

Лекция Описательная статистика

- 1 Расчет параметров описательной статистики при качественной вариации признака
- 2 Расчет параметров описательной статистики при количественной вариации признака
- 3 Теоретические распределения случайных величин и их свойства (нормальное и биномиальное распределение, распределение Пуассона).

Совокупность

```
graph TD; A[Совокупность] --> B[Математическая]; A --> C[Статистическая]; C --> D[Генеральная]; C --> E[Выборочная]; E --> F[Большая (N>30)]; E --> G[Малая (N≤30)];
```

Математическая

Статистическая

Генеральная

Выборочная

Большая ($N > 30$)

Малая ($N \leq 30$)

Чтобы выборочная совокупность как можно полнее отражала генеральную, необходимо учитывать следующие основные положения.

- 1. Выборка должна быть вполне представительной, или типичной, т.е. в ее состав входят варианты, наиболее полно отражающие генеральную совокупность (рост деревьев в высоту - исключить деревья, сломанные бурей, поврежденные огнем).
- 2. Выборка должна быть объективной (случайный отбор).
- 3. Выборка должна быть качественно однородной (особи одного пола, вида, возраста).

Правила вычисления результатов

(по П.М. Крылову)

- ▶ **Правила 1-5.** При сложении и вычитании, умножении и делении, возведении в квадрат или куб, извлечении корня квадратного или кубического, при использовании логарифмов - в результате нужно сохранять столько десятичных знаков после запятой, сколько их имеет “слагаемое” с наименьшим количеством десятичных знаков.
- ▶ **Правило 6.** Для промежуточного результата, получаемого по правилам 1-5, необходимо сохранить одну дополнительную “запасную” цифру; в конечном результате ее отбрасывают.
- ▶ **Правило 7.** Если исходные данные имеют разное количество десятичных знаков или значащих цифр, то их надо предварительно округлить с сохранением одной “запасной” цифры.
- ▶ **Правило 8.** Если результаты должны быть получены с n -значащими цифрами, то исходные данные следует брать с $n+1$ значащей цифрой.

Статистические показатели разделяются на две группы:

- показатели, которые характеризуют центральную тенденцию, или уровень ряда,
- показатели, измеряющие степень вариации.

К первой группе относятся различные средние величины:

- ▶ мода, медиана,
- ▶ средняя арифметическая, средняя геометрическая.

Ко второй:

- ▶ вариационный размах, лимиты,
- ▶ коэффициент вариации,
- ▶ среднее квадратическое отклонение (варианса, дисперсия),
- ▶ коэффициенты асимметрии и эксцесса.

Средние величины

1. *Средняя арифметическая (M или $\bar{X}$)* - важная обобщающая характеристика вариационного ряда, сохраняющая свойства отдельной единицы совокупности:

$$\bar{X} = \sum x/n \quad \text{или} \quad \bar{X} = \sum fx/n$$

В нашем примере с определением массы бурозубок средняя величина $\bar{X} = 9.298412698$ г.

(следует оставлять на одну значащую цифру больше, чем имели значения вариантов, а оценки ошибок - на две значащих цифры. Масса тела бурозубок колебалась от 7.3 до 11.9 г, отсюда средняя с учетом округления должна иметь вид $\bar{X} = 9.29$ г.

1. **Средняя арифметическая ($\bar{M}$ или $\bar{X}$)** - важная обобщающая характеристика вариационного ряда, сохраняющая свойства отдельной единицы совокупности:

$$\bar{X} = \sum x/n \quad \text{или} \quad \bar{X} = \sum fx/n$$

В нашем примере с определением массы бурозубок средняя величина $\bar{X} = 9.298412698$ г.

(следует оставлять на одну значащую цифру больше, чем имели значения вариантов, а оценки ошибок - на две значащих цифры. Масса тела бурозубок колебалась от 7.3 до 11.9 г, отсюда средняя с учетом округления должна иметь вид $\bar{X} = 9.29$ г.

4. Средняя квадратическая - применяется для получения истинного значения площади сечения или объема признака

$$\bar{x}_{kv} = \sqrt{\frac{1}{n} \sum x_i^2 n_i}.$$

5. Средняя гармоническая - применяется для вычисления средней характеристики признаков, которые представляют собой отношение двух других варьирующих величин.

$$\bar{X}_h = N / (\sum n/x),$$

Таблица - Основные виды средних в биологии

(по К.Е. Никитину и А.З. Швиденко)

Средняя	Формулы для вычисления средних	
	Малая выборка (невзвешенные средние)	Большая выборка (взвешенные средние)
Арифметическая	$\bar{X} = \frac{1}{n} \sum_{i=1}^n x_i$	$\bar{X} = \frac{1}{n} \sum_{i=1}^k x_i n_i$
Геометрическая	$G = \sqrt[n]{x_1 x_2 \dots x_n}$ $\lg G = \frac{1}{n} \sum \lg x_i$	$G = \sqrt[n]{x_1^{n_1} x_2^{n_2} \dots x_k^{n_k}}$ $\lg G = \frac{1}{n} \sum n_i \lg x_i$
Гармоническая	$H = \frac{n}{\sum \frac{1}{x_i}}$	$H = \frac{n}{\sum n_i \frac{1}{x_i}}$
Квадратическая	$\bar{X}_g \sqrt{\frac{1}{n} \sum x_i^2}$	$\bar{X}_g \sqrt{\frac{1}{n} \sum x_i^2 n_i}$

5. **Квантили** являются обобщением понятия медианы.

Если разделить частоты распределения на k равных частей, то $k-1$ значение случайной величины, соответствующее точкам деления, называется **квантилями распределения**.

При $k=2$ единственный квантиль будет медианой, при $k=4$ средняя из точек деления будет медианой, а первая и третья - нижним и верхним квартилями.

ПОКАЗАТЕЛИ ВАРИАЦИИ

В зависимости от характера изучаемого признака различают варьирование:

количественным признакам

непрерывное,
прерывистое (дискретное)

качественным признакам

Атрибутивное (степень окраски, консистенции, форма, вид, а также альтернативное, т.е. противопоставления здоровые - больным, сильные - слабым, окрашенные - неокрашенным, присутствующие - отсутствующим и т.д.

1. Показатель (Lim) указывает фактические границы вариабельности признака.
2. Вариационный размах
3. Среднее линейное отклонение, может иметь значение только при условии, что отклонения вариант от средней арифметической берутся без учета знаков, так как в противном случае $\sum(x - \bar{X}) = 0$.

$$\Delta = \frac{\sum |x - \bar{X}|}{n}$$

4. Чтобы преодолеть недостатки линейного отклонения, принято отклонения вариант от средней арифметической возводить в квадрат и сумму квадратов отклонений относить к общему числу наблюдений, т.е. к объему выборки. Полученный таким образом показатель служит центральным моментом второго порядка, он характеризует рассеяние вариант около средней арифметической. Этот показатель, называемый **дисперсией**, или **вариансой**, обозначается греческой буквой σ^2 (сигма малая в квадрате) и выражается следующей формулой:

$$\sigma^2 = \frac{\sum (x_i - \bar{x})^2}{n}$$

5. Среднее квадратическое отклонение. Среднее квадратическое отклонение, как и средняя арифметическая, относится к величинам именованным и выражается в тех же величинах, что и признак. Выборка, в которой рассеяние вариантов около средней арифметической больше, характеризуется и большей величиной среднего квадратического отклонения и, наоборот, при меньшей вариабельности признака среднее квадратическое отклонение оказывается меньшим.

$$\sigma = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n}}$$

7. Коэффициент вариации представляет собой показатель изменчивости изучаемого признака, выраженный в относительных единицах, обычно в процентах. Он определяется по формуле

$$V = \frac{\sigma}{X} \cdot 100\%$$

где V - коэффициент вариации;

σ - среднееквадратическое отклонение;

X - среднее значение.

Коэффициент вариации применяется при сравнительной оценке варьирования различных признаков. При $<30\%$ - выборка имеет большую степень концентрации вариант возле величины. При $30\% \leq \leq 100\%$ - степень концентрации допустимая. При $>100\%$ - делается вывод о неоднородности выборки.

8. Показатель асимметрии

9. Показатель эксцесса

Рабочие формулы

$$A = \sum f(x_i - X)^3 / n \cdot \sigma^3$$

$$E = \sum f(x_i - X)^4 / n \cdot \sigma^4 - 3$$

ФУНКЦИИ РАСПРЕДЕЛЕНИЯ. НОРМАЛЬНОЕ РАСПРЕДЕЛЕНИЕ

- ▶ Наиболее универсальной в биологии является уравнение кривой нормального распределения или функция Гаусса-Лапласа.
- ▶ Ее суть заключается в том, что частота отклонения отдельных вариантов от средней арифметической данной совокупности есть функция их величины. Вероятность частоты той или иной варианты в генеральной совокупности и определяется этой функцией.

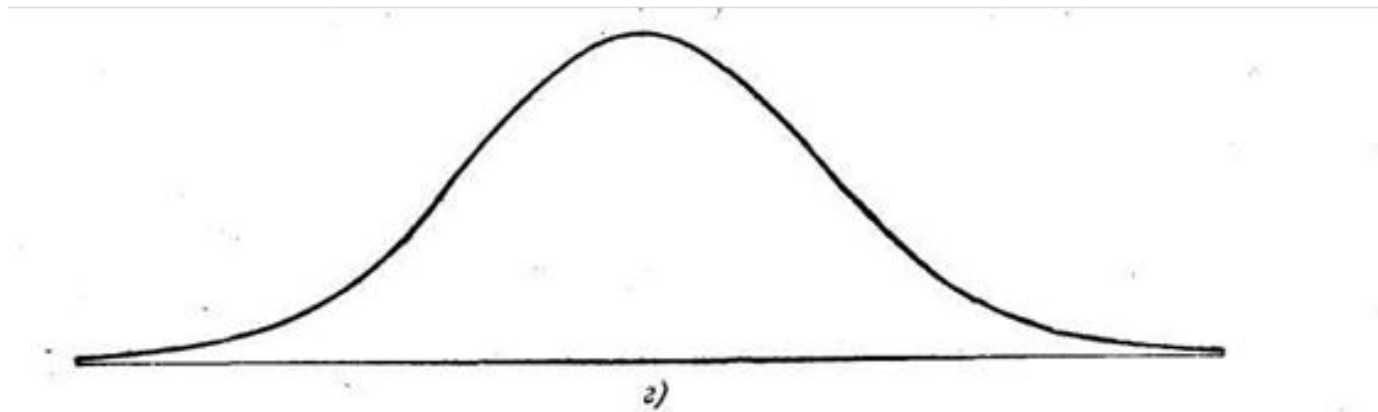


Рисунок 1- Кривая нормального распределения

Уравнение кривой нормального распределения выражает зависимость теоретических численностей $f(x)$ или y от значений x , являющейся непрерывно распределяющейся случайной величиной.

$$y = \frac{1}{\sigma \sqrt{2\pi}} \cdot e^{-\frac{1}{2} \left(\frac{x_i - \bar{x}}{\sigma} \right)^2}$$

Здесь y - ордината или высота кривой на любом расстоянии от X , т. е. центра распределения. Вправо от этого центра случайная величина x_i имеет положительные, а влево - отрицательные значения.

Так как π - 3,14593 и e - основание натуральных логарифмов, равное 2,7183, являются постоянными величинами, следовательно величина t - нормированное отклонение.

$$\frac{x_i - \bar{X}}{\sigma} = t$$

Кривая нормального распределения обладает следующими свойствами:

- Однозначно определяется двумя параметрами: $\bar{X}$ - средним значением и σ – среднеквадратическим отклонением.
- Кривая симметрична относительно среднего значения ($\bar{X}$) и имеет колоколообразную форму, которая зависит от величины σ , являющейся параметром масштаба, а положение определяется $\bar{X}$.
- Кривая имеет один максимум, равный $\frac{1}{\sigma\sqrt{2\pi}}$ и две точки перегиба на расстоянии $\pm\sigma$ от $\bar{X}$.
- Ветви кривой асимптотически приближаются к оси абсцисс на расстояние $\pm\infty$.

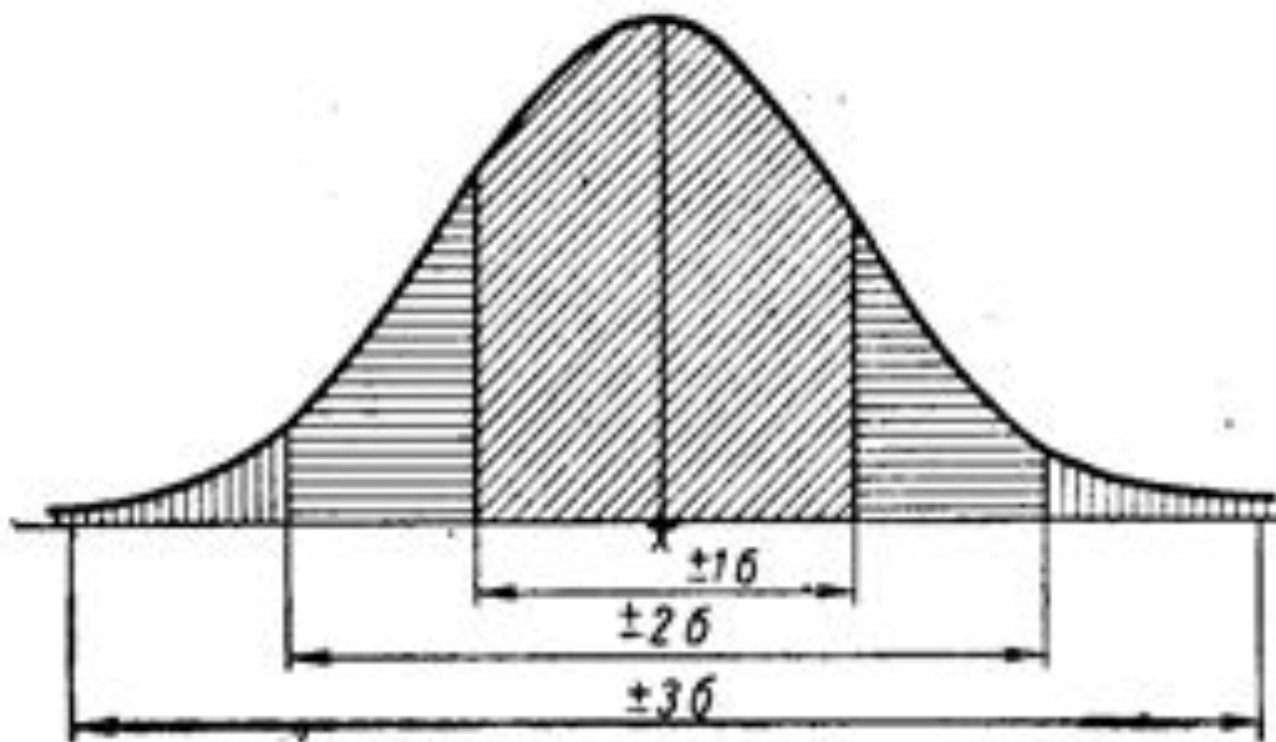


Рисунок 2 - Графическая иллюстрация “правила плюс - минус трех сигм”

«Правило трех сигм»

- ▶ в пределах $\pm 1\sigma$ находится 68,28% всех вариантов эмпирической совокупности, распределяющейся по нормальному закону;
- ▶ в пределах $\pm 2\sigma$ заключено 95,45%,
- ▶ в пределах $\pm 3\sigma$ содержится 99,73% всех вариантов совокупности.

Биномиальное распределение

Если значения частот вариант выразить в виде распределения вероятностей, получим биномиальное распределение $(a+b)^n$ - бином Ньютона

$$(a + b)^2 = a^2 + 2ab + b^2$$

$$(a+b)^n = a^n + C_n^1 a^{n-1} b + C_n^2 a^{n-2} b^2 + \dots + C_n^k a^{n-k} b^k + \dots + C_n^{n-1} a b^{n-1} + C_n^n b^n$$

где C_n^k - число сочетаний из n элементов по k .

$$C_n^k = \frac{A_n^k}{P_k} = \frac{n(n-1)(n-2)\dots(n-k+1)}{1*2*3\dots k} = C_n^{n-k}$$

Например, при $n=8$ и $k=5$ уравнение (7.2) будет иметь вид

$$C_8^5 = \frac{A_8^5}{P_5} = \frac{8 * 7 * 6 * 5 * 4}{1 * 2 * 3 * 4 * 5} = 56$$

$$C_8^5 = C_8^3 = \frac{8 * 7 * 6}{1 * 2 * 3} = 56$$

$$a_{n,m} = C_n^m a^m b^{n-m} = \frac{n!}{m!(n-m)!} a^m b^{n-m}$$

- Формула конечного члена (формула Бернули):

$$a_{n,m} = C_n^m a^m b^{n-m} = \frac{n!}{m!(n-m)!} a^m b^{n-m}$$

$a=1-b$; C_n^m - число сочетаний из n элементов по m ; ! – знак факториала, т.

е. $a, b, c, \dots, n! = a * b * c * \dots * n!$ например $5! = 1 * 2 * 3 * 4 * 5 = 120$

Теорема верна для $m=0$ или $m=n$.

Распределение Пуассона.

Распределение Пуассона, или пуассоново распределение, подобно биномиальному, относится к дискретной или прерывистой изменчивости. Оно имеет самостоятельное значение, хотя его можно рассматривать и как предельный случай биномиального. При биномиальном распределении значения a и b могут быть близки друг к другу, при пуассоновом же a очень мало, т.е. события осуществляются очень редко, а b приближается к единице.

$$y = C_n^m p^m q^{n-m} = \frac{n(n-1) \dots (n-m+1)}{m!} * \frac{\lambda}{n^m} \left(1 - \frac{\lambda}{n}\right)^{n-m}$$

где $\lambda = np$; $p = \lambda / n$.

Так как числитель первой дроби имеет m сомножителей, а в знаменателе стоит n^m каждый из сомножителей можно разделить на n . Получим:

$$y = \left(1 - \frac{1}{n}\right) \left(1 - \frac{2}{n}\right) \dots \left(1 - \frac{m-1}{n}\right) * \frac{\lambda^m}{m!} \left(1 - \frac{\lambda}{n}\right)^{n-m}$$

В биологии к распределению Пуассона прибегают, когда исследуют мутации при проведении генетико-селекционных исследований, при анализе появления альбиносов среди лесных зверей и птиц, для оценки выживаемости самосева сосны под пологом леса и т.д., при анализе других редких событий.