## 1 Генеральная совокупность, выборка, выборочный метод

Математической статистикой называется наука, занимающаяся разработкой методов получения, описания и обработки опытных данных с целью изучения закономерностей случайных массовых явлений.

Все задачи математической статистики касаются вопросов обработки наблюдений над массовыми случайными явлениями, но в зависимости от характера измеряемой величины (или системы величин), цели измерения при обработке результатов измерений эти задачи могут принимать ту или иную форму. Типичными задачами, важными по своим практическим применениям, являются:

1. Оценка на основании результатов измерений неизвестных законов распределений измеряемых СВ и случайных ошибок измерений.

2. Определение по результатам измерений приближенного значения *МОЖ, Д(х),* $\sigma(x)$, $K_{xy}$ измеряемой величины и оценка её точности.

3. Статистическая проверка гипотез. Задача ставится так:
на основании некоторых соображений можно считать, что функция распределения исследуемой *СВ X* есть *F(x).* Совместимы ли наблюдаемые значения с гипотезой, что *СВ X* действительно имеет функцию распределения *F(x),* или не опровергают ли опытные данные ту гипотезу, что параметры закона распределения имеют предположенные значения.

В основе математической статистики лежит ряд исходных понятий.

В практике статистических исследований различают два вида наблюдений: ***сплошное,*** когда изучаются все объекты (элементы, единицы) совокупности, и несплошное, ***выборочное,*** когда изучается часть объектов.

Сплошное наблюдение - перепись населения страны, выборочное - проводимые социологические исследования, охватывающие часть населения страны.

Вся подлежащая изучению совокупность объектов (наблюдений) называется ***генеральной совокупностью***. Число членов (число $N$), образующих генеральную совокупность, называется ***объемом генеральной совокупности***.

Так как в большинстве случаев можно произвести сколько угодно измерений, то генеральная совокупность бесконечна. Однако на практике сплошное обследование применяется сравнительно редко, т.к. проводить сплошное обследование не имеет смысла, В таких случаях случайно отбирают из всей совокупности ограниченное число объектов и подвергают их исследованию.

***Выборочной совокупностью***, или просто ***выборкой,*** называется совокупность случайно отобранных объектов из генеральной совокупности. Число членов (число $n$), образующих выборку является её объемом. (Причем, $n << N$).

***Сущность выборочного метода*** состоит в том, чтобы по некоторой части генеральной совокупности (по выборке) выносить суждение о её свойствах в целом.

**Преимущества выборочного метода** наблюдения по сравнению со сплошным:

- позволяет существенно экономить затраты ресурсов (материальных, трудовых, временных);
- является единственно возможным в случае бесконечной генеральной совокупности или в случае, когда исследование связано с уничтожением наблюдаемых объектов (например, предельных режимов работы приборов);
- при тех же затратах ресурсов дает возможность проведения углубленного исследования за счет расширения программы исследования;
- позволяет снизить ошибки ***регистрации,*** т.е. расхождения между истинным и зарегистрированным значениями признака.

Основной **недостаток выборочного метода**

- ошибки исследования, называемые ***ошибками репрезентативности*** (представительства)

***Чтобы по данным выборки иметь возможность судить о. генеральной совокупности, она должна быть отобрана случайно.***

Случайность отбора элементов в выборку достигается соблюдением ***принципа равной возможности*** всем элементам генеральной совокупности быть отобранными в выборку. На практике это достигается путем жеребьевки (лотереи) или с помощью случайных чисел.

***Выборка называется репрезентативной (представительной),*** если она достаточно хорошо воспроизводит генеральную совокупность.

Различают следующие виды выборок:

- ***собственно-случайная выборка***, образованная случайным выбором элементов без расчленения на части или группы;
- ***механическая выборка,*** в которую элементы из генеральной совокупности отбираются через определенный интервал;
- ***типическая (стратифицированная) выборка***, в которую случайным образом, отбираются элементы из типических групп, на которые по некоторому признаку разбивается генеральная совокупность;
- ***серийная (гнездовая) выборка,*** в которую случайным образом отбираются не элементы, а целые группы совокупности (серии), а сами серии подвергаются сплошному наблюдению.

Используют **два способа** образования выборки:

1. ***повторный отбор***, когда каждый элемент, случайно отобранный и обследованный, возвращается в общую совокупность и может быть повторно отобран (по схеме возвращенного шара),
2. ***бесповторный отбор,*** когда отобранный элемент не возвращается в общую совокупность (по схеме невозвращенного шара).

Важнейшей **задачей выборочного метода** является оценка параметров (характеристик) генеральной совокупности по данным выборки.

## 2. Представление статистических данных и оценивание закона распределения генеральной совокупности

Пусть исследуется некоторая дискретная или непрерывная *СВ X*, закон распределения которой известен. С этой целью над *СВ X*, проводится ряд независимых испытаний. Результаты измерений представляют в виде таблицы, состоящей из двух строк, в первой - указываются номера измерений $i$, а во второй - результаты измерения $x_i$ называемые вариантами.

| $i$ | 1 | 2 | … | $n$ |
|---|---|---|---|---|
| $x_i$ | $x_1$ | $x_2$ | … | $x_i$ |

Такую таблицу в математической статистике называют статистическим рядом. Статистический ряд представляет собой первичную форму записи статистического материала и может быть обработан различными способами.

Одним из способов такой обработки является построение статистического распределения *СВ X*.

Статистическим рядом распределения *СВ* называется таблица, в первой строке которой указываются полученные в результате наблюдения значения *СВ X*, а во второй - соответствующие им частоты

| $x_i$ | $x_1$ | … | $x_k$ |
|---|---|---|---|
| $m_i$ (частоты) | $m_1$ | … | $m_k$ |

| $x_i$ | $x_1$ | … | $x_k$ |
|---|---|---|---|
| $p_i^* = \frac{m_i}{n}$ (относит.част.) | $p_1^* = \frac{m_1}{n}$ | … | $p_k^* = \frac{m_k}{n}$ |

$$\left( \sum_{i=1}^{k} p_i^* = \sum_{i=1}^{k} \frac{m_i}{n} = 1 \right)$$

В теории вероятностей под законом распределения ДСВ понимают всякое соотношение, устанавливающее связь между возможными значениями *СВ* и соответствующими вероятностями, а в математической статистике ***статистический закон распределения устанавливает соответствие между наблюдаемыми значениями СВ и соответствующими им частотами.***

Статистический ряд распределения можно построить как *ДСВ*, так и для *НСВ*. Однако согласно теореме Бернулли при неограниченном увеличении числа опытов $n$ частоты наблюдаемых значений сходятся по вероятности к вероятностям этих значений. Это значит, что для *ДСВ X* при неограниченном увеличении числа опытов статистический ряд распределения сходится по вероятности к ряду распределениях *СВ X*, а для *НСВ X* - частоты сходятся по вероятности к нулю.

При большом числе опытов над *НСВ X* (или *ДСВ*, которая имеет счетное множество возможных значений производят подсчет результатов наблюдений, попадающих в определенные группы (интервалы), и составляют

таблицу, в которой указываются концы интервалов (группы) и частота получения результатов наблюдения в каждом интервале.

| интерв. (группы) | $(a_0;a_1)$ | $(a_0;a_1)$ | … | $(a_{k-1};a_k)$ |
|---|---|---|---|---|
| частоты $p_i^* = \frac{m_i}{n}$ | $p_1^* = \frac{m_1}{n}$ | $p_2^* = \frac{m_2}{n}$ | … | $p_k^* = \frac{m_k}{n}$ |

$$\sum_{i=1}^{k} p_i^* = 1$$

Такое распределение называют ***статистической совокупностью.***

Если при группировке наблюдаемых значений имеем значение, которое в точности лежит на границе двух групп, то необходимо прибавить к числам $m_i$ одного и другого интервала по 1/2. Число групп выбирают порядка 10 -20. Длины интервалов могут быть как одинаковыми, так и разными. Однако при оформлении данных о *СВ* распределение крайне неравномерно, иногда бывает удобно выбирать в области наибольшей плотности распределения интервалы узкие, чем в области малой плотности.

## 3 Эмпирические распределения

Эмпирический функцией распределения *СВ X* называют функцию *F(x),* определяющую для каждого значения аргумента *x* частоту события *X<x*

$$F^*(x) = p^*(X < x) \qquad \text{(I)}$$

Чтобы найти значение эмпирической функции распределения при данном *x,* надо подсчитать число опытов, в которых *СВ X* приняла значения, меньшее, чем *x* и разделить его на общее число произведенных опытов.

Эмпирическая функция распределения любой *СВ* (*ДСВ* или *НСВ* имеет вид

$$F^*(x) = \sum_{X_i < x} p^*(X = x_i), \qquad (2)$$

где символ $X_i < x$ под знаком суммы обозначает, что суммирование распространяется на все те наблюдаемые в результате опытов значения *СВ X*, которые по своей величине меньше аргумента *x*.

Согласно выражению (2) $F^*(x)$, любой *СВ* всегда ***разрывна*** и возрастает скачками при переходе через точки, которые соответствуют наблюдаемым значениям *СВ*, причем величина скачка равна частоте соответствующего значения.

Согласно теореме Бернулли при неограниченном увеличении числа опытов $n$ $p^*(X < x) \xrightarrow[n\to\infty]{p} p(X < x)$.

Это значит, что $F^*(x) \xrightarrow[n\to\infty]{p} F(x)$.

***Основное значение*** эмпирической функции распределения *F*(x)* состоит в том, что она ***используется в качестве оценки неизвестной вероятностной функции распределения F(x).***

Из определения эмпирической функции распределения следует,

что:

1. $0 \le F^*(x) \le 1$.

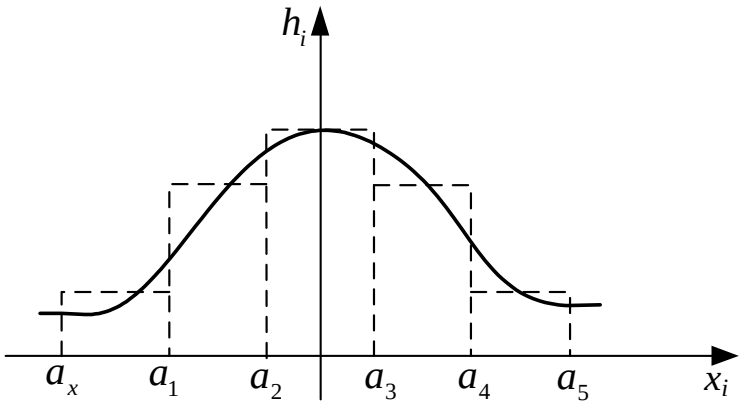
2. Эмпирическая функция распределения есть неубывающая функция, т.е. при $x_2 > x_1$ $F^*(X_2) \le F^*(X_1)$.

В целях наглядности статистические распределения представляют графиками. Наиболее распространенными являются полигон и гистограмма.

***Полигоном*** частот (относительных частот) называется графическое изображение статистического ряда распределения.

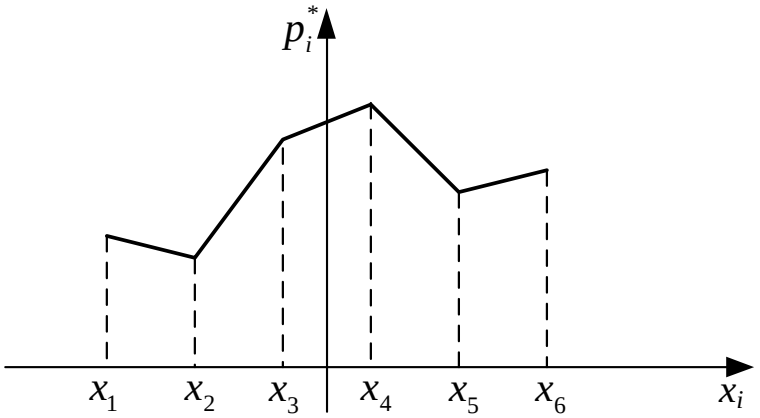
Полигон относительных частот строится следующим образом:

на оси абсцисс откладываются наблюдаемые значения $X_i$, а на оси ординат соответствующие относительные частоты $p_i^*$ (частоты $m_i$). Точки $(X_i, p_i^*)$ соединяют отрезками прямых и получают полигон. Полигон строится аналогично многоугольника распределения для *ДСВ*. Следовательно, согласно теореме Бернулли при увеличении числа опытов *n* ***полигон частот для ДСВ будет всё более приближаться к. многоугольнику распределения.***




***Гистограммой*** называется графическое изображение статистической совокупности.

Гистограмма строится следующим образом: на оси абсцисс откладываются интервалы и на каждом из них, как на основании, строится прямоугольник, площадь которого равна частоте данного интервала, т.е. $p_i^* = h_i l$, где $l$ - длина интервала, $h_i$ - высота. Из способа построения следует, что её полная площадь равна единице.



Если точки гистограммы (например, середины верхних оснований прямоугольников) соединить главной линией, то она в первом приближении будет представлять график плотности вероятности *СВ X*, т.е, ***гистограмма является оценкой неизвестной вероятностной функции -плотности вероятности.***

***Кумулятивная кривая*** (кривая накопленных частот или накопленных относительных частот (частостей)) строится следующим образом. В системе координат строят точки ( $X_i$ , $p_x^{*нак}$ ) или ( $X_i$ , $m_x^{нак}$ ), где $X_i$ - наблюдаемые значения, $p_x^{*нак}$ - соответствующие накопленные относительные частоты (частоты). Полученные точки соединяют отрезками.