Выборки

<u>*Замечание:*</u>
Отличие теории вероятности от математической статистики – в статистике неизвестна функция распределения (например, задача оценить вес кильки в балтийском море, пытаемся спрогнозировать вес рыб путём вылавливания).

Пусть X_1, X_2, … - независимые одинаково распределённые случайные величины с функцией распределения F. Функция распределения F неизвестная.

Опр.
F – генеральная совокупность (вес всех рыб в Балтийском море, оценки всех студентов БФУ, параметры всех деталей, производимых на каком-то предприятии). Генеральная совокупность считается бесконечной.

Опр.
$X_1,\ldots,X_n$ – выборка из генеральной совокупности (например из моря вылавливается 1000 рыб, тогда $n=1000$).
n – размер выборки.
$g(X_1,\ldots,X_n)$ – статистика – есть функция выборки (необходима случайная величина).

Примеры статистик:
$M=max(X_1,\ldots,X_n)$, $m=min(X_1,\ldots,X_n)$;
$R_n=M-m$ – размах выборки;
$X_1+\ldots+X_n$, $X_1+X_3-X_n$.

Опр.
$X_{(1)}\le\ldots\le X_{(n)}$ порядковые статистики $X_{1,n}\le\ldots\le X_{n,n}$ получены из величины $X_1,\ldots,X_n$ путем расположения их в порядке возрастания.
(два варианта обозначения, в скобках или через запятую)

Порядковые статистики – зависимые величины и не одинаково распределённые.
$max(X_1,\ldots,X_n)=X_{(n)}=X_{1,n}$;
$min(X_1,\ldots,X_n)=X_{(1)}=X_{n,n}$.

Основные задачи статистики:
- По $X_1,\ldots,X_n$ найти функцию распределения F.

- Оценка параметров $F(x|a)$ (в статистике функция распределения, как правило, зависит от параметров) (ВАЖНО ДЛЯ СТАТИСТИКИ, при различных параметрах свойства функций отличаются).
- Проверка статистических гипотез (“Проверить, что доля брака в выбранной партии не превосходит 0.5%“, так как есть задачи, в которых нужно выбрать четкий, конкретный ответ, к примеру - “Зависимы величины или нет?” “Бракованная партия или нет?“).
- Предсказание на основе известных данных новых данных (регрессионный анализ) (зависимость одной переменной от другой - “Как производительность труда влияет на зарплату?“).
- Определение атипичных наблюдений (робастный анализ).

Опр.

Эмпирическая (выборочная) функция распределения:

$$F_n^*(x)=\frac{1}{n}\sum_{i=1}^{n} I(X_i\leq x)$$

Где $I(A)=\begin{Bmatrix}1, A\\ 0, \bar{A}\end{Bmatrix}$, $i=(\overline{1,2,\ldots,n})$

Свойства:

(1) $F_n^*(-\infty)=0$; $F_n^*(\infty)=1$.

(2) неубывающая, форма лесенкой.

(3) $0\leq F_n^*(x)\leq 1$.

Основное отличие выборочной функции распределения от обычной дискретной – выборочная функция распределения это случайная величина.

Теорема (Гливенко-Картелли):

$$\sup_x |F_n^*(x)-F(x)| \xrightarrow{n.н} 0$$

Это значит, что при больших n выборочная функция распределения очень точно описывает $F(x)$.

Чем больше n (чем больше наблюдений), тем точнее выборка похожа на генеральную совокупность, однако дольше надо проводить наблюдения.

Замечание:

3-5 наблюдений это не статистика, минимально 20, чем больше тем лучше.

Принципы статистики:

- Выборка должна быть большой – первый принцип статистики

- Презентативная выборка (пример, узнать какие доходы в семье, социологи ездят по особнякам и спрашивают какие доходы, такой подход односторонний, берётся высший класс и спрашиваются доходы, необходимо спрашивать и средний класс и бедный, смотреть пропорции) – второй принцип статистики (разносторонняя статистика).

Опр.

Порядковые статистики: **дискретный случай**, $X_{(1)} \le X_{(2)} \le \ldots \le X_{(n)}$

Располагаем в возрастающем порядке. Некоторые статистики могут совпадать. Пусть n_k – количество совпадений $X_{(k)}$ (так как нет смысла писать одинаковые).

$X_{(1)}$	$X_{(2)}$	$\ldots$	$X_{(k)}$
n_1	n_2	...	n_k

Тогда:

$$n_1 + n_2 + n_3 + \ldots + n_k = n$$

n_j– частоты.

$F_n^*(x) = \dfrac{n_x}{n}$ – число частот меньших или равных x.

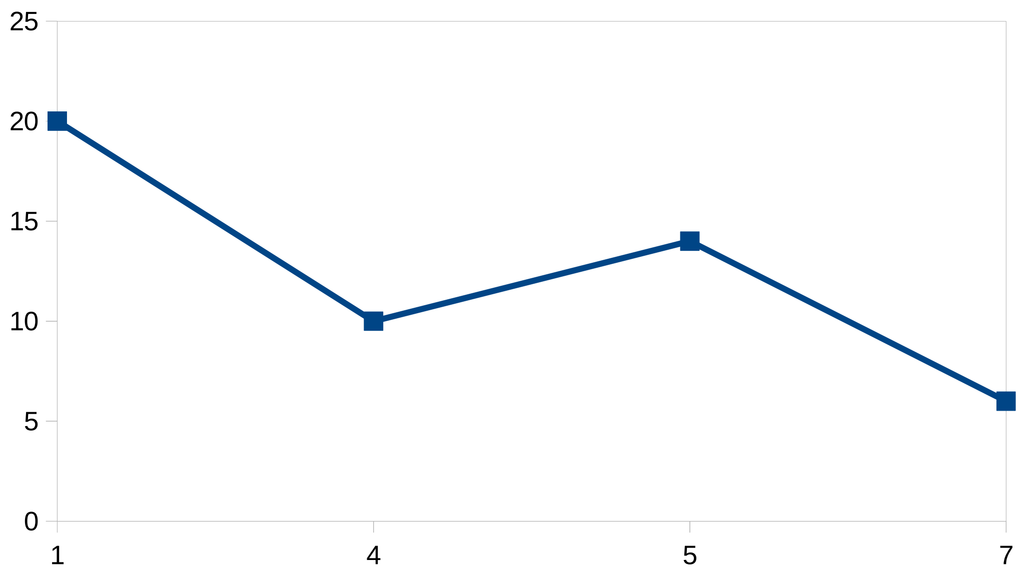
$\varpi_i = \dfrac{n_i}{n}$ – относительные частоты, $\sum_{i=1}^{n} \varpi_i = n$.

Опр.

Полигоном частот (в дискретном случае) называют ломаную, соединяющую точки $(x_1, n_1), \ldots, (x_k, n_k)$.

Полигоном относительных частот называют ломаную, соединяющую точки $(x_1, \varpi_1), \ldots, (x_k, \varpi_k)$.

x_i	1	4	5	7
n_i	20	10	14	6

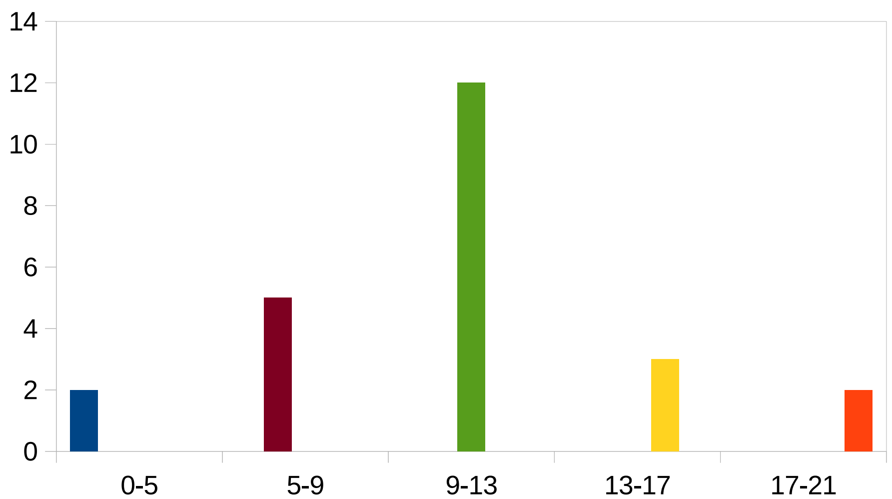


Опр.

В непрерывном случае часто строят гистограмму частот, гистограмму относительных частот. В непрерывном случае n_k – количество наблюдений, попадающих в k-ый интервал.

Высотой сколько людей попало в интервал строим прямоугольник.

Пример:



Замечание:
Пусть $f_n(x)$ – кривая, аппроксимирующая гистограмму.
$f_n(x) \to f(x)$ – где $f(x)$ - плотность генеральной совокупности (при больших n будет достаточно близка к неизвестной плотности).

Экспоненциальное семейство распределений

Опр.
Пусть:
$f(x|a)$ – плотность, либо функция вероятности.
$f(x) = F'(x)$ – плотность.
$f(x) = P(X=x)$ – функция вероятности (дискретный случай).

Если выполняется $f(x|a) = h(x)c(a)e^{\sum_{i=1}^{n} w_i(a)t_i(x)}$, где $c, h > 0$, то говорят, что распределение принадлежит экспоненциальному семейству распределений.

Пример:
Нормальное, гамма, бета, биномиальное, Пуассона, отрицательное биномиальное принадлежат экспоненциальному семейству распределений.

(1) Биномиальное:

$$f(x|p) = C_n^x p^x (1-p)^{(n-x)} = C_n^x (1-p)^n e^{x ln\left(\frac{p}{1-p}\right)}$$

$$h(x) = C_n^x, c(p) = (1-p)^n, w_1 = \ln\left(\frac{p}{1-p}\right), t_1 = x$$

=> *биномиальное распределение* $\in$ *экспоненциальному семейству распределений* .

(2) Нормальное:

$$f(x|a,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-a)^2}{2\sigma^2}}$$

$$c(a,\sigma^2) = e^{-\frac{a^2}{2\sigma^2}}, h=1, w_1 = \frac{-1}{2\sigma^2}, w_2 = \frac{a}{\sigma^2}, t_1 = x^2, t_2 = x$$

=> *нормальное распределение* $\in$ *экспоненциальному семейству распределений*.

(3) Распределение Коши – нет. Не разделить x и a:

$$f(x|a,\sigma)=\frac{1}{\left(a\sigma\left(1+\left(\frac{x-a}{\sigma}\right)^2\right)\right)}$$

=> *распределение Коши* $\notin$ *экспоненциальному семейству распределений*.

Основная идея экспоненциального семейства распределений:
Всегда есть что-то, что зависит только от x, и оно умножается на что-то, что зависит от только параметра.

Для чего нужны?
Например, для нахождения достаточной статистики, идея нахождения у распределений одного семейства одинаковая. Также сильно облегчает поиск полной статистики.

Выборочное среднее, выборочная дисперсия:

Опр.
Пусть:

$$EX=a,\ DX=\sigma^2$$

Тогда:

$\bar{X}=\frac{1}{n}\sum_{i=1}^{n}X_i$ – выборочное среднее (аналог мат. ожидания).

$S^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2$ (аналог дисперсии, делится на $n-1$ потому что так надо).

Важное свойство:

$$(n-1)S^2=\sum_{i=1}^{n}(X_i^2-2\bar{X}X_i+\bar{X}^2)=$$

$$\sum_{i=1}^{n}X_i^2-2\bar{X}\sum_{i=1}^{n}X_i+n\bar{X}^2=\sum_{i=p}^{n}X_i^2-n\bar{X}^2(*)$$

Справедливы и следующие свойства:

(1) $E\,\bar{X} = a$

(2) $D\,\bar{X} = D\left(\dfrac{\sum\limits_{i=1}^{n} X_i}{n}\right) = \dfrac{\sigma^2}{n}$

$$ES^2 = (*) = \frac{1}{(n-1)}\left(E\left(\sum_{i=1}^{n} X_i^2\right) - nE\,\bar{X}^2\right) =$$

(3) $$\frac{1}{n-1}\left(\left(\sum_{i=1}^{n} DX_i + (EX_i)^2\right) - nE\,\bar{X}^2\right) =$$

$$\frac{1}{n-1}\left(\left(n\sigma^2 + na^2\right) - n\left(D\,\bar{X} + \left(E\,\bar{X}\right)^2\right)\right) = \sigma^2$$