# ПРАКТИЧЕСКОЕ ЗАНЯТИЕ № 4

## ОЦЕНКА ПАРАМЕТРОВ ГЕНЕРАЛЬНОЙ СОВОКУПНОСТИ ПО ВЫБОРОЧНЫМ ДАННЫМ

**Цель работы**

1. Изучить основы первичной обработки данных;
2. Научиться строить гистограмму;
3. Освоить вычисление точечных и интервальных оценок выборочных данных.

Основными понятиями математической статистики являются: **генеральная совокупность, выборка, теоретическая функция распределения**.

***Генеральная совокупность*** – это множество всех мыслимых значений наблюдений (объектов), однородных относительно некоторого признака, которые могли быть сделаны. Число всех наблюдений, составляющих генеральную совокупность, называется ее объемом N. Например, популяция представляет собой множество индивидуумов. Изучение целой популяции трудоемко и дорого и, может быть, просто невозможно. Поэтому собирают данные по выборке индивидуумов, которых считают представителями этой популяции, позволяющими сделать вывод относительно этой популяции.

***Выборка*** – это совокупность случайно отобранных наблюдений (объектов) для непосредственного изучения из генеральной совокупности. Объем выборки $n$. Выборка обязательно должна удовлетворять условию **репрезентативности**, т.е. давать обоснованное представление о генеральной совокупности. Как сформировать репрезентативную (представительную) выборку? В идеале стремятся получить случайную (**рандомизированную)** выборку. Для этого составляют список всех индивидуумов в популяции и случайно их отбирают. Но иной раз затраты при составлении списка могут оказаться недопустимыми и тогда берут приемлемую выборку, например, одну клинику, больницу и исследуют всех пациентов в этой клинике с данным заболеванием.

Каждый элемент выборки $x_i$ называется **вариантой**. Число наблюдений варианты $n_i$ называется **частотой встречаемости**. Последовательность вариант, записанных в **<u>возрастающем порядке</u>**, называется **вариационным рядом.**

***Гистограмма частот*** – это ступенчатая фигура, состоящая из смежных прямоугольников, построенных на одной прямой, основания которых одинаковы и равны ширине класса, а высота равна или частоте попадания в интервал $n_i$ или относительной частоте $\frac{n_i}{n}$.

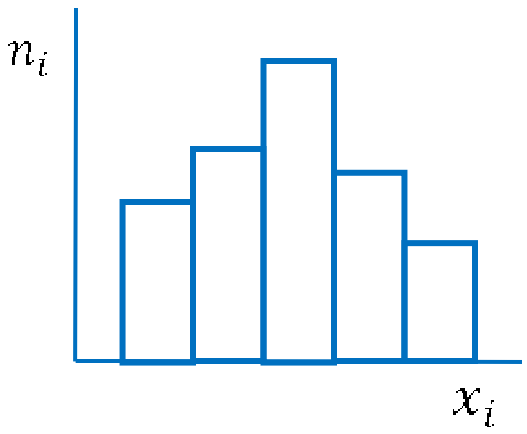
Ширину интервала $i$ можно определить по формуле Стерджеса:

$$i = \frac{x_{\max} - x_{\min}}{1 + 3.32\lg n},$$

где $x_{\max}$ – максимальное,

а $x_{\min}$ – минимальное значения вариант, а их разность носит название вариационный размах.

$n$ – объем выборки.




**Пример 4.1.** Наблюдения за числом частиц, попавших в счетчик Гейгера, в течение минуты дали следующие результаты:
21 30 39 31 42 34 36 30 28 30 33 24 31 40 31 33 31 27 31 45 31 34 27 30 48 30 28 30 33 46 43 30 33 28 31 27 31 36 51 34 31 36 34 37 28 30 39 31 42 37.

Построить по этим данным интервальный вариационный ряд с равными интервалами (I интервал 20-24; II интервал 24-28 и т.д.) и начертить гистограмму.

Р е ш е н и е .

| Интервал | 20-24 | 24-28 | 28-32 | 32-36 | 36-40 | 40-44 | 44-48 | 48-52 |
|---|---|---|---|---|---|---|---|---|
| Частота | 1 | 4 | 22 | 8 | 7 | 4 | 2 | 2 |

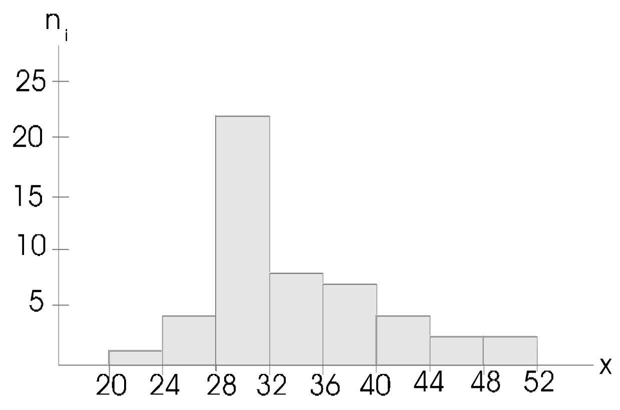



Рис.4.1. Гистограмма распределения

### *Оценка параметров генеральной совокупности по ее выборке*

Смысл статистических методов заключается в том, чтобы по выборке ограниченного объема n, т.е. по некоторой части генеральной совокупности, высказать **обоснованное суждение о свойствах генеральной совокупности.**

Числовые значения, характеризующие генеральную совокупность, называются параметрами. Одна из задач математической статистики – определение параметров большого массива по исследованию его части.

Статистическое оценивание может выполняться двумя способами:

1) ***Точечная оценка*** – оценка, которая дается для некоторого параметра одним значением.

2) ***Интервальная оценка*** – по данным выборки оценивается интервал, в котором лежит истинное значение параметра с заданной вероятностью.

## *Точечная оценка параметров генеральной совокупности*

***Точечная оценка*** – это оценка, которая определяется одним числом. И это число определяется по выборке. Это функция результатов выборки, и она является точечной оценкой генерального параметра, т.е. принимает только одно значение.

Качество оценки устанавливается по трем свойствам: быть состоятельной, эффективной и несмещенной.

Точечная оценка называется ***состоятельной,*** если при увеличении объема выборки выборочная характеристика стремится к соответствующей характеристике генеральной совокупности.

Точечная оценка называется ***эффективной,*** если она имеет наименьшую дисперсию выборочного распределения по сравнению с другими аналогичными оценками.

Точечную оценку называют ***несмещенной***, если ее математическое ожидание равно оценивающему параметру при любом объеме выборки.

***Несмещенной оценкой генеральной средней*** (математического ожидания) служит выборочная средняя $\bar{x}_в$:

$\bar{x}_в = \frac{1}{n}\sum_{i=1}^{K} x_i n_i$ , где

$x_i$ - варианта выборки;

$n_i$ - абсолютная частота встречаемости варианты $x_i$;
$n$ - объем выборки.

**Выборочная средняя является несмещенной оценкой генеральной средней,** так как $M(\bar{x}_в) = \bar{x}_{ген}$, то есть она эквивалентна истинному среднему в генеральной совокупности.

**Выборочная дисперсия** $S_в^2$ не обладает свойством несмещенности. Это **смещенная оценка генеральной дисперсии** $\sigma_{ген}^2$.

$$\sigma^2_{\text{ген}} \neq M(S^2_{\text{в}})$$

На практике используют **исправленную дисперсию** $S^2$, которая является несмещенной оценкой дисперсии генеральной совокупности:

$$S^2 = \frac{n}{n-1} S_в^2(x) = \frac{n}{n-1} \frac{\sum_{i=1}^{K} \left(x_i - \bar{x}_в\right)^2 \cdot n_i}{n};$$

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{K} \left(x_i - \bar{x}_в\right)^2 \cdot n_i.$$

$$\sigma^2_{\text{ген}} = M(S^2)$$

Кроме того, в расчетах используют ***S*** **- исправленное среднее квадратическое отклонение,** называемое **стандартным отклонением** в выборке и ошибку выборочной средней $m_{\bar{x}}$

$m_{\bar{x}} = \dfrac{S}{\sqrt{n}}$, которая отражает точность оценки

Стандартная ошибка уменьшится, т.е. оценка станет более точной, если объем выборки *n* увеличится и данные имеют небольшое рассеяние *S*.

Рассмотрим разницу между *S* – стандартным отклонением в выборке и *m* – стандартной ошибкой среднего. На первый взгляд они очень схожи, но их используют в разных целях.

Среднее квадратическое отклонение *S* отражает вариабельность в значениях данных, и его указывают, если надо пояснить изменчивость в наборе данных, разброс данных.

Ошибка выборочной средней $m_{\bar{x}}$ характеризует точность выборочного среднего $\bar{x}_в$ и должна быть указана, если интерес представляет среднее значение выборки.

**Пример 4.2.** Из генеральной совокупности извлечена выборка объема n = 50.

| $x_i$ | 2 | 5 | 10 | 7 |
|---|---|---|---|---|

| | | | | |
|---|---|---|---|---|
| $n_i$ | 16 | 12 | 8 | 14 |

Найти несмещенную оценку генеральной средней.

Р е ш е н и е .

$$\bar{x}_в = \frac{1}{n}\sum_{i=1}^{K} x_i n_i;$$

$$\bar{x}_в = \frac{16 \cdot 2 + 5 \cdot 12 + 7 \cdot 8 + 10 \cdot 14}{50} = 5{,}76.$$

**Пример 4.3.** По выборке объема 30 найдено значение выборочной дисперсии $S_в^2 = 3$. Найти несмещенную оценку дисперсии генеральной совокупности.

Р е ш е н и е .

Эта несмещенная оценка равна исправленной дисперсии:

$$S^2 = \frac{n}{n-1} S_в^2, \qquad S^2 = \frac{30}{29} \cdot 3 = 3{,}1.$$

**Пример 4.4.** Найти несмещенную оценку генеральной средней, дисперсии генеральной совокупности и стандартное отклонение по выборке объема 12, описывающую продолжительность в секундах физической нагрузки до развития приступа стенокардии.

289, 203, 359, 243, 232, 210, 251, 251, 246, 224, 239, 220, 211.

Р е ш е н и е .

$$\bar{x}_в = \frac{289 + 203 + ... + 211}{12} = 244;$$

$$S^2 = \frac{1}{11}\left[(289-244)^2 + (203-244)^2 + ... + (211-244)^2\right] = 1849;$$

$$S^2 = 1849;$$

$$S = 43.$$

## *Интервальная оценка параметров генеральной совокупности*

Точечные оценки параметров распределения не дают информации о степени близости к соответствующему теоретическому параметру. Поэтому построение интервала, в котором с заданной степенью достоверности будет находиться оцениваемый параметр, является более информативным способом оценивания неизвестных параметров.

***Интервальная оценка*** **– это числовой интервал, который определяется двумя числами – границами интервала, содержащий неизвестный параметр генеральной совокупности**.

***Доверительный интервал*** – это интервал, в котором с той или иной заранее заданной вероятностью находится неизвестный параметр генеральной совокупности.

***Доверительная вероятность*** $p$ – это такая вероятность, что событие вероятности $1-p$ можно считать невозможным. $\alpha = 1-p$ – это уровень значимости.

(**Обозначения доверительной вероятности и уровня значимости могут отличаться от приведенных выше**). Обычно в качестве доверительных вероятностей используют **<u>вероятности, близкие к 1.</u>** Тогда событие, что интервал накроет характеристику, будет практически достоверным. Это $p \geq 0{,}95$, $p \geq 0{,}99$, $p \geq 0{,}999$.

Эти вероятности признаны достаточными для уверенного суждения о генеральных параметрах на основании известных выборочных показателей. Обычно указывают 95 % доверительный интервал.

Для выборки малого объема ($n$< 30) нормально распределенного количественного признака $x$ доверительный интервал может иметь вид:

$$\bar{x}_в - m_{\bar{x}} t \leqslant \mu \leqslant \bar{x}_в + m_{\bar{x}} t$$

где, $\mu$ – генеральное среднее;

$\bar{x}_в$ – выборочное среднее;

$t$ – нормированный показатель распределения Стъюдента, с $(n-1)$ степенями свободы, который определяется вероятностью попадания генерального параметра в данный интервал. Термин "степени свободы" означает, что их можно вычислить как объем выборки минус число ограничивающих условий.

$m_{\bar{x}}$ – ошибка выборочной средней. $m_{\bar{x}} = \dfrac{S}{\sqrt{n}}$

Для интерпретации доверительного интервала в клинических работах следует помнить, что ширина доверительного интервала зависит от $m_{\bar{x}}$ – средней ошибки выборочной средней, которая в свою очередь зависит от объема выборки ($n$) и от изменчивости данных ($S$). Если выборка небольшая, то доверительный интервал более широкий, чем в случае выборки большого объема. Широкий доверительный интервал указывает на неточную оценку, а узкий – на точную оценку.

Верхний и нижние пределы доверительного интервала показывают, будут ли результаты клинически значимы.

**Пример 4.5.** Количественный признак $x$ генеральной совокупности распределен нормально. По выборке объема $n$ = 16 найдены среднее выборочное $=20{,}2$ и среднее квадратическое отклонение $S$ = 0,8. Определить неизвестное математическое ожидание при помощи доверительного интервала при $p \geq 0{,}95$.

Р е ш е н и е .

$$\bar{x}_в - m_{\bar{x}} t \leqslant \mu \leqslant \bar{x}_в + m_{\bar{x}} t$$

Найдем $t$ из таблицы распределения Стъюдента при уровне значимости $\alpha \leq 0.05$ и числе степеней свободы $f = n$-1; $f$ = 16-1 = 15.

$t(\alpha \leq 0{,}05, f = 15) = 2{,}13$.

Запишем:

$$20{,}2 - \frac{0.8}{\sqrt{16}} \cdot 2{,}13 \leq \mu \leq 20{,}2 + \frac{0.8}{\sqrt{16}} \cdot 2{,}13 (p \geq 0{,}95)$$

$$19{,}8 \leq \mu \leq 20{,}6 \qquad (p \geq 0{,}95).$$

**Пример 4.6.** Имеется выборка объема $n$ = 11 – это значение систолического давления у мужчин в начальной стадии шока.

$x$: 127, 124, 155, 129, 77, 147, 65, 109, 145,141.

С помощью пакета прикладных программ на ЭВМ провести статистическую обработку данных выборки и определить доверительный интервал для генеральной средней при $p \geq 0{,}95$.

Р е ш е н и е . Пусть расчет на ЭВМ дал выборочное среднее $\bar{x}_в = 122{,}01$; $m_{\bar{x}} = 8{,}59$. По таблице распределения Стъюдента найдем:

$t(\alpha \leq 0{,}05, f = 11-1 = 10) = 2{,}23$; $\mu = \bar{x}_в \pm m_{\bar{x}} t$

$$\mu = 122{,}01 \pm 8{,}59 \cdot 2{,}23 \qquad (p \geq 0{,}95);$$

$$\mu = 122 \pm 19; \qquad (p \geq 0{,}95).$$