Финансовый университет
при правительстве Российской Федерации

Шамраева
Виктория Викторовна

кандидат физико-математических наук,
доцент кафедры
математики и анализа данных

Теория вероятностей и математическая статистика

НАПРАВЛЕНИЕ ПОДГОТОВКИ: «Прикладная математика - ПМ»

КВАЛИФИКАЦИЯ (СТЕПЕНЬ): бакалавр

Раздел 1. Оценки параметров

Эмпирические начальный и центральный моменты порядка *k* ($k = 0, 1, 2, \ldots$) — это статистики

$$\hat{\nu}_k = \overline{X^k} = \frac{1}{n}\sum_{i=1}^{n} X_i^k$$

и

$$\hat{\mu}_k = \overline{(X - \bar{X})^k} = \frac{1}{n}\sum_{i=1}^{n} (X_i - \bar{X})^k.$$

соответственно.

Свойства эмпирических моментов аналогичны свойствам соответствующих **теоретических моментов**.

Раздел 1. Оценки параметров

Упражнение. Показать, что

1) $\widehat{\boldsymbol{\mu}}_{\boldsymbol{3}}(\bar{X}) = \dfrac{\boldsymbol{\mu}_{\boldsymbol{3}}(X)}{n^2}$;

2) $\widehat{\boldsymbol{\mu}}_{\boldsymbol{4}}(\bar{X}) = \dfrac{\boldsymbol{\mu}_{\boldsymbol{4}}(X)}{n^3} + \dfrac{3(n-1)}{n^3}\boldsymbol{\mu}_{\boldsymbol{2}}^{\boldsymbol{2}}(X)$.

Раздел 1. Оценки параметров

Эмпирический коэффициент асимметрии -

$$\hat{A}_X = \frac{\hat{\mu}_3}{\hat{\sigma}^3}.$$

Для симметричных распределений (вкл. нормальное распределение) варианты (x_i), равноудаленные от $\bar{x}$, имеют одинаковую частоту, потому

$$\hat{A}_X = 0.$$

Эмпирический коэффициент эксцесса -

$$\hat{E}_X = \frac{\hat{\mu}_4}{\hat{\sigma}^4} - 3.$$

Раздел 2. Статистика конечной совокупности

Для вычисления эмпирических коэффициентов асимметрии и эксцесса по *конкретной выборке* **генеральной совокупности** в **Microsoft Excel** используются функции

$\hat{A}_X$ =**СКОС.Г**(число1;[число2];...)

$\tilde{E}_X$ =**ЭКСЦЕСС**(число1;[число2];...)

$$\hat{A}_X = \frac{\hat{\mu}_3}{\hat{\sigma}^3}$$

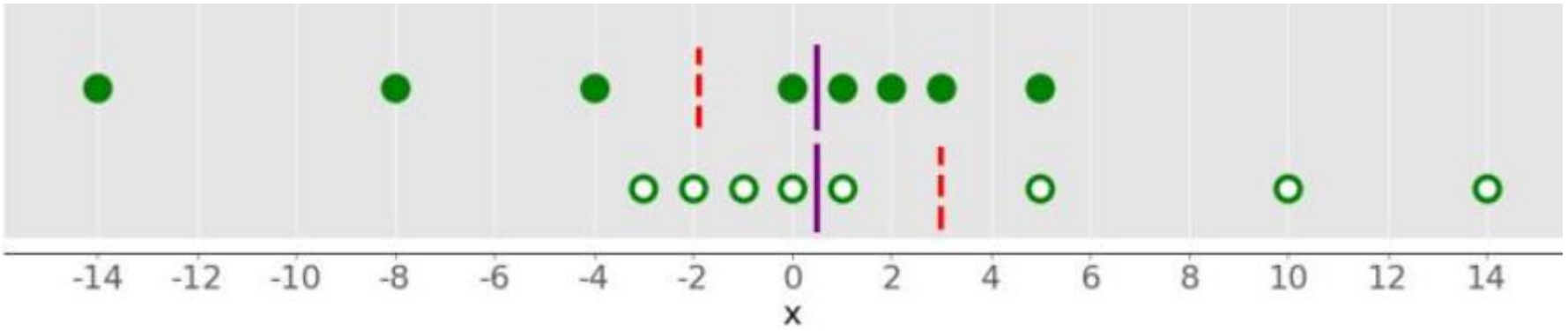
$$\hat{E}_X^{o} = \frac{(n-1)[(n+1)\hat{E}_X + 6]}{(n-2)(n-3)}.$$

В **Python** вычисляет эмпирических коэффициентов асимметрии и эксцесса функции

skew( ) и **kurtosis( )**

из библиотеки **SciPy**.

Раздел 2. Статистика конечной совокупности




```
import pandas as pd
```

```
xi = pd.DataFrame([-14, -8, -4, 0,1,2,3,5])[0]
yi = pd.DataFrame([-3, -2, -1, 0,1,5,10,14])[0]
```

```
xi.mean(), yi.mean()
```

```
(-1.875, 3.0)
```

```
xi.median(), yi.median()
```

```
(0.5, 0.5)
```

```
xi.skew(), yi.skew()
```

```
(-1.0836532081173458, 1.0214178945965033)
```

Раздел 2. Статистика конечной совокупности

Пример 1. По данной реализации выборки X: $10, 20, 30, 40, 50$ определить **эмпирический коэффициент асимметрии и исправленный коэффициент эксцесса**

$$\hat{A}_X = \frac{\hat{\mu}_3}{\hat{\sigma}^3}, \qquad \hat{E}_X^{0} = \frac{(n-1)[(n+1)\hat{E}_X + 6]}{(n-2)(n-3)}. \qquad (\hat{E}_X = \frac{\hat{\mu}_4}{\hat{\sigma}^4} - 3)$$

 # MS Excel:

	A	B	C	D	E	F	G
1	**X:**	10	20	30	40	50	
2							
3	**Эмпирический коэффициент асимметр**				0,00	=СКОС.Г(B1:F1)	
4	**Выборочный коэффициент эксцесса X:**				-1,2	=ЭКСЦЕСС(B1:F1)	

 # Python:

```
import scipy.stats as sts

X = [10,20,30,40,50]

print('Эмпирический коэффициент асимметрии:',sts.skew(X))
print('Выборочный коэффициент эксцесса:',sts.kurtosis(X,bias= False))
```

```
Эмпирический коэффициент асимметрии: 0.0
Выборочный коэффициент эксцесса: -1.20000000
```

Раздел 1. Оценки параметров

Эмпирический квантиль $\widehat{X}_\alpha$ порядка α определяется формулой

$$\hat{X}_\alpha = \begin{cases} X_{([n\alpha]+1)}, & \text{если } n\alpha \text{ дробное,} \\ X_{(n\alpha)}, & \text{если } n\alpha \text{ целое} \end{cases}.$$

Здесь $[x]$ –целая часть числа x.

Раздел 1. Оценки параметров

Эмпирические квартили (Q, quartile) – это такие числа, которые разбивают (делят) упорядоченное от минимума к максимуму множество данных на четыре равные по численности части, то есть **равновеликие между собой по объёму наблюдений**.

Вариационный ряд делится тремя квартилями Q_1, Q_2, Q_3 на 4 равные части.

Q_2 – медиана.

В целом, понятия **квантиль** и **процентиль** взаимозаменяемы, так же, как и шкалы исчисления вероятностей — абсолютная и процентная.

Процентили также называются **перцентилями** или **центилями**.

Раздел 1. Оценки параметров

Всё упорядоченное множество объектов выборки можно разбить на четыре равновеликие части, прибегая к использованию **трёх эмпирических квартилей**. Поэтому выделяют нижнюю (первую), среднюю (вторую) и верхнюю (третью квартиль).

Раздел 1. Оценки параметров

Эмпирический первый квартиль $\widehat{Q}_1$ есть выборочный квантиль порядка $\alpha = \frac{1}{4}$:

$$\hat{Q}_1 = \hat{X}_{1/4} = \begin{cases} X_{\left(\left[\frac{n}{4}\right]+1\right)}, \text{если } \frac{n}{4} \text{ дробное,} \\ X_{\left(\frac{n}{4}\right)}, \quad \text{если } \frac{n}{4} \text{ целое} \end{cases}.$$

Эмпирический третий квартиль $\widehat{Q}_3$ определяется формулой:

$$\hat{Q}_3 = \hat{X}_{(n-i+1)},$$

где i — порядковый номер выборочного первого квартиля в вариационном ряде.

Выборочные первый и третий квартили расположены симметрично относительно крайних элементов вариационного ряда.

Раздел 1. Оценки параметров

Квантиль в математической статистике — значение, которое заданная случайная величина не превышает с фиксированной вероятностью.

Если вероятность задана в процентах, то квантиль называется **процентилем** или **перцентилем**.

- 0,25-квантиль называется **первым (или нижним) квартилем** (от лат. quarta — четверть);
- 0,5-квантиль называется **медианой** (от лат. Mediāna — середина) или **вторым квартилем**;
- 0,75-квантиль называется **третьим (или верхним) квартилем**.

Раздел 1. Оценки параметров

Квартиль – значение признака, делящее совокупность на четыре равные части,

квантиль – на пять равных частей,

дециль – на десять равных частей,

перцентиль – на сто равных частей.

При вычислении квартильного, квантильного, децильного, или перцентильного значения признака важно не забывать соблюдать **условие ранжированности** (упорядоченности по возрастанию или убыванию) элементов изучаемой совокупности.

Раздел 1. Оценки параметров

Разность между третьим и первым квартилями называется **межквартильным размахом** (или **интерквартильной широтой**) выборки:

$$RQ = \hat{Q}_3 - \hat{Q}_1.$$

Раздел 1. Оценки параметров

Эмпирическая медиана (эмпирический второй квартиль $\widehat{Q_2}$) определяется формулой:

$$\widehat{Me} = \widehat{\mathrm{Q}}_2 = \begin{cases} X_{(\frac{n+1}{2})}, \text{если } n \text{ нечётное,} \\ X_{(\frac{n}{2})}, \quad \text{если } n \text{ чётное} \end{cases}.$$

Раздел 1. Оценки параметров

Пример. Для статистического ряда

$x_{(i)}$	1	5	6	8	9	10	15
n_i	1	2	3	1	2	3	1

вычислить эмпирический квантиль порядка $\alpha = 0{,}2$, эмпирическую медиану, первый и третий квартили и межквартильный размах.

Решение.

$$\hat{X}_\alpha = \begin{cases} X_{([n\alpha]+1)}, & \text{если } n\alpha \text{ дробное,} \\ X_{(n\alpha)}, & \text{если } n\alpha \text{ целое} \end{cases}$$

Раздел 1. Оценки параметров

Решение (продолжение).

	A	B	C	D	E	F	G
1	X						
2	1		x_0,2=	5,4	=ПРОЦЕНТИЛЬ.ВКЛ(A2:A14;0,2)		
3	5		Q1=	6	=КВАРТИЛЬ.ВКЛ(A2:A14;1)		
4	5		Q2=Me=	8	=КВАРТИЛЬ.ВКЛ(A2:A14;2)		
5	6		Q3=	10	=КВАРТИЛЬ.ВКЛ(A2:A14;3)		
6	6		R=	4	=D5-D3		
7	6						
8	8						
9	9						
10	9						
11	10						
12	10						
13	10						
14	15						

Вариационный ряд: 1; 5; 5; 6; 6; 6; 8; 9; 9; 10; 10; 10; 15

$\hat{Q}_1 = 6;\ \widehat{Me} = \hat{Q}_2 = 8;\ \hat{Q}_3 = 10;\ RQ = 4;\ \hat{x}_{0,2} = 5.$

Раздел 1. Оценки параметров

Решение (продолжение).

```
Ввод [1]: import numpy as np

          X = [1,5,5,6,6,6,8,9,9,10,10,10,10,15]

          print(f'Выборочный первый квартиль (Q1): {np.quantile(X,0.25)}\n')
          print(f'Перцентиль 25%: {np.percentile(X,25)}\n')

          print(f'Выборочный второй квартиль (Q2): {np.quantile(X,0.5)}\n')
          print(f'Выборочная медиана: {np.median(X)}\n')

          print(f'Выборочный третий квартиль (Q3): {np.quantile(X,0.75)}\n')

          print(f'Выборочный квантиль порядка 0.2: {np.quantile(X,0.2)}\n')
```

```
Выборочный первый квартиль (Q1): 6.0

Перцентиль 25%: 6.0

Выборочный второй квартиль (Q2): 8.5

Выборочная медиана: 8.5

Выборочный третий квартиль (Q3): 10.0

Выборочный квантиль порядка 0.2: 5.6
```

Вариационный ряд: 1; 5; 5; 6; 6; 6; 8; 9; 9; 10; 10; 10; 15

$\hat{Q}_1 = 6;\ \widehat{Me} = \hat{\mathrm{Q}}_2 = 8;\ \hat{Q}_3 = 10;\ RQ\ = 4;\ \hat{x}_{0,2} = 5.$

Раздел 1. Оценки параметров

Эмпирической модой вариационного ряда называется варианта, которой соответствует наибольшая частота.

Если интервальный ряд имеет одинаковую ширину интервалов, то за приближенное значение моды берут середину модального интервала, т.е. интервала с наибольшей частотой.

Точное значение моды можно получить по формуле:

$$\tilde{M}o = x_{м} + h \cdot \frac{n_{м} - n_{м-1}}{(n_{м} - n_{м-1}) + (n_{м} - n_{м+1})}$$

Раздел 1. Оценки параметров

В **MS Excel** формула **МОДА.ОДН** выводит в ячейку то число из набора, которое встречается чаще всего.

В старых версиях Excel существовала функция **МОДА**, но в более поздних она была разбита на две: **МОДА.ОДН** (для отдельных чисел) и **МОДА.НСК**(для массивов).

Впрочем, старый вариант тоже остался в отдельной группе, в которой собраны элементы из прошлых версий программы для обеспечения совместимости документов.

Раздел 1. Оценки параметров

Пример. По данной выборке X: 4, 1, 2, 2, 3,5,4 определить **выборочную моду.**

 # MS Excel:

	A	B	C	D	E	F	G	H
1	X	4	1	2	2	3	5	4
2								
3	**Мода X:**		4		**{=МОДА.НСК(B1:H1)}**			
4			2					

 # Python:

```
Ввод [1]: import statistics as sts

          X=[4, 1, 2, 2, 3, 5, 4]

          sts.multimode(X)

 Out[1]: [4, 2]
```

Эмпирические интервальные характеристики

Раздел 1. Оценки параметров

Рассмотрим, **таблицу интервальных частот**

a_1	b_1	n_1
a_2	b_2	n_2
…	…	…
a_s	b_s	n_s

в которой n_i ($\sum_{i=1}^{s} n_i = n$) — частота интервала (a_i, b_i) , $i = 1, 2, \ldots, s$.

Обозначим середину i-го интервала группировки через

$$x_i^* = \frac{a_i + b_i}{2}.$$

Раздел 1. Оценки параметров

К эмпирическим интервальным характеристикам относятся:

- **интервальное среднее**

$$\bar{x}^* = \frac{1}{n}\sum_{i=1}^{s} x_i^* n_i;$$

- **интервальная дисперсия**

$$(\hat{\sigma}_X^2)^* = \frac{1}{n}\sum_{i=1}^{s} (x_i^* - \bar{x}^*)^2 n_i;$$

- **интервальное стандартное отклонение** $\hat{\sigma}_X = \sqrt{(\hat{\sigma}_X^2)^*}$

$$x_i^* = \frac{a_i + b_i}{2}$$

$$\sum_{i=1}^{s} n_i = n$$

a_1	b_1	n_1
a_2	b_2	n_2
…	…	…
a_s	b_s	n_s

Раздел 1. Оценки параметров

Отметим, что эти и другие эмпирические интервальные характеристики вычисляются как характеристики эмпирического распределения

X	x_1^*	x_2^*	…	x_s^*
w_i	$\frac{n_1}{n}$	$\frac{n_2}{n}$	…	$\frac{n_s}{n}$

$$x_i^* = \frac{a_i + b_i}{2} \qquad \sum_{i=1}^{s} n_i = n$$

Раздел 1. Оценки параметров

В типичном случае, когда концы интервалов группировки $\Delta_i = (a_i, b_i)$ $(i = 1, 2, \ldots, s)$ образуют арифметическую прогрессию с шагом h,

$$\Delta_1 = (a_1, a_1 + h), \Delta_2 = (a_1 + h, a_1 + 2h), \ \ldots$$

для приближенного вычисления эмпирической дисперсии $\hat{\sigma}_X^2$ по интервальному распределению применяется **поправка Шеппарда**:

$$\hat{\sigma}_X^2 = (\hat{\sigma}_X^2)^* - \frac{1}{12}h^2.$$

$$x_i^* = \frac{a_i + b_i}{2}$$

$$\sum_{i=1}^{s} n_i = n$$

a_1	b_1	n_1
a_2	b_2	n_2
...	...	...
a_s	b_s	n_s

Описательная статистика

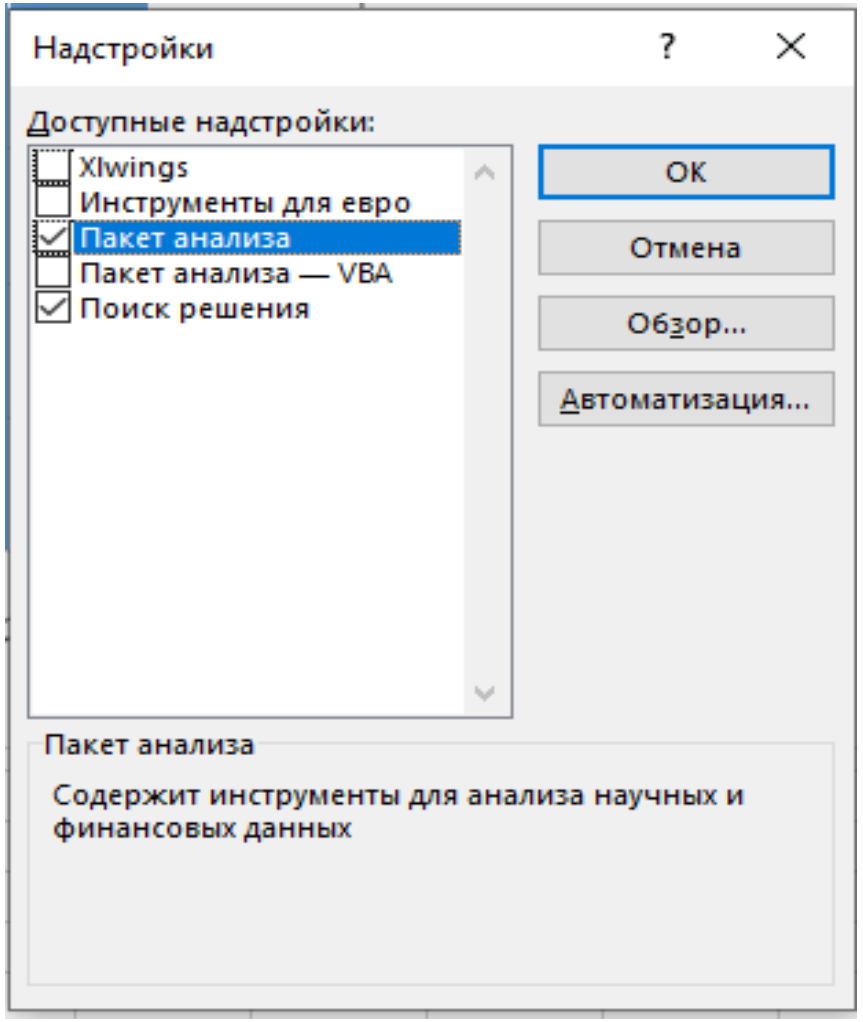
Раздел 1. Оценки параметров

В пакете **Microsoft Excel** есть надстройка «**Пакет анализа**», которая позволяет автоматизировать многие из задач статистического анализа данных.

Если в правой части вкладки «**Данные**» есть кнопка «**Анализ данных**», то это означает, что данная надстройка в **Microsoft Excel** установлена.

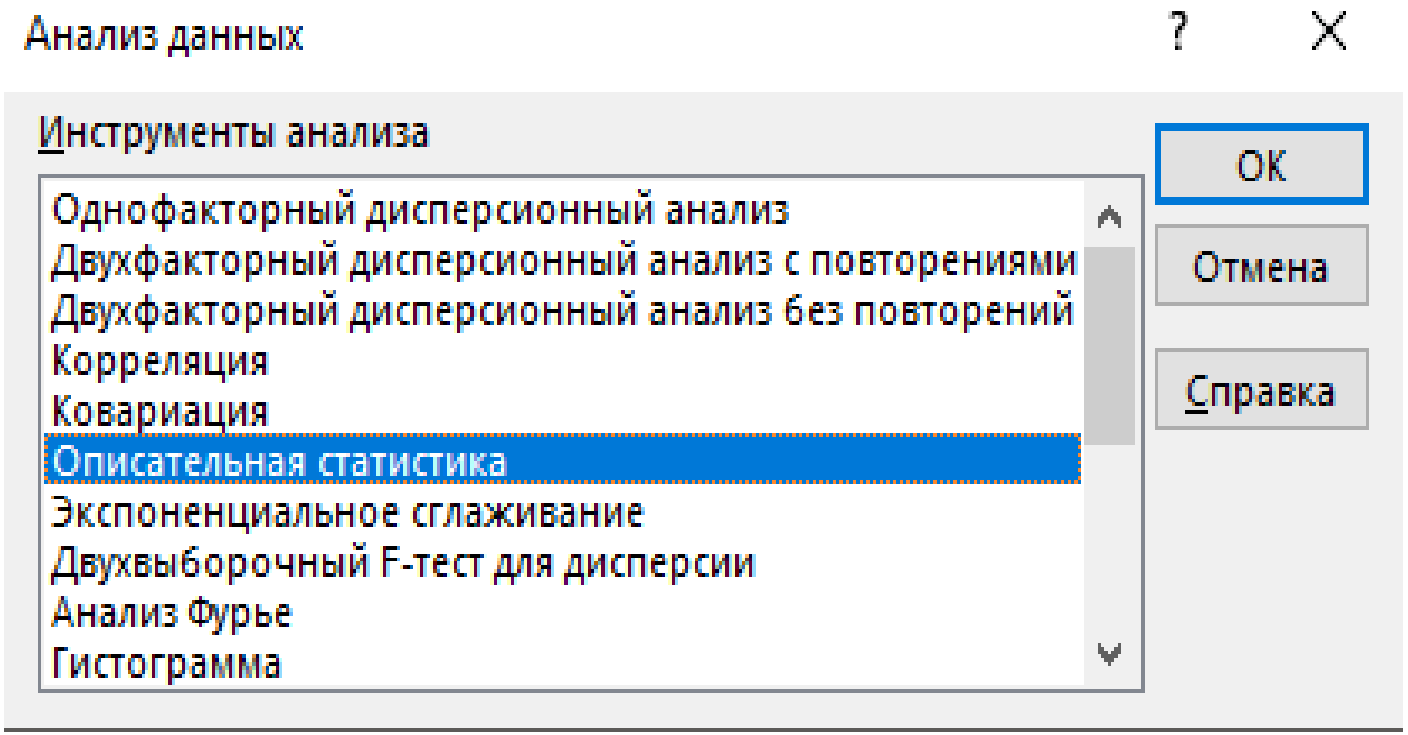
Раздел 1. Оценки параметров

Если же кнопка «**Анализ данных**» отсутствует, то нужно выбрать в окне управления параметрами **Excel** (меню «**Файл**» — «**Параметры**») пункт «**Надстройки**», далее в нижней строке «**Управление**» выбрать «**Надстройки Excel**», нажать кнопку «**Перейти…**», в появившемся окне отметить флажок «**Пакет анализа**» и нажать «**ОК**» — после этого кнопка «**Анализ данных**» должна появиться на вкладке «**Данные**».



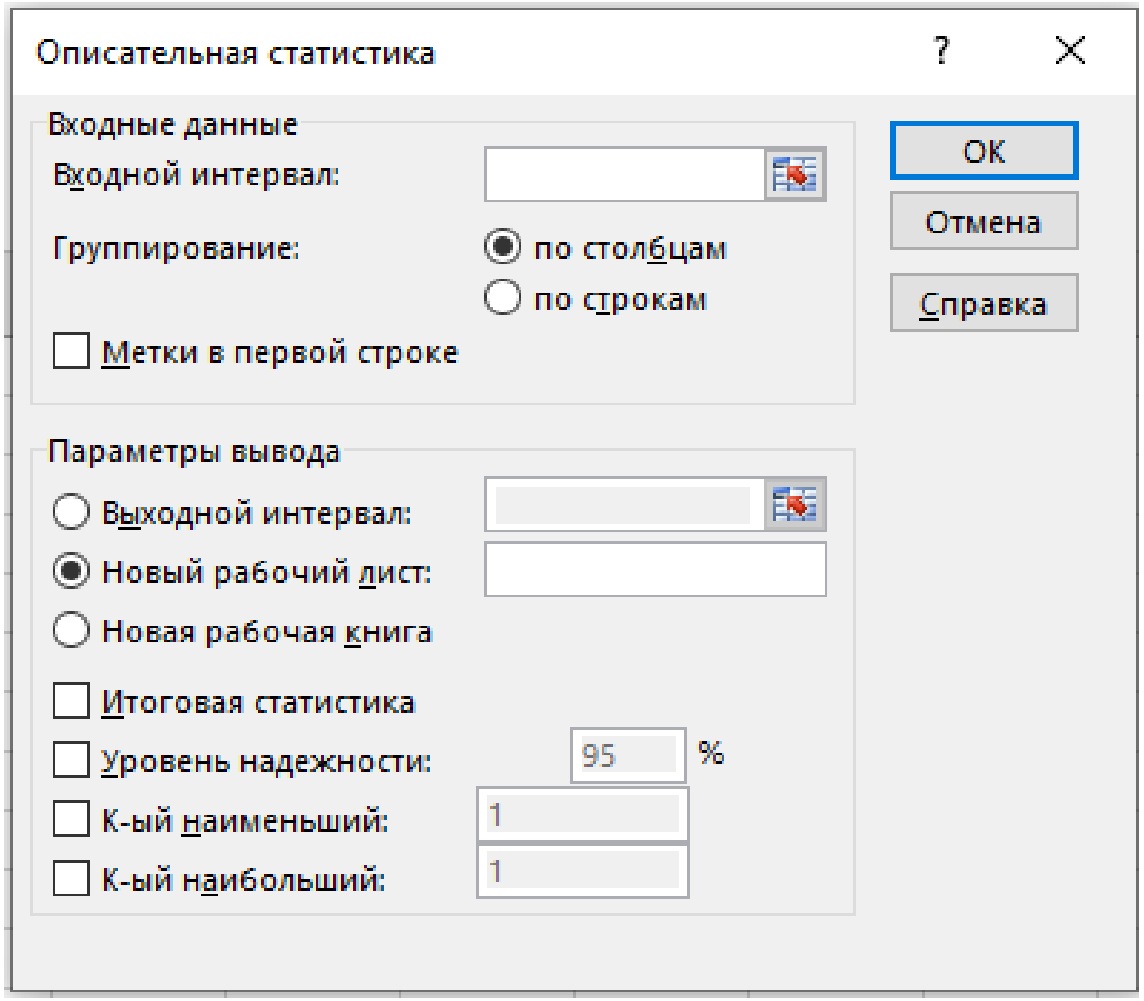
Раздел 1. Оценки параметров

Надстройка Пакет анализа вызывается командой главного меню **Данные → Анализ данных**. В появившемся окне **Анализ данных** выбираем пункт **Описательная статистика**.



Раздел 1. Оценки параметров

Далее откроется окно **Описательная статистика**, в котором необходимо сделать нужные установки.



Раздел 1. Оценки параметров

Описательную статистику в Python можно вывести с помощью

scipy.stats.describe( )

или с помощью метода **.describe()** библиотеки **Pandas**.

- **nobs** — количество наблюдений или элементов в вашем наборе данных;
- **minmax** — кортеж с минимальными и максимальными значениями;
- **mean** — среднее значение;
- **variance** — дисперсия;
- **skewness** — асимметрия;
- **kurtosis** — эксцесс вашего набора данных.

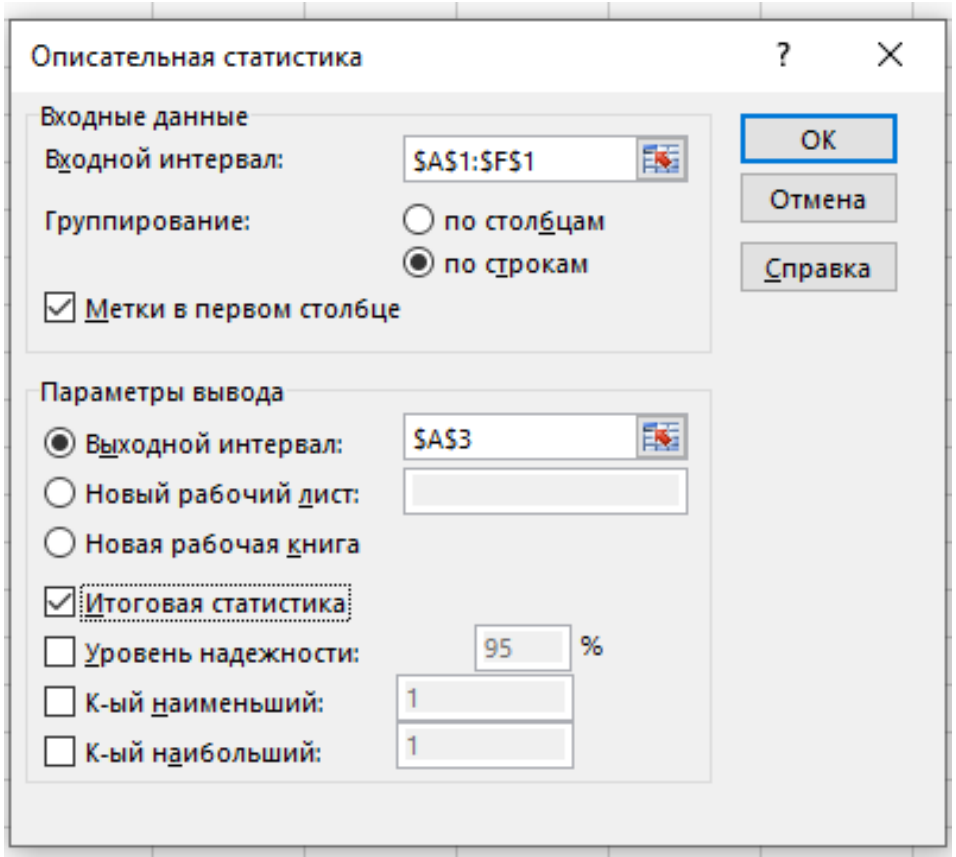
Аргументы **ddof** (при расчете **дисперсии**), **bias** - (при расчёте **асимметрии** и **эксцесса**) можно опускать.

Раздел 1. Оценки параметров

Пример. По данной выборке X: $10, 20, 30, 40, 50$ определить основные статистические характеристики воспользовавшись инструментом «**Описательная статистика**».

	A	B	C	D	E	F
1	**X:**	10	20	30	40	50

MS Excel:



Раздел 1. Оценки параметров

 # MS Excel:

	A	B	C	D	E	F
1	**X:**	10	20	30	40	50
2						
3	*X:*					
4						
5	Среднее	30				
6	Стандартная ошибка	7,071067812				
7	Медиана	30				
8	Мода	#Н/Д				
9	Стандартное отклонение	15,8113883				
10	Дисперсия выборки	250				
11	Эксцесс	-1,2				
12	Асимметричность	0				
13	Интервал	40				
14	Минимум	10				
15	Максимум	50				
16	Сумма	150				
17	Счет	5				

Пример. По данной выборке X: $10, 20, 30, 40, 50$ определить основные статистические характеристики воспользовавшись инструментом «**Описательная статистика**».

Раздел 1. Оценки параметров

Python:

Ввод [1]:
```
import scipy.stats as sts

X = [10,20,30,40,50]

#Сводка описательной статистики
sts.describe(X,ddof=0,bias=False)
```

Out[1]:
```
DescribeResult(nobs=5, minmax=(10, 50), mean=30.0,
variance=200.0, skewness=0.0, kurtosis=-1.200000000
0000004)
```

Пример. По данной выборке X: $10, 20, 30, 40, 50$ определить основные статистические характеристики воспользовавшись инструментом «**Описательная статистика**».

Другие виды эмпирических средних

Раздел 1. Оценки параметров

Средняя величина – это обобщающая характеристика единиц совокупности по какому–либо варьирующему признаку.

Средние величины позволяют сравнивать уровни одного и того же признака в различных совокупностях и находить причины этих расхождений.

Абсолютная величина, характеризующая уровень признака отдельной единицы совокупности, не позволяет сравнивать значения признака у единиц, относящихся к разным совокупностям.

Сравнивать можно лишь средние показатели.

Раздел 1. Оценки параметров

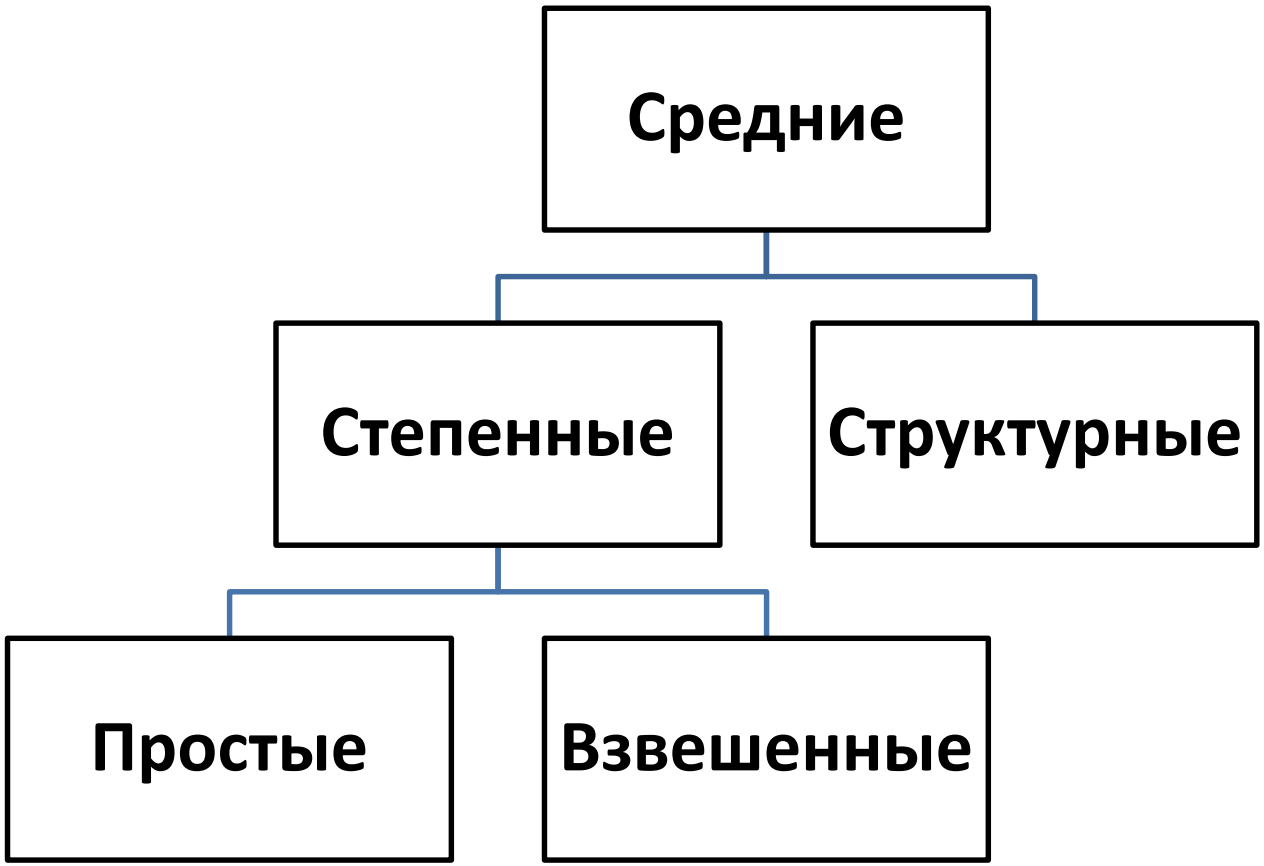
Средняя, рассчитываемая для совокупностей, состоящих из качественно **однородных единиц**, называется **типической средней**.

Средние величины, используемые в качестве характеристик для **неоднородных совокупностей**, называются **системными средними**.

Средняя должна вычисляться для совокупностей, состоящих из достаточно большого числа единиц.

Соблюдение этого условия необходимо для того, чтобы вошел в силу закон больших чисел, в результате действия которого случайные отклонения индивидуальных величин от общей тенденции взаимно погашаются.

Раздел 1. Оценки параметров



Раздел 1. Оценки параметров

Важной характеристикой центра распределения, являются, так называемые **структурные средние** – **мода** (Mo) и **медиана** (Me).

В отличие от средней арифметической, на которую влияют все значения изучаемого признака x_i, структурные средние не зависят от крайних значений признака.

Потому они являются лучшей, чем среднее арифметическое, характеристикой центра распределения для рядов с неопределенными границами (например, для рядов с открытыми крайними границами интервалов)

Раздел 1. Оценки параметров

По определению **мода** это значение признака наиболее часто встречающееся в вариационном ряду.

А **медианой** называют значение признака, которое делит упорядоченную последовательность x_i на две равные по численности части. В итоге у одной половины единиц совокупности значение признака не превышает медианный, а у другого – превышает медианный уровень.

Для дискретного ряда мода и медиана находятся непосредственно по определению.

Раздел 1. Оценки параметров

Пример. В таблице представлено распределение футбольных матчей по числу забитых за матч мячей обеими командами (Чемпионат мира 2018):

Число забитых мячей в матче	Число матчей
0	21
1	46
2	34
3	14
4	9

67 (= 21 + 46)

Определим, что типичным исходом футбольного матча во время Чемпионата мира 2018 г. был один забитый гол, т.е. **мода** равна 1 (наибольшее число матчей – 46, это матчи с одним забитым 21 голом, 1 – значение признака которое встречается чаще всего, это и есть мода).

Раздел 1. Оценки параметров

На практике встречаются многовершинные распределения, т.е. распределения в которых несколько максимумов частот, **несколько мод**. Наличие нескольких вершин (нескольких мод) является признаком того, что изучаемая совокупность состоит из неоднородных, с точки зрения изучаемого признака, единиц.

Раздел 1. Оценки параметров

Пример. Изучая спрос на мероприятия, проводимые на детских праздниках, было получено многовершинное распределение, распределение с двумя модами.

Наиболее предпочитаемая опция	Количество респондентов, чел.
Фокусы	57
Клоуны	26
Театр кукол	19
Костюмированный балл	22
Квест по мотивам мультфильма / фильма	21
Конкурсы и музыкальная программа	14
Мастер классы (рисование, кулинария)	34
Участие в дне рождения слона	57

Раздел 1. Оценки параметров

Пример. Для списка банков Санкт-Петербурга, упорядоченных по размеру собственного капитала, **медианным** банком будет банк «Петровский», а **медианой** – 268 млн руб., именно это значение признака делит упорядоченную (в данном случае по возрастанию) последовательность значений (169, 237, 268, 290, 1 007) на две равные по численности части (169, 237) и (290, 1 007)

Название	Собственный капитал, млн. руб.
Балтонэксимбанк	169
Банк «Санкт – Петербург»	237
Петровский	268
Балтийский	290
Промстройбанк	1 007

Раздел 1. Оценки параметров

Степенные средние

Пусть $x_1, \dots, x_n$ - все значения признака в совокупности объёма n.

В описательной статистике для вычисления среднего значения положительного признака X помимо уже рассмотренного среднего арифметического и медианы также применяются следующие эмпирические характеристики:

среднее геометрическое

$$G(X) = \sqrt[n]{x_1 \cdot \dots \cdot x_n};$$

среднее гармоническое

$$H(X) = \frac{n}{\frac{1}{x_1} + \dots + \frac{1}{x_n}};$$

среднее степенное $\sqrt[a]{\overline{X^a}}$ с показателем $a \neq 0$

$$\sqrt[a]{\overline{X^a}} = \left(\frac{x_1^a + \dots + x_n^a}{n}\right)^{1/a}.$$

Раздел 1. Оценки параметров

При использовании одних и тех же исходных данных, чем больше показатель степени a в вышеприведенной формуле, тем больше значение средней величины:

$$\bar{x}_{гар} \leq \bar{x}_{г} \leq \bar{x}_{ар} \leq \bar{x}_{кв} \leq \bar{x}_{куб}$$

Это свойство степенных средних возрастать с повышением показателя степени определяющей функции называется

правилом мажорантности средних.

Раздел 1. Оценки параметров

Из **свойств мажорантности** следует, что выбор формулы для расчета средней не может быть произвольным.

Он должен основываться на смысловом содержании исходных данных и на условиях применения конкретной формулы для вычисления средней.

Раздел 1. Оценки параметров

Пример. Студент получил в течение семестра две оценки: 3 и 2. Требуется рассчитать степенные средние всех видов и с их помощью проверить действие правила мажорантности.

Решение:

$$\bar{x}_{\text{ар}} = \frac{3+2}{2} = 2{,}5;$$

$$\bar{x}_{\text{кв}} = \sqrt{\overline{X^2}} = \left(\frac{x_1^2 + x_2^2}{2}\right)^{1/2} = \sqrt{\frac{3^2 + 2^2}{2}} = 2{,}55;$$

$$\bar{x}_{\text{гар}} = H(X) = \frac{2}{\frac{1}{x_1} + \frac{1}{x_2}} = \frac{2}{\frac{1}{2} + \frac{1}{3}} = 2{,}41;$$

$$\bar{x}_{\text{геом}} = G(X) = \sqrt{x_1 \cdot x_2} = \sqrt{3 \cdot 2} = 2{,}45.$$

$$2{,}55 > 2{,}50 > 2{,}45 > 2{,}41$$

Вывод.

Раздел 1. Оценки параметров

Для вычисления **степенных средних** по *конкретной выборке* в **Microsoft Excel** используются функции

$G(X) = \sqrt[n]{x_1 \cdot \ldots \cdot x_n} =$ **СРГЕОМ**(<выборка $x_1, x_2, \ldots, x_n$ >)

- среднее геометрическое

$H(X) = \frac{n}{\frac{1}{x_1}+\cdots+\frac{1}{x_n}} =$ **СРГАРМ**(<выборка $x_1, x_2, \ldots, x_n$ >)

- среднее гармоническое .

Раздел 1. Оценки параметров

В **Python** вычисляет **среднее геометрическое** $(G(X) = \sqrt[n]{x_1 \cdot \ \ldots \cdot x_n})$ значений элементов массива функция из библиотеки **scipy.stats**

gmean( )

среднее гармоническое $(H(X) = \frac{n}{\frac{1}{x_1}+\cdots+\frac{1}{x_n}})$ значений элементов массива функция из библиотеки **scipy.stats**

hmean( )

Раздел 1. Оценки параметров

Случаи использования различных средних величин

1. **Средняя арифметическая простая** используется в том случае, если числитель и знаменатель исследуемой системы приведен в исходных данных.
2. **Средняя арифметическая взвешенная** используется в том случае, если знаменатель исследуемой системы (логической схемы) известен, а числитель – нет.
3. **Средняя гармоническая** используется в том случае, если числитель исследуемой схемы приведен в исходных данных, а знаменатель – нет.
4. **Средняя квадратическая** используется только лишь при определении показателей вариации.
5. **Средняя геометрическая** используется только лишь при расчете среденегодового темпа роста.
6. **Структурные средние** используются, преимущественно при определении спроса и предложения.

Раздел 1. Оценки параметров

Следует учесть, что разные виды средних величин на одном и том же исходном материале имеют неодинаковые значения.

Раздел 1. Оценки параметров

Пример (средняя арифметическая простая). Измерив рост всех студентов в группе, получили следующие данные: 1,64 м, 1,86 м, 1,72 м, 1,95 м, 1,76 м, 1,65 м, 1,79 м, 1,82 м, 1,92 м. Найти средний рост студентов в группе.

Решение. Для определения среднего роста студентов в группе необходимо суммарный рост всех студентов в группе разделить на количество студентов.

Всего в группе 9 студентов, обозначим рост каждого студента x_i , где i принимает значения от 1-го до 9-ти.

Тогда, средний рост определяется по формуле **средней арифметической простой**:

$$\bar{X} = \frac{\sum_{i=1}^{9} x_i}{9} = \frac{1{,}64 + 1{,}86 + 1{,}72 + 1{,}95 + 1{,}76 + 1{,}65 + 1{,}79 + 1{,}82 + 1{,}92}{9}$$

$$= \frac{16{,}11}{9} = 1{,}79 \text{ м}$$

Раздел 1. Оценки параметров

Пример (средняя арифметическая взвешенная). В течение учебного года первые четыре месяца студент не получал стипендию, следующие шесть месяцев размер стипендии составил 2,5 тыс. руб., в оставшиеся два месяца – 3,3 тыс. руб. Найти среднюю стипендию студента в рассматриваемом году.

Решение. Запишем данные о размере стипендии студента в виде таблицы:

Размер стипендии в месяц, тыс. руб. (x_i – значения осредняемого признака)	Число месяцев, в течение которых стипендия составляла данную сумму (f_i - частота, показывающая сколько раз в рассматриваемом периоде (год) встречается i-е значение осредняемого признака)
0	4
2,5	6
3,3	2

Раздел 1. Оценки параметров

Решение (продолжение).

$$\bar{X} = \frac{\sum xf}{\sum f} = \frac{0 \cdot 4 + 2{,}5 \cdot 6 + 3{,}3 \cdot 2}{4 + 6 + 2} = \frac{21{,}6}{12} = 1{,}8 \text{ тыс. руб.}$$

Размер стипендии в месяц, тыс. руб. (x_i – значения осредняемого признака)	Число месяцев, в течение которых стипендия составляла данную сумму (f_i - частота, показывающая сколько раз в рассматриваемом периоде (год) встречается i-е значение осредняемого признака)
0	4
2,5	6
3,3	2

Раздел 1. Оценки параметров

Пример (средняя квадратическая простая). Были проведены испытания точности спортивной винтовки, а именно, произведены 5 выстрелов, в каждом из них пуля отклонилась от цели на:

Номер испытания по порядку	Отклонение от цели, в мм
1	0 (точно в цель)
2	20
3	–16
4	0 (точно в цель)
5	–4

Найти среднюю величину отклонения от цели при стрельбе из спортивной винтовки.

Раздел 1. Оценки параметров

Решение. В данном случае суммируя все значения осредняемого признака (x_i – отклонение от цели при стрельбе) получаем нулевую сумму:

$$\sum_{i=1}^{5} x_i = 0 + 20 + (-16) + 0 - 4 = 0.$$

В этом случае, используя простую среднюю арифметическую для расчета среднего значения отклонения от цели, мы получили бы нулевой результат. Что неверно, так как по данным испытай нельзя сказать, что винтовка бьет без промаха.

Потому, для расчета средней величины отклонения от цели воспользуемся **формулой средней квадратической простой**:

$$\bar{X} = \sqrt{\frac{\sum x^2}{n}} = \sqrt{\frac{\sum_{i=1}^{5} x_i^2}{n}} = \sqrt{\frac{0^2 + 20^2 + (-16)^2 + 0^2 + (-4)^2}{5}} \approx 11{,}59 \text{ мм}$$

Характеристики связи двух признаков

Раздел 1. Оценки параметров

Пусть $x_i = X(\omega_i)$ и $y_i = Y(\omega_i)$, $\omega_i \in \Omega$ — значения признаков X и Y на совокупности $\Omega = \{\omega_1, \dots, \omega_n\}$.

Эмпирическая ковариация определяется формулой

$$\widehat{cov(X,Y)} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}).$$

Раздел 1. Оценки параметров

Таблицей сопряженности или **совместным частотным распределением** признаков X и Y называется следующая таблица:

$X \setminus Y$	$Y=y_1$	$Y=y_2$	…	$Y=y_j$	…	$Y=y_s$
$X=x_1$	n_{11}	n_{12}	…	n_{1j}	…	n_{1s}
$X=x_2$	n_{21}	n_{22}	…	n_{2j}	…	n_{2s}
…	…	…	…	…	…	…
$X=x_i$	n_{i1}	n_{i2}	…	n_{ij}	…	n_{is}
…	…	…	…	…	…	…
$X=x_r$	n_{r1}	n_{r2}	…	n_{rj}	…	n_{rs}

где n_{ij} — частота пары (x_i, x_j), т.е. число элементов $\omega \in \Omega$, для которых $X(\omega) = x_i$, а $Y(\omega) = y_j$.

Раздел 1. Оценки параметров

$X \setminus Y$	$Y=y_1$	$Y=y_2$	…	$Y=y_j$	…	$Y=y_s$	
$X=x_1$	n_{11}	n_{12}	…	n_{1j}	…	n_{1s}	n_{1x}
$X=x_2$	n_{21}	n_{22}	…	n_{2j}	…	n_{2s}	n_{2x}
…	…	…	…	…	…	…	…
$X=x_i$	n_{i1}	n_{i2}	…	n_{ij}	…	n_{is}	n_{ix}
…	…	…	…	…	…	…	…
$X=x_r$	n_{r1}	n_{r2}	…	n_{rj}	…	n_{rs}	n_{rx}
	n_{y1}	n_{y2}	…	n_{yj}	…	n_{ys}	n

Раздел 1. Оценки параметров

X	x_1	x_2	…	x_i	…	x_r	
n_{ix}	n_{1x}	n_{2x}	…	n_{ix}	…	n_{rx}	n

Y	y_1	y_2	…	y_j	…	y_s	
n_{yj}	n_{y1}	n_{y2}	…	n_{yj}	…	n_{ys}	n

Раздел 1. Оценки параметров

На основе **таблицы сопряженности** **эмпирическая ковариация** находится по формуле:

$$\widehat{cov(X,Y)} = \frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{s}(x_i - \bar{x})(y_i - \bar{y})n_{ij}.$$

$X \setminus Y$	$Y=y_1$	$Y=y_2$	…	$Y=y_j$	…	$Y=y_s$
$X=x_1$	n_{11}	n_{12}	…	n_{1j}	…	n_{1s}
$X=x_2$	n_{21}	n_{22}	…	n_{2j}	…	n_{2s}
…	…	…	…	…	…	…
$X=x_i$	n_{i1}	n_{i2}	…	n_{ij}	…	n_{is}
…	…	…	…	…	…	…
$X=x_r$	n_{r1}	n_{r2}	…	n_{rj}	…	n_{rs}

Раздел 1. Оценки параметров

Теорема. Для эмпирической ковариации справедлива формула

$$\widehat{cov(X,Y)} = \overline{XY} - \bar{X} \cdot \bar{Y},$$

$$\overline{XY} = \frac{\sum_{i=1}^{r}\sum_{j=1}^{s} n_{ij} X_i X_j}{n},$$

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{r} n_{ix} X_i, \quad \bar{Y} = \frac{1}{n}\sum_{i=1}^{s} n_{yj} Y_i.$$

$$\widehat{cov(X,Y)} = \frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{s} (x_i - \bar{x})(y_i - \bar{y}) n_{ij}$$

X	x_1	x_2	…	x_i	…	x_r	
n_{ix}	n_{1x}	n_{2x}	…	n_{ix}	…	n_{rx}	n

Y	y_1	y_2	…	y_j	…	y_s	
n_{yj}	n_{y1}	n_{y2}	…	n_{yj}	…	n_{ys}	n

$X \setminus Y$	$Y=y_1$	$Y=y_2$	…	$Y=y_j$	…	$Y=y_s$
$X=x_1$	n_{11}	n_{12}	…	n_{1j}	…	n_{1s}
$X=x_2$	n_{21}	n_{22}	…	n_{2j}	…	n_{2s}
…	…	…	…	…	…	…
$X=x_i$	n_{i1}	n_{i2}	…	n_{ij}	…	n_{is}
…	…	…	…	…	…	…
$X=x_r$	n_{r1}	n_{r2}	…	n_{rj}	…	n_{rs}

Раздел 1. Оценки параметров

Эмпирический коэффициент корреляции вычисляется по формуле:

$$\hat{\rho}_{XY} = \frac{\widehat{cov(X,Y)}}{\hat{\sigma}_X \hat{\sigma}_Y}.$$

Раздел 1. Оценки параметров

Пример. В совокупности 16 студентов определены два признака: X — оценка по математике и Y — оценка по иностранному языку. Совместное частотное распределение оценок задано таблицей:

$X \setminus Y$	$X=2$	$X=3$	$X=4$	$X=5$
$Y=3$	1	0	1	0
$Y=4$	2	4	4	2
$Y=5$	0	1	0	1

Требуется найти эмпирический коэффициент корреляции $\hat{\rho}_{XY}$.

Решение.

Раздел 1. Оценки параметров

Генеральная совокупность, выборка и основные способы организации выборки.

Раздел 1. Оценки параметров

Предположим, что из генеральной совокупности Ω объема N извлекается выборка $\widehat{\Omega}$ объема n.

Пусть X — некоторый **признак** на Ω.

Поскольку все элементы $\widehat{\Omega}$, независимо от вида выборки, являются также элементами Ω, признак X определен и на совокупности $\widehat{\Omega}$.

Обозначим $x_{01}, x_{02}, \ldots, x_{0N}$ значения признака X в генеральной совокупности и $X_1, X_2, \ldots, X_n$ — значения X в выборке.

Далее значения $x_{01}, x_{02}, \ldots, x_{0N}$ рассматриваются как числа, а $X_1, X_2, \ldots, X_n$ — как случайные величины (с.в.).

Раздел 1. Оценки параметров

Напомним, что **Генеральными (соответственно выборочными) характеристиками** признака X называют эмпирические характеристики признака X в генеральной (соответственно выборочной) совокупности.

Например:

$\bar{x}_0 = \frac{1}{N}(x_{01} + x_{02} + \ldots + x_{0N})$ — **генеральное среднее** (число);

$\bar{X} = \frac{1}{n}(X_1 + X_2 + \ldots + X_n)$ — **выборочное среднее** (с.в.);

$\widetilde{Var}(X) = \sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_{0i} - \bar{x}_0)^2$ — **генеральная дисперсия** (число);

$\widehat{Var}(X) = \hat{\sigma}_X^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$ — **выборочная дисперсия** (с.в.).

Раздел 1. Оценки параметров

Теорема 1. Пусть $X_1, X_2, \ldots, X_n$ — значения признака X в выборке, $\bar{x}_0$ — **генеральное среднее**, а $\sigma_X^2 = \widetilde{Var}(X)$ — **генеральная дисперсия**. Тогда для выборочного среднего $\bar{X}$ имеем:

- в случае **повторной или бесповторной выборки**

$$E(\bar{X}) = \bar{x}_0;$$

- в случае **повторной выборки**

$$\sigma_{\bar{X}}^2 = \widetilde{Var}(\bar{X}) = \frac{\widetilde{Var}(X)}{n} = \frac{\sigma_X^2}{n};$$

- в случае **бесповторной выборки**

$$\sigma_{\bar{X}}^2 = \widetilde{Var}(\bar{X}) = \frac{\widetilde{Var}(X)}{n} \cdot \frac{N-n}{N-1} = \frac{\sigma_X^2}{n} \cdot \frac{N-n}{N-1},$$

где N — объем генеральной совокупности.

Раздел 1. Оценки параметров

Следствие 1. Пусть $X_1, X_2, \ldots, X_n$; $Y_1, Y_2, \ldots, Y_n$ — значения признаков X и Y в выборочной совокупности объема n, $\widetilde{Cov}(X,Y)$ — **ковариация признаков X и Y в генеральной совокупности** объема N. Тогда для ковариации выборочных средних справедливы соотношения:

- в случае **повторной выборки**

$$\widetilde{Cov}(\bar{X},\bar{Y}) = \frac{\widetilde{Cov}(X,Y)}{n};$$

- в случае **бесповторной выборки**

$$\widetilde{Cov}(\bar{X},\bar{Y}) = \frac{\widetilde{Cov}(X,Y)}{n} \cdot \frac{N-n}{N-1},$$

где N — объем генеральной совокупности.

Доказательство следствия.

Действительно, с учетом равенства $\overline{X \pm Y} = \bar{X} \pm \bar{Y}$ в случае **повторной выборки** имеем

$$\widetilde{Cov}(\bar{X}, \bar{Y}) = \frac{1}{4}\left(\widetilde{Var}(\bar{X} + \bar{Y}) - \widetilde{Var}(\bar{X} - \bar{Y})\right) =$$

$$= \frac{1}{4n}\left(\widetilde{Var}(X + Y) - \widetilde{Var}(X - Y)\right) = \frac{\widetilde{Cov}(X, Y)}{n}.$$

Для **бесповторной выборки** доказательство аналогично.

□

Формула для дисперсии суммы произвольных с.в.

Дисперсия суммы произвольных (зависимых или независимых) с.в. X и Y рассчитывается по формуле

$$\boldsymbol{Var}(X + Y) = \boldsymbol{Var}(X) + \boldsymbol{Var}(Y) + 2cov(X, Y).$$

Раздел 1. Оценки параметров

Утверждение теоремы можно распространить и на некоторые другие моменты выборочного среднего $\bar{X}$.

Следующая теорема аналог этой теоремы для третьего центрального момента.

Раздел 1. Оценки параметров

Теорема 2. Пусть $X_1, X_2, \ldots, X_n$ — значения признака X на случайной выборке объёма n, полученной из генеральной совокупности объёма $N \geq 2n$; $\mu_3(X)$ — **генеральный центральный момент третьего порядка**. Тогда для выборочного среднего $\bar{X}$ имеем:

- в случае **повторной выборки**

$$\mu_3(\bar{X}) = \frac{\mu_3(X)}{n^2};$$

- в случае **бесповторной выборки**

$$\mu_3(\bar{X}) = \frac{\mu_3(X)}{n^2} \cdot \frac{N-n}{N-1} \cdot \frac{N-2n}{N-2},$$

где N — объем генеральной совокупности.

Раздел 1. Оценки параметров

Пример. Признак $X(k)$ задан на множестве

$$\Omega = \{1, 2, \ldots, 10\}$$

следующей таблицей:

$\boldsymbol{k}$	1	2	3	4	5	6	7	8	9	10
$\boldsymbol{X(k)}$	1	4	3	2	2	1	4	2	4	2

Из Ω извлекается случайная **бесповторная выборка** объема 5. Найдите математическое ожидание и дисперсию среднего значения $\bar{X}$ признака X в выборке.

Решение.

Раздел 1. Оценки параметров

Пример. Признак $X(k)$ задан на множестве

$$\Omega = \{1, 2, \ldots, 10\}$$

следующей таблицей:

$\boldsymbol{k}$	1	2	3	4	5	6	7	8	9	10
$\boldsymbol{X(k)}$	3	3	3	3	1	3	3	2	1	2

Из Ω извлекается случайная повторная выборка объема 5. Найдите математическое ожидание и дисперсию среднего значения $\bar{X}$ признака X в выборке.

Решение.

Раздел 1. Оценки параметров

Пример. Итоговое распределение баллов на некотором письменном экзамене задано таблицей

Оценка работы	2	3	4	5
Число работ	10	18	23	39

Работы проверяли 6 преподавателей, которые разделили все работы между собой поровну случайным образом. Предполагая независимость оценки от личности проверяющего, найдите математическое ожидание и дисперсию среднего балла по результатам одного преподавателя.

Решение.

Раздел 1. Оценки параметров

Пример. Две игральные кости, красная и синяя, подбрасываются до тех пор, пока не выпадет 19 различных с учетом цвета комбинаций очков. Пусть S— сумма очков на красной и синей кости в i-той комбинации, $\bar{S}$— среднее арифметическое всех этих сумм, $i = 1,\ldots,19$. Найдите математическое ожидание и дисперсию среднего значения $\bar{S}$.

Решение.

Раздел 1. Оценки параметров

Пример. Две игральные кости, красная и синяя, подбрасываются до тех пор, пока не выпадет 10 различных (с учетом цвета) комбинаций очков. Пусть R_i – число очков на красной кости, а B_i – число очков на синей кости в комбинации с номером i. Случайные величины Xi задаются соотношениями: $X_i = 7R_i - 5B_i, i = 1, \dots, 10$. Среднее арифметическое этих величин обозначим $\bar{X} = \frac{1}{10}\sum X_i$.

Требуется найти:

1) математическое ожидание $E(\bar{X})$;

2) стандартное отклонение $\sigma(\bar{X})$.

Решение.

Раздел 1. Оценки параметров

Пример. Имеется 10 пронумерованных монет. Монеты подбрасываются до тех пор, пока не выпадет 20 различных (с учетом номера монеты) комбинаций орел-решка. Пусть X_i – число орлов в комбинации с номером i; а $\bar{X} = \frac{1}{20}\sum X_i$ – среднее число орлов в полученных таким образом комбинациях. Требуется найти:

1) математическое ожидание $E(\bar{X})$;

2) дисперсию $Var(\bar{X})$.

Решение.

Раздел 1. Оценки параметров

Пример. Значения признаков X и Y заданы на множестве $\Omega = \{1, 2, \ldots, 115\}$ таблицей частот

	$\boldsymbol{Y = 7}$	$\boldsymbol{Y = 8}$	$\boldsymbol{Y = 10}$
$\boldsymbol{X = 250}$	23	20	15
$\boldsymbol{X = 460}$	15	17	25

Из Ω **без возвращения** извлекаются 9 элементов. Пусть $\overline{X}$ и $\overline{Y}$ — средние значения признаков в выборочной совокупности. Найдите $\widetilde{Cov}(\overline{X}, \overline{Y})$.

Решение.

Раздел 1. Оценки параметров

Пример. Значения признаков X и Y заданы на множестве $\Omega = \{1, 2, \ldots, 2000\}$ таблицей частот

	$\boldsymbol{Y = 2}$	$\boldsymbol{Y = 4}$	$\boldsymbol{Y = 6}$
$\boldsymbol{X = 7}$	100	400	200
$\boldsymbol{X = 10}$	300	100	900

Из Ω **с возвращением** извлекаются 800 элементов. Пусть $\overline{X}$ и $\overline{Y}$ — средние значения признаков в выборочной совокупности. Найдите $\widetilde{Cov}(\overline{X}, \overline{Y})$.

Решение.

Теория вероятностей и математическая статистика

Конец лекции