 \cdot 2 + 9 \cdot 1{,}1 \cdot 3 = 196{,}1$

Выборочный коэффициент корреляции $\hat{r} = \frac{69 \cdot 196{,}1 - 353 \cdot 33{,}7}{\sqrt{69 \cdot 2053 - 353^2}\sqrt{69 \cdot 20{,}13 - 33{,}7^2}} = \frac{1634{,}8}{130{,}568 \cdot 15{,}915} = 0{,}787.$ $\hat{r}^2 = 0{,}619.$

Коэффициент детерминации показывает, что 61,9% *различий в объеме выпуска продукции объясняется стоимостью основных фондов.*

Проверим значимость $\hat{r}$ *на уровне* $\alpha = 0{,}05$. *Для этого вычислим статистику* $t_{расчет} = \frac{0{,}787 \cdot \sqrt{67}}{\sqrt{1 - 0{,}787^2}} = 10{,}441.$ *По таблицам распределения Стьюдента при* $\nu = n - 2 = 67$ *находим критическое значение* $t_{крит.} = t_{\,0.05;\,67} = 2{,}00$. *Расчетное значение критерия* $t_{расчет} = 10{,}441 > t_{крит.} = 2{,}00$, *поэтому считаем выборочное значение коэффициента корреляции значимым.*

Определим уравнение регрессии $y - \bar{y} = \hat{r}\frac{s_y}{s_x}(x - \bar{x})$:

$\bar{y}_x - 0{,}49 = 0{,}787 \cdot \frac{0{,}23}{1{,}88}(x - 5{,}12)$ *или* $\bar{y}_x = 0{,}0963x - 0{,}00296.$

Контрольные вопросы к модулю 3.

1. Определите разницу в понятиях: «статистика» (оценка) и параметр распределения.
2. Каким требованиям должны удовлетворять выборки из генеральной совокупности?
3. Какие требования предъявляются к оценкам параметров распределения?
4. Определите понятия «точечная», «интервальная» оценка.
5. Определите понятия «гипотеза», «критерий проверки гипотезы», «ошибка первого и второго рода».
6. Влияет ли выбор уровня значимости α на надежность оценки?
7. Как изменение уровня значимости влияет на вероятность совершения ошибки первого или второго рода?
8. Какие критерии проверки гипотез Вам известны?
9. Каким образом могут быть определены критические точки в разных ситуациях?

10. Опишите наиболее употребляемые распределения статистик, используемых для проверки гипотез.

11. Каким образом можно быстро сделать приблизительное заключение, подчиняется ли случайная величина нормальному закону распределения?

12. Для проверки каких гипотез применяются критерии значимости?

13. Для проверки каких гипотез применяются критерии согласия?

14. В каких случаях применяются односторонние и двусторонние критерии?

15. Определите понятие «мощность критерия».

16. Каким образом формируются гипотезы о вероятностях, о средних, о дисперсиях, о законах распределения?

17. Определите понятие «стандартное отклонение». В каких формулах для вычисления значений критериев присутствует эта величина?

18. Опишите суть метода однофакторного дисперсионного анализа.

19. Определите понятие «уравнение регрессии».

20. Что имеется в виду, когда говорится «регрессионная модель линейна»?

21. Для каких целей может быть использованы уравнения регрессии?

22. Опишите процедуру оценивания «метод наименьших квадратов».

23. Опишите процедуру оценивания «метод правдоподобия».

24. Что такое «система нормальных уравнений»?

25. Что является решением системы нормальных уравнений?

26. Как по графику линии регрессии определить коэффициенты a, b для линейной регрессии $Y=a+b\,X$?

27. Как проверяется «статистическая значимость» выборочного коэффициента корреляции?

28. Какова связь корреляционного отношения и коэффициента корреляции?

29. Каков смысл коэффициента детерминации?

30. Каково расположение графиков линейных регрессий Y на X и X на Y при приближении случайных величин к функциональной линейной зависимости?

Индивидуальное задание № 3

Номер варианта совпадает с последней цифрой номера зачетной книжки.

Задача № 1

Вероятность того, что расход электроэнергии в некотором учреждении окажется нормальным (не превысит определенного числа кВт/час в сутки) равна p. Построить ряд распределения случайной величины X – количество дней, для которых расход электроэнергии окажется нормальным в течение n суток. Найти математическое ожидание, дисперсию и среднее квадратическое отклонение этой СВ X.

№ варианта	0	1	2	3	4	5	6	7	8	9
p	0.8	0.7	0.6	0.8	0.5	0.6	0.9	0.7	0.8	0.6
n	4	6	5	6	6	5	5	4	4	6

Задача № 2

Банк обслуживает N вкладчиков. Для определения средней суммы вкладов в банке произведено n вкладов. По данным бесповторной выборки найти доверительный интервал для генерального среднего, который можно было бы гарантировать с точностью до $p\%$.

№ варианта	0		1		2	
N	7320		4300		5200	
$p\%$	98		99		93	
	Сумма вкладов	Число вкладов	Сумма вкладов	Число вкладов	Сумма вкладов	Число вкладов
	2-8	41	4-9	51	3-7	41
	8-14	32	9-14	73	7-11	23
	14-20	16	14-19	19	11-15	2
	20-26	8	19-24	21	15-19	14
	26-32	77	24-29	2	19-23	73
	32-38	49	29-34	41	23-27	17

№ варианта	3		4	
N	2400		2500	
p %	98		99	
	Сумма вкладов	Число вкладов	Сумма вкладов	Число вкладов
	3-8	21	4-8	9
	8-13	9	8-12	3
	13-18	17	12-16	7
	18-23	43	16-20	68
	23-28	17	20-24	29
	28-33	41	24-28	13

№ варианта	5		6		7	
N	4135		1000		3120	
p %	90,01		95,3		96,7	
	Сумма вкладов	Число вкладов	Сумма вкладов	Число вкладов	Сумма вкладов	Число вкладов
	10-17	14	2-12	29	3-8	93
	17-24	14	12-22	125	8-13	13
	24-31	25	22-32	71	13-18	51
	31-38	75	32-42	4	18-23	17
	38-45	140	42-52	59	23-28	47
	45-52	54	52-62	97	28-32	51

№ варианта	8		9	
N	2000		3000	
p %	97		95	
	Сумма вкладов	Число вкладов	Сумма вкладов	Число вкладов
	12-14	3	7-13	23
	14-16	4	13-19	10
	16-18	2	19-25	15
	18-20	12	25-31	12
	20-22	29	31-37	49
	22-24	13	37-43	41

Задача № 3

По данному статистическому материалу опыта требуется:

1. составить статистический ряд распределения;

2. составить интервальный вариационный ряд относительный частот, разбив размах варьирования на k интервалов;

3. построить гистограмму и полигон относительных частот;

4. найти эмпирическую функцию распределения и построить ее график;

5. вычислить числовые характеристики: среднее арифметическое выборки x, выборочную дисперсию $S^2_{выб}$, выборочное среднее квадратическое отклонение $S_{выб}$, коэффициент вариации $V_{выб}$;

6. по виду гистограммы и полигона относительных частот, а также по значению $V_{выб}$ сделать предварительный выбор закона распределения;

7. Найти точечные оценки параметров распределения и функцию распределения СВX.

8. Найти теоретические частости распределения, проверить согласие эмпирической функции распределения $\hat{F}(x)$ с теоретической $F(x)$ при помощи критерия согласия χ^2. Проверить гипотезу о нормальном распределении.

В случае нормального распределения по заданному уровню значимости α:

9. найти интервальные оценки параметров распределения;

10. проверить нулевую гипотезу H_0: $a=a_0$ о математическом ожидании при альтернативной гипотезе H_1: $a=a_0$ ($a>a_0$, $a<a_0$);

11. проверить нулевую гипотезу H_0: $\sigma^2=\sigma_0^2$ о дисперсии против альтернативной H_1: $\sigma^2=\sigma_0^2$ ($\sigma^2>\sigma_0^2$, $\sigma^2<\sigma_0^2$).

Статистические данные:

37	49	43	31	44	38	40	31	28	43	32	44
47	29	51	25	43	38	41	32	38	24	46	49
32	34	31	28	37	46	41	35	43	25	37	46
38	24	41	50	38	29	41	32	34	49	44	37
31	47	50	34	25	37	40	32	35	28	44	43
46	35	41	35	29	43	38	31	26	34	49	32
46	26	38	35	40	51	37	46	38	25	40	34

№ варианта	0	1	2	3	4	5
k	7	8	9	11	10	7
α- уровень значимости значимости	0,01	0,01	0,05	0,05	0,01	0,01
a_0	a_1	a_2	a_2	a_1	a_2	a_2
σ_0^2	S_2^2	S_1^2	S_2^2	S_2^2	S_1^2	S_2^2
гипотезы H_1	$a \neq a_0$ $\sigma^2 < \sigma_0^2$	$a > a_0$ $\sigma^2 \neq \sigma_0^2$	$a < a_0$ $\sigma^2 > \sigma_0^2$	$a \neq a_0$ $\sigma^2 < \sigma_0^2$	$a < a_0$ $\sigma^2 < \sigma_0^2$	$a > a_0$ $\sigma^2 \neq \sigma_0^2$
i	1	15	11	20	26	34

№ варианта	6	7	8	9
k	9	8	10	11
α- уровень значимости значимости	0,05	0,05	0,05	0,01
a_0	a_1	a_2	a_1	a_1
σ_0^2	S_2^2	S_2^2	S_2^2	S_1^2
гипотезы H_1	$a \neq a_0$ $\sigma^2 > \sigma_0^2$	$a \neq a_0$ $\sigma^2 < \sigma_0^2$	$a \neq a_0$ $\sigma^2 \neq \sigma_0^2$	$a < a_0$ $\sigma^2 < \sigma_0^2$
i	30	7	11	15

a_1, S_1 – соответственно значения доверительного интервала a и σ в левом конце, a_2, S_2 – в правом конце. Объем выборки n=50; i – порядковый номер x_i, от которого ведется отсчет значений случайной величины X, считая по строке.

Задача № 4

В течение пяти лет использовались три различные технологии по изготовлению продукции. Необходимо установить влияние различных технологий на продуктивность и степень этой значимости по данным таблицы:

№ варианта	0			1			2			3			4		
год	Технология (фактор A)			Технология (фактор A)			Технология (фактор A)			Технология (фактор A)			Технология (фактор A)		
	A_1	A_2	A_3	A_1	A_2	A_3	A_1	A_2	A_3	A_1	A_2	A_3	A_1	A_2	A_3
1	1,2	0,9	1,7	3,1	3,5	2,5	1,1	1,5	1,1	1,3	1,2	0,9	5,3	6,1	6,3
2	1,4	0,8	1,8	3,2	3,7	2,9	1,3	0,8	1,2	1,4	1,3	1,1	5,2	6,1	6,4
3	1,5	1,2	1,0	3,1	3,8	3,1	1,2	1,2	1,0	1,3	1,2	1,3	5,1	6,3	6,5
4	1,1	0,9	1,3	3,0	3,9	3,4	1,0	1,4	1,3	1,2	1,4	1,5	5,6	6,4	6,6
5	1,3	1,2	1,2	3,0	2,8	3,3	1,3	1,2	1,1	1,5	1,1	1,4	5,7	6,7	6,9

№ варианта	5			6			7			8			9		
год	Технология (фактор A)			Технология (фактор A)			Технология (фактор A)			Технология (фактор A)			Технология (фактор A)		
	A_1	A_2	A_3	A_1	A_2	A_3	A_1	A_2	A_3	A_1	A_2	A_3	A_1	A_2	A_3
1	2,0	3,1	1,2	5,3	3,1	7,0	3,1	4,2	5,2	1,3	2,3	1,2	4,2	3,1	5,3
2	2,1	3,1	1,3	5,4	3,2	7,2	3,6	4,0	5,3	1,4	2,1	1,1	4,5	3,7	5,4
3	2,2	3,5	1,1	5,2	3,7	7,1	3,8	4,3	5,1	1,6	2,6	1,3	4,3	4,5	5,2
4	2,4	3,6	1,5	5,7	3,8	7,9	3,4	4,5	5,4	1,5	2,0	1,5	4,4	4,8	4,9
5	2,7	3,7	1,6	4,9	4,5	7,0	3,5	4,8	5,2	1,6	2,2	1,4	4,1	4,3	5,7

Уровень значимости принять 0,05; 0,01.

Задача № 5

Даны распределения 100 фирм по производственным средствам X (млн. руб.) и суточной выработке Y(т). Известно, что между случайными величинами существует линейная корреляционная зависимость. По заданной корреляционной таблице определить:

- Числовые характеристики случайных величин X и Y; Коэффициент корреляции r;
- Уравнение прямой регрессии Y на X; Проверить значимость коэффициента корреляции при уровне значимости $\alpha = 0{,}05$.
- Построить корреляционное поле и график уравнения регрессии Y на X; Отклонения между теоретическими значениями $\bar{Y}_x$ и экспериментальными $\bar{y}_x$.
- Составить уравнение и построить график уравнения регрессии X на Y.

Вариант 0

Y \ X	1,5-2,9	2,9-4,3	4,3-5,7	5,7-7,1	7,1-8,5	8,5-9,9	9,9-11,3	11,3-12,7	$m_{.j}$
120 – 280		4	3	5					12
280 – 440		6	7	8					21
440 – 600			10	12	11				33
600 – 760					5	4	3		12
760 – 920						6	8		14
920 – 1080							3	5	8
$m_{i.}$	0	10	20	25	16	10	14	5	100

Вариант 1

Y \ X	50-62	62-74	74-86	86-98	98-110	110-122	122-134	134-146	$m_{.j}$
7 – 11	2	3	5						10
11 – 15		6	3	5					14
15 – 19			5	8	15				28
19 – 23				6	9	10			25
23 – 27					1	6	8		15
27 - 31						3	4	1	8
$m_{i.}$	2	9	13	19	25	19	12	1	100

Вариант 2

Y \ X	5,5-18,5	18,5-31,5	31,5-44,5	44,5-57,5	57,5-70,5	70,5-83,5	83,5-96,5	96,5-109,5	$m_{.j}$
20 – 520	4	2	5						11
520 – 1020			7	5	2				14
1020 – 1520				9	14	6			29
1520 – 2020				7	8	6			21
2020 – 2520					4	5	7		16
2520 - 3520						3	2	4	9
$m_{i.}$	4	2	12	21	28	20	9	4	100

Вариант 3

Y \ X	12,5-15,5	15,5-18,5	18,5-21,5	21,5-24,5	24,5-27,5	27,5-30,5	30,5-33,5	33,5-36,5	$m_{\cdot j}$
15-21	2	4	6						12
21-27		2	7	6					15
27-33			6	8	5				19
33-39				8	14	4			26
39-45					3	6	8		17
45-51							5	6	11
$m_{i\cdot}$	2	6	19	22	22	10	13	6	100

Вариант 4

Y \ X	100-120	120-140	140-160	160-180	180-200	200-220	220-240	240-260	$m_{\cdot j}$
8,5 – 11,5	1	3	4						8
11,5 – 14,5		5	6	5					16
14,5 – 17,5			4	8	6				18
17,5 – 20,5			6	15	9				30
20,5 – 23,5					5	6	7		18
23,5 – 26,5						1	7	2	10
$m_{i\cdot}$	1	8	20	28	20	7	14	2	100

Вариант 5

Y \ X	20-22	22-24	24-26	26-28	28-30	30-32	32-34	34-36	$m_{\cdot j}$
82,5 – 97,5	1	3	2						6
97,5 – 112,5		4	2	3					9
112,5 – 127,5			5	7	6				18
127,5 – 142,5				6	14	9			29
142,5 – 157,5					7	6	7		20
157,5 – 172,5						6	7	5	18
$m_{i\cdot}$	1	7	9	16	27	21	14	5	100

Вариант 6

Y \ X	15-17	17-19	19-21	21-23	23-25	25-27	27-29	29-31	$m_{.j}$
2,1 - 2,5	3	2	4						9
2,5 - 2,9		5	6	1					12
2,9 – 3,3			6	9	4				19
3,3 – 3,7				8	16	7			31
3,7 – 4,1					8	6	5		19
4,1 – 4,5						4	5	1	10
$m_{i.}$	3	7	16	18	28	17	10	1	100

Вариант 7

Y \ X	60-68	68-76	76-84	84-92	92-100	100-108	108-116	116-124	$m_{.j}$
0,85 – 1,15	2	3	5						10
1,15 – 1,45		6	3	5					14
1,45 – 1,75			5	8	15				28
1,75 – 2,05				6	9	10			25
2,05 – 2,35					1	6	8		15
2,35 – 2,65						3	4	1	8
$m_{i.}$	2	9	13	19	25	19	12	1	100

Вариант 8

Y \ X	1,5-3,5	3,5-5,5	5,5-7,5	7,5-9,5	9,5-11,5	11,5-13,5	13,5-15,5	15,5-17,5	$m_{.j}$
145 – 275	5	3	4						12
275 – 405		7	8						15
405 – 535			9	10	14				33
535 – 665				8	7	6			21
665 – 795					2	3	2		7
795 - 925							6	6	12
$m_{i.}$	5	10	21	18	23	9	8	6	100

Вариант 9

Y \ X	140-180	180-220	220-260	260-300	300-340	340-380	380-420	420-460	$m_{.j}$
8,5 – 11,5	1	4	5						10
11,5 – 14,5		6	7	2					15
14,5 - 17,5			5	8	6				19
17,5 - 20,5				9	13	6			28
20,5 - 23,5					7	8	4		19
23,5 - 26,5							6	3	9
$m_{i.}$	1	10	17	19	26	14	10	3	100

Решение задач параллельно возможно и с использованием Пакета анализа табличного процессора Excel.

ЛИТЕРАТУРА:

1. Белько И.В., Кузьмич К.К. Высшая математика для экономистов. III семестр: Экспресс-курс. – М.: Новое знание, 2002. – 144 с.

2. Гмурман В.Е. Теория вероятностей и математическая статистика. – М.: Высш. школа, 1997.

3. Гмурман В.Е. Руководство к решению задач по теории вероятностей и математической статистике: учеб. пособие для студентов вузов. Изд. 4-е стер. – М.: Высш. шк., 1998. – 400 с.: ил.

4. Горелова Г.В., Кацко И.А. Теория вероятностей и математическая статистика в примерах и задачах с применением Excel: учебное пособие для вузов. Изд. 2-е, испр. и доп. – Ростов н/Д: Феникс, 2002. – 400 с.: ил.

5. Калинина В.Н., Панкин В.Ф. Математическая статистика: учеб. для студ. сред. спец. учеб. заведений. – 3-е изд., испр. – М.: Высш. шк., 2001. – 336 с.: ил.

6. Кремер Н.Ш. Теория вероятностей и математическая статистика. – М.: ЮНИТИ, 2001.

СОДЕРЖАНИЕ

Учебное издание

Сазонова Алла Михайловна

ТЕОРИЯ ВЕРОЯТНОСТИ И МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

Модуль 3. АНАЛИЗ ВАРИАЦИОННЫХ РЯДОВ

Методические рекомендации

Технический редактор *А.Н. Гладун*
Компьютерная верстка *А.Л. Позняков*

Подписано в печать 7.09. 2010 г.
Формат 60x84/16. Гарнитура Times New Roman Cyr.
Усл.-печ. л. 4,2. Уч.-изд. л. 4,0. Тираж 70 экз. Заказ № 368.

Учреждение образования "Могилевский государственный университет им. А.А. Кулешова", 212022, Могилев, Космонавтов, 1
ЛИ № 02330/278 от 30.04.2004 г.

Отпечатано на ризографе отдела оперативной полиграфии МГУ им. А.А. Кулешова. 212022, Могилев, Космонавтов, 1