# Теория вероятностей и статистика

## Тема 5. Выборочный метод

Белов А.И.

Уральский федеральный университет

Екатеринбург, 2020

# Задачи статистики

1. Указать методы сбора и группировки статистических данных.

2. Предоставить методы анализа статистических данных в зависимости от цели исследования:
   - оценка неизвестной вероятности события;
   - аппроксимация неизвестной функции распределения;
   - оценка параметров распределения, вид которого известен;
   - оценка зависимости случайных величин друг от друга;
   - проверка статистических гипотез о величине параметров распределения, вид которого известен;
   - проверка статистических гипотез о виде неизвестного распределения.

# Генеральная и выборочная совокупности

**Определение**

Выборочной совокупностью (выборкой) называется множество объектов, отобранных для изучения.

Генеральной совокупностью называется абстракция, которая представляет собой конечное или бесконечное множество всех объектов, из которых производится выборка.

Объемом совокупности (выборочной или генеральной) называется количество элементов этой совокупности.

# Репрезентативные выборки

**Определение**

Выборка называется репрезентативной, если она правильно представляет пропорции генеральной совокупности.

Из закона больших чисел следует, что выборка будет репрезентативной, если объекты выбираются из генеральной совокупности в достаточном количестве, случайно и равновозможно.

# Повторные и бесповторные выборки

**Определение**

Выборка называется повторной, если объект после выбора возвращается в генеральную совокупность, и бесповторной, если он не возвращается.

Если объем генеральной совокупности велик, а объем выборки ничтожно мал по сравнению с объемом генеральной совокупности, то различие между повторной и бесповторной выборками стирается.

В случае бесконечной генеральной совокупности это различие исчезает.

# Способы отбора

1. Отбор без разбиения генеральной совокупности на части:
   - простой случайный бесповторный отбор;
   - простой случайный повторный отбор.

2. Отбор с разбиением генеральной совокупности на части:
   - типический отбор, когда отбор производится из каждой «типичной» части (его применяют, когда исследуемый признак значительно колеблется в различных типичных частях генеральной совокупности);
   - механический отбор, когда генеральная совокупность делиться на несколько частей «механически», после чего выбирается по одному объекту из каждой части;
   - серийный отбор, когда объекты выбираются сериями (его применяют, когда исследуемый признак незначительно изменяется от серии к серии).

На практике эти методы комбинируют.

# Вариационный ряд

Предположим, что мы изучаем количественный признак дискретной случайной величины и $x_1, \ldots, x_n$ — выборка.

**Определение**

Различные значения, наблюдаемые в выборке называются вариантами.
Варианты, расположенные в порядке возрастания:

$$\widetilde{x}_1 < \widetilde{x}_2 < \ldots < \widetilde{x}_m,$$

называются вариационным рядом.
Величина $\widetilde{x}_m - \widetilde{x}_1$ называется размахом выборки.

Если мы изучаем качественный признак, то вариантами будут служить условные обозначения значений признака, которые располагаются в вариационном ряде некоторым образом.

# Частоты

**Определение**

Пусть варианта $\widetilde{x}_k$ наблюдалась в выборке $n_k$ раз $(k = 1, \ldots, m)$. Числа $n_k$ называются частотами выборки.

Сумма всех частот $n = \sum_{k=1}^{m} n_k$ называется объемом выборки.

Числа $w_k = \dfrac{n_k}{n}$ называются относительными частотами выборки.

Очевидно, что $\sum_{k=1}^{m} w_k = 1$.

# Статистическое распределение

**Определение**

Статистическим распределением называют таблицу, содержащую вариационный ряд и значения частот или относительных частот.

Графически статистическое распределение представляют в виде гистограммы частот (относительных частот).

Статистическое распределение также может быть представлено в виде полигона частот — ломаной линии, соединяющей точки $(\widetilde{x}_1, n_1), (\widetilde{x}_2, n_2), \ldots, (\widetilde{x}_m, n_m)$ либо в виде полигона относительных частот — ломаной линии, соединяющей точки $(\widetilde{x}_1, w_1), (\widetilde{x}_2, w_2), \ldots, (\widetilde{x}_m, w_m)$.

# Пример

Выборка 3, 2, 2, 3, 2, 3, 3, 3, 2, 2, 2, 3, 3, 2, 1, 1, 1, 2, 3, 3.

Вариационный ряд выборки — значения 1, 2, 3. Объем выборки равен 20.

Статистическое распределение

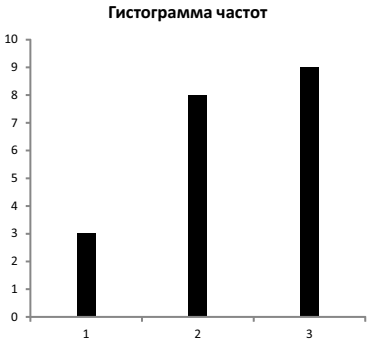
| $\widetilde{x}_k$ | 1 | 2 | 3 |
|---|---|---|---|
| $n_k$ | 3 | 8 | 9 |

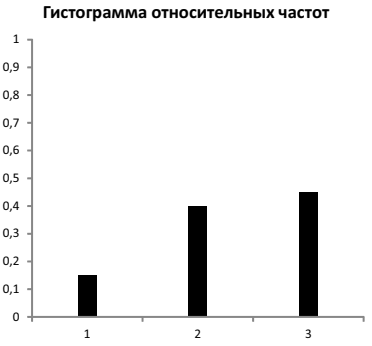
или

| $\widetilde{x}_k$ | 1 | 2 | 3 |
|---|---|---|---|
| $w_k$ | 0,15 | 0,40 | 0,45 |

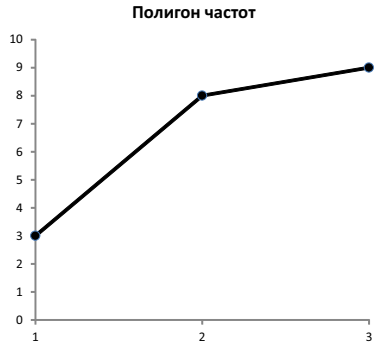
.

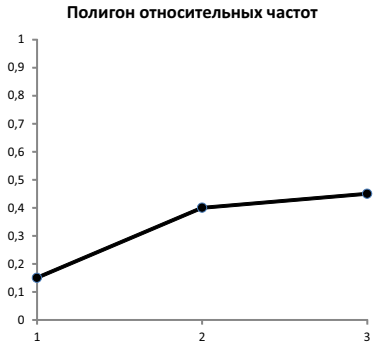
# Гистограммы



# Полигоны


Полигон частот
10
9
8
7
6
5
4
3
2
1
0
1
2
3


Полигон относительных частот
1
0,9
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0
1
2
3

# Групповые частоты

Если распределение непрерывное, то варианты в выборке практически не повторяются.
В этом случае (а также в случае дискретного распределения и большого числа вариант — больше 20–25) варианты группируются по интервалам.
Отрезок $[\widetilde{x}_1, \widetilde{x}_m]$ размещают в отрезке $[x'_0, x'_N]$, где $x'_0 < \widetilde{x}_1$, $\widetilde{x}_m < x'_N$.
Отрезок $[x'_0, x'_N]$ разбивают на $N$ частей точками $x'_k$: $x'_0 < x'_1 < \ldots < x'_N$..
Для каждого промежутка группировки подсчитывается число $n_i$ вариант, принадлежащих промежутку $[x'_{i-1}, x'_i)$, где $i = 1, \ldots N$.
В программе MS Excel подсчитывается число $n_i$ вариант, принадлежащих промежутку $(x'_{i-1}, x'_i]$.

**Определение**

Числа $n_i$ называются групповыми частотами, а числа $w_i = \dfrac{n_i}{n}$ называются относительными групповыми частотами.

# Число интервалов группировки

Для определения числа интервалов группировки как правило используются эмпирические функции.

Microsoft Excel:

$$N = \lfloor \sqrt{n} + 1 \rfloor,$$

где $n$ — объем выборки, а $\lfloor x \rfloor$ — целая часть числа $x$.

Формула Стэрджеса

$$N = \lfloor \log_2 n \rfloor.$$

Другие формулы

$$N = \lfloor \sqrt{n} - 0{,}013n + 0{,}5 \rfloor \qquad \text{или} \qquad N = \lfloor 1{,}72 \sqrt[3]{n} \rfloor.$$

# Метод равных интервалов

Отрезок $[x'_0, x'_N]$ разбивается на $N$ равных частей.

Длина каждого интервала группировки составит $\Delta = \dfrac{x'_N - x'_0}{N}$.

Точки разбиения будут $x'_i = x'_0 + i\Delta$.

Гистограмма частот строится на интервалах разбиения так, чтобы прямоугольники гистограммы имели высоту $h_i = \dfrac{n_i}{\Delta}$ (плотность частоты), тогда сумма площадей этих прямоугольников будет равна объему выборки $n$.

Гистограмма относительных частот состоит из прямоугольников высотой $h_i = \dfrac{w_i}{\Delta}$ (плотность относительной частоты). В этом случае сумма площадей прямоугольников будет равна 1.

# Метод равных частот

Отрезок изменения вариант разбивается на неравные по длине промежутки так, чтобы в каждый из них входило по возможности одинаковое число значений выборки.

Желаемое число значений в одном интервале: $\ell = \left\lfloor \frac{n}{N} \right\rfloor$.

Если значения в выборке не повторяются, то в качестве промежуточных точек разбиения берут $x'_i = \frac{x_{i\ell-1} + x_{i\ell}}{2}$.

В случае, если значения в выборке повторяются и получится так, что $x_{i\ell-1} = x_{i\ell}$, то все равные значения относят одному интервалу группировки.

Гистограмма частот: $h_i = \frac{n_i}{\Delta x'_i}$, где $\Delta x'_i = x'_i - x'_{i-1}$.

Гистограмма относительных частот: $h_i = \frac{w_i}{\Delta x'_i}$.

# Пример

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 2,29 | 2,75 | 1,70 | 1,63 | 1,38 | 2,22 | 2,12 | 2,64 | 2,10 | 2,47 |
| 1,96 | 1,46 | 2,18 | 1,51 | 1,57 | 2,38 | 1,48 | 2,25 | 2,07 | 2,07 |
| 1,59 | 2,05 | 1,78 | 1,26 | 2,07 | 1,90 | 1,14 | 1,84 | 2,48 | 2,06 |
| 1,74 | 1,84 | 2,04 | 1,61 | 1,51 | 1,96 | 1,67 | 2,67 | 1,11 | 1,98 |
| 2,67 | 1,38 | 1,68 | 1,99 | 1,99 | 2,25 | 1,36 | 1,95 | 1,61 | 1,43 |
| 1,73 | 1,77 | 1,97 | 1,70 | 2,31 | 1,95 | 1,93 | 1,32 | 2,03 | 1,49 |
| 2,61 | 2,82 | 1,24 | 1,20 | 2,62 | 3,34 | 1,65 | 1,47 | 1,30 | 1,45 |
| 1,96 | 1,70 | 1,48 | 1,33 | 2,23 | 2,37 | 1,38 | 1,64 | 3,02 | 1,58 |
| 1,61 | 1,99 | 1,08 | 2,32 | 2,72 | 1,37 | 2,71 | 1,67 | 2,37 | 2,00 |
| 2,13 | 2,11 | 2,22 | 1,47 | 2,11 | 1,25 | 1,95 | 2,00 | 1,24 | 2,49 |

Объем выборки $n = 100$. Количество интервалов группировки $N = \lfloor \sqrt{n} - 0{,}013n + 0{,}5 \rfloor = \lfloor 10 - 1{,}3 + 0{,}5 \rfloor = 9$.

Минимальная варианта 1,08; максимальная варианта 3,34.

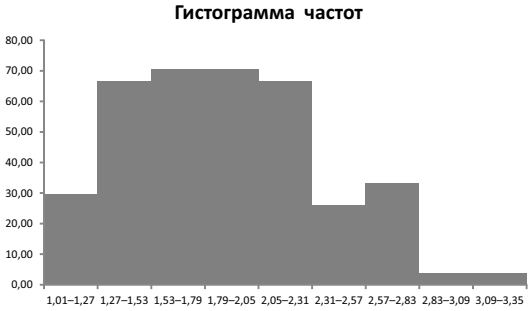
Размах выборки 2,27.

Возьмем $x'_0 = 1,01$, а $x'_9 = 3{,}35$.

Длина интервала группировки $\Delta = \dfrac{2{,}43}{9} = 0{,}26$.

# Пример (окончание)

| Интервал | $n_i$ | $n_i/\Delta$ |
|---|---|---|
| 1,01–1,27 | 8 | 29,63 |
| 1,27–1,53 | 18 | 66,67 |
| 1,53–1,79 | 19 | 70,37 |
| 1,79–2,05 | 19 | 70,37 |
| 2,05–2,31 | 18 | 66,67 |
| 2,31–2,57 | 7 | 25,93 |
| 2,57–2,83 | 9 | 33,33 |
| 2,83–3,09 | 1 | 3,70 |
| 3,09–3,35 | 1 | 3,70 |



# Эмпирическая функция распределения дискретной случайной величины

**Определение**

Эмпирической функцией распределения называется функция

$$F^*(x) = \frac{n_x}{n},$$

где $n_x$ — число значений выборки, меньших $x$, а $n$ — объем выборки.

Эмпирическую функцию распределения называют также функцией распределения выборки.

Она служит аппроксимацией функции распределения случайной величины, т. к. из закона больших чисел следует, что

$$\lim_{n\to\infty} P\left(|F(x) - F^*(x)| < \varepsilon\right) = 1.$$

# Свойства эмпирической функции распределения

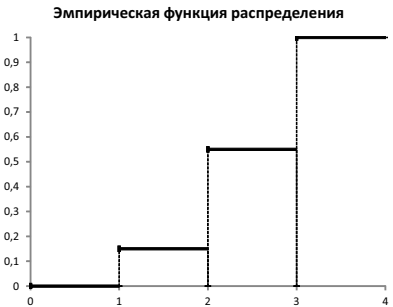
1. $0 \leqslant F^*(x) \leqslant 1$.
2. $F^*(x)$ — неубывающая ступенчатая функция.
3. $F^*(x) = 0$ для $x \leqslant \widetilde{x}_1$ и $F^*(x) = 1$ для $x > \widetilde{x}_m$.

Эти свойства очевидно следуют из определения $F^*(x)$.

# Пример

| Варианты $\widetilde{x}_k$ | 1 | 2 | 3 |
|---|---|---|---|
| Относительные частоты $w_k$ | 0,15 | 0,40 | 0,45 |
| Накопленные отн. частоты $\sum_{i=1}^{k} w_i$ | 0,15 | 0,55 | 1 |

$$F^*(x) = \begin{cases} 0 & \text{при } x \leqslant 1; \\ 0{,}15 & \text{при } 1 < x \leqslant 2; \\ 0{,}55 & \text{при } 2 < x \leqslant 3; \\ 1 & \text{при } x > 3. \end{cases}$$



# Эмпирическая функция распределения непрерывной случайной величины

Эмпирическую функцию распределения непрерывной случайной величины не имеет смысла определять как ступенчатую функцию.

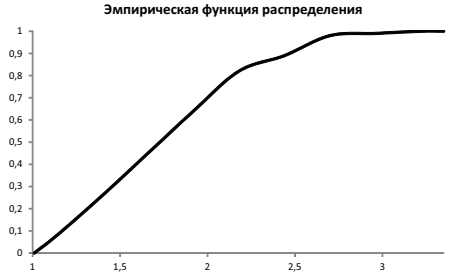
Ее интерполируют по точкам $(x'_0, 0)$, $(x'_N, 1)$ и $\left(\overline{x}_i, \dfrac{n_{x'_i}}{n}\right)$, где $\overline{x}_i = \dfrac{x'_{i-1} + x'_i}{2}$ — середина интервала группировки, $i = 1, \ldots, N$.

Если используется линейная интерполяция, то график такой функции — ломаная линия, состоящая из отрезков прямых.

Если используется сглаживание, например, кубическими сплайнами, то график — гладкая линия.

# Пример

| $k$ | Интервал | $\overline{x}_k$ | $w_k$ | $\sum_{i=1}^{k} w_i$ |
|---|---|---|---|---|
| 1 | 1,01–1,27 | 1,14 | 0,08 | 0,08 |
| 2 | 1,27–1,53 | 1,40 | 0,18 | 0,26 |
| 3 | 1,53–1,79 | 1,66 | 0,19 | 0,45 |
| 4 | 1,79–2,05 | 1,92 | 0,19 | 0,64 |
| 5 | 2,05–2,31 | 2,18 | 0,18 | 0,82 |
| 6 | 2,31–2,57 | 2,44 | 0,07 | 0,89 |
| 7 | 2,57–2,83 | 2,70 | 0,09 | 0,98 |
| 8 | 2,83–3,09 | 2,96 | 0,01 | 0,99 |
| 9 | 3,09–3,35 | 3,22 | 0,01 | 1 |



# Эмпирическая плотность распределения вероятностей

Теоретическая плотность вероятности:

$$f(x) = F'(x) = \lim_{\Delta x \to 0} \frac{P(x < X < x + \Delta x)}{\Delta x}.$$

Аналогично можно определить значения эмпирической плотности распределения вероятностей в точках $\overline{x}_i$ как

$$f^*(\overline{x}_i) = \frac{\overline{P}_n(x'_{i-1} < X < x'_i)}{\Delta x'_i} = \frac{w_i}{\Delta x'_i},$$

где $\overline{P}_n$ — функция статистической вероятности (частоты появления события).

Эмпирическую плотность распределения интерполируют по точкам $\left(\overline{x}_i, \dfrac{w_i}{\Delta x'_i}\right)$.

# Пример

| $k$ | Интервал | $\overline{x}_k$ | $w_k$ | $w_i/\Delta$ |
|---|---|---|---|---|
| 1 | 1,01–1,27 | 1,14 | 0,08 | 0,31 |
| 2 | 1,27–1,53 | 1,40 | 0,18 | 0,69 |
| 3 | 1,53–1,79 | 1,66 | 0,19 | 0,73 |
| 4 | 1,79–2,05 | 1,92 | 0,19 | 0,73 |
| 5 | 2,05–2,31 | 2,18 | 0,18 | 0,69 |
| 6 | 2,31–2,57 | 2,44 | 0,07 | 0,27 |
| 7 | 2,57–2,83 | 2,7 | 0,09 | 0,35 |
| 8 | 2,83–3,09 | 2,96 | 0,01 | 0,04 |
| 9 | 3,09–3,35 | 3,22 | 0,01 | 0,04 |