Федеральное государственное образовательное бюджетное учреждение высшего образования
«ФИНАНСОВЫЙ УНИВЕРСИТЕТ ПРИ ПРАВИТЕЛЬСТВЕ РОССИЙСКОЙ ФЕДЕРАЦИИ»
(Финансовый университет)

Департамент математики

Газарян В.А. Компанцева Е.И.

СБОРНИК ЗАДАНИЙ ПО СТАТИСТИЧЕСКОМУ АНАЛИЗУ ДАННЫХ. ЧАСТЬ 1

Учебное пособие

для студентов, обучающихся по направлению подготовки 38.03.01. "Экономика"
(программа подготовки бакалавров)

Москва 2021

Федеральное государственное образовательное бюджетное учреждение высшего образования
«ФИНАНСОВЫЙ УНИВЕРСИТЕТ ПРИ ПРАВИТЕЛЬСТВЕ РОССИЙСКОЙ ФЕДЕРАЦИИ»
(Финансовый университет)

Департамент математики

Газарян В.А. Компанцева Е.И.

СБОРНИК ЗАДАНИЙ ПО СТАТИСТИЧЕСКОМУ АНАЛИЗУ ДАННЫХ. ЧАСТЬ 1

Учебное пособие

для студентов, обучающихся по направлению подготовки 38.03.01. "Экономика"
(программа подготовки бакалавров)

Одобрено на заседании Совета департамента математики
(протокол № 16 от 26.05.2021)

Москва 2021

ББК 22.17я73

Газарян В. А. Компанцева Е.И.
Сборник заданий по статистическому анализу данных. Часть 1

Сборник заданий представляет собой краткое изложение таких разделов анализа данных, как двумерные дискретные случайные векторы, эмпирические характеристики признака, точечные и интервальные оценки параметров распределений. В каждом разделе рассмотрены примеры решения задач анализа данных, а также предлагается большое количество вариантов заданий для самостоятельной работы студентов с использованием Excel. Сборник заданий предназначен для студентов второго курса Финансового университета, изучающих курс «Анализ данных».

Авторы:
Газарян В.А. – к.ф.-м.н., доцент департамента математики Финансового университета при Правительстве РФ, научный сотрудник кафедры математического моделирования и информатики ФФ МГУ имени М.В. Ломоносова.
Компанцева Е.И. – д.ф.-м.н., профессор департамента математики Финансового университета при Правительстве РФ, профессор МПГУ.

Рецензент:
Плохотников К.Э. - д.ф.-м.н., профессор департамента математики Финансового университета при Правительстве РФ, в.н.с. физического факультета МГУ имени М.В.Ломоносова

УДК 519.2(072)

ББК 22.17я73

Содержание

Тема 1. Двумерные дискретные случайные векторы. Условные числовые характеристики. Формулы полного математического ожидания и полной дисперсии

Закон распределения двумерного дискретного случайного вектора (*X*,*Y*) описывает совместное распределение дискретных случайных величин *X* и *Y*, представленное в виде таблицы сопряженности:

	$Y=y_1$	$Y=y_2$	$\ldots$	$Y=y_n$
$X=x_1$	p_{11}	p_{12}		p_{1n}
$X=x_2$	p_{21}			p_{2n}
$\ldots$				
$X=x_m$	p_{m1}	p_{m2}		p_{mn}

Здесь $p_{ij}=P(X=x_i,Y=y_j),\, i=1,\ldots,m,\; j=1,\ldots,n$. $\sum_{i=1}^{m}\sum_{j=1}^{n}p_{ij}=1$.

Если задан закон распределения двумерного дискретного случайного вектора (*X*,*Y*), можно найти законы распределения его компонент - дискретных случайных величин *X* и *Y*: случайная величина *X* принимает значения $x_1,x_2,\ldots,x_m$ с вероятностями $P(X=x_i)=\sum_{j=1}^{n}p_{ij}$, случайная величина *Y* принимает значения $y_1,y_2,\ldots,y_n$ с вероятностями $P(Y=y_j)=\sum_{i=1}^{m}p_{ij}$.

Определим условное математическое ожидание случайной величины *X* при условии, что случайная величина *Y* принимает значение y_j, как $E(X\,|\,Y=y_j)=\sum_{i=1}^{m}x_iP(X=x_i\,|\,Y=y_j)$. Аналогично $E(Y\,|\,X=x_i)=\sum_{j=1}^{n}y_jP(Y=y_j\,|\,X=x_i)$.

Здесь $P(X=x_i\,|\,Y=y_j)$ и $P(Y=y_j\,|\,X=x_i)$ вычисляются по формулам условной вероятности: $P(X=x_i\,|\,Y=y_j)=\dfrac{P(X=x_i,Y=y_j)}{P(Y=y_j)}$, $P(Y=y_j\,|\,X=x_i)=\dfrac{P(Y=y_j,X=x_i)}{P(X=x_i)}$.

Условное математическое ожидание случайной величины X относительно случайной величины Y является случайной величиной $E(X\,|\,Y)$, заданной законом распределения:

$E(X\,\|\,Y)$	$E(X\,\|\,Y=y_1)$	$E(X\,\|\,Y=y_2)$	$\dots$	$E(X\,\|\,Y=y_n)$
P	$P(Y=y_1)$	$P(Y=y_2)$	$\dots$	$P(Y=y_n)$

Тогда по данной таблице распределения случайной величины $E(X\,|\,Y)$ можно найти её числовые характеристики – математическое ожидание и дисперсию: $E(E(X\,|\,Y))=\sum_{j=1}^{n}E(X\,|\,Y=y_j)P(Y=y_j)$, $Var(E(X\,|\,Y))=Var(Z)=E(Z^2)-E^2(Z)$, где $Z=E(X\,|\,Y)$.

Согласно формуле полного математического ожидания (ПМ) $E(E(X\,|\,Y))=E(X)$, $E(E(Y\,|\,X))=E(Y)$.

Условная дисперсия случайной величины X относительно случайной величины Y также является случайной величиной $Var(X\,|\,Y)$, заданной законом распределения

$Var(X\,\|\,Y)$	$Var(X\,\|\,Y=y_1)$	$Var(X\,\|\,Y=y_2)$	$\dots$	$Var(X\,\|\,Y=y_n)$
P	$P(Y=y_1)$	$P(Y=y_2)$	$\dots$	$P(Y=y_n)$

Здесь $Var(X\,|\,Y=y_j)=E(X^2\,|\,Y=y_j)-E^2(X\,|\,Y=y_j),\ j=1,\dots,n$, где $E(X^2\,|\,Y=y_j)=\sum_{i=1}^{m}x_i^2P(X=x_i\,|\,Y=y_j)$.

Тогда $E(Var(X\,|\,Y))=\sum_{j=1}^{n}Var(X\,|\,Y=y_j)P(Y=y_j)$.

Согласно формуле полной дисперсии (ПД) $Var(X)=E(Var(X\,|\,Y))+Var(E(X\,|\,Y))$, $Var(Y)=E(Var(Y\,|\,X))+Var(E(Y\,|\,X))$.

Пример

Задано распределение двумерного дискретного случайного вектора *(X,Y)*

	$Y=2$	$Y=4$	$Y=9$
$X=60$	0,3	0,1	0

$X=90$	0,1	0,2	0,3

1. Найти коэффициент корреляции случайных величин *X* и *Y*.
2. Проверить формулы ПМ и ПД для случайной величины *X*.

Решение

1. Коэффициент корреляции равен $r(X,Y)=\dfrac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$. Чтобы его вычислить, найдём числовые характеристики компонент *X* и *Y:*

$E(X)=60(0,3+0,1)+90(0,1+0,2+0,3)=78.$

$E(Y)=2(0,3+0,1)+4(0,1+0,2)+9(0+0,3)=4,7.$

$E(X^2)=3600(0,3+0,1)+8100(0,1+0,2+0,3)=6300.$

$E(Y^2)=4(0,3+0,1)+16(0,1+0,2)+81(0+0,3)=30,7.$

$Var(X)=E(X^2)-E^2(X)=6300-78^2=216.$

$Var(Y)=E(Y^2)-E^2(Y)=30,7-4,7^2=8,61.$

$E(XY)=60\cdot 2\cdot 0,3+60\cdot 4\cdot 0,1+60\cdot 9\cdot 0+90\cdot 2\cdot 0,1+90\cdot 4\cdot 0,2+90\cdot 9\cdot 0,3=393.$

$Cov(X,Y)=E(XY)-E(X)E(Y)=363-78\cdot 4,7=-3,6.$

$$r(X,Y)=\frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}=\frac{-3,6}{\sqrt{216\cdot 8,61}}\approx -0,08348.$$

2. Согласно формуле полного математического ожидания (ПМ) для случайной величины *X* получаем $E(E(X\,|\,Y))=E(X)$. Чтобы проверить формулу ПМ, вычислим $E(E(X\,|\,Y))$ и сравним с найденным в пункте 1 значением $E(X)$. Для этого построим вначале распределение случайной величины $E(X\,|\,Y)$:

$E(X\,|\,Y)$	$E(X\,|\,Y=2)$	$E(X\,|\,Y=4)$	$E(X\,|\,Y=9)$
P	$P(Y=2)$	$P(Y=4)$	$P(Y=9)$

$$E(X\,|\,Y=2)=60P(X=60\,|\,Y=2)+90P(X=90\,|\,Y=2)=60\frac{P(X=60,Y=2)}{P(Y=2)}+90\frac{P(X=90,Y=2)}{P(Y=2)}=$$

$$=60\frac{0,3}{0,3+0,1}+90\frac{0,1}{0,3+0,1}=67,5.$$

$$E(X\mid Y=4)=60P(X=60\mid Y=4)+90P(X=90\mid Y=4)=60\frac{P(X=60,Y=4)}{P(Y=4)}+90\frac{P(X=90,Y=4)}{P(Y=4)}=$$

$$=60\frac{0,1}{0,1+0,2}+90\frac{0,2}{0,1+0,2}=80.$$

$$E(X\mid Y=9)=60P(X=60\mid Y=9)+90P(X=90\mid Y=9)=60\frac{P(X=60,Y=9)}{P(Y=9)}+90\frac{P(X=90,Y=9)}{P(Y=9)}=$$

$$=60\frac{0}{0+0,3}+90\frac{0,3}{0+0,3}=90.$$

$P(Y=2)=0,3+0,1=0,4,\ P(Y=4)=0,1+0,2=0,3,\ \ P(Y=9)=0+0,3=0,3$.

Заполним таблицу:

$E(X\mid Y)$	67,5	80	90
P	0,4	0,3	0,3

Вычислим математическое ожидание случайной величины $Z=E(X\mid Y)$:

$E(E(X\mid Y))=E(Z)=67,5\cdot 0,4+80\cdot 0,3+90\cdot 0,3=78.$ Полученное значение совпадает с найденным в пункте 1 значением $E(X)$.

Согласно формуле полной дисперсии (ПД) $Var(X)=E(Var(X\mid Y))+Var(E(X\mid Y))$.

Вычислим вначале дисперсию случайной величины $Z=E(X\mid Y)$:

$$Var(E(X\mid Y))=Var(Z)=E(Z^2)-E^2(Z)=67,5^2\cdot 0,4+80^2\cdot 0,3+90^2\cdot 0,3-78^2=88,5.$$

Построим распределение случайной величины $Var(X\mid Y)$:

$Var(X\mid Y)$	$Var(X\mid Y=2)$	$Var(X\mid Y=4)$	$Var(X\mid Y=9)$
P	$P(Y=2)$	$P(Y=4)$	$P(Y=9)$

$$Var(X\mid Y=2)=E(X^2\mid Y=2)-E^2(X\mid Y=2)=$$

$$=60^2P(X=60\mid Y=2)+90^2P(X=90\mid Y=2)-67,5^2=3600\frac{P(X=60,Y=2)}{P(Y=2)}+8100\frac{P(X=90,Y=2)}{P(Y=2)}-$$

$$-67,5^2=3600\frac{0,3}{0,3+0,1}+8100\frac{0,1}{0,3+0,1}-67,5^2=168,75.$$

$$Var(X\mid Y=4)=E(X^2\mid Y=4)-E^2(X\mid Y=4)=$$

$$=60^2P(X=60\mid Y=4)+90^2P(X=90\mid Y=4)-80^2=3600\frac{P(X=60,Y=4)}{P(Y=4)}+8100\frac{P(X=90,Y=4)}{P(Y=4)}-$$

$$-80^2=3600\frac{0,1}{0,1+0,2}+8100\frac{0,2}{0,1+0,2}-80^2=200.$$

$$Var(X\mid Y=9)=E(X^2\mid Y=9)-E^2(X\mid Y=9)=$$

$$=60^2P(X=60\mid Y=9)+90^2P(X=90\mid Y=9)-90^2=3600\frac{P(X=60,Y=9)}{P(Y=9)}+8100\frac{P(X=90,Y=9)}{P(Y=9)}-$$

$$-90^2 = 3600\frac{0}{0+0,3} + 8100\frac{0,3}{0+0,3} - 90^2 = 0.$$

Заполним таблицу:

$Var(X \mid Y)$	$168,75$	200	0
P	0,4	0,3	0,3

Вычислим математическое ожидание случайной величины $Var(X \mid Y)$:

$$E(Var(X \mid Y)) = 168,75 \cdot 0,4 + 200 \cdot 0,3 + 0 \cdot 0,3 = 127,5.$$

Согласно формуле полной дисперсии (ПД)

$Var(X) = E(Var(X \mid Y)) + Var(E(X \mid Y)) = 127,5 + 88,5 = 216.$ Полученное значение совпадает с найденным в пункте 1 значением $Var(X)$.

Задания для самостоятельного решения

Задание 1.

Задано распределение двумерного дискретного случайного вектора (X,Y).

1. Найти $E(X)$, $E(Y)$, $E(XY)$, $Var(X)$, $Var(Y)$, $Cov(X,Y)$, $r(X,Y)$.
2. Построить закон распределения случайной величины $E(X \mid Y)$.
3. Найти $E(E(X \mid Y))$, $Var(E(X \mid Y))$.
4. Построить закон распределения случайной величины $Var(X \mid Y)$.
5. Найти $E(Var(X \mid Y))$.
6. По формуле полной дисперсии найти $Var(X)$.
7. Найти заданное в варианте условное математическое ожидание.

Варианты

В-1

	$X=-1$	$X=0$	$X=2$
$Y=-2$	0,2	0,1	0
$Y=2$	0,1	?	0,4

$E(Y \mid X \ge 0) - ?$

В-2

	$X=-3$	$X=-1$	$X=0$
$Y=-2$	?	0,1	0,1
$Y=2$	0,1	0,2	0,3

$E(X \mid Y=-2)-?$

В-3

	$X=2$	$X=4$	$X=6$
$Y=-6$	0	0	?
$Y=-2$	0,1	0,2	0,4

$E(Y \mid X<6)-?$

В-4

	$X=-2$	$X=-1$	$X=0$
$Y=-1$	0,2	0,1	0,1
$Y=1$	0,1	?	0,3

$E(X \mid Y\neq -1)-?$

В-5

	$X=-2$	$X=0$	$X=6$
$Y=-1$	0,1	0,2	0
$Y=2$	?	0,2	0,4

$E(Y \mid X=-2)-?$

В-6

	$X=3$	$X=4$	$X=6$
$Y=-6$	0	0,1	?
$Y=-3$	0,2	0,2	0,2

$E(X \mid Y=-3)-?$

B-7

	$X=1$	$X=3$	$X=6$
$Y=-6$	0	0	?
$Y=-1$	0,2	0,2	0,4

$E(Y \mid X>2)-?$

B-8

	$X=-2$	$X=-1$	$X=1$
$Y=-2$	?	0,2	0
$Y=2$	0,1	0,1	0,3

$E(X \mid Y<0)-?$

B-9

	$X=-3$	$X=-2$	$X=5$
$Y=-2$	0,3	?	0,1
$Y=1$	0,1	0,1	0,3

$E(Y \mid X<0)-?$

B-10

	$X=-3$	$X=-1$	$X=0$
$Y=-1$	?	0,1	0,1
$Y=1$	0	0,1	0,5

$E(X \mid Y>0)-?$

B-11

	$X=2$	$X=1$	$X=3$
$Y=0$	0,3	0,2	0,1

$Y=1$	0,1	?	0,1

$E(Y \mid X \neq 1)-?$

В-12

	$X=-2$	$X=1$	$X=6$
$Y=-1$	0,2	0,2	0
$Y=1$	?	0,2	0,3

$E(X \mid Y=1)-?$

В-13

	$X=-3$	$X=-2$	$X=0$
$Y=-1$	0,1	?	0,1
$Y=1$	0	0,1	0,6

$E(Y \mid X \geq -2)-?$

В-14

	$X=-1$	$X=1$	$X=2$
$Y=-1$	0,3	0,2	0,1
$Y=2$	0	0,1	?

$E(X \mid Y<1)-?$

Задание 2.

Задано распределение двумерного дискретного случайного вектора (X,Y).

1. Найти $E(X)$, $E(Y)$, $E(XY)$, $Var(X)$, $Var(Y)$, $Cov(X,Y)$, $r(X,Y)$.
2. Построить закон распределения случайной величины $E(Y \mid X)$.
3. Найти $E(E(Y \mid X))$, $Var(E(Y \mid X))$.
4. Построить закон распределения случайной величины $Var(Y \mid X)$.
5. Найти $E(Var(Y \mid X))$.
6. По формуле полной дисперсии найти $Var(Y)$.

В-15

	$Y=-1$	$Y=0$	$Y=2$
$X=-2$	0,2	0,1	0
$X=2$	0,1	?	0,4

В-16

	$Y=-3$	$Y=-1$	$Y=0$
$X=-2$	?	0,1	0,1
$X=2$	0,1	0,2	0,3

В-17

	$Y=2$	$Y=4$	$Y=6$
$X=-6$	0	0	?
$X=-2$	0,1	0,2	0,4

В-18

	$Y=-2$	$Y=-1$	$Y=0$
$X=-1$	0,2	0,1	0,1
$X=1$	0,1	?	0,3

В-19

	$Y=-2$	$Y=0$	$Y=6$
$X=-1$	0,1	0,2	0
$X=2$	?	0,2	0,4

В-20

	$Y=3$	$Y=4$	$Y=6$
$X=-6$	0	0,1	?
$X=-3$	0,2	0,2	0,2

В-21

	$Y=1$	$Y=3$	$Y=6$
$X=-6$	0	0	?
$X=-1$	0,2	0,2	0,4

В-22

	$Y=-2$	$Y=-1$	$Y=1$
$X=-2$	?	0,2	0
$X=2$	0,1	0,1	0,3

В-23

	$Y=-3$	$Y=-2$	$Y=5$
$X=-2$	0,3	?	0,1
$X=1$	0,1	0,1	0,3

В-24

	$Y=-3$	$Y=-1$	$Y=0$
$X=-1$	?	0,1	0,1
$X=1$	0	0,1	0,5

В-25

	$Y=2$	$Y=1$	$Y=3$
$X=0$	0,3	0,2	0,1
$X=1$	0,1	?	0,1

В-26

	$Y=-2$	$Y=1$	$Y=6$

$X=-1$	0,2	0,2	0
$X=1$	?	0,2	0,3

В-27

	$Y=-3$	$Y=-2$	$Y=0$
$X=-1$	0,1	?	0,1
$X=1$	0	0,1	0,6

В-28

	$Y=-1$	$Y=1$	$Y=2$
$X=-1$	0,3	0,2	0,1
$X=2$	0	0,1	?

Тема 2. Эмпирические характеристики признака

В прикладной статистике вводится понятие признака, аналогичное понятию случайной величины в теории вероятностей. В то время как случайная величина ξ – действительная функция, заданная на пространстве элементарных событий Ω, признак ξ задаётся на статистической совокупности Ω. В статистике признаки принято обозначать либо греческими буквами $\xi, \eta, \ldots$, как и случайные величины, либо заглавными латинскими: $X, Y, \ldots$. В данном сборнике будут использоваться оба типа обозначений.

Пример. Пусть статистическая совокупность $\Omega = (\omega_1, \omega_2, \ldots, \omega_N)$ - множество жителей Москвы, на данной статистической совокупность заданы два признака ξ и η, где ξ- рост, а η - вес жителя. Тогда $x_i = \xi(\omega_i)$ - значение признака ξ объекта ω_i, то есть рост *i*-го жителя, $y_i = \eta(\omega_i)$ - его вес, $i = 1, \ldots, N$, N - объем статистической совокупности, то есть число жителей Москвы.

В то время как все элементы $\omega_1, \omega_2, ..., \omega_N$ статистической совокупности Ω различны, значения $x_1, x_2, ..., x_N$ признака ξ могут повторятся (например, одинаковый рост у разных жителей).

Проводя дальнейшую аналогию с теорией вероятностей, представим признак ξ как случайную величину, определённую на вероятностном пространстве $\Omega = (\omega_1, \omega_2, ..., \omega_N)$ с равновероятными элементарными исходами: $P(\omega_i) = \frac{1}{N}, i = 1, ..., N$. В нашем примере это означает, что выбор любого жителя Москвы для измерения роста или веса случаен и равновероятен по отношению к другим жителям. Тогда эмпирические характеристики признака совпадают с их теоретико-вероятностными аналогами:

- **эмпирическое среднее**: $\bar{\xi} = \frac{1}{N}\sum_{i=1}^{N} x_i = \sum_{i=1}^{N} x_i \hat{p}_i,\ \hat{p}_i = \frac{1}{N},\ i = 1, ..., N$; (1)

- **эмпирическая дисперсия**: $\hat{D}\xi = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{\xi})^2 = \overline{\xi^2} - (\bar{\xi})^2 = \sum_{i=1}^{N}(x_i - \bar{\xi})^2 \hat{p}_i$; (2)

- **эмпирическое стандартное отклонение**: $\hat{\sigma} = \sqrt{\hat{D}\xi}$; (3)

- **эмпирическая функция распределения**: $\hat{F}(x) = \sum_{i: x_i < x} \hat{p}_i$; (4)

- **эмпирические начальный $\hat{v}_k$ и центральный $\hat{\mu}_k$ моменты *k*-го порядка**:

$$\hat{v}_k = \frac{1}{N}\sum_{i=1}^{N} x_i^k = \sum_{i=1}^{N} x_i^k \hat{p}_i, \qquad \hat{\mu}_k = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{\xi})^k = \sum_{i=1}^{N}(x_i - \bar{\xi})^k \hat{p}_i \ ;$$

- **эмпирический коэффициент асимметрии**: $As\xi = \frac{\hat{\mu}_3}{\hat{\sigma}^3}$; (5)

- **эмпирический коэффициент эксцесса** (островершинности распределения): $Ex\xi = \frac{\hat{\mu}_4}{\hat{\sigma}^4} - 3$; (6)

- **эмпирическая мода** *Mo* – наиболее часто встречающееся значение признака;

- **эмпирическая медиана** *Me* – значение признака, лежащее в середине вариационного (ранжированного) ряда и делящее его на две равные части, то есть эмпирический квантиль порядка 0,5: $\hat{F}(Me) = \sum_{i: x_i < Me} \hat{p}_i = 0,5$.

На практике эмпирическая медиана вычисляется следующим образом: строится **вариационный (ранжированный) ряд** $x_{(1)}, x_{(2)}, ..., x_{(N)}$ значений признака ξ, где $x_{(i)}, i = 1,..., N$, - значение признака, имеющее *i*-й порядковый номер в упорядоченном по неубыванию ряду значений признака, причём одинаковые значения признака повторяются в вариационном ряду столько же раз, сколько они наблюдаются в исходном (неупорядоченном) ряду $x_1, x_2, ..., x_N$. Тогда $Me = \begin{cases} x_{(k+1)}, N = 2k+1, \\ \frac{1}{2}(x_{(k)} + x_{(k+1)}), N = 2k. \end{cases}$ Здесь $x_{(k)}$ - значение признака, имеющее *k*-й порядковый номер в ранжированном ряду;

- **эмпирические квартили** Q_1, Q_2, Q_3, делящие вариационный ряд на четыре равные части: $\hat{F}(Q_1) = \sum_{i: x_i < Q_1} \hat{p}_i = 0,25$, $\hat{F}(Q_2) = \hat{F}(Me) = 0,5$, $\hat{F}(Q_3) = \sum_{i: x_i < Q_3} \hat{p}_i = 0,75$.

В курсе математической статистики была построена несмещённая оценка дисперсии случайной величины ξ, имеющей нормальное распределение, и принимающей значения $x_1, x_2, ..., x_N$, которая имеет вид

$$s^2 = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \bar{\xi})^2 = \frac{N}{N-1} \hat{D}\xi \ (2). \qquad (7)$$

Кроме дисперсии, которая характеризует степень изменчивости признака, в прикладной статистике часто используется относительный показатель изменчивости статистической совокупности и типичности средней величины – **эмпирический коэффициент вариации:** $\hat{V}_\sigma = \frac{\hat{\sigma}}{\bar{\xi}} \cdot 100\%$. Совокупность Ω считается однородной, если коэффициент вариации не превышает 33%.

Пример. Признак ξ, заданный на статистической совокупности Ω объёма 20, принимает следующие значения:

11, 10, 12, 17, 17, 19, 11, 15, 5, 11, 11, 14, 6, 7, 14, 11, 13, 13, 8, 5.

Для данной выборки найти в Excel:

эмпирические характеристики: среднее, моду, медиану, дисперсию, стандартное отклонение, размах вариации, квартили, коэффициент асимметрии, коэффициент эксцесса,

несмещённые оценки дисперсии и стандартного отклонения;

вариационный ряд.

Решение

Скопируем данные из условия задачи в ячейку A1, разобьём значения признака по столбцам, выполнив: Данные/Текст по столбцам/С разделителями/Запятая/Общий (формат данных), затем транспонируем данные, представленные строкой, в столбец:

11
10
12
17
17
19
11
15
5
11
11
14
6
7
14
11
13
13

8

5

Большинство эмпирических характеристик удобно получить, воспользовавшись надстройкой Анализ данных: выделим массив данных, выполним: Данные/Анализ данных/Описательная статистика/Итоговая статистика, укажем входной интервал (массив данных), выходной интервал – любую свободную ячейку для ввода результатов анализа, получим таблицу с результатами:

Среднее	11,5
Стандартная ошибка	0,884069538
Медиана	11
Мода	11
Стандартное отклонение	3,953679166
Дисперсия выборки	15,63157895
Эксцесс	-0,499038047
Асимметричность	-0,017032254
Интервал	14
Минимум	5
Максимум	19
Сумма	230
Счет	20

Таким образом, получим эмпирические характеристики:

среднее: 11,5

мода: 11

медиана: 11

дисперсия: ДИСП.Г(массив) = 14,85

стандартное отклонение: СТАНДОТКЛОН.Г(массив) = 3,854

размах вариации: $\hat{V}_\sigma = \frac{\hat{\sigma}}{\overline{\xi}} \cdot 100\% = 33{,}5\%$

первая квартиль: КВАРТИЛЬ.ВКЛ(массив;1) = 9,5

вторая квартиль: КВАРТИЛЬ.ВКЛ(массив;2) =11

третья квартиль: КВАРТИЛЬ.ВКЛ(массив;3) = 14
коэффициент асимметрии: -0,017
коэффициент эксцесса: -0,499
несмещённая оценка дисперсии: «дисперсия выборки» 15,63157895
несмещённая оценка стандартного отклонения: 3,953679166.

Чтобы построить вариационный ряд, следует вначале скопировать массив данных в другой столбец, затем, выделив в нём массив, выполнить: Главная/Сортировка и фильтр/Сортировка по возрастанию и получить следующий вариационный ряд:

5
5
6
7
8
10
11
11
11
11
11
12
13
13
14
14
15
17
17
19

Задания для самостоятельного решения

Для заданных в варианте значений признака X найти:

- вариационный ряд, частотное распределение;
- эмпирические характеристики: среднее, дисперсию, ст. отклонение, коэффициент вариации, моду, медиану, квартили, используя для расчёта:

  а) математические формулы,

  б) функции Excel
- эмпирическую функцию распределения, её график.

Варианты

В.0. 10; -10; 2; -6; -3; -10; -6; -6; 10; -6

В.1. -6; 8; 1; -9; -2; -2; -6; -2; -9; 8

В.2. -5; -2; -8; -10; 4; 4; 4; -5; 4

В.3. 1; 9; 4; -3; -9; -3; -9; -9

В.4. -10; -11; -12; 7; 6; -11; -11

В.5. -2; 5; 8; 12; 1; 12

В.6. -1; -2; 0; -6; -10; 0; -2; -6; -10; -2

В.7. 6; -12; 8; 11; 7; 8; 8; 8; 8

B.8. 7; 8; -6; -5; -11; -5; 7; -6

B.9. -10; 9; 1; 3; -12; 3; -12

B.10. 12; 5; 2; -12; 9; 5

B.11. -9; -4; -3; 7; 6; -3; 6; 6; 6; 7

B.12. -11; -8; -4; -5; 12; -5; -11; -5; -5

B.13. 1; -3; -4; 2; 5; 5; 1; 2

B.14. -8; 5; 10; 8; 1; 10; 8

B.15. -6; 8; 1; -9; -2; -2; -6; -2; -9; 8

B.16. -5; -2; -8; -10; 4; 4; 4; -5; 4

B.17. 1; 9; 4; -3; -9; -3; -9; -9

B.18. -10; -11; -12; 7; 6; -11; -11

B.19. -2; 5; 8; 12; 1; 12

B.20. -1; -2; 0; -6; -10; 0; -2; -6; -10; -2

B.21. 6; -12; 8; 11; 7; 8; 8; 8; 8

B.22. 7; 8; -6; -5; -11; -5; 7; -6

B.23. -10; 9; 1; 3; -12; 3; -12

B.24. 12; 5; 2; -12; 9; 5

B.25. -9; -4; -3; 7; 6; -3; 6; 6; 6; 7

B.26. -11; -8; -4; -5; 12; -5; -11; -5; -5

B.27. 1; -3; -4; 2; 5; 5; 1; 2

B.28. -8; 5; 10; 8; 1; 10; 8

B.29. -6; 8; 1; -9; -2; -2; -6; -2; -9; 8

B.30. -5; -2; -8; -10; 4; 4; 4; -5; 4

Тема 3. Эмпирическая ковариация и эмпирический коэффициент корреляции

В курсе теории вероятностей мы познакомились с понятиями ковариации (корреляционный момент) и коэффициента корреляции, отражающими линейную зависимость изучаемых признаков. В прикладной статистике вводятся аналогичные эмпирические характеристики.

Если на статистической совокупности Ω объёма N определены два признака ξ и η, $\Omega = (\omega_1, \omega_2, ..., \omega_N)$, $x_i = \xi(\omega_i)$ - значение признака ξ у объекта ω_i, $y_i = \eta(\omega_i)$ - значение признака η у объекта ω_i, $i = 1, ..., N$, то **эмпирическая ковариация** представима в виде

$$\hat{\mathrm{cov}}(\xi,\eta) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{\xi})(y_i - \overline{\eta}) = \overline{\xi\eta} - \overline{\xi}\cdot\overline{\eta} \qquad (11)$$

и характеризует тесноту линейной связи признаков ξ и η, а также их рассеяние около средних значений. Чтобы учесть только линейную зависимость ξ и η, вводится **эмпирический коэффициент корреляции Пирсона**

$$\hat{r} = \frac{\hat{\mathrm{cov}}(\xi,\eta)}{\sqrt{\hat{D}\xi \hat{D}\eta}}, \ |\hat{r}| \le 1, \qquad (12)$$

где $\hat{D}\xi$, $\hat{D}\eta$ - эмпирические дисперсии признаков ξ и η соответственно (2). Если $\hat{r} = 0$, линейной связи между ξ и η нет. Если $\hat{r} = \pm 1$, линейная связь между ξ и η носит не случайный, а функциональный характер: $\eta = a\xi + b$. При положительном значении $\hat{r}$ связь называется прямой, и с возрастанием значений признака ξ значения η в среднем также возрастают. При отрицательном значении $\hat{r}$ связь называется обратной, тогда с возрастанием значений признака ξ значения η в среднем убывают. В прикладной статистике принято использовать следующую таблицу тесноты линейной связи:

- $|r| \le 0,3$ - связь практически отсутствует;

- $0,3 \le |r| < 0,5$ - связь слабая;
- $0,5 \le |r| < 0,7$ - связь умеренная;
- $0,7 \le |r| < 1$ - связь сильная.

В курсе теории вероятностей при изучении двумерных дискретных случайных величин мы задавали распределение двумерного дискретного случайного вектора в виде таблицы сопряжённости. Совместное частотное распределение признаков ξ и η в ряде задач прикладной статистики имеет аналогичное представление (табл. 6). Таблицы сопряжённости принято использовать в тех случаях, когда некоторые пары значений признаков (x_i, y_i) встречаются неоднократно, то есть у нескольких объектов Ω.

Таблица 6

	$\eta = y_1$	$\eta = y_2$	$\dots$	$\eta = y_s$
$\xi = x_1$	n_{11}	n_{12}	$\dots$	n_{1s}
$\xi = x_2$	n_{21}	n_{22}	$\dots$	n_{2s}
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
$\xi = x_r$	n_{r1}	n_{r2}	$\dots$	n_{rs}

В таблице 6 частота n_{ij} - это число элементов ω статистической совокупности Ω, для которых $\xi(\omega) = x_i$, $\eta(\omega) = y_j$, $i = 1,\dots,r$, $j = 1,\dots,\mathrm{s}$, $\sum_{i=1}^{r}\sum_{j=1}^{s} n_{ij} = N$. Если исходные данные представлены в задаче в виде таблицы сопряжённости, эмпирическую ковариацию удобно вычислять по следующей формуле:

$$\hat{\mathrm{cov}}(\xi,\eta) = \frac{1}{N}\sum_{i=1}^{r}\sum_{j=1}^{s}(x_i - \bar{\xi})(y_i - \bar{\eta})n_{ij} = \overline{\xi\eta} - \bar{\xi}\bar{\eta}\,. \tag{13}$$

Пример. Для группы, состоящей из 16 студентов, определены два признака: ξ - оценка по математике, η - оценка по иностранному языку (табл. 7). Найти эмпирический коэффициент корреляции.

Таблица 7

№ студента в списке		
	оценка по матем. ξ	оценка по ин. яз. η
1	2	3
2	4	3
3	2	4
4	3	4
5	2	4
6	3	4
7	4	4
8	3	4
9	4	4
10	3	4
11	5	4
12	4	4
13	5	4
14	4	4
15	3	5
16	5	5

Решение. Применив формулы (11) и (12), получим следующее значение эмпирического коэффициента корреляции: $r = 0{,}25$. Такое же значение можно получить, заполнив таблицу сопряжённости (табл. 8) и применив для вычисления ковариации формулу (13).

Таблица 8. Совместное статистическое распределение оценок

оценка по ин. яз. η	оценка по математике ξ			
	2	3	4	5
3	1	0	1	0
4	2	4	4	2
5	0	1	0	1

В Excel эмпирическая ковариация $\hat{\text{cov}}(\xi,\eta)$ и эмпирический коэффициент корреляции $\hat{r}$ находятся значительно проще: $\hat{\text{cov}}(\xi,\eta) = КОВАРИАЦИЯ.Г(массив1; массив2) = 0{,}125$, $\hat{r} = КОРРЕЛ(массив1; массив2) = 0{,}25$. Полученное значение эмпирической корреляции позволяет сделать вывод о том, что линейной зависимости оценок по данным предметам не наблюдается.

Задания для самостоятельного решения

Признаки X и Y заданы на стат. совокупности объёма n.

1) Для признаков X и Y найти эмпирический коэффициент корреляции $r(X,Y)$ с помощью:

а) математических формул для расчёта ковариации и коэфф. корреляции, записать формулы и подстановку чисел;

б) функции Excel

2) Оценить тесноту и направление связи.

Варианты

В.1 $n = 9$

X: 1; 9; 4; -3; -9; -3; -9; -9; 1

Y: -5; -2; -8; -10; 4; 4; 4; -5; 4

В.2. $n = 8$

X: -10; -11; -12; 7; 6; -11; -11; 7

Y: -5; -2; -8; -10; 4; 4; 4; -5

В.3. $n = 7$

X: -2; 5; 8; 12; 1; 12; 1

Y: -5; -2; -8; -10; 4; 4; 4

В.4. $n = 6$

X: -1; -2; 0; -6; -10; 0

Y: -5; -2; -8; -10; 4; 4

B.5. $n = 10$

X: 6; -12; 8; 11; 7; 8; 8; 8; 8; 6

Y: -5; -2; -8; -10; 4; 4; 4; -5; 4; -5

B.6. $n = 9$

X: 7; 8; -6; -5; -11; -5; 7; -6; 8

Y: -5; -2; -8; -10; 4; 4; 4; -5; 4

B.7. $n = 8$

X: -10; 9; 1; 3; -12; 3; -12; -10

Y: -5; -2; -8; -10; 4; 4; 4; -5

B.8. $n = 7$

X: 12; 5; 2; -12; 9; 5; 5

Y: -5; -2; -8; -10; 4; 4; 4

B.9. $n = 6$

X: -9; -4; -3; 7; 6; -3

Y: -5; -2; -8; -10; 4; 4

B.10. $n = 10$

X: -11; -8; -4; -5; 12; -5; -11; -5; -5; -11

Y: -5; -2; -8; -10; 4; 4; 4; -5; 4; -5

B.11. $n = 9$

X: 1; -3; -4; 2; 5; 5; 1; 2; -3

Y: -5; -2; -8; -10; 4; 4; 4; -5; 4

B.12. $n = 8$

X: -8; 5; 10; 8; 1; 10; 8; 1

Y: -5; -2; -8; -10; 4; 4; 4; -5

B.13. $n = 7$

X: -10; -11; -12; 7; 6; -11; -11

Y: 1; 9; 4; -3; -9; -3; -9

B.14. $n = 6$

X: -2; 5; 8; 12; 1; 12

Y: 1; 9; 4; -3; -9; -3

B.15 $n = 9$

X: 1; 9; 4; -3; -9; -3; -9; -9; 1

Y: -5; -2; -8; -10; 4; 4; 4; -5; 4

B.16. $n = 8$

X: -10; -11; -12; 7; 6; -11; -11; 7

Y: -5; -2; -8; -10; 4; 4; 4; -5

B.17. $n = 7$

X: -2; 5; 8; 12; 1; 12; 1

Y: -5; -2; -8; -10; 4; 4; 4

B.18. $n = 6$

X: -1; -2; 0; -6; -10; 0

Y: -5; -2; -8; -10; 4; 4

B.19. $n = 10$

X: 6; -12; 8; 11; 7; 8; 8; 8; 8; 6

Y: -5; -2; -8; -10; 4; 4; 4; -5; 4; -5

B.20. $n = 9$

X: 7; 8; -6; -5; -11; -5; 7; -6; 8

Y: -5; -2; -8; -10; 4; 4; 4; -5; 4

B.21. $n = 8$

X: -10; 9; 1; 3; -12; 3; -12; -10

Y: -5; -2; -8; -10; 4; 4; 4; -5

B.22. $n = 7$

X: 12; 5; 2; -12; 9; 5; 5

Y: -5; -2; -8; -10; 4; 4; 4

B.23. $n = 6$

X: -9; -4; -3; 7; 6; -3

Y: -5; -2; -8; -10; 4; 4

B.24. $n = 10$

X: -11; -8; -4; -5; 12; -5; -11; -5; -5; -11

Y: -5; -2; -8; -10; 4; 4; 4; -5; 4; -5

B.25. $n = 9$

X: 1; -3; -4; 2; 5; 5; 1; 2; -3

Y: -5; -2; -8; -10; 4; 4; 4; -5; 4

B.26. $n = 8$

X: -8; 5; 10; 8; 1; 10; 8; 1

Y: -5; -2; -8; -10; 4; 4; 4; -5

B.27. $n = 7$

X: -10; -11; -12; 7; 6; -11; -11

Y: 1;   9;   4;   -3; -9; -3;   -9

В.28. $n = 6$

X: -2;   5; 8;   12;   1; 12

Y: 1;   9; 4;   -3; -9; -3

Тема 4. Интервальные оценки параметров нормального распределения

Пусть некоторый признак *X*, заданный на генеральной совокупности Ω, подчиняется определённому закону распределения. Рассмотрим *n* выборочных значений $X_1, X_2, ..., X_n$ из генеральной совокупности Ω и назовём их **выборкой из распределения**. Например, можно считать, что рост жителей большого города подчиняется нормальному закону распределения, тогда $X_1, X_2, ..., X_n$ - рост *n* выбранных наугад жителей данного города.

Пусть $X_1, X_2, ..., X_n$ - выборка объёма *n* из некоторого распределения, зависящего от параметра θ, $F(x,\theta)$ - функция данного распределения. Например, если имеется выборка из нормального распределения $N(\mu,\sigma^2)$, то в качестве параметра θ можно рассматривать μ или σ^2. Задача интервального оценивания заключается в том, чтобы найти так называемый **доверительный интервал** (θ_1, θ_2), в который попадает значение оцениваемого параметра θ с заданной вероятностью $\gamma = 1-\alpha$, которая называется **уровнем доверия** или **надёжностью** оценки: $P(\theta_1 < \theta < \theta_2) = 1-\alpha$.

1) Интервальная оценка математического ожидания при известной дисперсии

Пусть $X_1, X_2, ..., X_n$ - <u>повторная</u> выборка объёма *n* из нормального распределения $N(\mu,\sigma^2)$. При повторном отборе случайные величины

$X_1, X_2, ..., X_n$ являются независимыми. Предположим, что генеральная дисперсия $Var(X)=\sigma^2$ известна, а генеральное среднее (математическое ожидание) $E(X)=\mu$ неизвестно и необходимо построить интервальную оценку параметра μ, то есть для заданного $\gamma=1-\alpha$ найти доверительные границы (θ_1, θ_2): $P(\theta_1<\mu<\theta_2)=1-\alpha$. Рассмотрим выборочное среднее $\bar{X}=\frac{1}{n}\sum_{i=1}^{n}X_i$. Поскольку $X_i \sim N(\mu,\sigma^2)$ и независимы, $\bar{X}\sim N(\mu,\frac{\sigma^2}{n})$. Тогда $\frac{\bar{X}-\mu}{\sqrt{\frac{\sigma^2}{n}}}=\frac{(\bar{X}-\mu)\sqrt{n}}{\sigma}\sim N(0,1)$ и для заданного $\gamma=1-\alpha$ можно найти $z_{1-\frac{\alpha}{2}}$ - **квантиль** стандартного нормального распределения порядка $1-\frac{\alpha}{2}$. Напомним, что **квантиль порядка** $1-\frac{\alpha}{2}$ - такое значение случайной величины, для которого выполняется равенство $F(z_{1-\frac{\alpha}{2}})=P(X\le z_{1-\frac{\alpha}{2}})=1-\frac{\alpha}{2}$. $z_{1-\frac{\alpha}{2}}$ называется также **процентной точкой** стандартного нормального распределения **порядка** $\frac{\alpha}{2}$ или $\frac{\alpha}{2}100\%$-процентной точкой, поскольку $P(X>z_{1-\frac{\alpha}{2}})=\frac{\alpha}{2}$. В данном случае

$$F(z_{1-\frac{\alpha}{2}})=\Phi(z_{1-\frac{\alpha}{2}})=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{z_{1-\frac{\alpha}{2}}}e^{-\frac{x^2}{2}}dx=1-\frac{\alpha}{2},$$

где $\Phi(z_{1-\frac{\alpha}{2}})$ - функция стандартного нормального распределения. Таким образом $P(-z_{1-\frac{\alpha}{2}}<\frac{(\bar{X}-\mu)\sqrt{n}}{\sigma}<z_{1-\frac{\alpha}{2}})=1-\alpha$. Решая неравенство относительно оцениваемого параметра μ, получаем $P(\bar{X}-\frac{z_{1-\frac{\alpha}{2}}\sigma}{\sqrt{n}}<\mu<\bar{X}+\frac{z_{1-\frac{\alpha}{2}}\sigma}{\sqrt{n}})=1-\alpha$.

Следовательно, доверительные границы равны

$$\theta_1 = \bar{X} - \frac{z_{1-\frac{\alpha}{2}}\sigma}{\sqrt{n}},\ \theta_2 = \bar{X} + \frac{z_{1-\frac{\alpha}{2}}\sigma}{\sqrt{n}},$$

Таким образом, $P(\bar{X} - \Delta x < \mu < \bar{X} + \Delta x) = 1-\alpha$, где $\Delta x = \frac{z_{1-\frac{\alpha}{2}}\sigma}{\sqrt{n}}$

и называется ошибкой выборки, а $\frac{\sigma}{\sqrt{n}}$ - стандартной ошибкой выборки. Поскольку $z_{1-\frac{\alpha}{2}}$ - квантиль стандартного нормального распределения, то есть $\Phi(z_{1-\frac{\alpha}{2}}) = 1-\frac{\alpha}{2}$ (25а), то $z_{1-\frac{\alpha}{2}} = \Phi^{-1}(1-\frac{\alpha}{2})$ - обратная функция, значит $z_{1-\frac{\alpha}{2}}$ можно найти с помощью функции НОРМ.СТ.ОБР($1-\frac{\alpha}{2}$) в Excel.

Пример. Определяется средний рабочий стаж μ большой группы рабочих. Произведена случайная повторная выборка 900 личных листков. Средний рабочий стаж в выборке оказался равным 15,5 годам, а среднее квадратическое отклонение 4,8 года. Найти доверительные границы при оценке среднего стажа большой группы рабочих, которые можно гарантировать с вероятностью 0,95.

Решение. По условию $\bar{X} = 15{,}5,\ \sigma = 4{,}8,\ 1-\alpha = 0{,}95,\ n = 900.$ Найти (θ_1, θ_2).

$\theta_1 = \bar{X} - \frac{z_{1-\frac{\alpha}{2}}\sigma}{\sqrt{n}},\ \ \theta_2 = \bar{X} + \frac{z_{1-\frac{\alpha}{2}}\sigma}{\sqrt{n}}.$ $z_{1-\frac{\alpha}{2}}$=НОРМ.СТ.ОБР(1-0,05/2)=1,96. Интервальная оценка:

$$15{,}5 - 0{,}31 \le \mu \le 15{,}5 + 0{,}31,$$

$$15{,}19 \le \mu \le 15{,}81.$$

2) Интервальная оценка математического ожидания при неизвестной дисперсии

Пусть $X_1, X_2, ..., X_n$ - выборка независимых случайных величин, имеющих нормальное распределение $N(\mu, \sigma^2)$, причём дисперсия σ^2 считается неизвестной. Тогда при решении задачи интервального оценивания математического ожидания μ на уровне доверия $\gamma = 1-\alpha$ получаются следующие доверительные границы:

$$\theta_1 = \bar{X} - \frac{t_{1-\frac{\alpha}{2},n-1}\, s}{\sqrt{n}},\ \theta_2 = \bar{X} + \frac{t_{1-\frac{\alpha}{2},n-1}\, s}{\sqrt{n}},\ s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

где s^2 - несмещённая оценка дисперсии (7), $t_{1-\frac{\alpha}{2},n-1}$ - квантиль распределения Стьюдента с n-1 степенью свободы порядка $1-\frac{\alpha}{2}$. $F(t_{1-\frac{\alpha}{2},n-1}) = 1-\frac{\alpha}{2}$.

Если выборка $X_1, X_2, ..., X_n$ из нормального распределения $N(\mu,\sigma^2)$ является бесповторной, но $\frac{n}{N}100\% < 5\%$, то при решении задачи интервального оценивания математического ожидания μ при неизвестной дисперсии σ^2 на уровне доверия γ используют формулы для повторной выборки.

3) Интервальная оценка дисперсии при известном математическом ожидании

Пусть $X_1, X_2, ..., X_n$ - выборка объёма n независимых случайных величин, имеющих нормальное распределение $N(\mu,\sigma^2)$. Предположим, что генеральное среднее (математическое ожидание) $E(X) = \mu$ известно, а генеральная дисперсия $Var(X) = \sigma^2$ неизвестна и необходимо построить интервальную оценку параметра σ^2, то есть для заданного γ найти доверительные границы (θ_1, θ_2): $P(\theta_1 < \sigma^2 < \theta_2) = \gamma$. Можно показать, что

$$\theta_1 = \frac{\sum_{i=1}^{n}(X_i - \mu)^2}{\chi^2_{1-\frac{\alpha}{2},n}},\ \theta_2 = \frac{\sum_{i=1}^{n}(X_i - \mu)^2}{\chi^2_{\frac{\alpha}{2},n}},$$

где $\chi^2_{1-\frac{\alpha}{2},n}$ - квантиль распределения Пирсона с n степенями свободы порядка $1-\frac{\alpha}{2}$, то есть $F(\chi^2_{1-\frac{\alpha}{2},n})=1-\frac{\alpha}{2}$, $\chi^2_{\frac{\alpha}{2},n}$ - квантиль распределения Пирсона с n степенями свободы порядка $\frac{\alpha}{2}$, то есть $F(\chi^2_{\frac{\alpha}{2},n})=\frac{\alpha}{2}$.

4) **Интервальная оценка дисперсии при неизвестном математическом ожидании**

Пусть $X_1, X_2, ..., X_n$ - выборка объёма n независимых случайных величин, имеющих нормальное распределение $N(\mu,\sigma^2)$. Предположим, что генеральное среднее (математическое ожидание) $E(X)=\mu$ неизвестно и необходимо снова построить интервальную оценку параметра σ^2, то есть для заданного γ найти доверительные границы (θ_1,θ_2): $P(\theta_1<\sigma^2<\theta_2)=\gamma$. Можно показать, что

$$\theta_1=\frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{\chi^2_{1-\frac{\alpha}{2},n-1}},\ \theta_2=\frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{\chi^2_{\frac{\alpha}{2},n-1}},$$

где $\chi^2_{1-\frac{\alpha}{2},n-1}$ - квантиль распределения Пирсона с n-1 степенью свободы порядка $1-\frac{\alpha}{2}$, то есть $F(\chi^2_{1-\frac{\alpha}{2},n-1})=1-\frac{\alpha}{2}$, $\chi^2_{\frac{\alpha}{2},n-1}$ - квантиль распределения Пирсона с n-1 степенью свободы порядка $\frac{\alpha}{2}$, то есть $F(\chi^2_{\frac{\alpha}{2},n-1})=\frac{\alpha}{2}$.

Для работы в Excel удобно представлять доверительные границы в виде

$$\theta_1=\frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{\chi^2_{1-\frac{\alpha}{2},n-1}}=\frac{s^2(n-1)}{\chi^2_{1-\frac{\alpha}{2},n-1}},\ \theta_2=\frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{\chi^2_{\frac{\alpha}{2},n-1}}=\frac{s^2(n-1)}{\chi^2_{\frac{\alpha}{2},n-1}},$$

где s^2 - несмещённая оценка дисперсии (исправленная дисперсия).

5) **Обратные задачи**

1. Чему равна вероятность того, что генеральное среднее отличается от выборочного среднего по модулю не более, чем на заданное значение ε?

Следует найти такое значение $1-\alpha$, чтобы $P(\left|\mu-\bar{X}\right|\leq\varepsilon)=P(-\varepsilon\leq\mu-\bar{X}\leq\varepsilon)=P(\bar{X}-\varepsilon<\mu<\bar{X}+\varepsilon)=1-\alpha$, где $\varepsilon=\Delta x=\frac{z_{1-\frac{\alpha}{2}}\sigma}{\sqrt{n}}$,

следовательно, $z_{1-\frac{\alpha}{2}}=\frac{\Delta x\sqrt{n}}{\sigma}$, тогда $\Phi(z_{1-\frac{\alpha}{2}})$ =НОРМ.СТ.РАСП($z_{1-\frac{\alpha}{2}}$;1)=$1-\frac{\alpha}{2}$,

отсюда можно найти $1-\alpha$.

Пример. Определяется средний рабочий стаж μ большой группы рабочих. Произведена случайная повторная выборка 900 личных листков. Средний рабочий стаж в выборке оказался равным 15,5 годам, а среднее квадратическое отклонение 4,8 года. С какой вероятностью можно утверждать, что отклонение выборочной средней от генеральной не превысит 0,5 года.

Решение. По условию $\bar{X}=15{,}5,\ \sigma=4{,}8,\ \varepsilon=0{,}5,\ n=900.$. Найти $\gamma=1-\alpha$.

Вычисляем согласно (27а) $z_{1-\frac{\alpha}{2}}=\frac{\Delta x\sqrt{n}}{\sigma}=\frac{0{,}5\sqrt{900}}{4{,}8}\approx 3{,}13$. Тогда

$1-\frac{\alpha}{2}=НОРМ.СТ.РАСП(3{,}13;1)$, и $1-\alpha=2НОРМ.СТ.РАСП(3{,}13;1)-1\approx 0{,}9983.$

2. Какой объём выборки должен быть для того, чтобы генеральное среднее отличалось от выборочного среднего не более, чем на заданное значение Δx с заданной вероятностью $1-\alpha$?

Поскольку $\Delta x=\frac{z_{1-\frac{\alpha}{2}}\sigma}{\sqrt{n}}$, то $\sqrt{n}=\frac{z_{1-\frac{\alpha}{2}}\sigma}{\Delta x}$, $n=\left(\frac{z_{1-\frac{\alpha}{2}}\sigma}{\Delta x}\right)^2$.

Пример. В условиях предыдущего примера определить необходимый объем выборки, при котором ошибка не превысит 0,5 с доверительной вероятностью 0,999.

Решение. По заданной вероятности $1-\alpha$, равной 0,999, находим в Excel $z_{1-\frac{\alpha}{2}} = НОРМ.СТ.ОБР\left(1-0,001/2\right) \approx 3,29$ и $n = \left(\dfrac{z_{1-\frac{\alpha}{2}}\sigma}{\Delta x}\right)^2 = \dfrac{3,29^2 \cdot 4,8^2}{0,5^2} \approx 998.$

Пример. Сгенерировать в Excel выборку объёма 10 значений признака, имеющего в генеральной совокупности нормальное распределение с мат. ожиданием 4 и дисперсией 6.

1) Найти интервальную оценку генеральной дисперсии признака на уровне доверия 0,95

а) при неизвестном мат. ожидании

б) при известном мат. ожидании.

Решение. Смоделируем выборку объёма 10 в Excel: Данные/Анализ данных/Генерация случайных чисел. Получим следующие значения признака:

-

5,63 4,91 5,05 3,11 1,51 6,15 2,89 6,65 2,15 2,51

а) Если мат. ожидание неизвестно, находим квантили распределения Пирсона с 9 степенями свободы: $\chi^2_{\frac{\alpha}{2},n-1} = ХИ2.ОБР(0,05/2;9) = 2,7$, $\chi^2_{1-\frac{\alpha}{2},n-1} = ХИ2.ОБР(1-0,05/2;9) = 19$. Несмещённая оценка дисперсии $s^2 = ДИСП.В(массив) = 5,95$. Таким образом, левая граница доверительного интервала $\theta_1 = \dfrac{s^2(n-1)}{\chi^2_{1-\frac{\alpha}{2},n-1}} = 2,82$, правая граница доверительного интервала $\theta_2 = \dfrac{s^2(n-1)}{\chi^2_{\frac{\alpha}{2},n-1}} = 19,8$. Ответ: $(2,82;19,8)$.

б) При известном мат. ожидании, равном 4, находим квантили распределения Пирсона с 10 степенями свободы: $\chi^2_{\frac{\alpha}{2},n} = ХИ2.ОБР(0,05/2;10) = 3,25$, $\chi^2_{1-\frac{\alpha}{2},n} = ХИ2.ОБР(1-0,05/2;10) = 20,5$. Необходимо также найти $\sum_{i=1}^{n}(X_i-\mu)^2 = 54,2$.

Таким образом, левая граница доверительного интервала $\theta_1 = \dfrac{\sum_{i=1}^{n}(X_i-\mu)^2}{\chi^2_{1-\frac{\alpha}{2},n}} = 2,65$, правая граница доверительного интервала $\theta_2 = \dfrac{\sum_{i=1}^{n}(X_i-\mu)^2}{\chi^2_{\frac{\alpha}{2},n}} = 16,7$. Ответ: $(2,65;16,7)$.

Задания для самостоятельного решения

Лист

Данные:

B1	B2	B3	B4	B5	B6	B7	B8	B9	B10
11,287	13,261	7,71	5,182	4,819	9,6003	7,235	9,9193	10,09	11,295
13,406	14,155	9,622	14,99	8,6716	9,9485	12,31	5,9154	11,203	2,8536
8,9688	8,9209	12,24	12,21	8,4938	13,843	10,93	6,0202	13,443	6,7712
9,8897	8,1971	5,66	6,658	7,5855	12,801	6,356	8,7519	9,6449	5,0428
4,4967	10,084	15,58	13,6	13,664	8,0724	10,48	7,4668	12,184	12,606
12,758	13,674	11,38	3,618	8,7819	9,8601	12,24	6,2526	17,104	11,273
4,9954	9,4928	7,532	9,836	8,1176	9,6837	10,81	8,5606	7,6649	9,5403
11,882	11,723	8,742	5,316	12,439	9,7137	11,25	9,9365	4,7882	12,485
11,791	10,456	10,18	13,62	10,376	8,2447	10,1	8,3608	12,139	11,403
6,803	6,1261	8,024	9,855	14,485	4,8241	8,108	10,798	5,8727	5,7259
13,155	9,3894	7,144	9,226	8,3274	8,8147	6,754	13,818	10,69	15,065
9,3514	15,469	3,96	12,92	10,87	11,064	14,47	5,9304	10,803	8,631
10,214	10,005	13,09	13,52	9,7174	9,8752	7,597	13,201	8,6244	10,302
4,8088	7,4293	9,544	14,44	8,1886	13,163	10,93	8,213	9,8615	11,994
9,0543	15,159	9,999	12,49	8,3405	11,44	8,402	8,541	9,8217	10,213
14,82	5,7623	16,06	12,81	12,797	7,1912	11,39	7,6434	12,642	12,955
12,194	8,7834	15,57	10,83	14,263	9,2753	12,61	7,5037	7,8447	13,601
7,9375	9,0866	9,815	6,357	11,429	8,2838	12,23	8,0056	4,0672	9,3215
6,3168	6,2923	9,075	1,568	10,942	9,1768	9,07	9,9035	10,932	8,7321
13,057	15,109	7,024	7,601	12,051	9,7471	10,39	9,3017	10,944	12,13

B11	B12	B13	B14	B15	B16	B17	B18	B19	B20
12,165	10,659	13,259	11,287	13,26	7,71	5,18	4,819	9,600301	7,235
11,383	8,1137	13,688	13,406	14,16	9,62	15	8,672	9,948476	12,31
9,4744	3,3105	8,0466	8,9688	8,921	12,2	12,2	8,494	13,84345	10,93
16,427	13,637	14,211	9,8897	8,197	5,66	6,66	7,585	12,80136	6,356
9,6491	11,728	8,3996	4,4967	10,08	15,6	13,6	13,66	8,072387	10,48
7,5379	12,041	10,279	12,758	13,67	11,4	3,62	8,782	9,860073	12,24
9,501	11,14	9,2096	4,9954	9,493	7,53	9,84	8,118	9,683745	10,81
7,9809	9,5179	10,87	11,882	11,72	8,74	5,32	12,44	9,713731	11,25
7,5635	9,9074	9,4355	11,791	10,46	10,2	13,6	10,38	8,244739	10,1
9,9542	9,943	14,963	6,803	6,126	8,02	9,85	14,48	4,824134	8,108
21,525	10,606	12,691	13,155	9,389	7,14	9,23	8,327	8,814706	6,754
7,1615	11,172	10,275	9,3514	15,47	3,96	12,9	10,87	11,0644	14,47
5,608	8,8028	7,2609	10,214	10	13,1	13,5	9,717	9,875233	7,597
11,008	9,5975	10,457	4,8088	7,429	9,54	14,4	8,189	13,16323	10,93
11,787	11,223	5,2684	9,0543	15,16	10	12,5	8,341	11,44044	8,402
10,275	15,397	3,8992	14,82	5,762	16,1	12,8	12,8	7,191176	11,39
16,502	7,0601	13,012	12,194	8,783	15,6	10,8	14,26	9,275263	12,61
10,943	15,151	8,1787	7,9375	9,087	9,81	6,36	11,43	8,283804	12,23
12,828	16,434	10,248	6,3168	6,292	9,08	1,57	10,94	9,176796	9,07
6,1356	4,1401	8,8708	13,057	15,11	7,02	7,6	12,05	9,747141	10,39

B21	B22	B23	B24	B25	B26	B27	B28	B29	B30
9,919	10,09	11,295	12,17	10,659	13,26	11,29	13,261	7,71	5,182
5,915	11,2	2,8536	11,38	8,1137	13,69	13,41	14,155	9,622	14,99
6,02	13,44	6,7712	9,474	3,3105	8,047	8,969	8,9209	12,24	12,21
8,752	9,645	5,0428	16,43	13,637	14,21	9,89	8,1971	5,66	6,658
7,467	12,18	12,606	9,649	11,728	8,4	4,497	10,084	15,58	13,6
6,253	17,1	11,273	7,538	12,041	10,28	12,76	13,674	11,38	3,618
8,561	7,665	9,5403	9,501	11,14	9,21	4,995	9,4928	7,532	9,836
9,937	4,788	12,485	7,981	9,5179	10,87	11,88	11,723	8,742	5,316
8,361	12,14	11,403	7,564	9,9074	9,435	11,79	10,456	10,18	13,62
10,8	5,873	5,7259	9,954	9,943	14,96	6,803	6,1261	8,024	9,855
13,82	10,69	15,065	21,53	10,606	12,69	13,15	9,3894	7,144	9,226

5,93	10,8	8,631	7,162	11,172	10,28	9,351	15,469	3,96	12,92
13,2	8,624	10,302	5,608	8,8028	7,261	10,21	10,005	13,09	13,52
8,213	9,861	11,994	11,01	9,5975	10,46	4,809	7,4293	9,544	14,44
8,541	9,822	10,213	11,79	11,223	5,268	9,054	15,159	9,999	12,49
7,643	12,64	12,955	10,28	15,397	3,899	14,82	5,7623	16,06	12,81
7,504	7,845	13,601	16,5	7,0601	13,01	12,19	8,7834	15,57	10,83
8,006	4,067	9,3215	10,94	15,151	8,179	7,938	9,0866	9,815	6,357
9,903	10,93	8,7321	12,83	16,434	10,25	6,317	6,2923	9,075	1,568
9,302	10,94	12,13	6,136	4,1401	8,871	13,06	15,109	7,024	7,601

Выборки из нормального распределения для каждого варианта представлены на листе "данные" в столбцах. Номер варианта k равен порядковому номеру в списке группы.

$\alpha = 0{,}005 \cdot k$.

Скопировать выборку своего варианта и решить следующие задачи:

Задача №1

Построить интервальную оценку дисперсии на уровне доверия $1-\alpha$:

1) если мат. ожидание равно 10,
2) если мат. ожидание неизвестно.

Записать в ответе: надёжность, границы доверительного интервала, ошибку выборки («точность»), длину интервала.

Задача №2

1) Построить интервальную оценку мат. ожидания на уровне доверия $1-\alpha$, если дисперсия неизвестна. Записать в ответе: надёжность, границы доверительного интервала, ошибку выборки («точность»), длину интервала.

2) С какой вероятностью мат. ожидание отклоняется от выборочного среднего по модулю не более, чем на $0{,}01 \cdot k+1$, если дисперсия неизвестна.

Задача №3

1) Построить интервальную оценку мат. ожидания на уровне доверия 1-α, если среднеквадратическое отклонение равно 3. Записать в ответе: надёжность, границы доверительного интервала, ошибку выборки («точность»), длину интервала.

2) С какой вероятностью мат. ожидание отклоняется от выборочного среднего по модулю не более, чем на 0,01·k+1, если среднеквадратическое отклонение равно 3.

3) Какой минимальный объём выборки необходимо взять, чтобы при среднеквадратическом отклонении, равном 3, мат. ожидание отклонялось от выборочного среднего по модулю не более, чем на 0,01·k+1 с вероятностью 1-α.

Тема 5. Асимптотическая интервальная оценка генеральной доли значения альтернативного признака (вероятности успеха в одном испытании в схеме Бернулли)

Рассмотрим в генеральной совокупности Ω объёма N так называемый альтернативный признак *X*. Пусть из генеральной совокупности Ω извлечена выборка $\hat{\Omega}$ объёма n.

Таблица 10

X	x_1	x_2
частота	N_1	$N-N_1$
относит. частота	$\frac{N_1}{N}$	$1-\frac{N_1}{N}$

X	x_1	x_2
частота	n_1	$n-n_1$
относит. частота	$\frac{n_1}{n}$	$1-\frac{n_1}{n}$

Назовём **генеральной долей** p значения x_1 признака X относительную частоту $p=\frac{N_1}{N}$, **выборочной долей** $\hat{p}$ значения x_1 признака X - относительную частоту $\hat{p}=\frac{n_1}{n}$.

В теории вероятностей аналогом генеральной доли p являлась вероятность «успеха» в одном испытании схемы Бернулли.

В нашем случае «успех» - $X(\omega)=x_1,\ \omega\in\Omega$.

n_1 - число успехов в серии из n независимых испытаний.

Пример. Рассмотрим партию, содержащую 1000 различных образцов, среди которых 50 бракованных. Для тестирования отбирают случайным образом 100 из них, проверяют качество и находят 4 бракованных.

X - качество изделия: x_1 – «брак», x_2 – «не брак». Тогда $N=1000$, $n=100$, $N_1=50$, $n_1=4$, $p=0{,}05$, $\hat{p}=0{,}04$.

Утверждения:

1) В случае повторной и бесповторной выборок $M\hat{p}=p$. (23)

2) В случае повторной выборки $D\hat{p}=\dfrac{pq}{n}$. (24)

3) В случае бесповторной выборки $D\hat{p}=\dfrac{pq}{n}\dfrac{N-n}{N-1}$. (25)

Предположим, что параметр p неизвестен. Ранее строили точечную оценку $\hat{p}=\dfrac{n_1}{n}$ генеральной доли p и находили среднеквадратическую ошибку такого оценивания: $\sqrt{M(\hat{p}-p)^2}=\sqrt{M(\hat{p}-M\hat{p})^2}=\sqrt{D\hat{p}}=\sigma(\hat{p})$ (24), (25) по аналогии с точечным оцениванием генерального среднего (мат. ожидания).

Пример. В выборах приняли участие 1 млн. избирателей. Предполагая, что за наиболее популярного кандидата проголосует ≈50% избирателей, найдите стандартное отклонение процента бюллетеней в его пользу среди первых 900000 обработанных бюллетеней.

Решение. Поскольку генеральная доля кандидата p ≈ 0,5, получим

$$\sigma_{\hat{p}}\approx\sqrt{\frac{0{,}5\cdot 0{,}5}{900\,000}\left(1-\frac{900\,000}{1\,000\,000}\right)}=\frac{5}{3}\cdot 10^{-4}\approx 0{,}017\%,$$

учитывая замечание 1 в п.3. Полученный результат показывает, что уже после обработки первых 900000 бюллетеней можно с достаточной степенью точности сделать вывод об итогах выборов.

Теперь построим **интервальную оценку генеральной доли**, то есть найдём **приближённый доверительный интервал**, в который генеральная доля p попадёт с заданной вероятностью $\gamma = 1 - \alpha$.

Замечание. При выводе асимптотической интервальной оценки генеральной доли p (вероятности успеха в одном испытании) применяется ЦПТ, поэтому на практике интервальное оценивание p можно проводить только при больших объёмах выборки n.

Доверительный интервал мат. ожидания при известной дисперсии для нормального распределения имел вид:

$$\theta_1 = \bar{X} - \frac{z_{1-\frac{\alpha}{2}}\sigma}{\sqrt{n}}, \; \theta_2 = \bar{X} + \frac{z_{1-\frac{\alpha}{2}}\sigma}{\sqrt{n}}.$$

Если в качестве выборочных значений X_i рассмотреть бернуллиевские случайные величины (число успехов в одном i-ом испытании в схеме Бернулли) и применить ЦПТ, получим приближённый доверительный интервал генеральной доли p (вероятности успеха в одном испытании):

$$\theta_1 = \hat{p} - \frac{z_{1-\frac{\alpha}{2}}\sqrt{\hat{p}\hat{q}}}{\sqrt{n}}, \; \theta_2 = \hat{p} + \frac{z_{1-\frac{\alpha}{2}}\sqrt{\hat{p}\hat{q}}}{\sqrt{n}}. \qquad (30)$$

Таким образом, $P(\hat{p} - \Delta p < p < \hat{p} + \Delta p) = 1 - \alpha$, где $\Delta p = \frac{z_{1-\frac{\alpha}{2}}\sqrt{\hat{p}\hat{q}}}{\sqrt{n}}$

Пример. *В результате проведенного социологического опроса* n = 1140 *человек рейтинг кандидата в президенты составил* 14%. *Найдите доверительный интервал для рейтинга кандидата с гарантированной надежностью* 95%.

Исходные данные: $n = 1140$, выборочная доля составляет $\hat{p} = 0{,}14$, доверительная вероятность $\gamma = 0{,}95$. Квантиль стандартного нормального распределения порядка 0,975 приближённо равен 1,96. Используя формулу (30), получаем приближённый доверительный интервал для рейтинга кандидата:

$$\left(0{,}14 - 1{,}96\sqrt{0{,}14 \cdot \frac{0{,}86}{1140}}; 0{,}14 + 1{,}96\sqrt{0{,}14 \cdot \frac{0{,}86}{1140}}\right) =$$

$$= (0{,}1199; 0{,}1601) = (11{,}99\%; 16{,}01\%).$$

Задания для самостоятельного решения

Номер варианта соответствует порядковому номеру в списке группы и обозначается k.

Сгенерировать выборку объёма 100+k значений признака, имеющего в генеральной совокупности нормальное распределение с мат. ожиданием μ, равным 2·k, и стандартным отклонением, равным k.

1) Найти интервальную оценку генеральной дисперсии признака на уровне доверия 0,7+k·0,01

а) при неизвестном мат. ожидании

б) при известном мат. ожидании

2) Найти интервальную оценку генеральной доли тех значений признака, которые оказались не больше, чем $\mu + s$ на уровне доверия 0,7+k·0,01, где s – несмещённая оценка стандартного отклонения.

Тема 6. Точечные оценки параметров распределения

Пусть некоторый признак X, заданный на генеральной совокупности Ω, подчиняется определённому закону распределения с функцией распределения $F(x, \theta)$, зависящей от неизвестного параметра θ. Рассмотрим n

выборочных значений $X_1, X_2, ..., X_n$ из генеральной совокупности Ω, которую мы называем выборкой из распределения. Задача точечного оценивания заключается в том, что на основании выборочных значений $X_1, X_2, ..., X_n$ следует оценить неизвестный параметр θ известного распределения $F(x,\theta)$. Например, если признак X подчиняется закону распределения Пуассона, то неизвестный параметр θ равен λ, поскольку $P\left(X = X_i\right) = \dfrac{\lambda^{X_i}}{X_i!}e^{-\lambda}$. Точечная оценка является функцией от выборочных значений и обозначается $\hat{\theta} = \hat{\theta}\left(X_{1,} X_2, ..., X_n\right)$, например, выборочное среднее $\bar{X} = \dfrac{1}{n}\sum_{i=1}^{n} X_i$ является точечной оценкой математического ожидания. Естественно предположить, что распределение $\hat{\theta} = \hat{\theta}\left(X_{1,} X_2, ..., X_n\right)$ должно в некотором смысле концентрироваться около оцениваемого параметра θ. Основными критериями качества построенной оценки являются такие свойства, как 1) несмещённость, 2) состоятельность и 3) эффективность оценки.

1) Точечная оценка $\hat{\theta} = \hat{\theta}\left(X_{1,} X_2, ..., X_n\right)$ называется несмещённой, если $E\left(\hat{\theta}\right) = \theta$.

2) Точечная оценка $\hat{\theta} = \hat{\theta}\left(X_{1,} X_2, ..., X_n\right)$ называется состоятельной, если $\hat{\theta} \to \theta$ при $n \to \infty$ по вероятности, то есть $\lim\limits_{n\to\infty} P\left(\left|\hat{\theta} - \theta\right| \le \varepsilon\right) = 1$.

Если точечная оценка $\hat{\theta} = \hat{\theta}\left(X_{1,} X_2, ..., X_n\right)$ является несмещённой, то достаточным условием её состоятельности является сходимость дисперсии к нулю при $n \to \infty$: $Var\left(\hat{\theta}\right) \to 0, n \to \infty$.

3) Точечная оценка $\hat{\theta} = \hat{\theta}\left(X_{1,} X_2, ..., X_n\right)$ называется эффективной в некотором классе оценок, если её дисперсия минимальна на множестве всех оценок данного класса:$Var\left(\hat{\theta}\right) \sim \min$.

Несмещённые оценки математического ожидания и дисперсии

1) Несмещенной оценкой генерального среднего μ является выборочное среднее $\bar{X}$, т.к. $E\left(\bar{X}\right)=E(X)=\mu$. Действительно, $E\left(\bar{X}\right)=E(\bar{X}=\frac{1}{n}\sum_{i=1}^{n}X_i)=\frac{1}{n}\sum_{i=1}^{n}E\left(X_i\right)=\frac{1}{n}nE\left(X\right)=E\left(X\right)/$

2) Если $\mu=E\left(X\right)$ известно, то несмещенной оценкой генеральной дисперсии $Var\left(X\right)$ является $s_0^2=\frac{1}{n}\left[\left(X_1-\mu\right)^2+\left(X_2-\mu\right)^2+...+\left(X_n-\mu\right)^2\right]$. Действительно,

$$E\left(s_0^2\right)=E\left(\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\mu\right)^2\right)=\frac{1}{n}\sum_{i=1}^{n}E\left(X_i-\mu\right)^2=\frac{1}{n}\sum_{i=1}^{n}Var\left(X_i\right)=\frac{1}{n}n\cdot Var\left(X\right)=Var\left(X\right)$$

3) Если $\mu=E\left(X\right)$ неизвестно, то несмещенной оценкой генеральной дисперсии $Var\left(X\right)$ является «исправленная дисперсия» $s^2=\frac{1}{n-1}\left[\left(X_1-\bar{X}\right)^2+\left(X_2-\bar{X}\right)^2+...+\left(X_n-\bar{X}\right)^2\right]$

Часто при решении задач оказывается полезной связь исправленной дисперсии с выборочной дисперсией $\hat{V}ar\left(X\right)=\hat{\sigma}^2$:

$\hat{V}ar\left(X\right)^2=\frac{1}{n}\left[\left(X_1-\bar{X}\right)^2+\left(X_2-\bar{X}\right)^2+...+\left(X_n-\bar{X}\right)^2\right]$ =>

$s^2=\frac{n}{n-1}\hat{V}ar\left(X\right)=\frac{n}{n-1}\hat{\sigma}^2$ (s^2 = ДИСП.В, $\hat{\sigma}^2$ = ДИСП.Г)

Пример. Дана выборка $X_1, X_2, ..., X_n$ из некоторого распределения. Найти точечную оценку квантили данного распределения порядка *p*.

Решение

Напомним, что квантилью порядка *p* некоторого распределения называется такое значение x_p случайной величины *X*, имеющей данное распределение, для которого выполняется равенство $F(x_p)=p=P(X\le x_p)$. Квантиль порядка *p* часто называют процентной точкой порядка 1 - *p*, например, квантиль порядка 0,95 является 5-ти процентной точкой. В данной задаче требуется найти точечную оценку квантили порядка p, то есть эмпирическую квантиль $\hat{x}_p$: $p \approx \dfrac{k}{n}$, где k – число выборочных значений X_i, для которых выполняется условие $X_i \le \hat{x}_p$. В Excel $\hat{x}_p = ПРОЦЕНТИЛЬ.ВКЛ(X_1, X_2, ..., X_n; p)$.

Одним из методов построения точечных оценок параметров распределения является **метод моментов** (ММ). В качестве моментов используются начальные и центральные моменты k-го порядка ν_k и μ_k. Наиболее часто применяют начальный момент 1-го порядка – математическое ожидание (генеральное среднее). С помощью закона больших чисел можно показать, что оценка параметра распределения, полученная методом моментов, всегда является состоятельной.

Пример Пусть признак *X* имеет показательное распределение с параметром λ, т.е. $f(x) = \begin{cases} 0, & x < 0 \\ \lambda e^{-\lambda x}, & x \ge 0 \end{cases}$ Оценить λ методом моментов.

Решение

$E(X) = \dfrac{1}{\lambda}$ => $\lambda = \dfrac{1}{E(X)}$, оценкой $E(X)$ является выборочное среднее $\bar{X}$ =>

$$\lambda_{MM} = \frac{1}{\bar{X}}$$

Пример

Пусть признак *X* имеет геометрическое распределение с параметром *p*. Дана выборка $X_1 = 4$, $X_2 = 3$, $X_3 = 2$, $X_4 = 4$, $X_5 = 2$. Оценить *p* методом моментов.

Решение

$P(X = X_i) = (1-p)^{X_i - 1} \cdot p$, например, $P(X = 3) = (1-p)^2 \cdot p$.

$E(X) = \dfrac{1}{p}$, $p = \dfrac{1}{E(X)}$, $p_{MM} = \dfrac{1}{\bar{X}} = \dfrac{1}{3}$, поскольку $\bar{X} = \dfrac{2 \cdot 2 + 3 + 4 \cdot 2}{5} = 3$.

Задания для самостоятельного решения

Смоделировать выборку объёма 1000 из заданного в варианте распределения. Построить:

а) точечную оценку заданного в варианте параметра распределения **методом моментов**;

б) точечную оценку **квантили** порядка, заданного в варианте;

в) дисперсию несмещённой оценки мат. ожидания

г) **несмещённую** оценку дисперсии.

Варианты

Вар.1 Биномиальное распределение, $p = 0{,}1$, $n = 200$.

а) Параметр - p

б) Порядок квантили - 0,72

Вар.2 Равномерное распределение на отрезке [a, b], a=1, b=8.

а) Параметр – a

б) Порядок квантили - 0,68

Вар.3 Распределение Пуассона, $\lambda = 3$.

а) Параметр - λ

б) Порядок квантили - 0,7

Вар.4 Нормальное распределение с мат. ожиданием 4, дисперсией 2.

а) Параметр – станд. отклонение

б) Порядок квантили 0,74

Вар.5 Равномерное распределение на отрезке [a, b], a = 2, b = 7.

а) Параметр – b

б) Порядок квантили - 0,76

Вар.6 Биномиальное распределение, $p = 0,2$, $n = 300$.

а) Параметр - n

б) Порядок квантили - 0,78

Вар.7 Равномерное распределение на отрезке [a, b], a=3, b=8.

а) Параметр – a

б) Порядок квантили - 0,8

Вар.8 Распределение Пуассона, $\lambda = 3,5$.

а) Параметр - λ

б) Порядок квантили - 0,82

Вар.9 Нормальное распределение с мат. ожиданием 2, дисперсией 9.

а) Параметр – станд. отклонение

б) Порядок квантили - 0,84

Вар.10 Равномерное распределение на отрезке [a, b], a = - 2, b = 3.

а) Параметр – b

б) Порядок квантили - 0,86

Вар.11 Биномиальное распределение, $p = 0,3$, $n = 400$.

а) Параметр - p

б) Порядок квантили - 0,58

Вар.12 Равномерное распределение на отрезке [a, b], a=4, b=8.

а) Параметр – a

б) Порядок квантили - 0,6

Вар.13 Распределение Пуассона, $\lambda = 4$.

а) Параметр - λ

б) Порядок квантили - 0,62

Вар.14 Нормальное распределение с мат. ожиданием -2, дисперсией 0,64.

а) Параметр – станд. отклонение

б) Порядок квантили - 0,64

Вар.15 Равномерное распределение на отрезке [a, b], a = - 4, b = 1.

а) Параметр – b

б) Порядок квантили - 0,65

Вар.16 Биномиальное распределение, p = 0,4, n = 500.

а) Параметр - n

б) Порядок квантили - 0,3

Вар.17 Равномерное распределение на отрезке [a, b], a= - 3, b = 2.

а) Параметр – a

б) Порядок квантили - 0,32

Вар.18 Распределение Пуассона, $\lambda = 5$.

а) Параметр - λ

б) Порядок квантили - 0,34

Вар.19 Нормальное распределение с мат. ожиданием 5, дисперсией 16.

а) Параметр – станд. отклонение

б) Порядок квантили - 0,36

Вар.20 Равномерное распределение на отрезке [a, b], a = - 7, b = - 2.

а) Параметр – b

б) Порядок квантили - 0,38

Вар.21 Биномиальное распределение, $p = 0{,}6$, $n = 600$.

а) Параметр - p

б) Порядок квантили - 0,4

Вар.22 Равномерное распределение на отрезке [a, b], a = 5, b = 9.

а) Параметр – a

б) Порядок квантили - 0,42

Вар.23 Распределение Пуассона, $\lambda = 6$.

а) Параметр - λ

б) Порядок квантили - 0,44

Вар.24 Нормальное распределение с мат. ожиданием 7, дисперсией 25.

а) Параметр – станд. отклонение

б) Порядок квантили - 0,46

Вар.25 Равномерное распределение на отрезке [a, b], a = - 9, b = - 4.

а) Параметр – b

б) Порядок квантили - 0,48

Вар.26 Биномиальное распределение, $p = 0{,}1$, $n = 200$.

а) Параметр - n

б) Порядок квантили - 0,22

Вар.27 Биномиальное распределение, $p = 0{,}1$, $n = 200$.

а) Параметр - p

б) Порядок квантили - 0,72

Вар.28 Равномерное распределение на отрезке [a, b], a=1, b=8.

а) Параметр – а

б) Порядок квантили - 0,68

Вар.29 Распределение Пуассона, $\lambda = 3$.

а) Параметр - λ

б) Порядок квантили - 0,7

Вар.30 Нормальное распределение с мат. ожиданием 4, дисперсией 2.

а) Параметр – станд. отклонение

б) Порядок квантили 0,74

Литература

1. Вуколов Э. Л. Основы статистического анализа. Практикум по статистическим методам и исследованию операции с использованием пакетов STATISTICA и EXCEL: учебное пособие. — 2-е изд.,— М.: ФОРУМ. 2008. — 464 с.

2. Газарян В.А., Терентьев Е.Н. Методические указания и задания для выполнения лабораторного практикума по дисциплине Статистика. Издательство Московского Гуманитарного Университета, Москва, 2011.

3. Боровиков В. STATISTICA. Искусство анализа данных на компьютере: Для профессионалов. 2 изд.: Питер, 2003. – 688 с.

4. К.Э. Плохотников, С.В. Колков. Статистика: учеб. пособие. – М.: Флинта: МПСИ, 2017. – 288 с.

5. А.В. Браилов, П.Е. Рябов. Теория вероятностей и математическая статистика. Методические рекомендации по самостоятельной работе Часть 5. Москва. 2011. Изд-во Финансового Университета при Правительстве РФ.

6. И.Е. Денежкина, М.Г. Орлова, Ю.Н. Швецов. Основы математической статистики. Учебно-методическое пособие для самостоятельной работы бакалавров. М.: Финуниверситет, кафедра «Теория вероятностей и математическая статистика», 2010.

7. В.А. Газарян. Выборочный метод в статистическом анализе данных. — Москва: Москва, 2019.
8. В. А. Газарян, Е.И. Компацева. Основы статистического анализа в пакете STATISTICA. — М.: М., 2017.