# Описательная статистика

# Содержание

1. Генеральная совокупность и выборка
2. Ранжирование данных эксперимента
3. Статистические ряды распределения.
   - Атрибутивный ряд распределения.
   - Вариационный ряд распределения.
4. Числовые характеристики случайных величин
5. Полигон частот
6. Проверка гипотез

# Генеральная совокупность и выборка

**Генеральная совокупность** *(популяция)* ***W*** *— полный набор объектов* ***w****, с которыми связана данная проблема. Эти объекты могут быть людьми, животными, изделиями и так далее. С каждым объектом связана величина (или величины), называемая* ***исследуемым признаком*** *(****xi****).*

**Основной целью статистического анализа** является выяснение некоторых свойств рассматриваемой генеральной совокупности.

Если генеральная совокупность конечна, то наилучшая процедура — рассмотрение каждого ее элемента. Однако в большинстве интересных задач используются либо бесконечные генеральные совокупности, либо конечные, но трудно обозримые. В этой ситуации необходимо отобрать из генеральной совокупности подмножество из ***n*** элементов, называемое ***выборкой объема n***, исследовать его свойства, а затем обобщить эти результаты на всю генеральную совокупность. Это обобщение называется **статистическим выводом**.

- **Признак**- свойство генеральной совокупности
- **Вариантами** считаются отдельные значения признака, которые он принимает в вариационном ряду.
- **Выборка** – это часть объектов, выбранных из генеральной совокупности для исследования**. Объем выборки -** число элементов в ней. Менее 30 элементов – малая выборка, выборки могут быть в 100 и более элементов в зависимости от задачи

## Пример

| **Генеральная совокупность** | **Все студенты МГТУ** | **Жилые дома г. Магнитогорска** |
|---|---|---|
| Выборка | Студенты 1 курса ИГО | Дома 127 микрорайона |
| Признак | Рост | Площадь квартир |
| Вариант | 1,75 | 64 $м^2$ |

# Статистические таблицы

**Статистическая таблица** – это особый способ краткой и наглядной записи сведений об изучаемых общественных явлениях.

Статистическая таблица позволяет охватить материалы статистической сводки в целом, она также является системой мыслей об исследуемом объекте, излагаемых цифрами на основе определенного порядка в расположении систематизированной информации.

# Частота распределения

**Частота** — количество элементов совокупности, которые имеют данное значение признака.

**Частость** — отношение **частоты** к общему количеству исследуемых элементов, т.е. объему совокупности. 70/120*100%=58%

| Категория | Частота | Частость, в % |
|---|---|---|
| Рабочие | 70 | 58% |
| Служащие | 20 | 17% |
| Инженеры | 15 | 13% |
| Прочие | 15 | 13% |
| Итого | 120 | 100% |

# Частота и частость для дискретного ряда

Контрольную по математике писали 15 человек. Из них двое получили 2, четверо – 3, пятеро – 4, четверо – 5. Построить дискретный вариационный ряд. Определить частости для каждой оценки.

| Оценки | Частота | Частость, в % |
|---|---|---|
| 2 | 2 | =2/15*100=13,3% |
| 3 | 4 | =4/15*100=26,7% |
| 4 | 5 | =5/15*100=33,3% |
| 5 | 4 | =4/15*100=26,7% |
| Итого | 15 | 100% |

**=Частота(массив данных; массив признаков)**

**Статистический ряд распределения** – это упорядоченное распределение значений признака или свойства генеральной совокупности на группы по определённому варьирующему признаку.

В зависимости от типа признака, положенного в основу образования ряда распределения, различают:

- **атрибутивные** ряды (качественный признак)
- **вариационные** ряды (количественный признак)

- **Атрибутивными** — называют ряды распределения, построенные по качественными признакам. Ряд распределения принято оформлять в виде таблиц.

| Образование | Количество работников | |
|---|---|---|
| | Абсолютное | В процентах |
| Среднее | 50 | 38,5 |
| Среднее специальное | 35 | 26,9 |
| Неполное высшее | 25 | 19,2 |
| Высшее | 20 | 15,4 |
| Всего работников | 130 | 100% |

Таблица 1 - Распределение видов юридической помощи, оказанной адвокатами гражданам одного из регионов РФ.

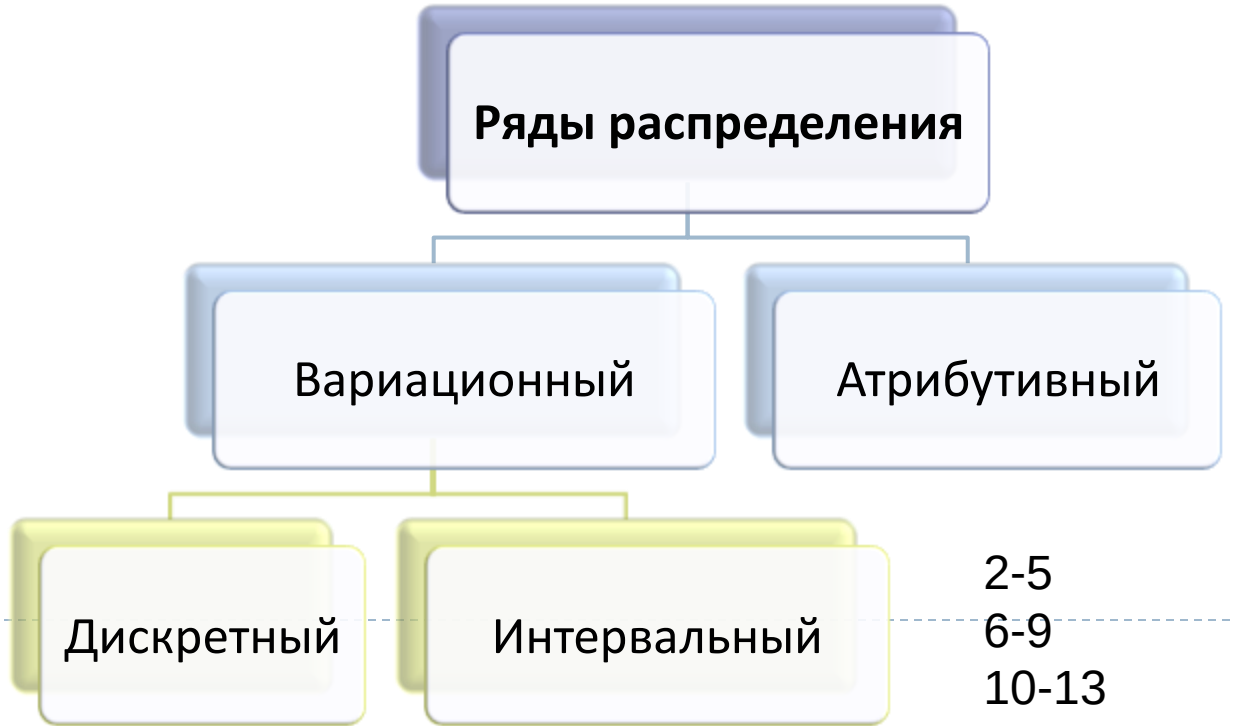
- **Вариационными** рядами называют ряды распределения, построенные по количественному признаку. Любой вариационный ряд состоит из двух элементов: вариантов и частот.

| **Варианты** | **Частоты** |
|---|---|

| Балл (оценка), $x_i$ | Число студентов, $f_i$ | Удельный вес в общей численности студентов, % |
|---|---|---|
| 5 | 34 | 29,8 |
| 4 | 37 | 32,5 |
| 3 | 33 | 28,9 |
| 2 | 10 | 8,8 |
| ИТОГО | 114 | 100 |

# Виды вариационных рядов

- В зависимости от характера вариации признака различают **дискретные** и **интервальные** вариационные ряды.
- Вариационный ряд называется **дискретным**, если любые его варианты отличаются на постоянную величину, и **интервальным**, если варианты могут отличаться один от другого на сколь угодно малую величину.




2
3
4

2-5
6-9
10-13

# Ряд распределения ДСВ

Дискретные величины могут принимать только конечное или счетное множество определенных значений.

Например, число очков при бросании игральной кости; число телефонных звонков, поступающих конкретному абоненту в течение суток.

Такие величины удобнее характеризовать указанием возможных значений и их вероятностей.

| Значения $X_i$ | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|---|
| Вероятности $P(x_i)$ | 0 | 1/6 | 1/6 | 1/6 | 1/6 | 1/6 | 1/6 |
| Кумулятивная вероятность | 0 | 1/6 | 2/6 | 3/6 | 4/6 | 5/6 | 1 |

# Пример дискретного вариационного ряда:

Таблица 2 - Распределение семей по числу занимаемых комнат в квартирах в 1989 г. в РФ.

| N П/п | Группы семей, проживающих в квартирах с числом комнат | Число семей | |
|---|---|---|---|
| | | всего, тыс.ед. | в % к итогу |
| 1 | 1 | 4064 | 16,3 |
| 2 | 2 | 12399 | 49,7 |
| 3 | 3 | 7659 | 30,7 |
| 4 | 4 и более | 832 | 3,3 |
| ВСЕГО | | 24954 | 100,00 |

В первой колонке таблицы представлены варианты дискретного вариационного ряда, во второй – помещены частоты вариационного ряда, в третьей – показатели частости.

# Интервальный вариационный ряд

**Интервальным вариационным рядом** называется упорядоченная совокупность интервалов варьирования значений случайной величины с соответствующими частотами или частостями попаданий в каждый из них значений величины.

| Кредитная ставка, % | Количество банков, ед. (частоты) | Накопленные частоты |
|---|---|---|
| 12,0–14,0 | 5 | 5 |
| 14,0–16,0 | 9 | 14 |
| 16,0–18,0 | 4 | 18 |
| 18,0–20,0 | 15 | 33 |
| 20,0–22,0 | 11 | 44 |
| 22,0–24,0 | 2 | 46 |
| 24,0–26,0 | 4 | 50 |
| Итого | 50 | — |

# Числовые характеристики статистического распределения

# Производство зерна в России

| **Показатель** | 2000 | 2001 | 2002 | 2003 | 2004 | 2005 | 2006 |
|---|---|---|---|---|---|---|---|
| Произ--во зерновых, млн. т | 65,5 | 85,2 | 86,6 | 67,2 | 78,1 | 78,2 | 78,6 |
| Урожайность, ц/га | 15,6 | 19,4 | 19,6 | 17,8 | 18,8 | 18,5 | 18,9 |
| Произ-во пшеницы, млн. т | 34,5 | 47,0 | 50,6 | 34,1 | 45,4 | 47,7 | 45,0 |

Найти наибольшее, наименьшее значение и размах (R):

а)произ-ва зерновых max= 86,6 min= 65,5 R= 21,1.

б)произ-ва пшеницы max= 50,6 min= 34,1 R= 16,5.

в)урожайности max =19,6 min = 15,6 R = 4.

# Числовые характеристики статистического распределения

В качестве характеристик измеримого признака вместо исходных значений величин или таблиц их частот используют числовые характеристики, называемые также **статистическими мерами.**

**Среднее арифметическое** - это отношение суммы чисел к их количеству

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

**Среднее арифметическое для дискретного ряда**

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

**Среднее арифметическое для непрерывного ряда**

При расчете средней арифметической в интервальном ряду за значение варианты принимается середина интервала. Середина интервала вычисляется как среднее арифметическое его границ

# Числовые характеристики статистического распределения

**Медиана** — срединное значение для ряда измерений ***n***. Для ее вычисления необходимо все наблюдения расположить в порядке возрастания или убывания результатов. Если ***n*** — нечетное число, то медиана просто является числом, находящимся в середине упорядоченной последовательности. При четном ***n*** равна среднему арифметическому двух расположенных в середине значений упорядоченной последовательности.

$$N_e = \frac{n+1}{2}$$

$$M_e = \frac{x_m + x_{m+1}}{2}$$

**Мода** — (наиболее вероятное значение) есть наиболее часто встречающаяся в данном измерении величина.

# Производство пшеницы в России в 1995-2001гг. Среднее арифметическое

| **Год** | 1995 | 1996 | 1997 | 1998 | 1999 | 2000 | 2001 |
|---|---|---|---|---|---|---|---|
| **Производство, млн. тонн** | 30,1 | 34,9 | 44,3 | 27,0 | 31,0 | 34,5 | 47,0 |

(30,1+34,9+44,3+27,0+31,0+34,5+47,0):7 ≈ 35,5.

Получаем, что среднее производство пшеницы в России за рассматриваемый период 1995-2001гг. Составляло приблизительно 35,5 млн. тонн в год.

# Урожайность зерновых культур в России в 1992-2001 гг.

| **Год** | 92 | 93 | 94 | 95 | 96 | 97 | 98 | 99 | 2000 | 01 |
|---|---|---|---|---|---|---|---|---|---|---|
| **Урожайность, ц/га** | 18,0 | 17,1 | 15,3 | 13,1 | 14,9 | 17,8 | 12,9 | 14,4 | 15,6 | 19,4 |

а)Средняя урожайность зерновых культур в России за 1992-1996гг.

(18,0+17,1+15,3+13,1+14,9):5 ≈ 15,68.

б)Средняя урожайность зерновых культур в России за 1997-2001гг.

(17,8+12,9+14,4+15,6+19,4):5 ≈ 16,02.

в)Средняя урожайность зерновых культур в России за 1992-2001гг.

(18,0+17,1+15,3+13,1+14,9+17,8+12,9+14,4+15,6+19,4):10 ≈ 15,85.

# Население шести крупнейших городов Московской области в разные годы, тыс. чел.

| Город | 1959 | 1970 | 1979 | 2002 | 2006 |
|---|---|---|---|---|---|
| Балашиха | 58 | 92 | 117 | 148 | 183 |
| Коломна | 118 | 136 | 147 | 150 | 148 |
| Люберцы | 95 | 139 | 154 | 157 | 159 |
| Мытищи | 99 | 119 | 141 | 159 | 162 |
| Подольск | 129 | 169 | 202 | 182 | 180 |
| Химки | 47 | 85 | 119 | 141 | 180 |

Среднее число жителей крупнейших городов Московской области

а)в 1959г. (58+118+95+99+129+47):6 ≈ 91.

б)в 1970г. (92+136+139+119+169+85):6 ≈ 123,3

в)в 1979г. (117+147+154+141+202+119):6 ≈ 146,6

г)в 2002г. (148+150+157+159+182+141):6 ≈ 156,7

д)в 2006г. (183+148+159+162+180+180):6 ≈ 168,6

# Медиана

## Производство пшеницы в России в 1995-2001гг.

| **Год** | 1995 | 1996 | 1997 | 1998 | 1999 | 2000 | 2001 |
|---|---|---|---|---|---|---|---|
| **Производство** | 30,1 | 34,9 | 44,3 | 27,0 | 31,0 | 34,5 | 47,0 |

Средний урожай 35,5 млн. тонн в год. Вычислим медиану. Упорядочим числа:

27,0; 30,1; 31,0; **34,5;** 34,9; 44,3; 47,0.

Медиана равна **Me=34,5** млн. тонн (урожай 2000г.)

# Медиана

**Медианой вариационного ряда является число**, которое делит ранжированную совокупность на две равные части: 50 % «нижних» членов ряда данных будут иметь значение признака не больше, чем медиана, а «верхние» 50 % — значения признака не меньше, чем медиана. Если в выборке количество элементов четное, то медиана будет средним арифметическим двух чисел, стоящих в середине этого ряда  иначе это число в середине отсортированного ряда.

**Пример 1.** Возьмём какой-нибудь набор различных чисел, например 1,4,7,9,11. Медианой в этом случае оказывается число, стоящее в точности посередине, m=7.

**Пример 2.** Рассмотрим набор 1,3,6,11. Медианой этого набора служит любое число, которое больше 3 и меньше 6. По определению в качестве медианы в таких случаях берут центр срединного интервала. В нашем случае это центр интервала (3,6). Это полусумма его концов  (3+6):2=4,5  Медианой этого набора считают число 4,5.

# Медиана

**Медиана для дискретного ряда**
с четным числом элементов

$$M_e = \frac{x_{n/2} + x_{n/2+1}}{2}$$

**Медиана для дискретного ряда**
с нечетным числом элементов

$$M_e = x_{\frac{n+1}{2}}$$

$$M_e = x_{Me} + h_{Me} \frac{1/2 \cdot \sum f - S_{fMe-1}}{f_{Me}}$$

где $x_{Me}$ – нижняя граница медианного интервала; $h_{Me}$–ширина медианного интервала; $f_{Me}$ – частота медианного интервала;

$S_{f\,Me-1}$ – накопленная частота интервала, предшествующего медианному

# Определение медианы

Найти медиану следующих наборов чисел

а) 2, 4, 8, 9 (4+8):2=6 m=6

б) 1, 3, 5, 7, 8, 9 (5+7):2=6 m=6

в) 10, 11, 11, 12, 14, 17, 18, 22

(12+14):2=13 m=13

# Числовые характеристики

- **Мода -** наиболее часто встречающаяся в ряду варианта. В интервальном вариационном ряду определяется модальный интервал. Мода используется для характеристики среднего уровня в неоднородных совокупностях, как и медиана.
- **Размах вариации** (R) - разность максимального и минимального значений периода в вариационном ряду.
- **R=max-min** Зависит от случайных колебаний выборки, т.е. для вычисления R используются лишь крайние значения варианты.

# Наибольшее и наименьшее значения. Размах вариации

Определение: Разность между наибольшим и наименьшим числом называется **размахом** набора чисел.

| **Год** | 1995 | 1996 | 1997 | 1998 | 1999 | 2000 | 2001 |
|---|---|---|---|---|---|---|---|
| **Произ-во, млн. тонн** | 30,1 | 34,9 | 44,3 | 27,0 | 31,0 | 34,5 | 47,0 |

Самый большой урожай пшеницы в эти годы был получен в 2001г. Он составил 47,0 млн. тонн. Самый маленький урожай 27,0 млн. тонн был собран в 1998г. Размах производства пшеницы в эти годы составил 20 млн. тонн. Это довольно большая величина по сравнению со средним значением производства в эти годы 35,5 млн. тонн.

# Числовые характеристики статистического распределения

**Среднее линейное отклонение d** — среднее арифметическое абсолютных величин отклонений вариантов от их средней арифметической, где $n_i$ — частота признака

$$d = \frac{\sum_{i=1}^{n} \left| x_i - \overline{x} \right| \cdot n_i}{n}$$

# Числовые характеристики статистического распределения

**Дисперсия D** — средняя арифметическая квадратов отклонений вариантов от их средней:

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

**Среднее квадратичное отклонение** — квадратный корень из дисперсии.

$$\delta = \sqrt{D}$$

# Отклонение

$$d = \frac{\sum_{i=1}^{n} \left| x_i - \overline{x} \right| \cdot n_i}{n}$$

**Среднее линейное отклонение** – среднее арифметическое абсолютных величин отклонений вариантов от их средней арифметической, где ni - частота признака

Пример: возьмём набор 1,6,7,9,12. Вычислим среднее арифметическое: (1+6+7+9+12)/5=7.

Найдём отклонение каждого числа от среднего и запишем в таблицу.

**Сумма отклонений чисел от среднего арифметического этих чисел равна нулю.**

| xi | 1 | 6 | 7 | 9 | 12 | d |
|---|---|---|---|---|---|---|
| Xi-Xср | 1 -7= -6 | 6 -7= -1 | 7 - 7=0 | 9-7=2 | 12 -7=5 | -6-1+ +0+2+5=0 |

# Дисперсия

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

**Определение:** среднее арифметическое квадратов отклонений от среднего значения называется в статистике **дисперсией** набора чисел.

Пример 1. Снова обратимся к таблице производства пшеницы в России. Мы нашли, что среднее производство пшеницы за период 1995-2001гг. составило 35,5 млн. тонн в год. Вычислим дисперсию. Составим таблицу, разместив данные по производству не в строке, а в столбце. Вычислим отклонения от среднего и их квадраты. Полученные числа занесём в два новых столбца.

# Таблица 8. Производство пшеницы в России в 1995-2001гг., млн. тонн.

| Год | Производство | Отклонение от среднего | Квадрат отклонения |
|---|---|---|---|
| 1995 | 30,1 | -5,4 | 29,16 |
| 1996 | 34,9 | -0,6 | 0,36 |
| 1997 | 44,3 | 8,8 | 77,44 |
| 1998 | 27,0 | -8,5 | 72,25 |
| 1999 | 31,0 | -4,5 | 20,25 |
| 2000 | 34,5 | -1,0 | 1,00 |
| 2001 | 47,0 | 11,5 | 132,25 |

Для расчета дисперсии следует сложить все значения в столбце «Квадрат отклонений» и разделить на количество слагаемых:

(29,16+0,36+77,44+72,25+20,25+1,00+132,25):7=47,53.

1.Для данных чисел **-1, 0, 4** вычислить среднее значение. Составить таблицу отклонений и квадратов отклонений от среднего и вычислить дисперсию:

| Число | Отклонение, $x_i - \overline{x}$ | Квадрат отклонения, $(x_i - \overline{x})^2$ |
|---|---|---|
| -1 | -2 | 4 |
| 0 | -1 | 1 |
| 4 | 3 | 9 |

**Ответ:** $\overline{x} = \frac{-1+0+4}{3} = 1$; **D=14.**

2.Для чисел **-1,-3,-2, 3, 3** вычислить среднее и дисперсию:

| Число | Отклонение, $x_i - \overline{x}$ | Квадрат отклонения, $(x_i - \overline{x})^2$ |
|---|---|---|
| -1 | -1 | 1 |
| -3 | -3 | 9 |
| -2 | -2 | 4 |
| 3 | 3 | 9 |
| 3 | 3 | 9 |

**Ответ:** $\overline{x} = \frac{-1-3-2+3+3}{5} = 0$; **D=32.**

# Числовые характеристики – функции Excel

| | |
|---|---|
| **max** | **=МАКС(A2:A100)** |
| **min** | **=МИН(A2:A100)** |
| **R** | **=max-min** |
| **k** | **=1+3,2·log10(n)** |
| $\overline{x}$ | **=СРЗНАЧ (A2:A100)** |
| **медиана** | =МЕДИАНА(A2:A10) |
| **мода** | **=МОДА (A2:A100)** |
| **дисперсия** | **=ДИСП (A2:A100)** |
| δ | **=СТАНДОТКЛОН (A2:A100)** |
| **эксцесс** | **=ЭКСЦЕСС(A2:A100)** |
| **ассиметрия** | **=СКОС (A2:A100)** |
| **C**(цена деления) | **=R/k** |

1) Рассчитываются числовые характеристики, с помощью функций Excel,
2) Представляют данные эксперимента в виде интервального ряда.
3) Рассчитывают частоты для каждого признака
4) Строят полигон и гистограмму частот.
5) Делают вывод о форме распределения ген. совокупности

# Ранжирование

- ***Ранжирование*** — процедура упорядочения любых объектов по возрастанию или убыванию некоторого их свойства при условии, что они этим свойством обладают. Например, можно ранжировать респондентов по степени: их удовлетворенности чем-то, их политической активности, отношения к чему-то и т. д.

Можно ранжировать информационные телепередачи по степени их информативности, профессии — по престижности, политических лидеров — по их влиянию на принятие решений президентом. Возможно ранжирование качеств человека по их важности в карьере, ранжирование товаров по предпочтению покупателей.

- ***Объекты ранжирования*** — это те объекты, которые упорядочиваются. Они могут быть самыми разными.

- ***Основание ранжирования*** — это то свойство, по которому объекты упорядочиваются. В результате упорядочения получаем ***ранжированный ряд.***

- В нем каждому объекту приписывается ***ранг*** — место в этом ряду. Число мест и, соответственно, число рангов равно числу объектов. Обратите вни-мание на различие между ранжированием и измерением по порядковой шкале.

# Ранжирование

- Объекты ранжирования могут быть либо все разными с точки зрения выраженности в Них заданного свойства, либо некоторые объекты могут быть неразличимыми, как в случае только что рассмотренных примеров измерения по порядковой шкале. В первом случае все ранги будут различны, а во втором случае появятся одинаковые ранги. Они называются ***связанными рангами.***
- В таблице рассмотрена именно такая ситуация. В первой строке этой таблицы приведены показатели качества жизни для произвольных 9-ти государств, обозначенных буквами А, Б, В, Г, Д, Е, К, Л, М. Во второй строке — результаты ранжирования, т. е. ранжированный ряд.

| **Государства** | **А** | **Б** | **В** | **Г** | **Д** | **Е** | **К** | **Л** | **М** |
|---|---|---|---|---|---|---|---|---|---|
| Качество жизни | 6,5 | 7 | 6,5 | 5,9 | 4,6 | 5,9 | 4,5 | 5,9 | 4,5 |
| Ранг | | | | | | | | | |

# Ранжирование

В этой таблице представлены результаты ранжирования в порядке убывания значения показателя качества жизни.

1) Таблица сортируется по убыванию или возрастанию
2) Для расчета рангов используется номер элемента в отсортированном ряде
3) Если элемент в выборке не повторяется, то его ранг равен его порядковому номеру
4) Для повторяющихся элементов ранг считается так

$$R=(N_1+N_2+N_3+\ldots)/kol$$

Складываются номера **всех повторяющихся элементов и делятся на количество повторений.** Например, страны **А и В** имеют одинаковое качество жизни 6,5.

Их ранг считается так: **(2+3)/2=2,5.**

Или страны **Г, Е, М** имеют качество жизни 5,9, тогда их **ранг: (4+5+6)/3=5.**

| **Государства** | **Б** | **А** | **В** | **Г** | **Е** | **М** | **Д** | **К** | **Л** |
|---|---|---|---|---|---|---|---|---|---|
| № элемента | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| Качество жизни | 7 | 6,5 | 6,5 | 5,9 | 5,9 | 5,9 | 4,6 | 4,5 | 4,5 |
| Ранг | 1 | 2,5 | 2,5 | 5 | 5 | 5 | 7 | 8,5 | 8,5 |

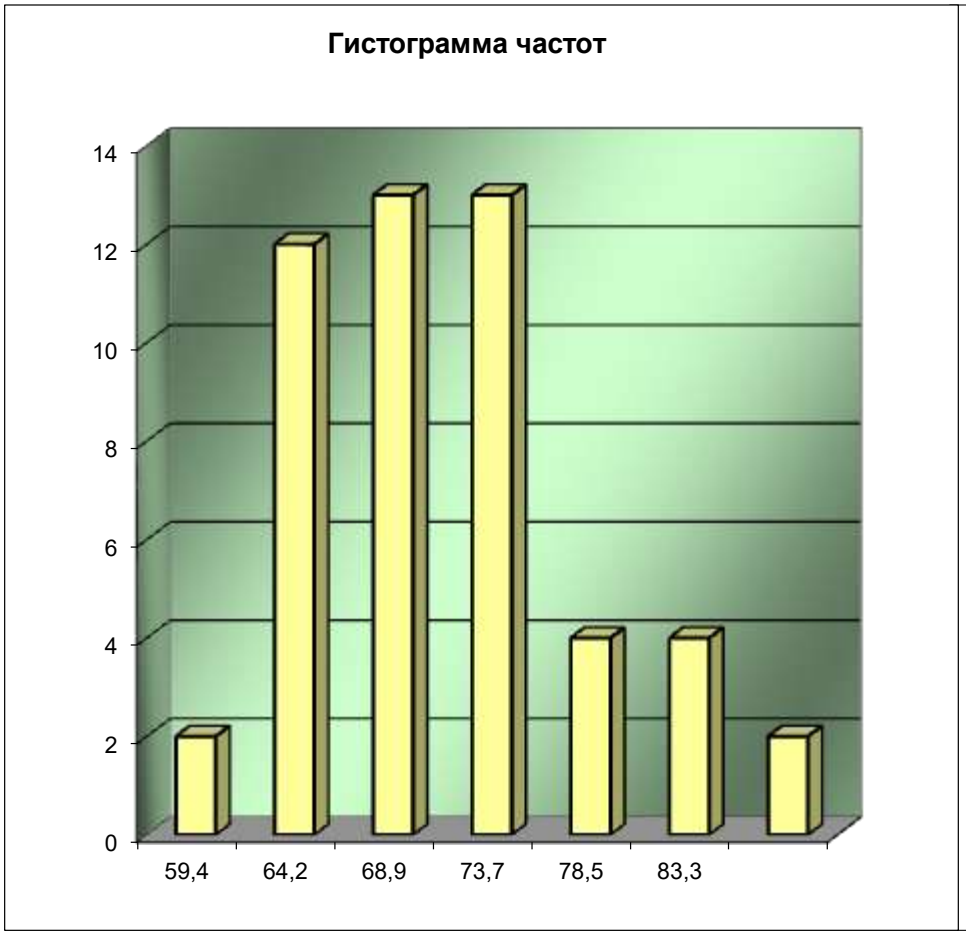
# Графическое изображение рядов распределения

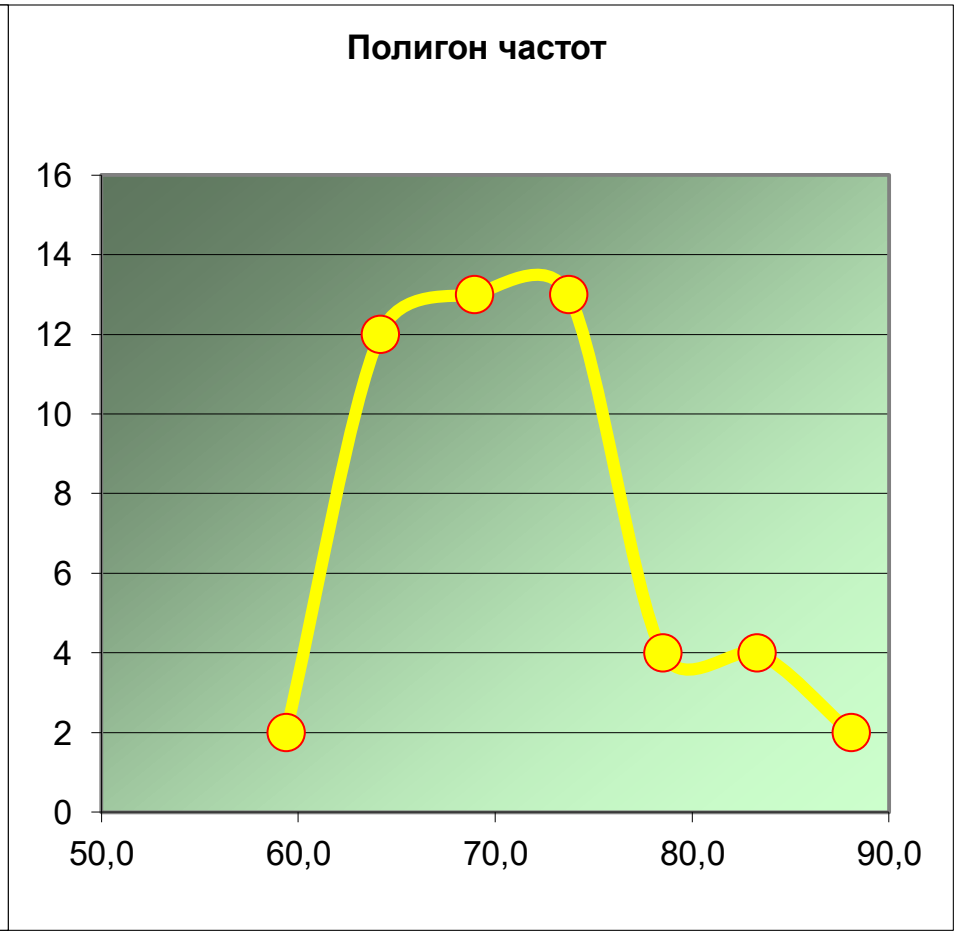
Наглядно ряды распределения представляются при помощи графических изображений.

Ряды распределения изображаются в виде:

- Полигона
- Гистограммы
- Кумуляты

# Полигон частот



# Результаты измерения веса 100 студентов МГТУ

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 61 | 57 | 61 | 85 | 48 | 41 | 73 | 66 | 91 | 70 |
| 50 | 45 | 64 | 46 | 55 | 82 | 69 | 75 | 82 | 72 |
| 68 | 43 | 81 | 71 | 47 | 50 | 54 | 75 | 81 | 68 |
| 80 | 67 | 64 | 76 | 61 | 57 | 62 | 57 | 66 | 53 |
| 79 | 56 | 63 | 88 | 65 | 74 | 67 | 54 | 65 | 80 |
| 86 | 40 | 59 | 64 | 65 | 71 | 72 | 78 | 70 | 61 |
| 39 | 63 | 89 | 59 | 61 | 75 | 67 | 51 | 65 | 55 |
| 62 | 60 | 75 | 73 | 91 | 72 | 54 | 46 | 52 | 55 |
| 78 | 67 | 94 | 60 | 44 | 49 | 88 | 74 | 44 | 60 |
| 52 | 61 | 66 | 74 | 56 | 52 | 71 | 73 | 75 | 60 |

## Выполнить статистический анализ данных выборки

# Выполнение статистического анализа

| | | | Признаки (интервалы) | Частоты |
|---|---|---|---|---|
| max | 85,4 | | | |
| min | 54,6 | | 59,4 | 2 |
| R | 30,8 | | 64,2 | 12 |
| k | 7 | | 68,9 | 13 |
| Срзнач. | 69,30 | | 73,7 | 13 |
| медиана | 68,53 | | 78,5 | 4 |
| мода | 68 | | 83,3 | 4 |
| дисперсия | 45,77 | | 88,1 | 2 |
| δ | 6,765 | | | |
| эксцесс | 0,057 | | | |
| скос | 0,385 | | | |
| C | =R/k | | | |

Числовые характеристики рассчитываются с помощью функций Excel,

Данные эксперимента представляют
в виде интервального ряда.

Рассчитывают признаки

- Определяют **частоты** для каждого признака
- Строят полигон и гистограмму частот.
- Делают вывод о форме распределения генеральной совокупности

# Пример задания на построение вариационного ряда

**Условие:** Приводятся данные о распределении 25 работников одного из предприятий по тарифным разрядам: 4; 2; 4; 6; 5; 6; 4; 1; 3; 1; 2; 5; 2; 6; 3; 1; 2; 3; 4; 5; 4; 6; 2; 3; 4

**Задача:** Построить дискретный вариационный ряд и изобразить его графически в виде полигона распределения.

**Решение:** В данном примере вариантами является тарифный разряд работника. Для определения частот необходимо рассчитать число работников, имеющих соответствующий тарифный разряд.

Для построения полигона распределения (рис 1) по оси абсцисс (X) откладываем количественные значения варьирующего признака — варианты, а по оси ординат — частоты или частости.

Для построения полигона распределения по оси абсцисс (X) откладываем количественные значения варьирующего признака — варианты, а по оси ординат — частоты или частости.

| Тарифный разряд $X_j$ | Число работников $f_j$ |
|---|---|
| 1 | 3 |
| 2 | 5 |
| 3 | 4 |
| 4 | 6 |
| 5 | 3 |
| 6 | 4 |
| Итого: | 25 |

# Гистограмма частот

- Если значения признака выражены в виде интервалов, то такой ряд называется интервальным.
- Интервальные ряды распределения изображают графически в виде гистограммы.
- Гистограмма – график, на котором ряд изображен в виде смежных друг с другом столбиков. Для построения гистограммы по оси абсцисс указывают значения границ интервалов и на их основании строят прямоугольники, высота которых пропорциональна частотам (или частостям).


Рис. 6.2. Распределение населения России по возрастным группам

| Все население | В том числе в возрасте | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | до 10 | 10-20 | 20-30 | 30-40 | 40-50 | 50-60 | 60-70 | 70 и старше | Всего |
| Численность населения | 12,1 | 15,7 | 13,6 | 16,1 | 15,3 | 10,1 | 9,8 | 7,3 | 100,0 |

# Пример задания на построение гистограммы частот

**Задача** : Приводится распределение 30 работников фирмы по размеру месячной заработной платы

**Цель**: Изобразить интервальный вариационный ряд графически в виде гистограммы и кумуляты.

**Решение**:

1. Неизвестная граница первого интервала определяется по величине второго интервала: 7000 — 5000 = 2000 руб. С той же величиной находим нижнюю границу первого интервала: 5000 — 2000 = 3000 руб.

2. Для построения гистограммы в прямоугольной системе координат по оси абсцисс откладываем отрезки, величины которых соответствуют интервалам варицонного ряда.

Эти отрезки служат нижним основанием, а соответствующая частота (частость) — высотой образуемых прямоугольников.

3. Построим гистограмму:

# Гистограмма частот

| Размер заработной платы руб. в месяц | Численность работников чел. |
|---|---|
| до 5000 | 4 |
| 5000 — 7000 | 12 |
| 7000 — 10000 | 8 |
| 10000 — 15000 | 6 |
| Итого: | 30 |

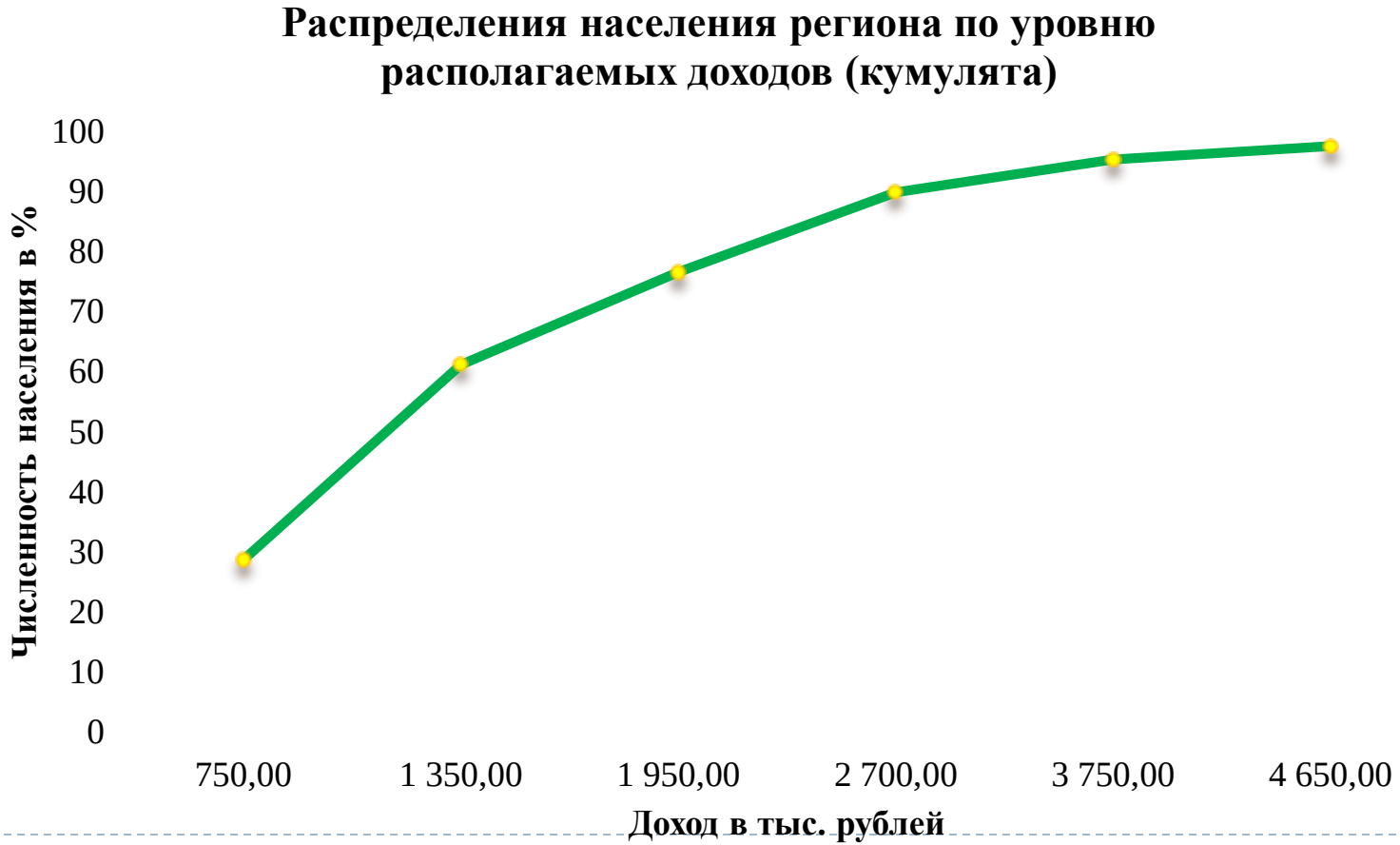
# Кумулята

- Распределение признака в вариационном ряду по накопленным частотам (частостям) изображается с помощью кумуляты.
- Кумулята или кумулятивная кривая в отличие от полигона строится по накопленным частотам или частостям. При этом на оси абсцисс помещают значения признака, а на оси ординат — накопленные частоты или частости .
- Для построения кумуляты необходимо рассчитать накопленные частоты (частости). Они определяются путем последовательного суммирования частот (частостей) предшествующих интервалов и обозначаются S. Накопленные частоты показывают, сколько единиц совокупности имеют значение признака не больше, чем рассматриваемое.

- Рассчитаем накопленные частоты:
- Наколенная частота первого интервала рассчитывается следующим образом: 0 + 4 = 4, для второго: 4 + 12 = 16; для третьего: 4 + 12 + 8 = 24 и т.д.

| Размер заработной платы руб в месяц $X_j$ | Численность работников чел. $f_j$ | Накопленные частоты S |
|---|---|---|
| до 5000 | 4 | 4 |
| 5000 — 7000 | 12 | 16 |
| 7000 — 10000 | 8 | 24 |
| 10000 — 15000 | 6 | 30 |
| Итого: | 30 | - |

При построении кумуляты накопленная частота (частость) соответствующего интервала присваивается его верхней границе:



# Построить интервальный вариационный ряд с равными интервалами

**Задача**: Приведены данные о размерах вкладов 20 физических лиц в одном банке (тыс. руб) 60; 25; 12; 10; 68; 35; 2; 17; 51; 9; 3; 130; 24; 85; 100; 152; 6; 18; 7; 42.

Решение:

1. Исходная совокупность состоит из 20 единиц (N = 20).
2. По формуле Стерджесса определим необходимое количество используемых групп: k=1+3,322*lg20=5
3. Вычислим величину равного интервала: c=(152 — 2) /5 = 30 тыс.руб
4. Расчленим исходную совокупность на 5 групп с величиной интервала в 30 тыс.руб.
5. Результаты группировки представим в таблице:

# интервальный вариационный ряд

| Размер вкладов тыс.руб $X_i$ | Число вкладов $f_i$ | Число вкладов в % к итогу $W_i$ |
|---|---|---|
| 2 — 32 | 11 | 55 |
| 32 — 62 | 4 | 20 |
| 62 — 92 | 2 | 10 |
| 92 — 122 | 1 | 5 |
| 122 — 152 | 2 | 10 |
| Итого: | 20 | 100 |

При такой записи непрерывного признака, когда одна и та же величина встречается дважды (как верхняя граница одного интервала и нижняя граница другого интервала), то эта величина относится к той группе, где эта величина выступает в роли верхней границы.

# Статистическая гипотеза

- **Статистическая гипотеза** – предположение о свойствах распределения вероятностей выборки или генеральной совокупности.

- Часто гипотеза формулируется как утверждение наличия или отсутствия связи между признаками (зависимыми и независимыми переменными).
- **Например**, проверяется гипотеза о том, что женщины тратят больше времени на разговоры по телефону, чем мужчины. Предположим, что в исследовании принимали участие 52 мужчины и 43 женщины. Среднее время разговора составило 25 мин в день у мужчин и 35 мин в день у женщин.

**Является ли это различие статистически значимым?**

# Примеры гипотез

**Примеры общих гипотез:**

- **гипотеза Демокрита** об атомистическом строении вещества,
- **гипотеза Канта-Лапласа** о происхождении небесных тел

Все в природе непрерывно изменяется, развивается — эволюционирует. И Земля и Солнце раньше не были такими, какие они сейчас; было время, когда составляющее их вещество существовало совсем в другом виде.

- **гипотеза А. И. Опарина** о возникновении жизни на Земле

Жизнь возникла естественным путем из неорганической материи. Биологической эволюции предшествовала химическая.

В результате исследования гипотезы либо опровергаются, либо подтверждаются и становятся положениями теории, истинность которой уже доказана. Общая гипотеза после ее доказательства становится научной теорией

# Нулевая гипотеза

Например, производительность труда в среднем по заводу составляет 50 деталей/час. Пусть выясняется, отличается ли производительность труда рабочих определенной возрастной группы от средних показателей по заводу.

Тогда нулевая гипотеза может выглядеть так

$$H_0: \mu = 50.$$

Такая гипотеза обозначает, что проверяется равенство среднего значения производительности по исследуемой группе рабочих (выборке) значению $\mu = 50$ (средней производительности труда по заводу - генеральной совокупности)

# Альтернативная гипотеза

Альтернативная гипотеза ($H_A$) –обычно противоречит нулевой и утверждает о наличии различий или существовании связи.

**Статистически значимое различие** – это различие, которое настолько велико, что вероятность его возникновения вследствие простой случайности крайне мала.

**$H_0$:** женщины тратят больше времени на разговоры по телефону, чем мужчины

**$H_A$:** различие во времени телефонных разговоров у мужчин и женщин незначительное

Например, $H_A$: $\mu \neq 50$. В таком случае мы имеем дело с ненаправленной альтернативной гипотезой.

# Альтернативная гипотеза

Символ строгого или нестрогого равенства всегда присутствует в формулировке

**нулевой гипотезы ≤, =, ≥.**

А в формулировке **альтернативной гипотезы только** символы: **<, >.**

**Например:**

**Нулевая гипотеза :** $H_0 : \mu \leq 50;$

**Альтернативная гипотеза :** $H_A : \mu > 0$

# Выдвижение и проверка гипотез

- Гипотеза в исследовании – это научно обоснованное предположение о структуре социальных объектов, о характере элементов и связей, образующих эти объекты, о механизме их функционирования и развития.
- Научная гипотеза может быть сформулирована только в результате предварительного анализа изучаемого объекта.
- В результате исследования гипотезы либо опровергаются, либо подтверждаются и становятся положениями теории, истинность которой уже доказана

# Методы проверки гипотез

Эмпирическое распределение выборки рассматривается в качестве оценки теоретической функции распределения F(x) генеральной совокупности.

Различают две группы математико-статистических методов:

- **Непараметрические методы** (статистические тесты).

Предположение, при котором вид распределения неизвестен называется **непараметрической гипотезой.**

- **Параметрические методы** (**оценка параметров распределения**).

Статистическая проверка гипотез предполагает выдвижение определенных допущений (гипотез) относительно неизвестных параметров F(x). Правильность этих гипотез проверяется затем по числовым значениям, полученным из выборки, и, в зависимости от результата проверки, гипотезы принимаются или отвергаются.

# Критерии согласия

**Критерии согласия** проверяют, согласуется ли заданная выборка с заданным фиксированным **распределением**, с заданным параметрическим **семейством распределений**, или с **другой выборкой**.

Критерий хи-квадрат (Пирсона)

Критерий Колмогорова-Смирнова

Критерий омега-квадрат (фон Мизеса)

# Критерий Стьюдента (t-критерий)

Критерий позволяет найти вероятность того, что оба средних арифметических в двух выборках относятся к одной и той же совокупности.

При использовании критерия можно выделить два случая.

1) **гипотеза о равенстве генеральных средних двух независимых, несвязанных** выборок - двухвыборочный t-критерий. В этом случае есть контрольная и экспериментальная группа.

2) **гипотез о равенстве средних в одной выборке в разные моменты времени -** парный t-критерий. Выборки при этом называют зависимыми, связанными.

# Критерий Стьюдента для независимых выборок

Статистика критерия для случая несвязанных, независимых выборок равна:

$$t_{эмп} = \frac{\bar{x} - \bar{y}}{\delta_{x-y}} \qquad (1)$$

В числителе средние арифметические в экспериментальной и контрольной групп,

$\delta_{x-y}$- стандартная ошибка разности средних арифметических. Находится из формулы:

$$\delta_{x-y} = \sqrt{\frac{\sum (x_i - \bar{x})^2 + \sum (y_i - \bar{y})^2}{n_1 + n_2 - 2} \cdot \left[\frac{1}{n_1} + \frac{1}{n_2}\right]} \qquad (2)$$

где n1 и n2 соответственно величины первой и второй выборки.

# Критерий Стьюдента для независимых выборок

**Пример 1.** В двух группах учащихся - экспериментальной и контрольной - получены следующие результаты по учебному предмету (тестовые баллы). Была выдвинута параметрическая гипотеза:

$H_0$: **учащиеся экспериментальной и контрольной группы имеют одинаковый уровень знаний;**

$H_1$ **(альтернативная гипотеза): учащиеся экспериментальной группы показывают в среднем более высокий уровень знаний**

| Первая группа экспериментальная $N_1$=11 человек | | | | | | | | | | | Вторая группа контрольная $N_2$=9 человек | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **12** | **14** | **13** | **16** | **11** | **9** | **13** | **15** | **15** | **18** | **14** | **13** | **9** | **11** | **10** | **7** | **6** | **8** | **10** | **11** |

Общее количество членов выборки: $n_1$=11, $n_2$=9.

Расчет средних арифметических: Xcp=13,636; Ycp=9,444

Стандартное отклонение: sx=2,460; sy=2,186

# Критерий Стьюдента

1) По формуле рассчитываем стандартную ошибку разности арифметических средних:

$$\delta_{xy} = \sqrt{\frac{60.545 + 38.222}{11 + 9 - 2} \cdot (\frac{1}{9} + \frac{1}{11})} = 1.053$$

2) Считаем статистику критерия Стьюдента:

$$t = \frac{13.636 - 9.444}{1.053} = 3.981$$

3) Подсчет **числа степеней свободы** осуществляется по формуле:

$$k = n1 + n2 - 2$$

При численном равенстве выборок $k = 2n - 2$.

# Критерий Стьюдента (t-критерий)

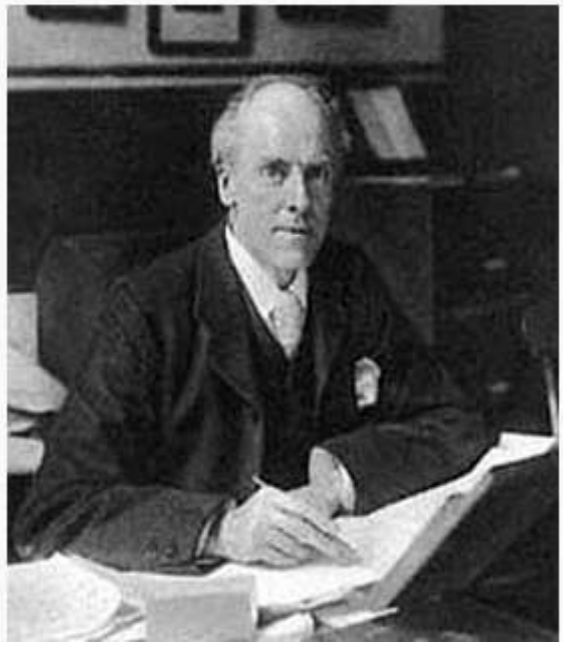
4) Далее необходимо сравнить полученное значение tэмп с теоретическим значением критерия Стьюдента (см. статистики).

Если tэмп<tкрит, то гипотеза $H_0$ принимается, в противном случае нулевая гипотеза отвергается и принимается альтернативная гипотеза.

**Тэкс>Ттаб**
**3,981>2,10**

**Табличное значение tт =2,1** при (уровень значимости=5 % или 0,05).

**Вывод $H_1$ учащиеся экспериментальной группы показывают в среднем более высокий уровень знаний**



# Критерий Пирсона

Карл Пирсон 1857 -1936- математик, статистик, биолог и философ; основатель математической статистики, один из основоположников биометрики.

Автор свыше 650 опубликованных научных работ.

Критерий согласия Пирсона наиболее часто употребляемый критерий для

проверки гипотезы о принадлежности наблюдаемой выборки объёмом n некоторому теоретическому закону распределения

# Критерий Пирсона

В качестве меры расхождения теоретического и эмпирического рядов частот возьмем величину: критерий согласия Пирсона.

$$\chi^2 = \sum_{i=1}^{m} \frac{(n_i - n_i^o)^2}{n_i^0}$$

Из выражения видно, что $\chi 2 = 0$ лишь при совпадении всех соответствующих эмпирических и теоретических частот:

$$n_i = n_i^0.$$

В противном случае $\chi 2$ отлично от нуля и тем больше, чем больше расхождение между указанными частотами.

Величина $\chi 2$ является случайной и имеет распределение **$\chi 2$-распределение.** Параметр ***k*** назван **числом степеней свободы**.

# Критерий Пирсона алгоритм расчета

Для построения расчетной таблицы нужно взять:

- **Частоты и интервалы экспериментального распределения**
- **Среднее арифметическое**
- **Среднее квадратичное отклонение**

Альфа и бетта в таблице – это границы интервала.

**Для первого интервала** альфа совпадает с минимумом

Для следующего интервала – это предыдущее бетта


| Интервал $\alpha_i$-$\beta_i$ | Частота | $\beta_i$-$x_{ср}$ | $\alpha_i$-хср | $\Phi(x_1)$ | $\Phi(x_2)$ | Теор частота | Разности $(n_i-n_i^0)$ | $\frac{(n_i-n_i^0)^2}{n_i^0}$ |
|---|---|---|---|---|---|---|---|---|
| -1,92 | 3,00 | -1,96 | -2,70 | 0,02 | 0,00 | 2,16 | 0,84 | 0,33 |
| -1,17 | 11,00 | -1,21 | -1,96 | 0,11 | 0,02 | 8,75 | 2,25 | 0,58 |

# Критерий Пирсона

Количество частот в группе должно быть ***больше 5***. Если количество меньше, то соседние группы следует объединить.

Если мы нашли — среднее арифметическое и $\sigma 2$ — дисперсию, используя данные опытного распределения и установили, сумма частот опытного распределения равна сумме частот теоретического распределения, то число связей $s = 3$.

Если же эмпирическое распределение не использовалось для нахождения параметров теоретического закона и теоретических частот, а эмпирические частоты не связаны никакими дополнительными соотношениями, то $k$ равно числу групп эмпирического распределения.

# Расчет критерия Пирсона

| Интервал $\alpha_i$-$\beta_i$ | Частота | $\beta_i$-$x_{ср}$ | $\alpha_i$-xcp | $\Phi(x_1)$ | $\Phi(x_2)$ | Теоретич частота | Разности $(n_i-n_i^0)$ | $\frac{(n_i-n_i^0)^2}{n_i^0}$ |
|---|---|---|---|---|---|---|---|---|
| -1,92 | 3,00 | -1,96 | -2,70 | 0,02 | 0,00 | 2,16 | 0,84 | 0,33 |
| -1,17 | 11,00 | -1,21 | -1,96 | 0,11 | 0,02 | 8,75 | 2,25 | 0,58 |
| -0,42 | 16,00 | -0,47 | -1,21 | 0,32 | 0,11 | 20,79 | -4,79 | 1,10 |
| 0,32 | 28,00 | 0,28 | -0,47 | 0,61 | 0,32 | 29,00 | -1,00 | 0,03 |
| 1,07 | 29,00 | 1,03 | 0,28 | 0,85 | 0,61 | 23,74 | 5,26 | 1,16 |
| 1,82 | 10,00 | 1,77 | 1,03 | 0,96 | 0,85 | 11,41 | -1,41 | 0,18 |
| 2,57 | 3,00 | 2,52 | 1,77 | 0,99 | 0,96 | 3,22 | -0,22 | 0,01 |
| | | | | | | | **χ2 Эксп** | **3,40** |

| **xcp** |
|---|
| 0,04 |
| **σ** |
| 0,98 |
| **$\chi^2$ табл** |
| **9,49** |

| **χ 2 табл** | **χ2 эксп** |
|---|---|
| **9,49** | **3,40** |

Вывод: экспериментальные частоты незначительно отличаются от теоретических, значит выборочное распределение имеет закон нормальный закон распределения

# Критерий Пирсона алгоритм расчета $Фx_1$ $Фx_2$

Значение $Ф(x_1)$ вычисляется

как функция нормального распределения,

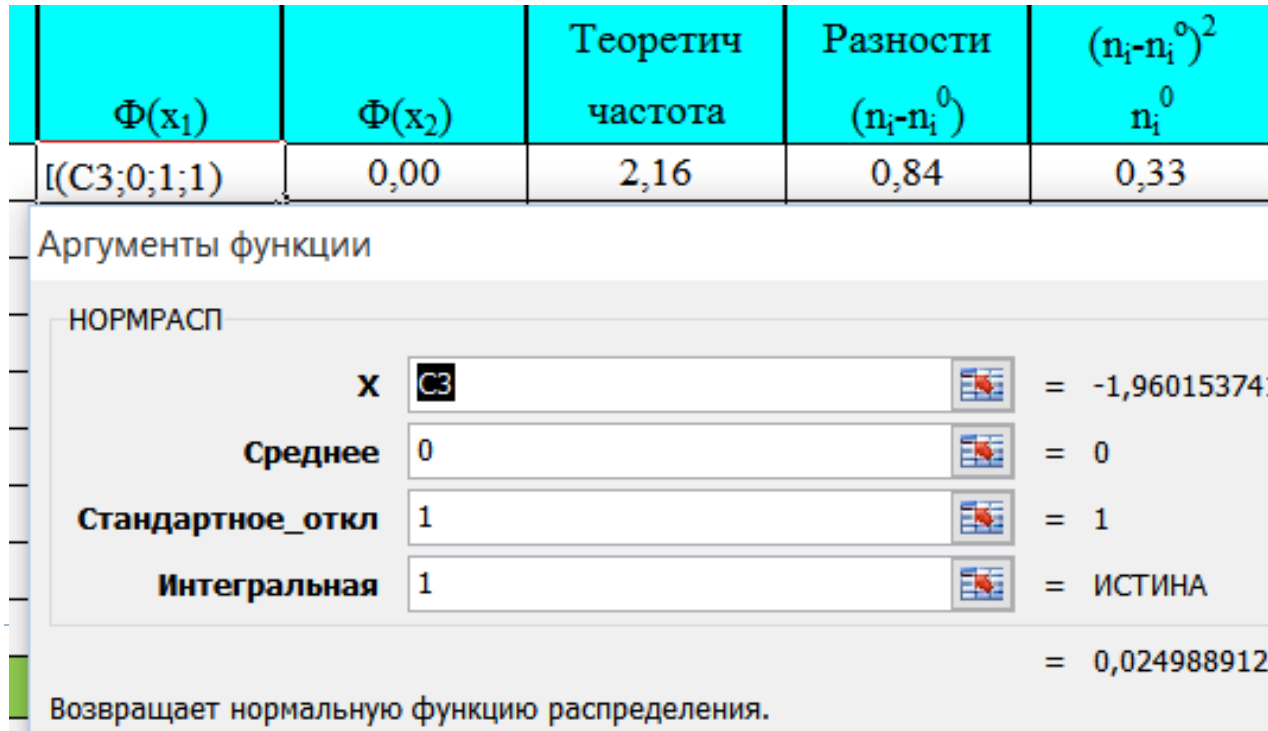
НОРМРАСП со средним 0, отклонением 1,

а аргумент X

считается так:

$$x_1 = \frac{\alpha_i - \bar{X}}{\sigma}$$

$$x_2 = \frac{\beta_i - \bar{x}}{\sigma}$$



# Критерий Пирсона алгоритм расчета теоретическая частота

Теоретическая частота рассчитывается как разность между $(\Phi(x_2) - \Phi(x_1))*n$,

Где n –это объем выборки, в нашем случае n=100

Затем считается разность между экспериментальными и теоретическими частотами, а потом разность возводится в квадрат и делится на теоретическую частоту. Находится сумма элементов последнего столбца –это и есть экспериментальный критерий Пирсона - $\chi 2$

# Критерий Пирсона

Далее необходимо сравнить полученное значение критерия Хи-квадрат с теоретическим значением взятым из таблиц или полученным с помощью функции Excel : =ХИОБР() (см. статистики).

Если **$\chi 2$ эксп** < **$\chi 2$** крит, то гипотеза $H_0$ принимается, в противном случае нулевая гипотеза отвергается и принимается альтернативная гипотеза.

**Табличное значение $\chi 2$ т =9,49 экспериментальное $\chi 2$ эксп = 3,40**

**3,40<9,49,** то есть $\chi 2$ эксп < $\chi 2$ при (уровне значимости 5 % или 0,05).

**Вывод: гипотеза $H_0$ принимается**, поскольку экспериментальные частоты незначительно отличаются от теоретических, **значит выборка имеет закон нормальный закон распределения**