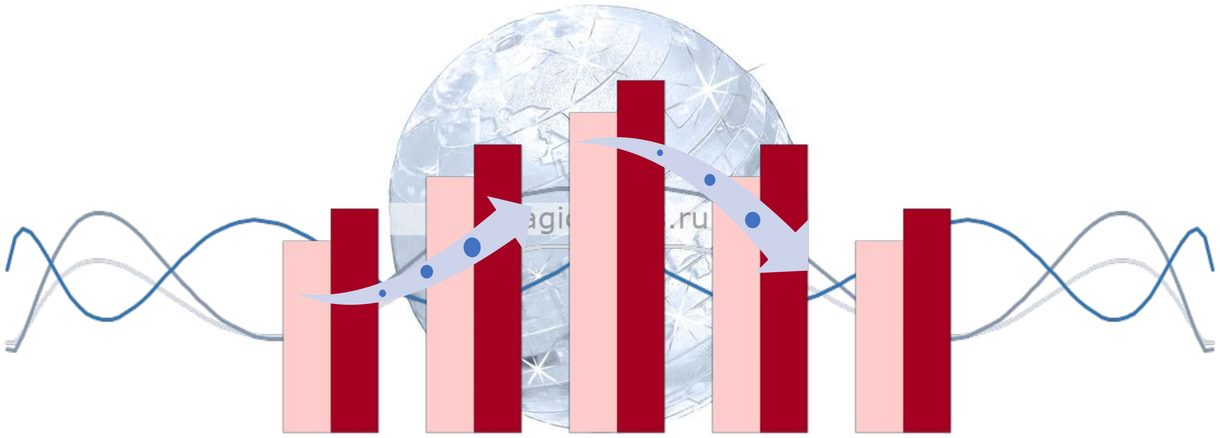


Математическая статистика

Лекция 1

Основные понятия математической статистики. Статистические оценки параметров распределения

Статистика как наука и отрасль практической деятельности

В настоящее время важную роль в механизме управления экономикой выполняет статистика. Она осуществляет сбор, научную обработку, обобщение и анализ информации, характеризующей социально-экономическое развитие страны.

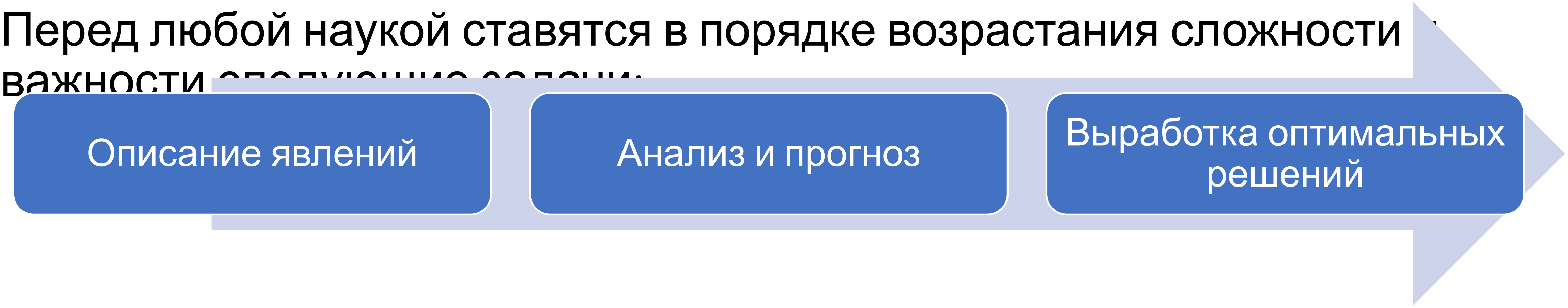
Термин «статистика» происходит от латинских «Status», что означает «определенное состояние явления, положение вещей», и «Stato» – «государство». Он был введен в научный оборот в 1749 году немецким ученым Готфридом Ахенвалем, опубликовавшем книгу под названием «Статистика», в которой приводилось описание политического устройства государств Европы.

Предмет и задачи математической статистики

Математическая статистика – это наука, изучающая случайные явления посредством обработки и анализа результатов наблюдений и измерений.

Любой результат можно представить как совокупность значений, принятых в результате *n* опытов над какой-то с.в. или системой случайных величин.

Перед любой наукой ставятся в порядке возрастания сложности важности следующие задачи:



Задачи математической статистики

- Указать способы получения, группировки и обработки статистических данных, собранных в результате наблюдений, специально поставленных опытов или произведенных измерений;
- Разработать методы анализа статистических сведений в зависимости от целей исследования.

- Оценка неизвестной вероятности события;
- Оценка параметров распределения с.в.
- Проверка гипотез о параметрах распределения или о виде неизвестного распределения

Источники информации

- Внутренние источники: финансовая и статистическая отчетность предприятия;
- Внешние источники: налоговая, банковская, таможенная статистика, платежный баланс и др.

Статистические службы международных организаций

- Статистические службы организаций системы ООН;
- Институт статистики ЮНЕСКО;
- Статистический директорат Организации экономического сотрудничества и развития;
- Статистическое бюро Европейских Сообществ и др.

Основные понятия математической статистики

- ***Статистическое наблюдение*** представляет собой планомерный, научно организованный и, как правило, систематический сбор данных о явлениях и процессах общественной жизни путем регистрации заранее намеченных существенных признаков с целью получения в дальнейшем обобщающих характеристик этих явлений и процессов.
- ***Статистическая совокупность*** – это множество единиц явления, объединенных в соответствии с задачей исследования единой качественной основой (однородностью), но отличающиеся друг от друга признаками.

Основные понятия математической статистики

Случайная величина X – генеральная совокупность X

Совокупность случайно отобранных объектов из генеральной совокупности называют ***выборочной совокупностью*** или ***выборкой.***

Количество элементов совокупности - ***объем совокупности***.

Способы отбора единиц совокупности:

- ✔ ***повторный*** – объект перед отбором следующего возвращается в генеральную совокупность;
- ✔ ***бесповторный*** – объект перед отбором следующего не возвращается в генеральную совокупность.

выборка должна быть репрезентативной – все объекты генеральной совокупности имеют одинаковую вероятность оказаться в выборке

Статистический ряд

В результате обработки и систематизации первичных данных статистического наблюдения получают группировки, называемые рядами распределения.

Статистические ряды распределения представляют собой упорядоченное расположение единиц изучаемой совокупности на группы по группировочному признаку.

Различают ***атрибутивные*** и ***вариационные*** ряды распределения.

- ***Атрибутивный*** – это ряд распределения, построенный по качественным признакам.
- По количественному признаку строится ***вариационный ряд распределения.***

В зависимости от характера вариации признака различают ***дискретные*** и ***интервальные*** (непрерывные) вариационные ряды распределения.

Статистический ряд

Пример дискретного ряда

№ п/п	Основные фонды, млн руб.
1	164
2	147
3	171
4	267
5	211
6	123
7	238
8	109
9	176
10	255

Пример интервального ряда

Основные фонды, млн. руб.	Число предприятий, n
До 150	3
150-200	6
200-250	7
250-300	3
300 и выше	1

n – частота выборки X

$w = \frac{n_j}{n}$ - относительная частота

Пример вариационного ряда распределения

$x_{max}=9$
$x_{min}=0$
Размах = $x_{max} - x_{min} + 1$ = 9-0+1 = 10
Размах небольшой, значит можно составить вариационный ряд по значениям.

x_i	n_i	w_i	Интервал	F_i^*
			$x \le 0$	0,0000
0	1	0,0116	$0 < x \le 1$	0,0116
1	11	0,1279	$1 < x \le 2$	0,1395
2	13	0,1512	$2 < x \le 3$	0,2907
3	14	0,1628	$3 < x \le 4$	0,4535
4	19	0,2209	$4 < x \le 5$	0,6744
5	13	0,1512	$5 < x \le 6$	0,8256
6	5	0,0581	$6 < x \le 7$	0,8837
7	6	0,0698	$7 < x \le 8$	0,9535
8	3	0,0349	$8 < x \le 9$	0,9884
9	1	0,0116	$x > 9$	1,0000
Σ	86	1,0000	–	–

n_i – частоты
w_i – относительные частоты
F_i^* - накопленные относительные частоты

Сводка и группировка данных статистического наблюдения

- ✔ Статистическая сводка – это приведение собранной информации к виду, удобному для проведения анализа.
- ✔ Группировка – это процесс образования однородных групп на основе расчленения статистической совокупности на части или объединения изучаемых единиц в частные совокупности по существенным признакам.

Оптимальное число групп можно определить по формуле Стерджесса: ***n = 1 + 3,322 x lg N***,

- где *n* - число групп
- *N* - число единиц совокупности.

Величина равного интервала определяется по следующей формуле:

$$i = \frac{x_{\max} - x_{\min}}{n}$$

Описательные статистики

Меры центральной тенденции: средние величины

Выделяют три основных класса средних:
- ✔ Средние степенные;
- ✔ Средние структурные;
- ✔ Средние хронологические.

Средние степенные

1. Средняя арифметическая.

Простая	*Взвешенная (по вариационному ряду)*

В случае, если исходные данные представлены в виде интервального ряда распределения, то в качестве вариантов усредняемого признака (x_i) принимают середины интервалов, вычисляемые по каждой группе.

Описательные статистики

Меры центральной тенденции: средние величины

Средние степенные

2. Средняя геометрическая.

Простая	*Взвешенная (по вариационному ряду)*

Средняя геометрическая обычно применяется в тех случаях, когда варианты ряда представлены относительными показателями динамики. Эта средняя выражает, как правило, средний темп относительного роста или спада.

Описательные статистики

Меры центральной тенденции: средние величи

Структурные средние

1. Мода - величина признака (варианта), наиболее часто повторяющаяся в изучаемой совокупности. Мода отражает типичный, наиболее распространенный вариант значения признака.

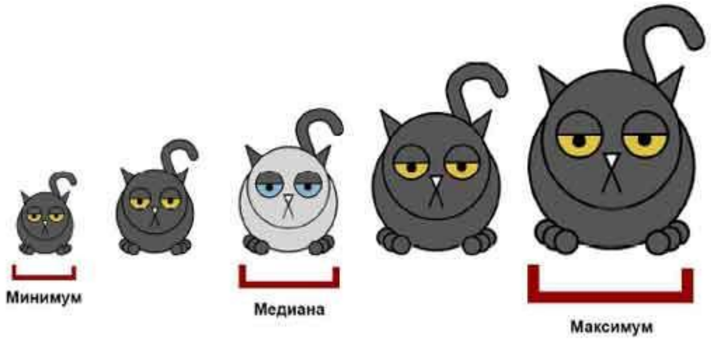
В **дискретном ряду** распределения мода – это варианта, которой соответствует наибольшая частота.
В **интервальном ряду** распределения сначала определяют модальный интервал (т.е. интервал, содержащий моду), которому соответствует наибольшая частота. Конкретное значение моды определяется формулой:

$$Mo = x_{Mo} + i \cdot \frac{n_{Mo} - n_{Mo-1}}{(n_{Mo} - n_{Mo-1}) + (n_{Mo} - n_{Mo+1})}$$

x_{Mo} – нижняя значение модального интервала;
i – величина модального интервала;
n_{Mo} – частота модального интервала;
n_{Mo-1} – частота интервала, предшествующего модальному;
n_{Mo+1} – частота интервала, следующего за модальным.

Описательные статисти

Меры центральной тенденции: средние величин



Структурные средние

2. Медиана - это варианта, находящаяся в середине ранжированного ряда (варианта, делящая ранжированный ряд пополам).

В **дискретном ряду** медианой является варианта, которой соответствует член кумулятивного ряда, впервые превысившая половину общей суммы частот.

В **интервальном ряду** распределения сначала необходимо определить медианный интервал (т.е. интервал, содержащий медиану). Медианным интервалом является тот, которому соответствует член кумулятивного ряда, впервые превысившая половину общей суммы частот. Затем работает формула:

$$Me = x_{Me} + i \cdot \frac{\frac{n}{2} - S_{Me-1}}{n_{Me}}$$

где x_{Me} – нижняя граница медианного интервала;
i – величина медианного интервала;
S_{Me-1} – член кумулятивного ряда, предшествующий медианному интервалу;
n_{Me} – частота медианного интервала.

Описательные статистики

Показатели вариации, характеристики диапазона и формы распределения статистических данных

1. Дисперсия

Простая	*Взвешенная (по вариационному ряду)*

Можно отметить следующий недостаток этого показателя вариации – если варианты x_i имеют некоторую размерность (метр, рубль, килограмм и т.д.), то дисперсия имеет размерность в квадрате, что затрудняет ее интерпретацию (например, если средняя зарплата составляет 18 тысяч рублей, то соответствующая дисперсия может составить 500 тысяч рублей в квадрате, что лишено экономическог

2. Среднее квадратическое отклонение

$$\sigma = \sqrt{D}$$

Описательные статистики

Показатели вариации, характеристики диапазона и формы распределения статистических данных

Относительные показатели вариации

Расчет относительных показателей вариации осуществляют как отношение абсолютного показателя вариации к средней арифметической. Как правило, они рассчитываются в процентах.

Относительный размах (коэффициент осцилляции):

$$v_R = \frac{R}{\bar{x}} \cdot 100\%$$

Коэффициент вариации:

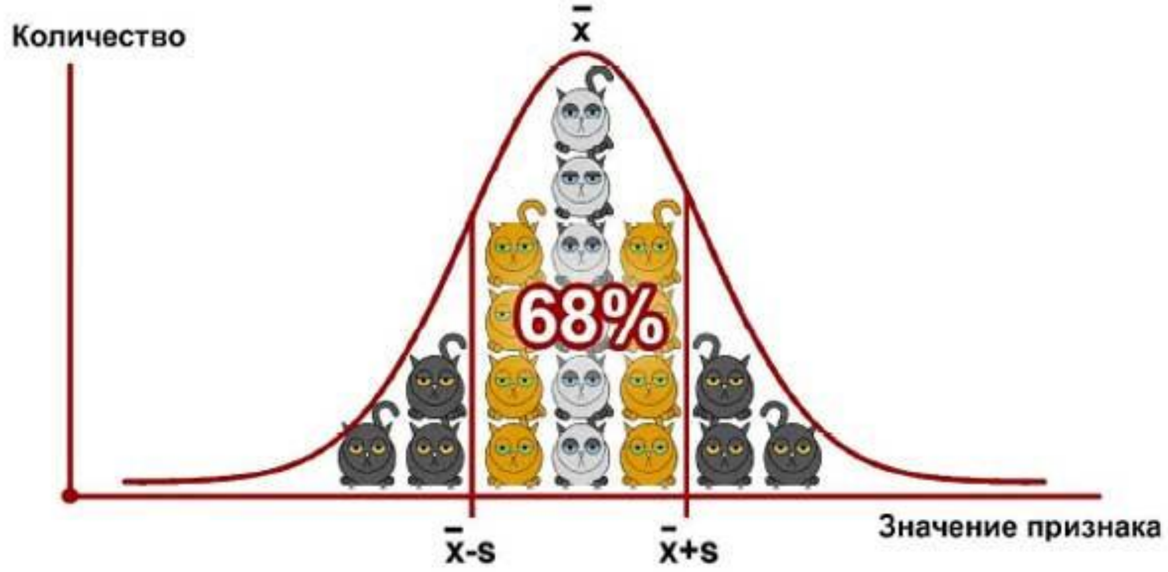
$$v_\sigma = \frac{\sigma}{\bar{x}} \cdot 100\%$$

Коэффициент вариации – это наиболее распространенный относительный показатель вариации. Считается, что если v>30%, то это говорит о большой вариации признака в изучаемой совокупности.

Характеристики и формы распределения

В статистике широко используются различные виды теоретических распределений:

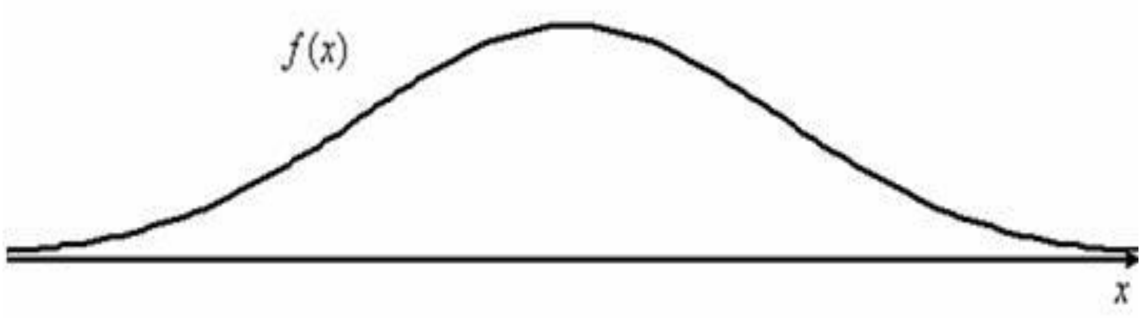
- ✔ распределение Стьюдента,
- ✔ Пуассона,
- ✔ нормальное распределени
- ✔ *хи-квадрат* распределение
- ✔ распределение Фишера,
- ✔ биномиальное (распределение Бернулли),
- ✔ равномерное распределение.



Нормальный закон распределения

Первым фундаментальным по значимости является нормальный закон распределения, часто называют – закон Гаусса (ЗНР). Подчиненность закону нормального распределения тем точнее, чем больше факторов действует вместе.

Нормальное распределение полностью определяется двумя входными параметрами: средней арифметической и среднеквадратическим отклонением (σ)



Вид функции плотности нормального распределения вероятностей

Например, рост, вес людей (и не только), их физическая сила, умственные способности и т.д. Существует «основная масса» (по тому или иному признаку) и существуют отклонения в обе стороны.

Точечные оценки параметров распределения

это вероятностные погрешности измерения, выраженные одним числом. Любая точечная оценка, вычисленная на основании опытных данных, является случайной величиной.

Качество оценки характеризуется следующими свойствами:

- ✔ ***Состоятельность:*** с ростом объема выборки оценка сходится по вероятности к параметру
- ✔ ***Несмещенность:*** математическое ожидание совпадает с истинным значение оцениваемого параметра
- ✔ ***Эффективность:*** дисперсия оценки меньше дисперсии любой другой оценки данного параметра

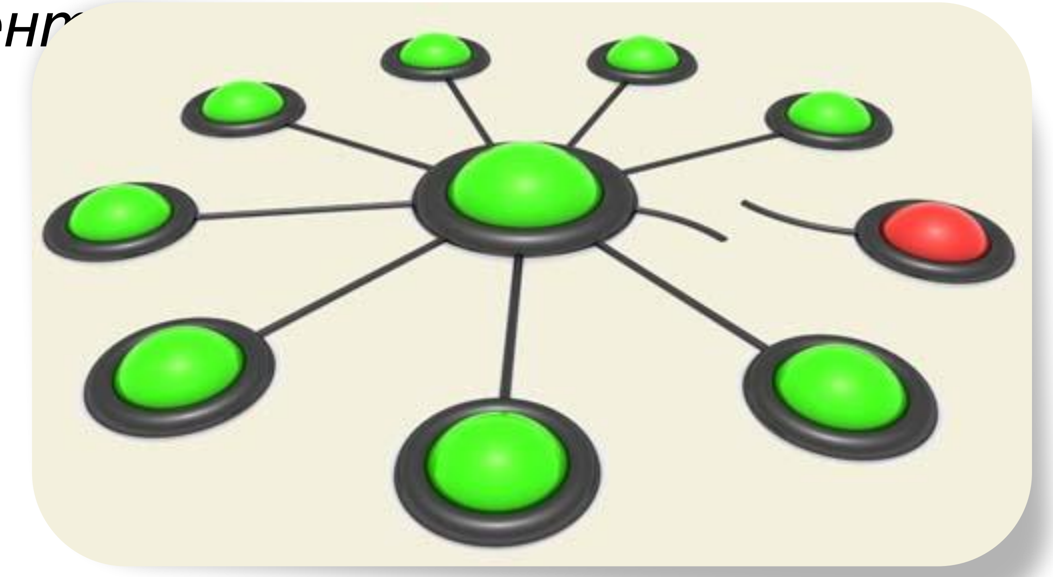
Среди всех нормально распределенных оценок наилучшей будет несмещенная эффективная оценка.

Точечные оценки параметров распределения

Теоретическим обоснованием возможности экспериментального определения вероятностных характеристик является закон больших чисел

ЗБЧ*: среднее значение конечной выборки из фиксированного распределения близко к математическому ожиданию этого распределения. Закон больших чисел важен, поскольку он гарантирует устойчивость для средних значений некоторых случайных событий при достаточно длинной серии эксперимент*

Смысл ЗБЧ: *совместное действие большого числа случайных факторов приводит к результату, почти не зависящему от случая*



Точечная оценка для математического ожидания

Математическое ожидание — среднее значение случайной величины при стремлении количества выборок или количества её измерений (иногда говорят — количества испытаний) к бесконечности.

Математическое ожидание случайной величины ***x*** *обозначается* ***M(x)***

Математическое ожидание – это сумма произведений всех возможных значений случайной величины на вероятности этих значений.

Математическое ожидание – это в теории азартных игр сумма выигрыша, которую может заработать или проиграть игрок, в среднем, по каждой ставке. На языке азартных игроков это иногда называется «преимуществом игрока» (если оно положительно для игрока) или «преимуществом казино» (если оно отрицательно для игрока).

Точечная оценка для математического ожидания

Пусть генеральная совокупность имеет MX=m и DX=D

Найдем оценку для MX=m: $\widetilde{m} = \bar{x} = \frac{1}{n}\sum_{i=1}^{m} n_i x_i$

- *Состоятельность*: при увеличении n среднее арифметическое наблюдаемых значений сходится по вероятности к математическому ожиданию с.в. (первая теорема Чебышева)

- *Несмещенность*: среднее арифметическое результатов наблюдений является несмещенной оценкой математического ожидания с.в., а следовательно, ее истинное значение совпадаете математическим ожиданием с.в.

$$M(\widetilde{m}) = M(\bar{x}) = M\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = \frac{1}{n}\sum_{i=1}^{n} M(x_i) = \frac{1}{n}\sum_{i=1}^{n} m = \frac{1}{n} mn = m$$

- *Эффективность*: так как среднее арифметическое результатов измерений получено в результате сложения случайных величин, то оно также является случайной величиной с дисперсией DX.

$$D(\widetilde{m}) = D(\bar{x}) = D\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} D(x_i) = \frac{1}{n^2}\sum_{i=1}^{n} D = \frac{1}{n^2} Dn = \frac{D}{n}$$

Т.е. точность результата измерения можно повысить при увеличении числа измерений.

Точечная оценка для дисперсии

Найдем оценку для DX=D

В качестве точечной оценки дисперсии выбирают среднее значение квадрата отклонения случайной величины от среднего значения: $D=\frac{1}{n}\sum_{i=1}^{m} n_i(x_i-\bar{x})^2$; $\mathrm{D}=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2$;

$$\mathrm{D}=\frac{1}{n}\sum_{i=1}^{n} x_i^2-(\bar{x})^2$$

- *Состоятельность*: эта оценка является состоятельной $\frac{1}{n}\sum_{i=1}^{n} x_i^2 \sim M(X^2)$, $\bar{x}\sim m$, $D\sim M(X^2)-m^2=D$, но смешенной, так как ее математическое ожидание равно $MD=\frac{n-1}{n^2}\sum_{i=1}^{n} D=\frac{n-1}{n}D$
- *Несмещенность*: среднее арифметическое имеет дисперсию, в n раз меньшую, чем дисперсия случайной погрешности. В связи с этим в качестве точечной оценки дисперсии среднего арифметического принимается выражение: $\widetilde{D}=\frac{n}{n-1}D=S^2=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$
- *Эффективность*: при n>20, оценка $\widetilde{D}$, независимо от закона распределения с.в. X, распределена приблизительно нормально.

Если с.в. X распределена нормально, то $D(\widetilde{D})=\frac{2}{n-1}D^2$

Если с.в. X распределена равномерно на интервале (a,b), то $D(\widetilde{D})=\frac{0{,}8n+1{,}2}{n(n-1)}D^2$

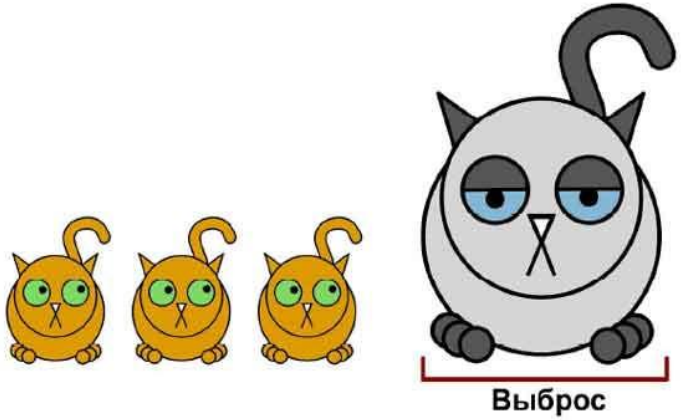
Точечная оценка для среднего квадратического отклонения

- *Выборочное среднее квадратическое отклонение* (смещенная оценка)

$$\sigma = \sqrt{D}$$

- Исправленное среднее квадратическое отклонение (несмещенная оценка)

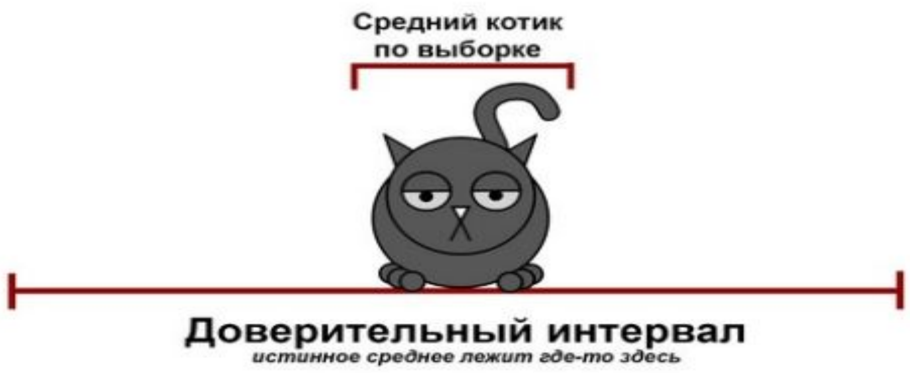
$$\tilde{\sigma} = S = \sqrt{\tilde{D}} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$



Доверительный интервал для математического ожидания

Пусть случайная величина X генеральной совокупности распределена нормально, учитывая, что дисперсия и среднее квадратическое отклонение этого распределения известны. Требуется оценить неизвестное математическое ожидание по выборочной средней. В данном случае задача сводится к нахождению доверительного интервала для математического ожидания с надежностью $\alpha = 1 - \gamma$. Если задаться значением доверительной вероятности (надежности), то можно найти вероятность попадания в интервал для неизвестного математического ожидания, используя формулу:

$$m \in \left(\tilde{m} - \frac{t \cdot \sigma}{\sqrt{n}}; \tilde{m} + \frac{t \cdot \sigma}{\sqrt{n}}\right)$$



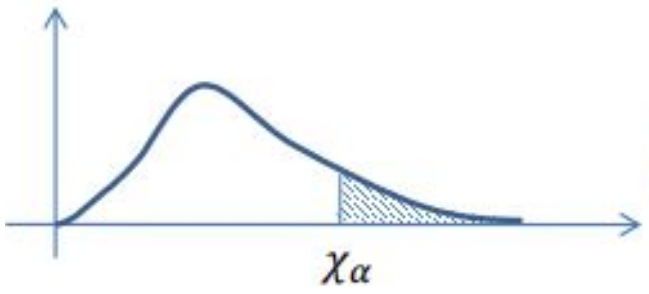
t находим по таблице Стьюдента при заданном уровне надежности $\boldsymbol{\alpha = 1 - \gamma}$, где $\boldsymbol{\gamma}$ – доверительная вероятность

Доверительный интервал для дисперсии и СКО

$$X \in \aleph(m, \sigma)$$

Рассмотрим с.в. $\chi^2 = \frac{(n-1)\cdot S^2}{\sigma^2}$ имеет распределение Пирсона с числом степеней свободы *k=n-2*

По таблицам χ^2 распределения можно найти x_α, удовлетворяющее условию: $P\left(\chi^2 > \chi_\alpha\right) = \alpha$



По таблицам χ^2 распределения можно найти такие два числа u_1 и u_2, которые удовлетворяют условию $P(u_1 \le \chi^2 \le u_2) = \gamma$ Пар u_1 и u_2, удовлетворяющих данному условию, существует бесконечное множество. Чтобы выбрать одну такую пару, введем дополнительное условие: $P(\chi^2 < u_1) = P(\chi^2 > u_2) = \frac{1-\gamma}{2}$

Таким образом: $P(\chi^2 > u_2) = \frac{1-\gamma}{2} \Rightarrow u_2$

$$P(\chi^2 > u_1) = \frac{1+\gamma}{2} \Rightarrow u_1$$

$$\frac{(n-1)\cdot s^2}{u_2} \le D \le \frac{(n-1)\cdot s^2}{u_1}$$

Доверительный интервал для среднего квадратического отклонения

$$\frac{\sqrt{n-1}\cdot s}{\sqrt{u_2}} \le \sigma \le \frac{\sqrt{(n-1)}\cdot s}{\sqrt{u_1}}$$