Основы математической статистики

Математическая статистика — раздел математики, в котором изучаются методы сбора, систематизации и обработки результатов наблюдений массовых случайных явлений для определения существующих закономерностей.

Задачи математической статистики:

- указать способы сбора и группировки статистических сведений, полученных в результате наблюдений или в результате специально поставленных экспериментов;
- разработать методы анализа статистических данных в зависимости от целей исследования.

Результаты исследования статистических данных методами математической статистики используются для принятия решений в задачах планирования, прогнозирования, организации производства, при контроле качества продукции.

Основные определения

- *Генеральная совокупность* — вся совокупность объектов, подлежащих изучению.

  Примеры генеральных совокупностей: население страны, месячная продукция завода, популяция рыб, живущих в водоёме и т.д.

  Иногда проводится *сплошное* обследование, тогда изучаются все объекты генеральной совокупности (пример сплошного обследования: перепись населения). Если интересующая совокупность объектов слишком многочисленна, или объекты труднодоступны, или имеются другие причины, не позволяющие изучить все объекты, прибегают к изучению какой-то части объектов.
- *Выборочной совокупностью или выборкой* называется совокупность объектов, отобранных случайным образом из генеральной совокупности.
- При составлении выборки можно поступать двумя способами: после того, как объект отобран и над ним произведено наблюдение, он может быть возвращен либо не возвращен в генеральную совокупность. Таким образом, выборки подразделяют на *повторные и бесповторные.*

- *Ранжированный ряд* — последовательность всех значений совокупности, записанных в возрастающем порядке.
- *Объём совокупности* — количество объектов в совокупности.

Обозначение: n

$n_г$ — объем генеральной совокупности

$n_в$ — объем выборочной совокупности

Функция Excel для расчета объема совокупности:

=СЧЁТ(ссылка на ранжированный ряд)

=СЧЁТЗ(ссылка на ранжированный ряд)

- Наблюдаемые значения называются *вариантами* — x_i

  n_i показывает сколько раз встречается соответствующий вариант в ряде наблюдений и называется *частотой*.

$$\sum n_i = n$$

Сумма частот равна объему совокупности

- *Относительной частотой* называется отношение частоты к объему *совокупности* $W_i = \frac{n_i}{n}$

- *Статистическим распределением совокупности или статистическим рядом* называют перечень вариант и соответствующих им частот или относительных частот.
- *Размахом ряда* называется разность между наибольшим и наименьшим значениями вариант.

$$R = x_{max} - x_{min}$$

Задача № 1: Известны зарплаты сотрудников фирмы (17 человек). Составить ранжированный ряд, таблицу частот, построить полигон частот.

Ранжированный ряд *— последовательность всех значений наблюдаемого признака, записанных в возрастающем порядке.*

Ранжированный ряд: 10, 10, 20, 20, 20, 25, 25, 25, 25, 25, 35, 35, 35, 60, 60, 90, 100.

№	Фамилия	Зарплата тыс. руб.
1	Андреев	25
2	Васильева	25
3	Гордеев	100
4	Шилов	60
5	Зорин	35
6	Иванова	10
7	Карпов	35
8	Николаева	20
9	Никонов	90
10	Орлов	25
11	Петрова	10
12	Рыбкин	25
13	Сидоров	20
14	Смирнова	20
15	Соколова	25
16	Фадеев	60
17	Федорова	35

Ранжированный ряд: 10, 10, 20, 20, 20, 25, 25, 25, 25, 25, 35, 35, 35, 60, 60, 90, 100.

Объем совокупности *(кол-во значений в совокупности)* **n** = 17

Таблица частот

x_i *(варианты)*	10	20	25	35	60	90	100	**Проверка** *(сумма $n_i = n$)*
n_i *(частоты)*	2	3	5	3	2	1	1	17

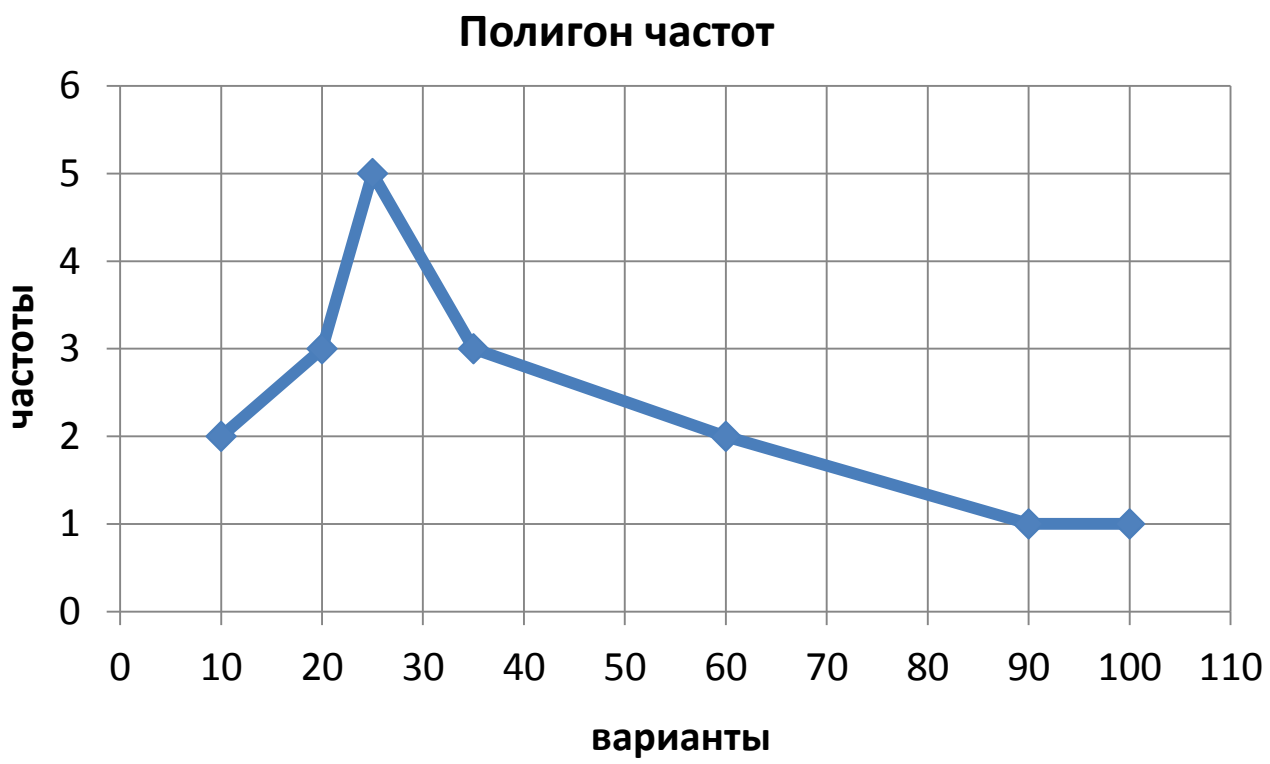
Таблица относительных частот

x_i *(варианты)*	10	20	25	35	60	90	100	**Проверка** *(сумма $W_i = 1$)*
$W_i=n_i/n$ *(относительные частоты)*	0,12	0,18	0,29	0,18	0,12	0,06	0,06	1

Полигон частот или относительных частот представляет собой ломаную, отрезки которой соединяют точки (x_i, n_i) или (x_i, W_i)

Таблица частот

x_i *(варианты)*	10	20	25	35	60	90	100	**Проверка** *(сумма $n_i = n$)*
n_i *(частоты)*	2	3	5	3	2	1	1	17



Характеристики ранжированного ряда

- Среднее арифметическое значение (или просто среднее)— центр рассеивания значений наблюдаемого признака.

$$\overline{x} = \frac{\sum_{i=1}^{k} x_i n_i}{n} = \frac{x_1 n_1 + \cdots + x_k n_k}{n}$$

Среднее бывает генеральным $\overline{x}_{\text{Г}}$, если оно найдено для генеральной совокупности, и выборочным $\overline{x}_{\text{В}}$, если для выборочной.

Функция Excel для расчета среднего значения:

=СРЗНАЧ(ссылка на ранжированный ряд)

Характеристики ранжированного ряда

- Медиана

Медиана позволяет уточнить информацию о ряде данных, которую даёт среднее арифметическое значение.

Медианой ранжированного ряда чисел с **нечетным** числом членов называется число, записанное **посередине**.

Медианой ранжированного ряда чисел с **четным** числом членов называется **среднее арифметическое двух чисел, записанных посередине**.

Обозначение: M_e

Вычисление медианы:

1) 2, 3, 4, 4, **5,** 6, 7, 8, 9 n=9 *(нечетное)* $M_e = 5$

2) 0, 1, 1, 2, **3, 4,** 5, 5, 6, 7 n=10 *(четное)* $M_e = \frac{3+4}{2} = 3{,}5$

Функция Excel для расчета медианы:

=МЕДИАНА(ссылка на ранжированный ряд)

Характеристики ранжированного ряда

- Мода — вариант (значение наблюдаемого признака), которому соответствует наибольшая частота.

Моду находят тогда, когда хотят выявить типичный показатель.

Обозначение: M_o

Вычисление моды:

1) Ранжированный ряд : 1, **2, 2, 2, 2,** 4, 5, 5, 7

Таблица частот

$\mathbf{x_i}$ *(варианты)*	1	2	4	5	7
$\mathbf{n_i}$ *(частоты)*	1	4	1	2	1

$M_o = 2$ *(частота данного варианта максимальна и равна 4)*

Функция Excel для расчета моды:

Excel 2007: =МОДА(ссылка на ранжированный ряд)

Excel 2010: =МОДА.ОДН(ссылка на ранжированный ряд)

Характеристики ранжированного ряда

2. В случае если все значения в совокупности встречаются **одинаково часто**, считают, что совокупность не имеет моды.

   Ранжированный ряд **5, 5, 6, 6, 7, 7** не имеет моды.

3. Когда два **соседних** значения ранжированного ряда имеют одинаковую частоту и их частота больше частот любых других значений, мода вычисляется как среднее арифметическое этих двух значений.

   Ранжированный ряд: **1, 2, 2, 2, 5, 5, 5, 6.** Частоты соседних значений 2 и 5 равны друг другу ($n_2=n_5=3$), следовательно $M_o = \frac{2+5}{2} = 3{,}5$

4. Когда два **не соседних** значения ранжированного ряда имеют одинаковую частоту и их частота больше частот любых других значений, то выделяют две моды.

   Ранжированный ряд **10, 11, 11, 11, 12, 13, 14, 14, 14, 17** имеет две моды: 11 и 14 (совокупность бимодальная).

   Могут существовать мультимодальные совокупности, имеющие более двух мод.

Характеристики ранжированного ряда

- Дисперсия характеризует разброс значений наблюдаемого признака вокруг среднего значения.

$$D = \frac{\sum_{i=1}^{k} n_i (x_i - \overline{x})^2}{n}$$

Дисперсия бывает генеральной $D_Г$, если она найдено для генеральной совокупности, и выборочной $D_В$, если для выборочной.

Функция Excel для расчета дисперсии:

Excel 2007: =ДИСПР(ссылка на ранжированный ряд)

Excel 2010: =ДИСП.Г(ссылка на ранжированный ряд)

Характеристики ранжированного ряда

- Стандартное отклонение (или стандарт)

  Для того, чтобы приблизить размерность дисперсии к размерности наблюдаемого признака, применяют операцию извлечения квадратного корня из дисперсии. Полученную величину называют генеральным (выборочным) стандартом.

  Обозначение: σ (сигма)

  Выборочный стандарт: $\boldsymbol{\sigma}_{\text{в}} = \sqrt{\boldsymbol{D}_{\text{в}}}$

  Генеральный стандарт : $\boldsymbol{\sigma}_{\text{г}} = \sqrt{\boldsymbol{D}_{\text{г}}}$

Функция Excel для расчета стандартного отклонения:

Excel 2007: =СТАНДОТКЛОНП(ссылка на ранжированный ряд)

Excel 2010: =СТАНДОТКЛОН.Г(ссылка на ранжированный ряд)

Задача № 1 (продолжение): Известны зарплаты сотрудников фирмы. Вычислить характеристики ранжированного ряда.

Ранжированный ряд: 10, 10, 20, 20, 20, 25, 25, 25, 25, 25, 35, 35, 35, 60, 60, 90, 100.

Объем совокупности *(кол-во значений в совокупности)* **n** = 17

Таблица частот

x_i *(варианты)*	10	20	25	35	60	90	100	**Проверка** *(сумма n_i = n)*
n_i *(частоты)*	2	3	5	3	2	1	1	17

$$\bar{x} = \frac{x_1 n_1 + \cdots + x_k n_k}{n} = \frac{10 \cdot 2 + 20 \cdot 3 + 25 \cdot 5 + 35 \cdot 3 + 60 \cdot 2 + 90 + 100}{17} \approx 36$$

$$D = \frac{\sum_{i=1}^{k} n_i (x_i - \bar{x})^2}{n}$$

$$D = \frac{2 \cdot (10-36)^2 + 3 \cdot (20-36)^2 + 5 \cdot (25-36)^2 + 3 \cdot (35-36)^2 + 2 \cdot (60-36)^2 + (90-36)^2 + (100-36)^2}{17} \approx 640$$

$$\sigma = \sqrt{D} = \sqrt{640} \approx 25$$

$$R = x_{max} - x_{min} = 100 - 10 = 90$$

$$M_o = 25$$

$$M_e = 25$$

Дискретные и непрерывные случайные величины

- *Дискретной (ДСВ)* называют случайную величину, которая принимает только определенные отдельные друг от друга значения, которые можно установить и перечислить. Число возможных значений дискретной случайной величины может быть конечным или бесконечным.

  Пример: Число студентов в аудитории — может быть только целым положительным числом: 0, 1, 2, …, 20. Эти значения отделены одно от другого промежутками, в которых нет возможных значений случайной величины. Таким образом, в этом примере случайная величина принимает отдельные, изолированные возможные значения.

Дискретные и непрерывные случайные величины

- *Непрерывной* называют случайную величину, которая может принимать любые значения внутри некоторого интервала, который может иметь резко выраженные границы, а если они не известны, то считают, что значения случайной величины лежат в интервале $(-\infty; \infty)$. Очевидно, что число возможных значений непрерывной случайной величины бесконечно.

Примеры: температура, давление, вес и рост людей.

В данных примера случайная величина могла принять любое из значений промежутка. Здесь нельзя отделить одно возможное значение от другого промежутком, не содержащим возможных значений случайной величины.

При больших объёмах совокупности ($n>50$) или когда случайная величина является непрерывной составляют *интервальный статистический ряд*.

Отрезок, содержащий совокупность, разбивают на несколько промежутков, как правило, одинаковой длины. Далее подсчитывают частоты попадания наблюдаемых значений в промежутки и относительные частоты.

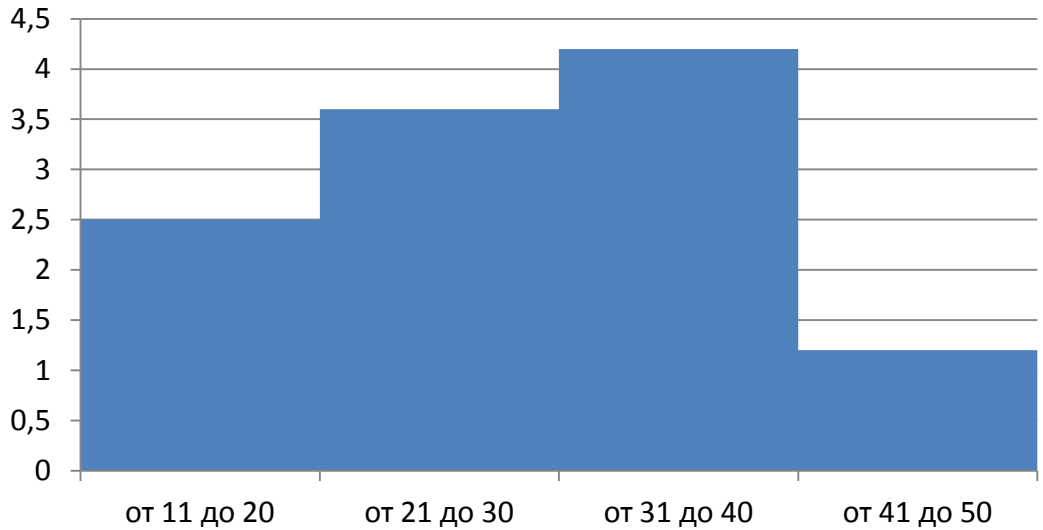
В случае непрерывного признака строят *гистограмму частот* — ступенчатую фигуру, состоящую из прямоугольников, основаниями которых служат частичные интервалы длиной h, а высотами n_i/h (данное отношение определяет *плотность частоты*). Площадь гистограммы частот равна сумме всех частот, т.е. объему совокупности.

Гистограмма относительных частот — ступенчатая фигура, состоящая из прямоугольников, основаниями которых служат частичные интервалы длиной h, а высотами W_i/h. Площадь гистограммы относительных частот равна сумме всех относительных частот, т.е. 1.

Задача № 2: Известны зарплаты сотрудников фирмы (115 человек). Составить интервальную таблицу частот, построить гистограмму частот.

Длина интервала h=10

интервальная таблица частот

Частичные интервалы (зарплата)	от 11 до 20 тыс. руб.	от 21 до 30 тыс. руб.	от 31 до 40 тыс. руб.	от 41 до 50 тыс. руб.	**Проверка** *(сумма $n_i = n$)*
Интервальные частоты (кол-во сотруд.)	25	36	42	12	115
Плотность частоты n_i/h	2,5	3,6	4,2	1,2	

Гистограмма частот