1. Основные понятия математической статистики.

Математическая статистика – это раздел математики, в котором изучаются методы обработки и анализа экспериментальных данных с помощью аппарата теории вероятностей. Предположим, что ведётся наблюдение над некоторой случайной величиной и снимаются показания, которые фиксируются. Статистический материал называется выборкой и обозначается $\underline{X}=(X_1,X_2,\ldots,X_n)$. Выборка представляет собой совокупный результат n независимых наблюдений над некоторой случайной величиной X (её называют генеральной случайной величиной). С точки зрения математической статистики выборка – это n-мерный случайный вектор с одинаково распределёнными независимыми компонентами. Величина n называется **объёмом выборки**. **Выборочное пространство** (иначе – **генеральная совокупность**) – это пространство, состоящее из реализаций вектора $\underline{X}$:

$$\aleph = \{\underline{x}=(x_1,x_2,\ldots,x_n)\},$$

где x_i – выборочное наблюдение ($i=1,2,\ldots,n$).

Если выборочные наблюдения в конкретной выборке представить в порядке возрастания числовых значений, то получим вариационный ряд $X_{(1)}\le X_{(2)}\le\ldots\le X_{(n)}$.

Как правило, при обилии данных проводят группировку данных. Число интервалов m рассчитывают по формуле Стерджеса:

$$m=1+3{,}322\lg n,$$

тогда ширина интервала (или величина интервала) равна

$$k=\frac{x_{\max}-x_{\min}}{1+3{,}322\lg n},$$

где x_{max}-x_{min} – **вариационный размах**;

x_{max} – значение X_{max}; x_{min} – значение X_{min};

$X_{max} = X_{(n)}$; $X_{min} = X_{(1)}$.

Сгруппированы данные записывают в виде таблицы:

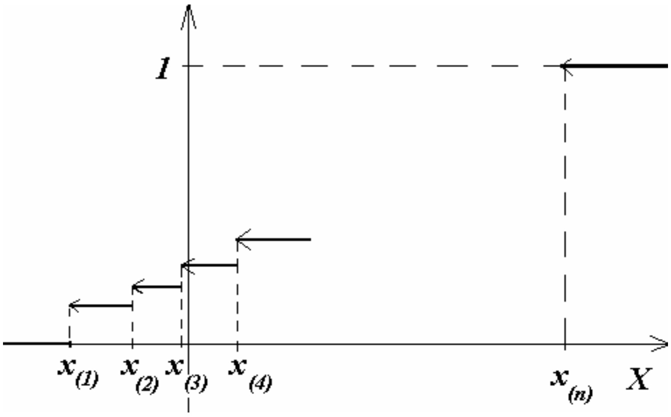
№ интервала	Интервал	Частоты
1	$(x_{(1)}, x_{(1)}+k)$	n_1
2	$(x_{(1)}+k, x_{(1)}+2k)$	n_2
…	…	…

Частота – количество данных, попавших в интервал.

Для графического изображения (представления) выборочных данных используют следующие характеристики: эмпирическую функцию распределения, гистограмму, полигон и камуляту.

Эмпирическая функция распределения имеет вид: $F_n(x)=\frac{\mu_n(x)}{n}$, где $\mu_n(x)$ – число элементов выборки, оказавшихся меньше x (-**накоплённая частота**).

При построении эмпирической функции распределения, как правило, используют не сгруппированные данные.



Воспользуемся следующей методикой, нанесём на ось ***OX*** члены вариационного ряда, затем построим ступенчатую функцию.

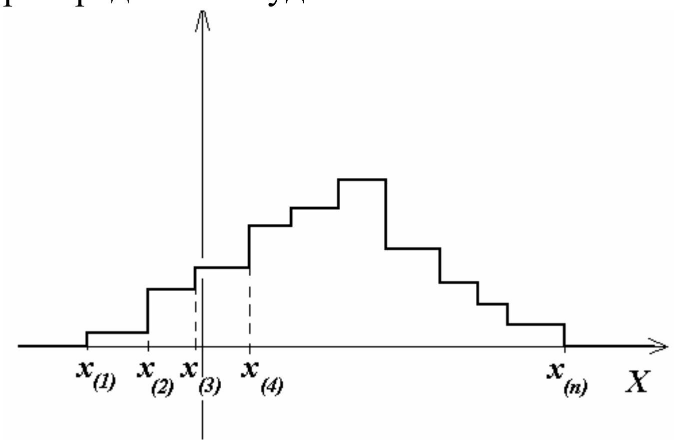
$$\text{Высота каждой ступеньки} = \begin{cases} \dfrac{n_i}{n}, \ åñëè\ äàííûå\ \ ñãðóïïèðîâàíû \\ \dfrac{1}{n}, \ åñëè\ äàííûå\ \ íå\ \ ñãðóïïèðîâàíû \end{cases}$$

На основе теоремы Бернулли для эмпирической функции распределения можно доказать **теорему Гливенко**: $\forall\, x$

$F_n(x) \xrightarrow[n\to\infty]{} F(x)$, где $F(x)$ – теоретическая функция распределения.

Также верна **теорема Колмогорова:** $\sup\limits_{x\in R}\left|F_n(x)-F(x)\right| = 0(\frac{1}{\sqrt{n}})$.

То есть, если вокруг эмпирической функции распределения построить узкую зону, то с большой вероятностью можно утверждать, что функция распределения будет лежать в этой зоне.

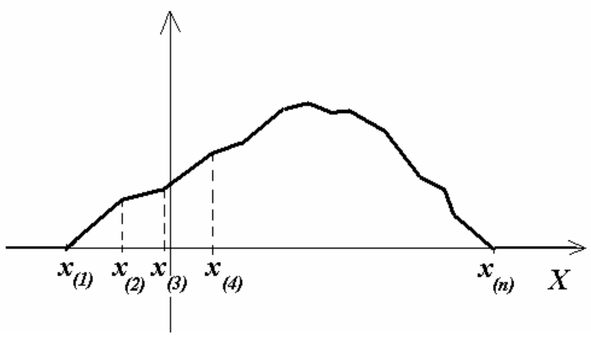


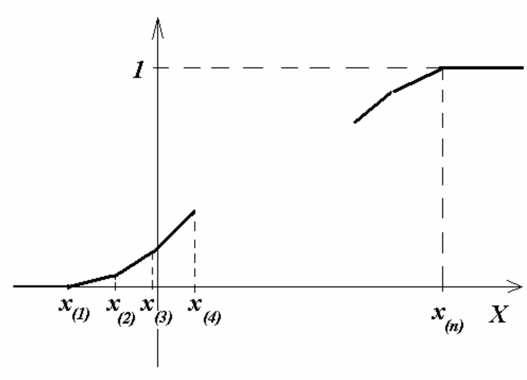
Гистограмма (аналог плотности).

При построении гистограммы используются сгруппированные данные. По оси ***OX*** откладывают интервалы шириной k от x_{min} до x_{max}. На каждом интервале строят прямоугольник площадью $p_i = \frac{n_i}{n}$ (= относительной частоте попадания в данный интервал), то есть высота прямоугольника $h_i = \frac{p_i}{k}$.

При $n\to\infty$ гистограмма приближается к плотности.

Полигон – представляет собой ломаную, концы отрезков прямой имеют координаты (x_i, n_i) {или$(x_i; \frac{n_i}{n})$}.





Кумулята (или **кривая накопленных частот**)- это ломаная, соединяющая точки $\boldsymbol{(x_i;\frac{n_i^{нак}}{n})}$ или $\boldsymbol{(x_i;n_i^{нак})}$.

Вернёмся к вариационному ряду. Вариационный ряд содержит достаточно полную информацию об изменчивости признака (с.в.$\boldsymbol{X}$), однако обилие числовых данных, с помощью которых он задаётся, усложняет их использование. На практике достаточно часто оказывается достаточным знание лишь сводных характеристик вариационного ряда: выборочных моментов, моды, медианы, асимметрии и эксцесса.

Расчёт статистических характеристик представляет собой второй после группировки данных этап обработки результатов наблюдений. Рассмотрим основные характеристики.

Если данные не сгруппированы, то **выборочная (эмпирическая) средняя**

$$\bar{x} = \frac{1}{n}\sum_{j=1}^{n} x_j .$$

Если данные приведены в сгруппированном виде, и $\boldsymbol{x_j}$ – середина интервала, n_j - соответствующая частота, то

$$\bar{x} = \frac{1}{n}\sum_{j=1}^{m} x_j n_j ,$$

где $\boldsymbol{m}$ – число интервалов группировки.

Выборочная медиана – значение признака, приходящееся на середину вариационного ряда:

$$\hat{\text{Ì}}\,\text{å} = \begin{cases} x_{\left(\frac{n+1}{2}\right)} & n - \text{íå÷¸òíî} \\ \frac{1}{2}\left[x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}\right] & n - \text{÷¸òíî} \end{cases}.$$

Медиану, как меру средней величины, используют в том случае, если крайние члены вариационного ряда по сравнению с остальными, оказались чрезмерно большими или малыми.

Выборочная мода – выборочное значение, которому соответствует наибольшая частота.

Проще всего моду найти графическим путем с помощью гистограммы (или полигона).

Выборочная (эмпирическая) дисперсия

$$S^2 = \frac{1}{\boldsymbol{n}} \sum_{\boldsymbol{j}=1}^{\boldsymbol{n}} (\boldsymbol{x_j} - \overline{\boldsymbol{x}})^2 .$$

(или для сгруппированного ряда)

$$S^2 = \frac{1}{\boldsymbol{n}} \sum_{\boldsymbol{j}=1}^{\boldsymbol{m}} (\boldsymbol{x_j} - \overline{\boldsymbol{x}})^2 \cdot \boldsymbol{n_j} .$$

Отметим важное свойство эмпирической дисперсии, известное в статистике как «*правило сложения дисперсий*». *Если выборка состоит из нескольких групп наблюдений, то общая дисперсия равна сумме средней арифметической групповых дисперсий и межгрупповой дисперсии*:

$$\mathbf{S^2} = \overline{S}^2 + \delta^2,$$

где $\overline{S}^2 = \dfrac{1}{n}\sum_{i=1}^{l} S_i n_i$ - средняя арифметическая групповых дисперсий;

$\delta^2 = \dfrac{1}{n}\sum_{i=1}^{l} (\overline{x}^{(i)} - \overline{x})^2 \cdot n_i$ – межгрупповая дисперсия;

$\mathbf{S^2} = \dfrac{1}{n_i}\sum_{j=1}^{m_i} (x_j^{(i)} - \overline{x}^{(i)})^2 \cdot n_j^{(i)}$ - общая дисперсия; $\boldsymbol{l}$ – количество групп.

Эмпирический коэффициент асимметрии

$$\hat{\boldsymbol{A}} = \frac{1}{\boldsymbol{n} \cdot \boldsymbol{S}^3} \sum_{\boldsymbol{j}=1}^{\boldsymbol{m}} (\boldsymbol{x_j} - \overline{\boldsymbol{x}})^3 \cdot \boldsymbol{n_j} .$$

Если $\hat{\boldsymbol{A}} = 0$, то распределение имеет симметричную форму. При $\hat{\boldsymbol{A}} > 0$ ($\hat{\boldsymbol{A}} < 0$) говорят о положительной (правосторонней) или отрицательной (левосторонней) асимметрии.

Эмпирический эксцесс

$$\hat{\boldsymbol{E}} = \frac{1}{\boldsymbol{n} \cdot \boldsymbol{S}^4} \sum_{\boldsymbol{j}=1}^{\boldsymbol{m}} (\boldsymbol{x_j} - \overline{\boldsymbol{x}})^4 \cdot \boldsymbol{n_j} - 3. .$$

Если $\hat{\boldsymbol{E}} > 0$ ($\hat{\boldsymbol{E}} < 0$), то полигон вариационного ряда имеет более крутую (пологую) вершину по сравнению с нормальной кривой.

Пример 1.1. Для определения заработной платы в отрасли было обследовано 100 человек.

Зар.пл.	190-192	192-194	194-196	196-198	198-200	200-202	202-204	204-206	206-208
Число чел.	1	5	9	22	28	19	11	4	1

Проведём первичную статистическую обработку данных.

$\boldsymbol{x_{min}}=191$; $\boldsymbol{x_{max}}=207$; $\overline{x}=199$, $\hat{\boldsymbol{Me}} = \hat{\boldsymbol{Mo}} = 199$; $\mathbf{S^2}=9{,}36$; $\mathbf{S}=3{,}06$;

$\hat{\boldsymbol{A}}=-0{,}039$, $\hat{\boldsymbol{E}}=-0{,}074$.

2. Теория оценивания.
2.1.Понятие оценки.

С помощью гистограммы мы можем приближенно построить график плотности (или многоугольник распределения). Вид этого графика часто позволяет высказать предположение о распределении случайной величины. В выражение плотности (или ряда) распределения обычно входят некоторые параметры, которые требуется определить (оценить) из опытных данных.

Поскольку все сведения о генеральной случайной величине сосредоточены в конкретной выборке $(X_1,X_2,\ldots,X_n)$, то естественно получать оценки числовых характеристик генеральной случайной величины в форме некоторых функций от выборочных данных.

Оценка – это некоторая функция выборки.

Оцениваемый параметр будем обозначать θ, а его оценку - $\hat{\theta}$.

$$\hat{\theta}=T(X_1, X_2,\ldots,X_n).$$

Поскольку $X_1,X_2,\ldots,X_n$ – случайные величины, то и оценка $\hat{\theta}$ (в отличие от оцениваемого параметра) является случайной величиной.

Один и тот же параметр можно оценивать различными способами, поэтому нужны критерии «качества» оценки. После введения критериев, можно ввести понятие оптимальной оценки (наилучшей по данному критерию). Каждый критерий задаётся соответствующей мерой близости оценки и оцениваемого параметра.

Оценка $\hat{\theta}$ называется **несмещённой**, если для любого $\theta\in\Theta$ верно:

$\mathbf{M}_\theta(\hat{\theta})=\theta$ (математическое ожидание совпадает с истинным значением.

Несмещенность означает, что оценка в среднем даёт желаемый результат. Требование несмещённости гарантирует отсутствие систематических ошибок при оценивании (если $\mathbf{M}_\theta(\hat{\theta})>\theta$, то оценка в среднем завышает параметр, если же $\mathbf{M}_\theta(\hat{\theta})<\theta$, то занижает его).

Оценка $\hat{\theta}$ называется **состоятельной**, если она сходится по вероятности к оцениваемому параметру:

Для любого $\varepsilon > 0$ $\quad P\{|\hat{\theta}-\theta| > \varepsilon\} \to 0$ $\quad$ (при $n\to\infty$).

В случае использования состоятельных оценок оправдывается увеличение объёма выборки, так как при этом становятся маловероятными значительные ошибки при оценивании.

Оценка называется **эффективной**, если она является несмещенной с минимальной дисперсией (заметим, что такая оценка существует не всегда).

Эффективные оценки являются наиболее точными, так как имеют наименьший разброс вокруг истинного значения.

В качестве статистических оценок желательно использовать те, которые обладают всеми перечисленными свойствами. Но, к сожалению, это не всегда возможно.

2.2. Методы нахождения оценок.

Имеется два подхода к оцениванию неизвестных параметров распределений: точечный и интервальный. Точечный указывает точку, около которой находится оцениваемый параметр; при интервальном находят интервал, который с некоторой вероятностью накрывает неизвестное числовое значение параметра.

2.2.1. Точечные оценки.

Суть **метода моментов (К. Пирсона)** – определённое количество эмпирических (выборочных) моментов приравнивается к соответствующим теоретическим моментам генеральной случайной величины.

Пример 2.1. Найдём методом моментов оценку параметра λ в распределении Пуассона.

Так как математическое ожидание случайной величины, имеющей распределение Пуассона, совпадает с параметром распределения, то

$$\hat{\lambda} = \overline{X} = \frac{1}{n}\sum_{j=1}^{n} X_j .$$

Оценки метода моментов обычно состоятельны, однако по эффективности они часто не являются наилучшими. Тем не менее, они активно применяются на практике из-за своей простоты.

Метод максимального правдоподобия (Р.Фишер).

В основе метода лежит понятие **функции правдоподобия**, которая выражается через плотность вероятности или (вероятность) распределения генеральной случайной величины:

$$\mathbf{L}(\underline{\mathbf{x}},\boldsymbol{\theta}) = \prod_{i=1}^{n} \varphi(x_i, \theta),$$

где $\varphi(x_i,\theta) = \begin{cases} P_\theta(X = x_i) & \text{для дискретного распределения} \\ f_\theta(x_i) & \text{для непрерывного распределения} \end{cases}$

Согласно методу максимального правдоподобия в качестве оценки неизвестного параметра θ принимается такое значение, которое максимизирует функцию правдоподобия. Для этого решают уравнение правдоподобия:

$$\frac{\partial \ \ln L(x,\theta)}{\partial \ \theta} = 0 .$$

Пример. Найдём методом максимального правдоподобия оценку для параметра показательного распределения. Запишем функцию правдоподобия:

$$L(\underline{x},\theta)=\prod_{i=1}^{n}\frac{1}{\theta}\exp\left\{-\frac{x_i}{\theta}\right\}=\frac{1}{\theta^n}\exp\left\{-\frac{1}{\theta}\sum_{i=1}^{n}x_i\right\};$$

прологарифмируем её:

$$\ln L(\underline{x},\theta)=\ln\frac{1}{\theta^n}+\ln\ \exp\left\{-\frac{1}{\theta}\sum_{i=1}^{n}x_i\right\}=-n\cdot\ln\theta-\frac{1}{\theta}\sum_{i=1}^{n}x_i\ .$$

Далее дифференцируем:

$$\frac{\partial\ \ln L(x,\theta)}{\partial\ \theta}=-n\cdot\frac{1}{\theta}+\frac{1}{\theta^2}\sum_{i=1}^{n}x_i$$

и приравнивая полученную производную к нулю, получаем:

$$\hat{\boldsymbol{\theta}}=\frac{1}{\boldsymbol{n}}\sum_{\boldsymbol{i}=1}^{\boldsymbol{n}}\boldsymbol{x_i}\ .$$

Аналогично можно поступить и для других известных распределений. Приведём таблицы с полученными результатами.

1) Для непрерывных распределений:

Распределение	Оцениваемый параметр и его оценка	Плотность распределения
Нормальное **N(θ,σ²)**	θ- математическое ожидание; $\hat{\boldsymbol{\theta}}=\tilde{\boldsymbol{O}}$ - выборочная средняя.	$\frac{1}{\sqrt{2\pi\sigma}}\exp\{-\frac{(x-\theta)^2}{2\sigma^2}\}$
Нормальное **N(m,θ²)**	θ^2- дисперсия; $\hat{\boldsymbol{\theta}}=\frac{1}{\boldsymbol{n}}\sum_{\boldsymbol{i}=1}^{\boldsymbol{n}}(\boldsymbol{X_i}-\boldsymbol{m})^2$.	$\frac{1}{\sqrt{2\pi\theta}}\exp\{-\frac{(x-m)^2}{2\theta^2}\}$
Равномерное **R(a,b)**	a, b – параметры распределения; $\hat{\boldsymbol{a}}=\boldsymbol{x}_{(1)},\hat{\boldsymbol{b}}=\boldsymbol{x}_{(\boldsymbol{n})}$- порядковые статистики (первый и последний члены вариационного ряда).	$\begin{cases}\frac{1}{b-a}, & \text{если } x_i\in[a,b]\\ 0, & \text{если } x_i\notin[a,b]\end{cases}$

2)Для дискретных распределений:

Распределение	Оцениваемый параметр и его оценка	Вероятности
Биномиальное **Bi(k,θ)**	θ-вероятность успеха в одном испытании; $\hat{\boldsymbol{\theta}}=\tilde{\boldsymbol{O}}/\boldsymbol{k}$	$C_k^x\theta^x(1-\theta)^{k-x}$, x=0,1, 2 ,…,k
Пуассона **П(θ)**	θ- параметр; $\hat{\boldsymbol{\theta}}=\tilde{\boldsymbol{O}}$ - выборочная средняя.	$e^{-\theta}\frac{\theta^x}{x!}$, θ>0; x=0,1, 2,…
Геометрическое **G(1-θ)**	$M(X)=\frac{1}{1-\theta}$ (математическое ожидание); $\hat{\boldsymbol{I}}\ (\tilde{\boldsymbol{O}})=\tilde{\boldsymbol{O}}$ - выборочная средняя.	**(1-θ)·θˣ**, x=0,1, 2 ,3,…

Рассмотрим подробно случай, когда в нормальной модели неизвестны 2 параметра - математическое ожидание и дисперсия. В этом случае оценки максимального правдоподобия и метода моментов совпадают:

$$\hat{\boldsymbol{m}} = \frac{1}{n}\sum_{j=1}^{n} X_j \ ; \quad \hat{\boldsymbol{\sigma}}^2 = \frac{1}{n}\sum_{j=1}^{n}(X_j - \overline{X})^2 \ .$$

Данная оценка для дисперсии является смещенной:

$$M_{\theta}[\hat{\boldsymbol{\sigma}}^2] = M_{\theta}\{\frac{1}{n}\sum_{j=1}^{n}(X_j - \overline{X})^2\} = \frac{n-1}{n}\sigma^2 \ ,$$

то есть, в среднем выборочная дисперсия немного занижает истинное значение дисперсии. Обычно, в качестве оценки дисперсии используют «подправленную» выборочную дисперсию:

$$\tilde{\boldsymbol{S}}^2 = \frac{1}{n-1}\sum_{j=1}^{n}(X_j - \overline{X})^2 \ .$$

эта оценка уже является несмещенной.

Оценки максимального правдоподобия, как правило, являются состоятельными, асимптотическими несмещенными и асимптотически эффективными (то есть несмещенными и эффективными для достаточно больших $\boldsymbol{n}$).

2.2.2. Интервальные оценки.

Доверительный интервал – это статистическая оценка параметра вероятностного распределения, имеющая вид интервала, границами которого служат функции от результатов наблюдений и который с высокой вероятностью «накрывает» неизвестный параметр.

Границы интервала зависят от выборки и поэтому являются случайными величинами, в отличии от оцениваемого параметра – величины неслучайной. Поэтому в определении говорится, что доверительный интервал «накрывает» оцениваемую величину, а не содержит её.

Интервал **(α,β)**, который накрывает оцениваемый параметр с вероятностью **γ** называется **γ-доверительный интервал**:

$$\mathbf{P\{\ \theta \in (\alpha,\beta)\ \} = \gamma.}$$

При этом вероятность **γ** называют **доверительной вероятностью**.

Выбор доверительной вероятности определяется конкретными условиями, обычно используют **γ** =0,9; 0,95; 0,99.

Для получения доверительного интервала наименьшей длины при заданном объёме выборки **n** и заданной доверительной вероятности **γ**, в качестве оценки параметра следует брать эффективную или асимптотически эффективную (т.е. становящуюся эффективной при достаточно больших **n**) оценку. Для построения доверительного интервала необходимо знать распределение статистики, которая является точечной оценкой данного параметра.

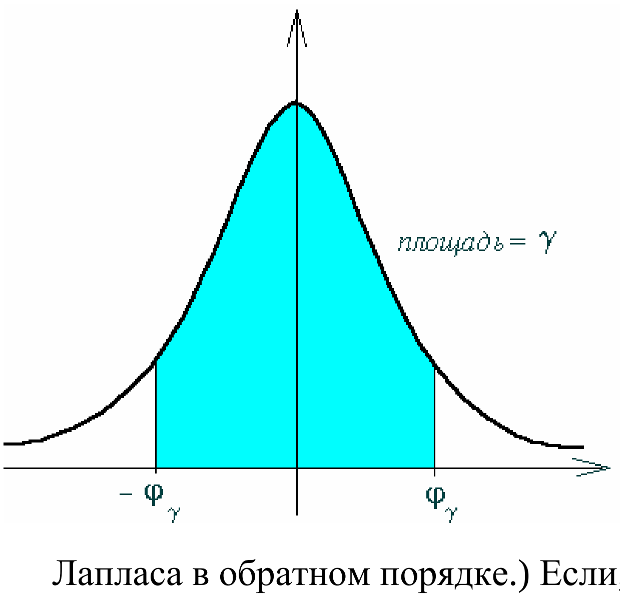
Если объём выборки **n** достаточно мал (**n**<40), то требуется знать вид распределения генеральной случайной величины **X**, если же **n**- велико (**n**>40), то можно строить асимптотические оценки, используя предельные теоремы теории вероятностей.

Интервальная оценка для математического ожидания при известной дисперсии.

Для использования этой оценки на практике требуется, чтобы распределение генеральной случайной величины было нормальным с параметрами **m** и $\boldsymbol{\sigma}^2$, либо, чтобы объём выборки был достаточно велик (**n**>40). Как известно точечная несмещённая оценка для математического ожидания – это выборочная средняя:

$$\hat{\boldsymbol{m}} = \overline{\boldsymbol{X}} = \frac{1}{n}\sum_{j=1}^{n} X_j .$$

В силу свойств нормального распределения, эта оценка, которая является случайной величиной, имеет нормальное распределение с параметрами **m** и $\dfrac{\boldsymbol{\sigma}^2}{\boldsymbol{n}}$. Тогда нормированное отклонение оценки от оцениваемой величины $\dfrac{\overline{X}-m}{\sigma/\sqrt{n}}$ имеет стандартно нормальное распределение. Зададимся доверительной вероятностью **γ** и постоим доверительный интервал, симметричный относительно точечной оценки:



$$\mathbf{P}\{-\varphi_\gamma < \frac{\overline{X}-m}{\sigma/\sqrt{n}} < \varphi_\gamma\} = \gamma$$

Используя свойства плотности распределения и определение квантиля имеем, что $\boldsymbol{\varphi_\gamma}$ - квантиль стандартно нормального распределения уровня **γ/2**. (Его находят по таблице функции Лапласа в обратном порядке.) Если, например, **γ=0,95** , то **φ_γ =1,96**.

Итак, *γ*-доверительный интервал для неизвестного математического ожидания при известной дисперсии имеет вид:

$$\left(\overline{X} - \varphi_\gamma \frac{\sigma}{\sqrt{n}}, \overline{X} + \varphi_\gamma \frac{\sigma}{\sqrt{n}}\right)$$

Пример. Фирма коммунального хозяйства желает на основе выборки оценить среднюю квартплату за квартиры определённого типа с надёжностью не менее 99% и погрешностью, меньшей 10 у.е. Предполагая, что квартплата имеет нормальное распределение с неизвестным математическим ожиданием и дисперсией 35^2 у.е., найти минимальный объём выборки.

Решение. Итак, нужно найти n:

$$P\{|\overline{X}-m|<10\}\geq 0{,}99.$$

Доверительный интервал имеет вид:

$$\left(\overline{X}-2{,}58\cdot\frac{35}{\sqrt{n}};\overline{X}+2{,}58\cdot\frac{35}{\sqrt{n}}\right)$$

В задаче требуется, чтобы половина длины интервала была10, то есть

$$2{,}58\cdot\frac{35}{\sqrt{n}}\leq 10,$$

отсюда $\boldsymbol{n>81{,}54}$ и, следовательно, $\boldsymbol{n_{min}=82.}$

Интервальная оценка для математического ожидания при неизвестной дисперсии.

Если дисперсия неизвестна, то её заменяют на «наилучшую» оценку

$$\tilde{S}^2=\frac{1}{n-1}\sum_{j=1}^{n}(X_j-\overline{X})^2.$$

В этом случае распределение статистики

$$T(X_1,X_2,...,X_n)=\frac{\overline{X}-m}{\tilde{S}/\sqrt{n-1}}$$

уже не является стандартно нормальным (так как в знаменателе корень квадратный из суммы квадратов случайных величин) – это распределение Стьюдента с **n-1** степенью свободы. Поэтому симметричный γ-доверительный интервал будет иметь вид:

$$\left(\overline{X}-t_{\gamma,n-1}\frac{\tilde{S}}{\sqrt{n-1}},\overline{X}+t_{\gamma,n-1}\frac{\tilde{S}}{\sqrt{n}-1}\right),$$

где $t_{\gamma,n-1}$ - квантиль уровня **γ/2** распределения Стьюдента с **n-1** степенью свободы.

Если **n**>30(40), то распределение Стьюдента близко к нормальному, и можно пользоваться таблицами нормального распределения.

Пример. Для отрасли, включающей 1200 фирм, составлена случайная выборка из 19 фирм. По выборке оказалось, что в среднем в фирме работает 77 человек, при среднеквадратичном отклонении 25 человек. Пользуясь 95% доверительным интервалом, оценить среднее число работающих в фирме и общее число работающих в отрасли. Считать, что число работающих имеет нормальное распределение.

Решение. По условию задачи **n**=19, по таблице находим $\mathbf{t_{0,95;\,18}=2{,}10.}$

Тогда доверительный интервал для среднего числа работающих в одной фирме

$$\left(77-2{,}1\cdot\frac{25}{\sqrt{18}};77+2{,}1\cdot\frac{25}{\sqrt{18}}\right)=(65;89),$$

а доверительный интервал для числа работающих во всей отрасли:

$$(1200 \cdot 65; 1200 \cdot 89) = (78000; 106800) .$$

<u>Интервальная оценка для дисперсии (среднеквадратичного отклонения) при известном математическом ожидании .</u>

В этом случае эффективной оценкой дисперсии является

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{j=1}^{n}(X_j - m)^2 .$$

Статистика

$$\chi^2(n) = \frac{n \cdot \hat{\sigma}^2}{\sigma^2}$$

имеет χ^2 распределение с n степенями свободы. Отсюда следует, что γ-доверительный интервал (он уже не будет симметричным, так как плотность распределения χ^2 не является симметричной):

$$\left(\sqrt{\frac{n}{g_{\gamma,n}}} \cdot \hat{\sigma}; \sqrt{\frac{n}{h_{\gamma,n}}} \cdot \hat{\sigma}\right),$$

где $g_{\gamma,n}$ - квантиль уровня $\frac{1-\gamma}{2}$ распределения χ^2 с n степенями свободы;

$h_{\gamma,n}$ - квантиль уровня $\frac{1+\gamma}{2}$ распределения χ^2 с n степенями свободы.

Если математическое ожидание – неизвестно, то количество степеней свободы уменьшается на 1, и доверительный интервал имеет вид:

$$\left(\sqrt{\frac{n-1}{g_{\gamma,n-1}}} \cdot \tilde{S}; \sqrt{\frac{n-1}{h_{\gamma,n-1}}} \cdot \tilde{S}\right),$$

где $\tilde{S}^2 = \frac{1}{n-1}\sum_{j=1}^{n}(X_j - \overline{X})^2$ - «подправленная» выборочная дисперсия.

Пример. В условиях предыдущей задачи, оценить среднеквадратическое отклонение работающих в фирме по всей отрасли, взяв доверительную вероятность, равную 0,9.

Решение. По условию задачи **n**=19, $\tilde{S}$ =25, γ=0.9.

Далее, $\frac{1-\gamma}{2}$ = 0,05; $\frac{1+\gamma}{2}$ = 0,95.

По таблице находим $g_{0,9;\,18} = 28,9$; $h_{0,9;\,18} = 9,39$.

Таким образом, доверительный интервал для среднеквадратического отклонения числа работающих в одной фирме в среднем по отрасли:

$$\left(\sqrt{\frac{18}{28,9}} \cdot 25; \sqrt{\frac{18}{9,39}} \cdot 25\right) = (19,7; 34,6).$$