Генеральная совокупность и случайная выборка

Статистическая устойчивость - в пределе бесконечно большом числе измерений

На практике исследователь - лишь с ограниченным числом наблюдений, а результаты в силу закона случая в общем случае могут не совпадать с теми же величинами, вычисленными по большому числу наблюдений, выполненных в тех же условиях.

В математической статистике - абстрактная *генеральная совокупность* (все допустимых значения) и *выборка* (совокупность ограниченного числа значений).

Выборочные характеристики по ограниченному числу наблюдений зависят от этого числа. Им соответствуют им характеристики генеральной совокупности как оценка соответствующих характеристик генеральной совокупности.

*Репрезентативная в*ыборка (представительная) - дает достаточное представление об особенностях генеральной совокупности, но она случайна и любое суждение о генеральной совокупности по ней тоже случайно. Поэтому…

Выборка $x_1, x_2, \ldots, x_n$ случайной величины X.

n_x - число выборочных значений **левее** x по числовой оси X.

n_x / n - частота появления значений X, меньших x и является функцией от x.

Эта функция, получаемая по выборке, называется *эмпирической* или *выборочной* функцией распределения (в отличие от распределения генеральной совокупности) и обозначается как:

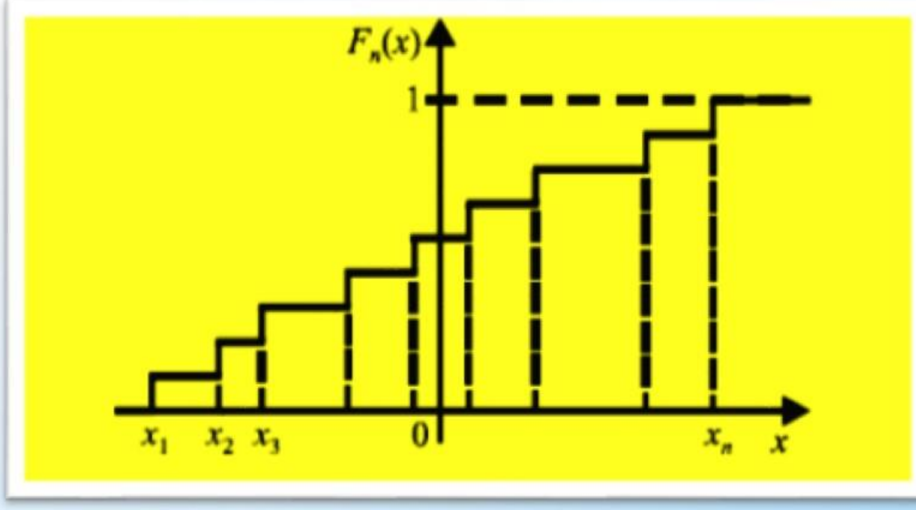
$$F_n(x) = n_x / n$$

Можно доказать, что с вероятностью, равной 1:

$$\lim_{n\to\infty} \{F_n(x) - F(x)\} = 0$$

Пусть $x_1, x_2, \ldots, x_n$ - упорядоченная по величине выборка, или вариационный ряд. Все элементы ее имеют одинаковую вероятность $1/n$. Поэтому

$$F_n(x) = 0, \qquad x < x_1$$
$$F_n(x) = k/n, \quad x_k \le x < x_{k+1}, \quad k = 1, 2, \ldots, n-1$$
$$F_n(x) = 1, \qquad x \ge x_n$$



Обычно метод «сгруппированных данных»:
выборка объема n преобразуется в статистический ряд:
Весь диапазон от $x_{\min}$ до $x_{\max}$ на k равных интервалов .
Их число интервалов можно выбирать произвольно или по эмпирическим формулам, например: **k =1+1.39 ln (n)** с округлением до ближайшего целого.
Длина интервала равна **h = ($x_{\min}$ - $x_{\max}$) / k**.

n_j - число элементов в j-ом интервале.

$p^*_j = n_j / n$ - относительная частота попадания величины в j-ый интервал.

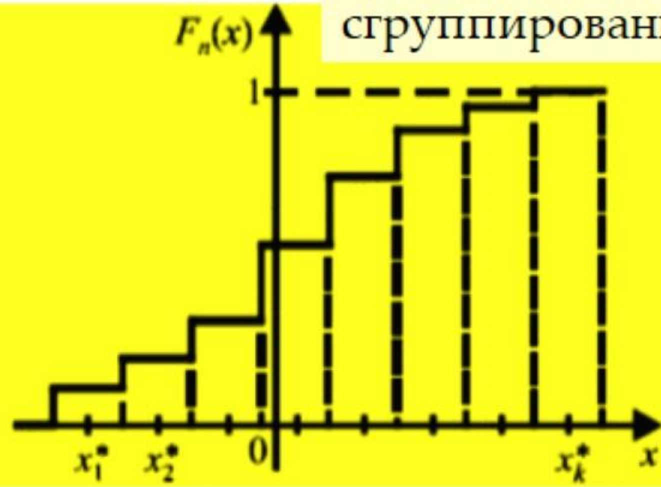
Все точки, попавшие в j-интервал, относят к его середине: $x_j^* = \dfrac{x_{j-1}+x_j}{2}$

Статистический ряд.

Интервал	Длина интервала	Середина интервала	Число точек в интервале	Относительная частота
1	(x_{min}, x_1)	x_1^*	n_1	p_1^*
2	(x_1, x_2)	x_2^*	n_2	p_2^*
…	…	…	…	…
k	(x_{k-1}, x_{max})	x_k^*	n_k	p_k^*
Σ			n	1

Гистограмма распределения

График функции $F_n(x)$, по сгруппированным данным



По выборкам могут быть рассчитаны выборочные статистические характеристики (выборочное среднее, дисперсия и т.д.), которые являются оценками соответствующих генеральных параметров.

Оценка $a^*(x_1, x_2, ..., x_n)$ называется ***состоятельной***, если с увеличением объема выборки ***n*** она стремится (по вероятности) к оцениваемому параметру *a*. ***Эмпирические (выборочные) моменты являются состоятельными оценками теоретических моментов***.

Оценка $a^*(x_1, x_2, ..., x_n)$ называется ***несмещенной***, если ее математическое ожидание при любом объеме выборки равно оцениваемому параметру а, т. е. ***M* [*a**] = *a*** **всегда.**

Важной характеристикой оценок генеральных параметров является также их ***эффективность***, которая для *различных несмещенных оценок одного и того же параметра при фиксированном объеме выборок обратно пропорциональна дисперсиям этих оценок.*

Метод максимального правдоподобия

Окружим каждую точку x_i окрестностью длины δ вероятность попадания в $(x_i - \delta / 2)$, $(x_i + \delta / 2)$ приближенно равна $f(x, a)\,\delta$. Если произведено n наблюдений, то вероятность того, что одновременно первое наблюдение попадет в первый интервал, второе — во второй и т.д., есть вероятность совместного осуществления всех этих независимых событий и равна:

$P(x, a) = f(x_1, a) \cdot f(x_2, a) \cdot \ldots \cdot f(x_n, a) \cdot \delta^n = = f(x_1) \cdot f(x_2) \cdot \ldots \cdot f(x_n) \cdot \delta^n$

$$L(x,a) = \ln \frac{P(x,a)}{\delta^n} = \sum_{i=1}^{n} \ln f(x_i, a)$$ - функция правдоподобия

Сущность метода заключается в нахождении таких оценок неизвестных параметров, для которых функция правдоподобия при случайной выборке объема n будет иметь максимальное значение. Т.е. найти такую совокупность допустимых значений параметров a_1^*, a_2^*, …, a_k^*, которая обращает функцию правдоподобия в максимум.

$$\left|\frac{\partial P(x,a)}{\partial a}\right|_{a=a^*} = 0, \qquad \frac{\partial^2 P(x,a)}{\partial a^2} < 0.$$

Оценка математического ожидания и дисперсии нормально распределенной случайной величины

Применим вышесказанное к нормальному распределению

$$f(x)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{(x-m)^2}{2\sigma^2}\right]$$

$$P(x,m,\sigma^2)\approx\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-m)^2\right]\cdot\delta^n$$

$$L(x,m,\sigma^2)=\ln\left(\frac{P}{\delta^n}\right)=-\frac{n}{2}\ln 2\pi-\frac{n}{2}\ln\frac{1}{2\sigma^2}-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-m)^2$$

$$\frac{\partial L}{\partial m}=\frac{2}{2\sigma^2}\sum_{i=1}^{n}(x_i-m)=0 \qquad \Rightarrow\sum_{i=1}^{n}(x_i-m)=0$$

$$\frac{\partial L}{\partial\sigma^2}=-\frac{n}{2\sigma^2}+\frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n}(x_i-m)^2=-\frac{1}{2\sigma^2}\left[n-\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-m)^2\right]=0\Rightarrow\left[n-\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-m)^2\right]=0$$

Тогда оценка для математического ожидания равна $m^*=\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$

Тогда оценка для дисперсии равна $(\sigma^2)^*=s_1^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2$

Метод максимального правдоподобия всегда приводит к состоятельным, хотя иногда и смещенным оценкам (зависящей от выборки), имеющим наименьшую возможную дисперсию при неограниченном возрастании объема выборки.
Так, выборочная дисперсия s_1^2 называется смещенной оценкой генеральной дисперсии:

$$M\left[s_1^2\right] = \frac{n-1}{n}\sigma^2$$

Для получения несмещенной оценки: $s^2 = \dfrac{n}{n-1}s_1^2 = \dfrac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}{n-1}$

Уменьшение знаменателя в на единицу непосредственно связано с тем, что величина $\bar{x}$, относительно которой берутся отклонения, сама *зависит от элементов выборки*. Каждая величина, зависящая от элементов выборки и входящая в формулу выборочной дисперсии, *называется связью*. Можно доказать, что знаменатель выборочной дисперсии всегда равен *разности между объемом выборки* $\boldsymbol{n}$ *и числом связей* l, наложенных на эту выборку.

Эта разность $f = n - 1$ называется *числом степеней свободы выборки*.

В практических вычислениях для выборочной дисперсии часто более удобна следующая формула, получаемая из путем арифметических преобразований:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i)^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}{n-1}$$

Можно показать, что дисперсия среднего в *n* раз меньше дисперсии единичного измерения, поэтому для стандартного отклонения

$$\sigma(\bar{x}) = \frac{\sigma}{\sqrt{n}}$$

Если принять $\boldsymbol{\sigma(\bar{x})}$ в качестве меры случайной ошибки среднего выборки, то *увеличение числа параллельных определений одной и той же величины снижает величину случайной ошибки. Это свойство случайной величины используют на практике для повышения точности результатов измерений.*

Доверительные интервалы и доверительная вероятность, уровень значимости.

Пусть для генерального параметра а получена из опыта несмещенная оценка *а**. Назначим достаточно большую вероятность β (такую, что событие с вероятностью β можно считать практически достоверным) и найдем такое значение $\varepsilon_\beta = f(\beta)$, для которого:

$$P\left(\left|a^*-a\right| \le \varepsilon_\beta\right)=\beta \quad \longrightarrow \quad a^*-\varepsilon_\beta \le a \le a^*-\varepsilon_\beta$$

Уровень значимости: $p = 1-\beta$

β - доверительная вероятность (характеризует надежность полученной оценки)

$\boldsymbol{I_\beta = a^* \pm \varepsilon\beta}$ - доверительный интервал

$\boldsymbol{a' = a^* - \varepsilon\beta}$ и $\boldsymbol{a'' = a^* + \varepsilon\beta}$ - доверительные границы

Увеличение числа опытов проявляется в сокращении доверительного интервала при постоянной доверительной вероятности или в повышении доверительной вероятности при сохранении доверительного интервала

При построении доверительного интервала решается задача об абсолютном отклонении:

$$P\left(\left|a^*-a\right|\le\varepsilon_\beta\right)=P\left(\left|\Delta a\right|\le\varepsilon_\beta\right)=F\left(\varepsilon_\beta\right)-F\left(-\varepsilon_\beta\right)=\int_{-\varepsilon_\beta}^{\varepsilon_\beta}f(a)da=\beta$$

Наилучшей оценкой для математического ожидания *m* является среднее выборки среднего со стандартным отклонением среднего

$$\sigma\left(\bar{x}\right)=\frac{\sigma}{\sqrt{n}}$$

Используя функцию Лапласа, получаем

$$P\left(\left|a^*-a\right|\le\varepsilon_\beta\right)=\beta=2\Phi\left(\frac{\varepsilon_\beta}{\sigma\left(\bar{x}\right)}\right)=2\Phi\left(k_\beta\right)$$

И задавшись доверительной вероятностью β, определяем по таблице для Φ (x):

$$\bar{x}-k_\beta\sigma\left(\bar{x}\right)\le m_x\le\bar{x}+k_\beta\sigma\left(\bar{x}\right)$$

Закон распределения оценки a^* зависит от закона распределения величины X и, в частности, от самого параметра a. Тогда при $n\ge 50$ заменяют $\sigma\left(\bar{x}\right)\approx s\left(\bar{x}\right)$ или переходят к другой величине, которая не зависит от этого параметра a.

Квантиль

Кванти́ль в математической статистике x_α — значение, которое заданная случайная величина не превышает с фиксированной вероятностью. Если вероятность задана в процентах, то квантиль называется *процентилем* или *перцентилем*

$$P(X \le x_\alpha) \ge \alpha$$
$$P(X \ge x_\alpha) \ge 1 - \alpha$$

Если распределение непрерывно, то квантиль однозначно задаётся уравнением

$$F(x_\alpha) = \alpha$$

$$P\left(x_{\frac{1-\alpha}{2}} \le X \le x_{\frac{1+\alpha}{2}}\right) = \alpha$$

0,25 - квантиль (**α = 0.25**) - первый (нижний) кварти́ль (лат. quarta — четверть)
0,50 - квантиль (**α = 0.50**) - второй (медиана) кварти́ль (лат. mediāna — середина)
0,75 - квантиль (**α = 0.75**) - третий (или верхний) кварти́ль

Уи́льям Си́ли Го́ссет (William Sealy Gosset, 13 июня 1876, Кентербери — 16 октября 1937, Беконсфильд) — британский учёный-статистик, более известный под своим псевдонимом Стьюдент (Student) благодаря своим работам по исследованию т. н. распределения Стьюдента. Исследования были обращены к нуждам пивоваренной компании и проводились на малом количестве наблюдений. Чтобы предотвратить раскрытие коммерческой информации Госсет вынужден был опубликовать свои работы под псевдонимом Стьюдент, чтобы скрыть себя от работодателя.