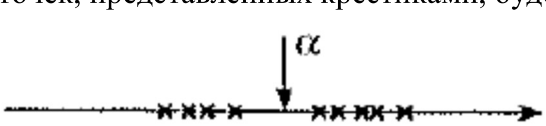
5. Оценивание
5.1. Несмещенная оценка
Одна из главных задач статистики состоит в том, чтобы иметь возможность оценить значение параметров генеральной совокупности, не проводя перепись генеральной совокупности.
Рассмотрим пример администратора больницы, который обеспокоен тем, что в больнице иногда не хватает коек для обслуживания всех своих пациентов. Она хочет оценитьμ - среднее время, в течение которого пациент остается на больничной койке в течение определенной недели. Она берет случайную выборку пациентов, которые были приняты в течение этой недели, и записывает продолжительность времени в днях, в течение которого каждый пациент выборки остается в больнице. Теперь она должна подумать, как использовать собранные ею данные для оценки μ. Например, предположим, что она собрала данные о времени пребывания в больнице 10-ти пациентов. В результаты может получиться что-то вроде:
7, 5, 5, 9, 3, 11, 6, 4, 2, 20
Какой расчет она должна провести по этим значениям, чтобы оценить μ ? Вы можете подумать, что ответ очевиден. Имеет смысл оценить μ , взяв среднее значение выборки. Таким образом, в этом случае оценка μ будет равна 72/10 = 7,2. Но является ли это единственной разумной оценкой, которую можно сделать на основе этих данных? Что было бы неправильно, например, если бы мы взяли первый раз, 7 дней, в качестве оценки μ ? Возможно, что реальное ее значение μ ближе к 7, чем к 7,2. Понятно, что вам нужна стратегия, чтобы решить, какой метод лучше.
Чтобы принять решение, вам нужно тщательно подумать о том, что пытается сделать администратор. Представьте себе, что она еще не построила выборку из 10-ти пациентов. Таким образом, в настоящее время продолжительность времен пациентов в больнице являются случайными величинами, потому что неизвестно, каких пациентов она выберет. Предположим, назовем эти случайные величины как X_1, X_2, …, X_{10}. Были предложены два возможных метода оценки μ .
Метод 1. Используйте арифметическое среднее этих 10-ти выборочных значений. Назовем это значение M_1. Тогда $M_1 = \frac{1}{10}(X_1 + X_2 + \cdots + X_{10})$. Напомним, что это также можно назвать $\bar{X}$.
Метод 2. Используйте первое выбранное значение. Назовем это значение M_2. Итак, $M_2 = X_1$.
Полученные таким образом значения сами по себе являются случайными величинами и обычно называются **оценочными функциями**. Значение, которое принимает оценочная функция для конкретной выборки, называется **оценкой**.
Должно быть ясно, что $X_2, X_3, ..., X_{10}$ имеют такое же распределение, как и X_1, поэтому математическое ожидание каждой случайной величины равна μ. Ранее мы видели, что если $X_1, X_2, ..., X_{10}$-это набор случайных величин с одинаковым математическим ожиданием равным μ , то $M(\bar{X}) = \mu$.
Следовательно, для обоих предложенных методов $M(M_1) = M(M_2) = \mu$
Что именно говорит вам это утверждение? В нем говорится, что если вы возьмете все возможные выборки размером 10 и вычислите значение M_1 для каждой выборки, то математическое ожидание этих значений будет равно μ, значению, которое пытается оценить администратор. То же самое относится и к M_2. Это означает, что обе оценочные функции **несмещенные** оценки.

Для неизвестного параметра α-генеральной совокупности, оценочная функция M является <u>несмещенной</u> оценкой, если $M(M) = \alpha$, где математическое ожидание берется по всем возможным выборкам заданного размера.
Оценочную функцию некоторого параметра генеральной совокупности называют <u>несмещенной</u>, если ее математическое ожидание равно этому параметру. В противном случае ее называют <u>смещенной</u>.

Следующий рисунок представляет способ понимания идеи несмещенной оценки. Прямая представляет собой ось действительных чисел, а крестики представляют значения, которые дает оценочная функция для каждой выборки генеральной совокупности. Если оценка несмещенная, то средним значением всех точек, представленных крестиками, будет параметр генеральной совокупности α.



Представление несмещенной оценки

Очевидно, что вы хотите, чтобы оценочная функция была несмещенной, потому что в этом случае интересующий параметр, α, в действительности является «фокусом» всех возможных значений, которые может принимать оценочная функция. Конечно, здесь это имеет только теоретическое значение, потому что администратор имеет только одну выборку, и, поскольку обе оценки несмещенные, ей нужны другие способы решить, какая оценка лучше.

Предположим, что дисперсия случайной величины X_1 равна σ^2. Тогда дисперсии случайных величин $X_2, X_3, \ldots, X_{10}$ также равны σ^2.

Если количество пациентов в стационаре в течение недели является большим числом, то можно предположить, что распределения случайных величин $X_1, X_2, \ldots, X_{10}$ являются независимыми. При этих условиях: $Д(\bar{X}) = \frac{1}{10}\sigma^2$.

Это означает, что оценки, найденные из $M_1 = \bar{X}$, в среднем будут ближе к μ, чем оценки, найденные из $M_2 = X_1$. На рисунке показано сравнение двух оценок.

Более близкие оценки находятся по более крупным выборкам

Вы можете видеть, что кресты сгруппированы более близко вокруг μ в первом случае. Это подтверждает то, что вы, возможно, догадались с самого начала: лучше использовать оценочную функцию $M_1 = \bar{X}$, потому что он использует всю информацию в выборке, тогда как использование только $M_2=X_1$ отбрасывает 90% информации, полученной в результате выборки.

В целом можно показать, что $\bar{X}$всегда является объективной оценкой μ. Другими словами, разумный метод оценки математического ожидания генеральной совокупности состоит в том, чтобы взять случайную выборку и использовать ее среднее значение в качестве оценки.

Так верно ли это для всех параметров генеральной совокупности? Чтобы оценить медиану генеральной совокупности, возьмем ли мы медиану случайной выборки, или чтобы оценить моду генеральной совокупности, возьмем ли мы моду случайной выборки? Предыдущий пример, казалось бы, предполагает, что это разумная процедура, но рассмотрим следующий пример.

Пример 5.1.1.

Сумка для хранения дисков содержит очень большое количество маленьких дисков, на 90% из которых написано число 0, а на остальных 10% - число 1. Люк не знает, какие числа записаны на дисках. Он берет из сумки наугад два диска. Люк хочет оценить наибольшее число, записанное на любом диске в сумке. Он решает, что его метод оценки максимума будет состоять в том, чтобы взять большее из чисел на двух дисках в его выборке. Если на обоих дисках записано одинаковое число, то он принимает это значение за свою оценку максимума. Решите, дает ли предложенный метод несмещенную оценку максимума генеральной совокупности.

Решение:

Пусть максимальное значение в выборке будет M.

Метод Люка - взять максимум выборки, M, в качестве оценки максимума генеральной совокупности. В таблице приведены все возможные выборки размера два и значение M для каждой выборки вместе с вероятностью выбора этой выборки.

Поскольку говорят, что сумка содержит большое количество дисков, разумно предположить, что шансы выбора 0 или 1 остаются постоянными для каждого выбора.

Значения выборки	m, Значения M	Вероятность
0, 0	0	$0.9 \cdot 0.9 = 0.81$
0, 1	1	$0.9 \cdot 0.1 = 0.09$
1, 0	1	$0.1 \cdot 0.9 = 0.09$
1, 1	1	$0.1 \cdot 0.1 = 0.01$

Возможные значения выборки для выборки размера два.

Закон распределения случайной величины M:

Значение, m	0	1
P(M=m)	0.81	0.19

Следовательно, $M(M) = 0 \cdot 0.81 + 1 \cdot 0.19 = 0.19$
Но максимум генеральной совокупности равен 1, и поэтому максимум выборки, который является оценкой Люка, недооценивает максимум генеральной совокупности в среднем. Следовательно, максимум выборки по методу Люка является смещенной оценкой максимума генеральной совокупности.
Теперь вы видели, что не всегда целесообразно использовать выборочную версию статистики в качестве оценки той же версии статистики генеральной совокупности. Это особенно важное предупреждение. Речь идет о оценке дисперсии генеральной совокупности. Рассмотрим следующий пример.
Пример 5.1.2.
Сумка для хранения дисков содержит большое количество маленьких дисков, 40% из которых пронумерованы цифрой 1, 40% пронумерованы цифрой 2 и 20% пронумерованы цифрой 3. Случайная выборка размера три взята из этой сумки. Рассматривая все возможные значения для выборок размера три, определите, является ли дисперсия выборки несмещенной оценкой дисперсии генеральной совокупности.
Решение:
Математическое ожидание генеральной совокупности:
$\mu = 1 \cdot 0.4 + 2 \cdot 0.4 + 3 \cdot 0.2 = 1.8$
Дисперсия генеральной совокупности:
$\sigma^2 = 1^2 \cdot 0.4 + 2^2 \cdot 0.4 + 3^2 \cdot 0.2 - 1.8^2 = 0.56$
Рассмотрим возможные значения выборки размера три. Одно из таких значений выборки составляет 1, 1, 2.
Дисперсия этих значений: $\frac{1}{3}(1^2 + 1^2 + 2^2) - \left(\frac{1+1+2}{3}\right)^2 = \frac{2}{9}$
Шанс выбора конкретно этой выборки: $0.4 \cdot 0.4 \cdot 0.4 = 0.064$
Вы можете повторить этот расчет для всех 27 возможных значений выборки. Нижеследующая таблица показывает все возможные значения выборки. Она также показывает выборочную дисперсию, V и вероятность выбора этой конкретной выборки.

Значение выборки	Дисперсия этой выборки	Вероятность выбора этой выборки	Значение выборки	Дисперсия этой выборки	Вероятность выбора этой выборки
1, 1, 1	0	$0.4^3 = 0.064$	3, 3, 1	$\frac{8}{9}$	$0.4 \cdot 0.2^2 = 0.016$
2, 2, 2	0	$0.4^3 = 0.064$	2, 2, 3	$\frac{2}{9}$	$0.4^2 \cdot 0.2 = 0.032$
3, 3, 3	0	$0.2^3 = 0.064$	2, 3, 2	$\frac{2}{9}$	$0.4^2 \cdot 0.2 = 0.032$
1, 1, 2	$\frac{2}{9}$	$0.4^3 = 0.064$	3, 2, 2	$\frac{2}{9}$	$0.4^2 \cdot 0.2 = 0.032$
2, 1, 1	$\frac{2}{9}$	$0.4^3 = 0.064$	2, 3, 3	$\frac{2}{9}$	$0.4 \cdot 0.2^2 = 0.016$
1, 2, 1	$\frac{2}{9}$	$0.4^3 = 0.064$	3, 2, 3	$\frac{2}{9}$	$0.4 \cdot 0.2^2 = 0.016$
1, 2, 2	$\frac{2}{9}$	$0.4^3 = 0.064$	3, 3, 2	$\frac{2}{9}$	$0.4 \cdot 0.2^2 = 0.016$
2, 1, 2	$\frac{2}{9}$	$0.4^3 = 0.064$	1, 2, 3	$\frac{6}{9}$	$0.4^2 \cdot 0.2 = 0.032$
2, 2, 1	$\frac{2}{9}$	$0.4^3 = 0.064$	1, 3, 2	$\frac{6}{9}$	$0.4^2 \cdot 0.2 = 0.032$
1, 1, 3	$\frac{8}{9}$	$0.4^2 \cdot 0.2 = 0.032$	2, 1, 3	$\frac{6}{9}$	$0.4^2 \cdot 0.2 = 0.032$
1, 3, 1	$\frac{8}{9}$	$0.4^2 \cdot 0.2 = 0.032$	2, 3, 1	$\frac{6}{9}$	$0.4^2 \cdot 0.2 = 0.032$
3, 1, 1	$\frac{8}{9}$	$0.4^2 \cdot 0.2 = 0.032$	3, 1, 2	$\frac{6}{9}$	$0.4^2 \cdot 0.2 = 0.032$

1, 3, 3	$\frac{8}{9}$	$0.4 \cdot 0.2^2 = 0.016$	3, 2, 1	$\frac{6}{9}$	$0.4^2 \cdot 0.2 = 0.032$
3, 1, 3	$\frac{8}{9}$	$0.4 \cdot 0.2^2 = 0.016$			
Возможные значения выборки и дисперсия выборки для выборки размера три					

Из этой таблицы вы можете построить закон распределения вероятностей случайной величины V:

Значение, *v*	0	$\frac{2}{9}$	$\frac{6}{9}$	$\frac{8}{9}$
P(V=v)	0.136	0.528	0.192	0.144

Теперь: $M(V) = 0 \cdot 0.136 + \frac{2}{9} \cdot 0.528 + \frac{6}{9} \cdot 0.192 + \frac{8}{9} \cdot 0.144 = 0.3733 \ldots$

Заметим, что $M(V) \neq \sigma^2 = 0.56$. Это означает, что выборочная дисперсия V, не является несмещенной оценкой дисперсии генеральной совокупности.

Однако можно изменить статистику выборки V, чтобы получить несмещенную оценку. Для начала, можете проверить, что в приведенном выше примере $M(V) = \frac{2}{3} \cdot 0.56$. Это пример общего правила.

Если V - выборочная дисперсия выборки размера n, взятой из генеральной совокупности с неизвестной дисперсией $\boldsymbol{\sigma^2}$, тогда $\boldsymbol{M(V) = \frac{n-1}{n} \cdot \sigma^2}$

Учитывая это общее правило и немного подумав о значении математического ожидания, вы сможете увидеть, что: $M\left(\frac{nV}{n-1}\right) = \sigma^2$

Это означает, что $V = \left(\frac{X_1^2 + X_2^2 + \cdots + X_n^2}{n}\right) - \bar{X}^2$ является <u>смещенной</u> оценкой σ^2, но:

$$\frac{nV}{n-1} = \frac{n}{n-1}\left(\left(\frac{X_1^2 + X_2^2 + \cdots + X_n^2}{n}\right) - \bar{X}^2\right)$$

является <u>несмещенной</u> оценкой σ^2.

Выражение $\frac{n}{n-1}\left(\left(\frac{X_1^2 + X_2^2 + \cdots + X_n^2}{n}\right) - \bar{X}^2\right)$ обозначается как S^2.

Или если записать выражение S^2 с использованием знака суммы $\sum$:

$$S^2 = \frac{n}{n-1}\left(\frac{\sum X^2}{n} - \bar{X}^2\right) = \frac{1}{n-1}\left(\sum X^2 - n\bar{X}^2\right)$$

Подставив: $\bar{X} = \frac{\sum X}{n}$, получим

$$S^2 = \frac{1}{n-1}\left(\sum X^2 - n\left(\frac{\sum X}{n}\right)^2\right) = \frac{1}{n-1}\left(\sum X^2 - \frac{(\sum X)^2}{n}\right)$$

Напомним, что: $\frac{\sum X^2}{n} - \bar{X}^2 = \frac{1}{n}\sum(X - \bar{X})^2$

Подставив в предыдущее выражение, получим:

$$S^2 = \frac{n}{n-1}\left(\frac{1}{n}\sum(X - \bar{X})^2\right) = \frac{1}{n-1}\sum(X - \bar{X})^2$$

Ниже приводится краткое изложение методов получения несмещенных оценок математического ожидания и дисперсии генеральной совокупности по выборочным значениям $x_1, x_2, \ldots, x_n$.

Если вы берете случайную выборку со значениями $x_1, x_2, \ldots, x_n$ из генеральной совокупности,

- ***несмещенная оценка математического ожидания μ, генеральной совокупности - это выборочное среднее значение по выборке.***
- ***несмещенная оценка дисперсии $\boldsymbol{\sigma^2}$, генеральной совокупности это:***

$$\boldsymbol{s^2 = \frac{1}{n-1}\left(\sum x^2 - \frac{(\sum x)^2}{n}\right) = \frac{1}{n-1}\sum(x - \bar{x})^2}$$

Обратите внимание, что заглавные буквы $X_1, X_2, \ldots, X_n$ и S здесь заменены строчными буквами $x_1, x_2, \ldots, x_n$ и s, поскольку определение относится к набору значений выборок, которые вы фактически получили.

Вы знали две разные формулы для вычисления дисперсии: формула $\frac{\sum x^2}{n} - \bar{x}^2$, или ее эквивалент

$\frac{1}{n}\sum(x-\bar{x})^2$ и формула, которая была только что получена:

$$s^2=\frac{1}{n-1}\left(\sum x^2-\frac{(\sum x)^2}{n}\right)=\frac{1}{n-1}\sum(x-\bar{x})^2$$

Нельзя однозначно сказать, какую формулу использовать в данной конкретной ситуации. Контекст вашей задачи определит, какую формулу вы должны использовать.

Если значения $x_1, x_2, \ldots, x_n$ представляют всю генеральную совокупность, то у вас есть все необходимые значения, и вам необходимо рассчитать дисперсию этой генеральной совокупности, используйте:

$$\boldsymbol{\sigma^2=\frac{\sum x^2}{n}-\bar{x}^2} \textbf{\textit{ или ее эквивалент }} \boldsymbol{\frac{1}{n}\sum(x-\bar{x})^2}$$

Если, с другой стороны, вы пытаетесь оценить дисперсию генеральной совокупности, из которой значения $x_1, x_2, \ldots, x_n$ являются случайной выборкой, используйте:

$$\boldsymbol{s^2=\frac{1}{n-1}\left(\sum x^2-\frac{(\sum x)^2}{n}\right)} \textbf{\textit{ или ее эквивалент }} \boldsymbol{s^2=\frac{1}{n-1}\sum(x-\bar{x})^2}$$

Большинство калькуляторов имеют отдельные команды для вычисления этих двух различных дисперсий.

Если значения выборки даны в виде сгруппированных данных, несмещенная оценка дисперсии генеральной совокупности определяется как:

$$s^2=\frac{1}{(\sum f)-1}\left(\sum x^2 f-\frac{(\sum xf)^2}{\sum f}\right)$$

Пример 5.1.3.

(а) Было воспроизведено девять компакт-дисков и записано время воспроизведения каждого компакт-диска. Время в минутах указано ниже.

49, 56, 55, 68, 61, 57, 61, 52, 63

Найдите среднее время воспроизведения девяти компакт-дисков и дисперсию времени воспроизведения девяти компакт-дисков.

(б) Студент делал проект изучающий время воспроизведения компакт-дисков. Она хотела бы оценить среднее время воспроизведения компакт-дисков, продаваемых по всей стране, и она также хотела оценить дисперсию времени воспроизведения компакт-дисков, продаваемых по всей стране. Она взяла выборку из девяти компакт-дисков и записала их время воспроизведения. Результаты приведены ниже.

49, 56, 55, 68, 61, 57, 61, 52, 63

(i) Используйте данные студента для оценки среднего времени воспроизведения компакт-дисков, продаваемых в стране.

(ii) Используйте данные студента, чтобы оценить дисперсию времени воспроизведения компакт-дисков, продаваемых в стране.

Решение:

Обратите внимание, что две части, (а) и (б) выглядят очень похожими. Однако в части (а) вас интересует время воспроизведения **только этих** девяти компакт-дисков. В части (b) вы хотите сделать оценку параметров генеральной совокупности (**всех** компакт-дисков, продаваемых в стране) на основе выборки из девяти компакт-дисков, которые вам были предоставлены.

(а) $\bar{x}=\frac{1}{9}(49+56+\cdots+63)=58$, таким образом, среднее время воспроизведения **только этих** девяти компакт-дисков равна 58 минут.

$$\text{Дисперсия}=\frac{\sum x^2}{n}-\bar{x}^2=\frac{1}{9}(49^2+56^2+\cdots+63^2)-58^2=30.4\ldots$$

то есть, дисперсия времени воспроизведения **только этих** девяти компакт-дисков равна 30.4 минут^2, с точностью до трех значащих цифр.

(б) В этой части вы оцениваете среднее время и дисперсию времени воспроизведения **всей** генеральной совокупности (**всех** компакт-дисков, произведенных в стране) на основе выборки из девяти компакт-дисков.

(i) Напомним, что несмещенная оценка математического ожидания μ - является средним значением для выборки: $\bar{x}$, поэтому оценка μ составляет 58 минут.

(ii) Чтобы получить несмещенную оценку дисперсии, σ^2, вам нужно использовать:

$$s^2 = \frac{1}{n-1}\left(\sum x^2 - \frac{(\sum x)^2}{n}\right)$$

$s^2 = \frac{1}{9-1}\left((49^2 + 56^2 + \cdots + 63^2) - \frac{(49+56+\cdots+63)^2}{9}\right) = 34.25 = 34.3$ минут^2, с точностью до трех значащих цифр.

В следующем примере вам не дают отдельные значения выборочных данных. Вместо этого данные сведены в сгруппированную таблицу частот.

Пример 5.1.4.

Рыболовецкая бригада зафиксировала массу в килограммах 200 рыб особого вида, которые были пойманы на их траулере. Результаты приведены в таблице ниже. Указанные веса являются значениями середины весового интервала.

Вес рыбы (кг)	0.5	1.25	1.75	2.25	2.75	3.5	4.5	5.5	7.0	10.5
Количество рыбы в этом весовом интервале	21	32	33	24	18	21	16	12	11	12

Предполагая, что эти рыбы являются случайной выборкой из генеральной совокупности этого вида, оцените:

(a) среднюю массу в килограммах рыбы этого вида,

(b) Дисперсия массы рыб этого вида.

Решение:

Вес рыбы, x (кг) (середина весового интервала)	Количество рыбы в этом весовом интервале, f	xf	x^2f
0.5	21	10.5	5.25
1.25	32	40	50
1.75	33	57.75	101.0…
2.25	24	54	121.5
2.75	18	49.5	136.1…
3.5	21	73.5	257.2…
4.5	16	72	324
5.5	12	66	363
7	11	77	539
10.5	12	126	1323
ИТОГО:	200	626.5	3220.1875

(a) Чтобы найти выборочное среднее выборки, используем формулу:

$\bar{x} = \frac{\sum xf}{\sum f} = \frac{626.25}{200} = 3.13125$, выборочное среднее 3.13125 кг.

Поскольку выборочное среднее является несмещенной оценкой среднего значения генеральной совокупности, вы можете принять это значение как оценку средней массы всех рыб этого вида. Поэтому оценка средней массы рыбы этого вида составляет 3,131 кг.

(б) Несмещенная оценка дисперсии генеральной совокупности:

$$s^2 = \frac{1}{(\sum f) - 1}\left(\sum x^2 f - \frac{(\sum xf)^2}{\sum f}\right) = \frac{1}{200-1}\left(3220.1875 - \frac{626.25^2}{200}\right) =$$

$= 6.327 = 6.33$, с точностью до трех значащих цифр.

5.2. Доказательство формулы несмещенной оценки дисперсии генеральной совокупности

Пусть сделано n независимых наблюдений случайной величины X со средним значением (математическим ожиданием) μ и дисперсией σ^2.

$$S^2 = \frac{\sum(X_i - \bar{X})^2}{n-1} = \frac{n}{n-1}\left(\frac{\sum X_i^2}{n} - \bar{X}^2\right) = \frac{1}{n-1}\left(\sum X_i^2 - n\bar{X}^2\right)$$

Рассмотрим математическое ожидание этой оценочной функции:

$$M\left(\sum X_i^2 - n\bar{X}^2\right) = M(X_1^2) + M(X_2^2) + \cdots + M(X_n^2) - nM(\bar{X}^2) =$$
$$= nM(X^2) - nM(\bar{X}^2)$$

Теперь рассмотрим: $Д(X) = \sigma^2 = M(X^2) - \left(M(X)\right)^2 = M(X^2) - \mu^2$, значит:

$M(X^2) = \sigma^2 + \mu^2$, и учитывая, что: $M(\bar{X}) = \mu$, и $Д(\bar{X}) = \frac{\sigma^2}{n}$, то

$Д(\bar{X}) = M(\bar{X}^2) - \left(M(\bar{X})\right)^2 \leftrightarrow \frac{\sigma^2}{n} = M(\bar{X}^2) - \mu^2$, получим:

$$M(\bar{X}^2) = \frac{\sigma^2}{n} + \mu^2$$

Подставляя в предыдущее выражение:

$$M\left(\sum X_i^2 - n\bar{X}^2\right) = n(\sigma^2 + \mu^2) - n\left(\frac{\sigma^2}{n} + \mu^2\right) = (n-1)\sigma^2$$

Следовательно: $M(S^2) = M\left(\frac{1}{n-1}(\sum X_i^2 - n\bar{X}^2)\right) = \sigma^2$.

Значит, оценочная функция: $S^2 = \frac{\sum(X_i-\bar{X})^2}{n-1}$, является несмещенной оценкой дисперсии σ^2

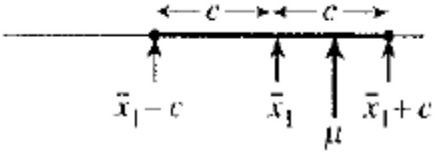
5.3. Понятие доверительного интервала

Вы узнали, что среднее значение $\bar{X}$,случайной выборки является несмещенной оценкой математического ожидания генеральной совокупности μ. Например, чтобы оценить среднее количество карманных денег, получаемых **всеми** детьми в начальной школе, вы можете взять случайную выборку детей и спросить каждого ребенка, сколько карманных денег он получает. Предположим, что $\sum x = 111.50$ для случайной выборки из 50 детей, где *x* измеряется в долларах. Тогда несмещенная оценка математическое ожидание генеральной совокупности, μ , определяется как:

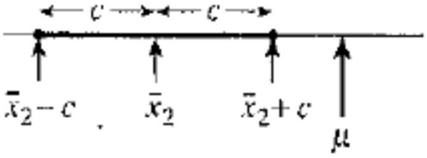
$$\bar{x} = \frac{\sum x}{n} = \frac{111.50}{50} = 2.23$$

Такое значение иногда называют **точечной оценкой**, потому что оно дает оценку математического ожидания генеральной совокупности в виде одного значения или "точки" на числовой линии. Такие значения полезны, например, при сравнении генеральных совокупностей. Однако, поскольку $\bar{X}$ является случайной величиной, значение, которое она принимает, будет варьироваться от выборки к выборке. В результате вы не имеете ни малейшего представления о том, насколько близка к фактическому значению математического ожидания генеральной совокупности будет ваша точечная оценка. Цель "доверительного интервала" состоит в том, чтобы дать оценку в форме, которая также указывает на вероятную точность оценки. **Доверительный интервал** для среднего - это диапазон значений, который имеет заданную вероятность "захвата" математического ожидания генеральной совокупности. Обычно считается, что он симметричен относительно среднего значения выборки. Таким образом, если выборочное среднее принимает значение$\bar{x}$, то соответствующий доверительный интервал будет равен $[\bar{x} - c, \bar{x} + c]$, где c-число, значение которого еще не найдено.

Обратите внимание на обозначение для используемого интервала.
Интервал $[\bar{x} - c, \bar{x} + c]$ *означает действительные числа от* $\bar{x} - c$ *до* $\bar{x} + c$*, включая конечные точки.*
Например, [-2.1,6.8] означает такие действительные числа y, что$-2.1 \le y \le 6.8$.

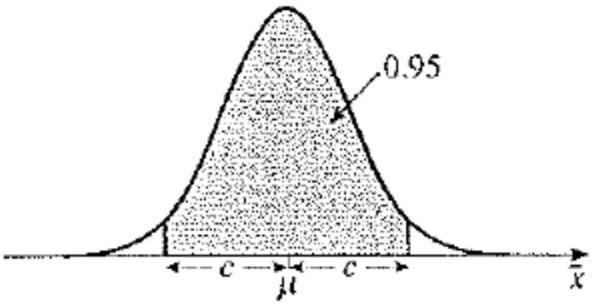


На этом рисунке выборочное среднее $\bar{X}$ принимает значение $\overline{x_1}$, а доверительный интервал охватывает диапазон значений $[\overline{x_1} - c, \overline{x_1} + c]$. В этом случае доверительный интервал «ловит» среднее значение генеральной совокупности, μ. Для другой выборки, с другим средним значением выборки, $\overline{x_2}$, доверительный интервал может не «ловить» μ. Эта ситуация проиллюстрирована на следующем рисунке:

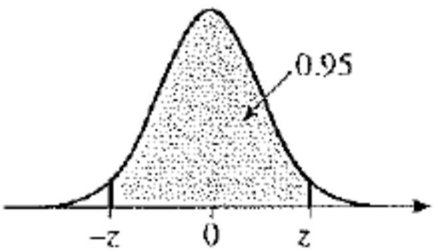


Границы доверительного интервала сами по себе являются случайными переменными, поскольку они варьируются от выборки к выборке. Их можно записать как$\bar{X} - c$ и $\bar{X} + c$. Из предыдущих рисунков видно, что доверительный интервал будет «улавливать» μ, если разница между средним значением выборки и средним значением генеральной совокупности меньше или равна c. Алгебраически это условие выражается как:
$|\bar{X} - \mu| \le c$.
В дальнейшем объясняется, как значение c выбирается таким образом, чтобы дать заданную вероятность того, что доверительный интервал «захватывает» μ.

5.4.Расчет доверительного интервала.
Рассмотрим следующую ситуацию. Известно, что массы таблеток, произведенных машиной, распределены по нормальному закону распределения со стандартным отклонением 0,012 г. Средняя масса полученных таблеток контролируется через регулярные промежутки времени путем взятия выборки из 25 таблеток и вычисления выборочного среднего $\bar{X}$.
Предположим, вы хотите найти интервал, который с вероятностью 95% «отловит» среднее значение генеральной совокупности, μ. Масса X (в граммах) одной таблетки распределена по нормальному закону распределения с неизвестным средним значением μ и стандартным отклонением 0,012; то есть, $X \sim N(\mu; 0.012^2)$. Тогда для выборки размером 25, $\bar{X} \sim N\left(\mu; \frac{0.012^2}{25}\right)$ (центральная предельная теорема)
Среднее значение генеральной совокупности μ будет находиться в интервале
$[\bar{X} - c, \bar{X} + c]$, где c – константа, если $|\bar{X} - \mu| \le c$. Среднее значение генеральной совокупности μ, будет лежать вэтом интервале с вероятностью 95% тогда и только тогда, когда: $P(|\bar{X} - \mu| \le c) = 0.95$.

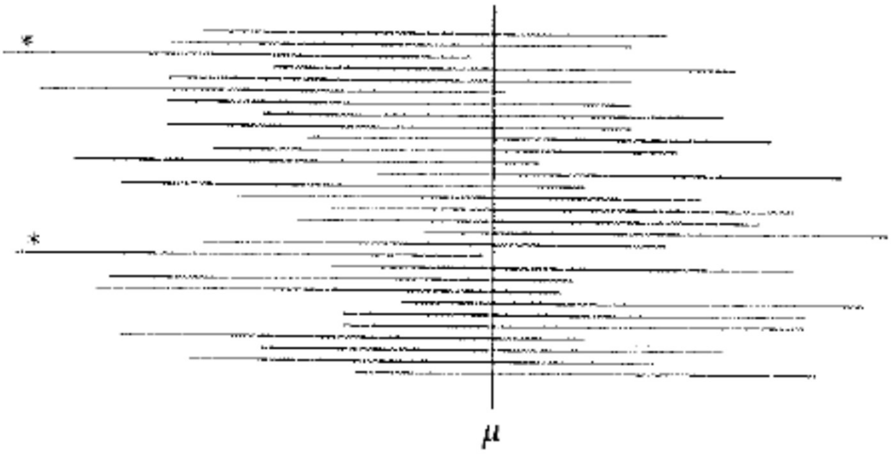


Этот рисунок показывает распределение выборочного среднего $\bar{X}$ с указанной вероятностью. Значение c определяется путем стандартизации:
$Z = \frac{\bar{X} - \mu}{\sqrt{\frac{0.012^2}{25}}}$, где Z~N(0;1).



На этом рисунке показано распределение Z с указанной вероятностью 0,95. Не заштрихованные области имеют вероятность $\frac{1}{2} \cdot 0.05 = 0.025$. Поэтому значение z на этой диаграмме определяется как Ф (z) = 0,975. Его можно найти с помощью таблиц нормального распределения. Получим z = 1.96.
Тогда: $1.96 = \frac{\bar{X} - \mu}{\sqrt{\frac{0.012^2}{25}}} = \frac{\bar{X} - \mu}{\frac{0.012}{\sqrt{25}}}$. Но $\bar{X} - \mu =$ c, значит: $1.96 = \frac{\text{c}}{\frac{0.012}{\sqrt{25}}}$. Следовательно: $\text{c} = 1.96 \cdot \frac{0.012}{\sqrt{25}} = 0.00470$, с точностью до трех значащих цифр.
Таким образом, интервал, в котором с вероятностью 95% находится значение
μ: $[\bar{X} - 0.0047, \bar{X} + 0.0047]$.
Предположим, что вы взяли выборку из 25 таблеток и обнаружили, что средняя масса составила,

например, 0.5642 г. Это среднее значение для выборки даст значение для интервала [0.5642 – 0.0047, 0.5642 + 0.0047], которое составит [0.5595, 0.5689]. Такой интервал называется 95% **доверительным интервалом для среднего значения генеральной совокупности.**
Важно понимать, что такой доверительный интервал может содержать или не содержать μ, в зависимости от значения $\bar{X}$. Предположим на мгновение, что вы знаете значение μ и берете несколько разных выборок по 25 таблеток. На следующем рисунке показаны доверительные интервалы, рассчитанные описанным выше способом для 30 различных выборок по 25 таблеток.



Большинство доверительных интервалов содержат μ, но некоторые (отмеченные *) – не содержат μ. В среднем доля доверительных интервалов, которые содержат μ, составляет 95%. На практике, конечно, вы не знаете μ и обычно берете только одну выборку и из нее вычисляете один доверительный интервал. В этой ситуации вы не можете знать, содержит ли ваш конкретный доверительный интервал значение μ : все, что вы можете сказать, это то, что в среднем 95 раз из 100 он будет содержать μ.
Метод, который был описан для нахождения 95% доверительного интервала для среднего значения, можно обобщить для выборки размера n из генеральной совокупности, распределенной по нормальному закону распределения со стандартным отклонением σ. Замена 25 на n и 0,012 на σ в этих уравнении: $1.96 = \frac{c}{\frac{\sigma}{\sqrt{n}}}$, отсюда $c = 1.96 \cdot \frac{\sigma}{\sqrt{n}}$.

Для выборки размера n из генеральной совокупности, распределенной по нормальному закону распределения с дисперсией $\boldsymbol{\sigma^2}$, 95% доверительный интервал для среднего значения μ этой генеральной совокупности, определяется как:
$\left[\boldsymbol{\bar{x} - 1.96 \cdot \frac{\sigma}{\sqrt{n}}; \bar{x} + 1.96 \cdot \frac{\sigma}{\sqrt{n}}}\right]$***, где $\boldsymbol{\bar{x}}$ – выборочное среднее.***

Деятельность на компьютере
Эта деятельность требует доступа к компьютеру.
Сделайте электронную таблицу, содержащую 500 выборок размером 25 из генеральной совокупности, распределенной по нормальному закону распределения: $N(0.56; 0{,}012^2)$. Используя возможности электронной таблицы, рассчитайте 95% доверительный интервал для среднего значения для каждой выборки. Сколько из этих доверительных интервалов содержит среднее значение 0,56? (Ваш ответ должен быть достаточно близок к 475. Почему?) Объясните, почему ваш ответ не обязательно должен быть точно 475.
Пример 5.4.1.
Известно, что длины гвоздей, изготавливаемых машиной, распределены по нормальному закону распределения со средним μ мм и стандартным отклонением 0,7 мм. Длина случайной выборки из 5 гвоздей составляет 107,29, 106,56, 105,94, 106,99, 106,47 мм.
(a) Рассчитайте симметричный 95% доверительный интервал для μ, указав границы интервала с точностью до 1 знака после запятой.
(б) Отбирают двести случайных выборок из 5 гвоздей и для каждой выборки рассчитывают симметричный 95% доверительный интервал для μ . Найдите ожидаемое количество интервалов, которые не содержат μ.
Решение:
(a) Выборочное среднее:

$$\bar{x} = \frac{107.29 + 106.56 + 105.94 + 106.99 + 106.47}{5} = 106.65$$

Подставляя значение $\bar{x} = 106.65$, $\sigma = 0.7$ и n=5 в формулы 95% доверительного интервала, получим: $\left[106.65 - 1.96 \cdot \frac{0.7}{\sqrt{5}}; 106.65 + 1.96 \cdot \frac{0.7}{\sqrt{5}}\right]$, которое после упрощения даст: [106.036…;107.263…].

Таким образом, симметричный 95% доверительный интервал для μ с точностью до 1 знака после запятой равен [106.0; 107.3] мм.

(б) В среднем, 95% доверительных интервалов содержит μ. Это означает, что 5% не содержит μ. Тогда из 200 доверительных интервалов около

$200 \cdot 5\% = 10$-ти интервалов не содержит μ.

Пример 5.4.2

Известно, что для метода измерения скорости звука в воздухе результаты повторных экспериментов распределяются по нормальному закону распределения со стандартным отклонением 6 м/с. С помощью этого метода проводится ряд измерений, и из этих измерений рассчитывается симметричный 95% доверительный интервал для скорости звука в воздухе. Найти ширину этого доверительного интервала для (а) 4, (б) 36 измерений.

Решение:

Симметричный 95% доверительный интервал простирается от $\bar{x} - 1.96 \cdot \frac{\sigma}{\sqrt{n}}$до $\bar{x} + 1.96 \cdot \frac{\sigma}{\sqrt{n}}$, значит его ширина равна:

$$\bar{x} + 1.96 \cdot \frac{\sigma}{\sqrt{n}} - \left(\bar{x} - 1.96 \cdot \frac{\sigma}{\sqrt{n}}\right) = 2 \cdot 1.96 \cdot \frac{\sigma}{\sqrt{n}}$$

В этом примере $\sigma = 6$, поэтому ширина доверительного интервала равна

$$2 \cdot 1.96 \cdot \frac{6}{\sqrt{n}} = \frac{23.52}{\sqrt{n}}$$

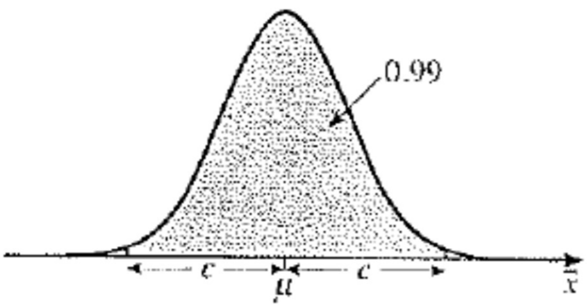
(а) Для n = 4, ширина доверительного интервала равна $\frac{23.52}{\sqrt{4}} = 11.76$ м/с

(б) Для n = 36, ширина доверительного интервала равна $\frac{23.52}{\sqrt{36}} = 3.92$ м/с

Обратите внимание, что требуется в девять раз больше измерений, чтобы уменьшить ширину доверительного интервала в 3 раза.

5.5.Различные уровни доверия

Как показывает предыдущий пример, существует вероятность 5 на 100, что 95% доверительный интервал не содержит μ. Существуют обстоятельства, при которых вы хотите быть более уверены, что рассчитанный вами доверительный интервал действительно содержит μ. Например, вы хотите быть уверены на 99%. На рисунке показана диаграмма для этой ситуации.



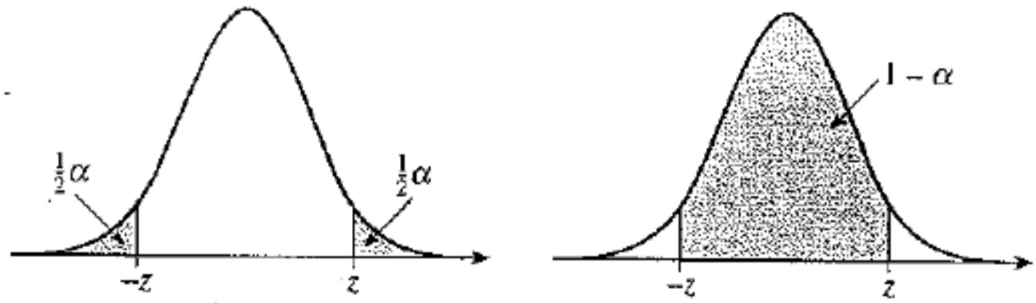
Единственная разница в расчете доверительного интервала состоит в том, что требуется другое значение z. В этом случае Ф (z) = 0,995, что дает z = 2,557, так что 99%доверительный интервал для среднего значения определяется как:

$$\left[\bar{x} - 2.576\frac{\sigma}{\sqrt{n}}; \bar{x} + 2.576\frac{\sigma}{\sqrt{n}}\right]$$

Обратите внимание, что этот 99% - ный доверительный интервал шире 95% - ного доверительного интервала: этого следует ожидать, поскольку он с большей вероятностью содержит μ, чем 95% - ный интервал. Если вам нужен 90% доверительный интервал, то подходящим значением z будет то значение, для которого Ф(z) = 0.95 дает z = 1.645 и доверительный интервал уже (меньше). Вы можете видеть, что существует баланс между точностью и достоверностью: если вы увеличиваете одно, вы уменьшаете другое. Для того чтобы увеличить как точность, так и достоверность доверительного интервала, необходимо увеличить размер выборки.

В общем случае, $100(1-\alpha)$%-ый доверительный интервал для среднего значения генеральной совокупности (μ) и для выборки размера n, взятой из этой генеральной совокупности распределенной по нормальному закону распределения с дисперсией σ^2, задается следующим образом:
$\left[\bar{x} - z\frac{\sigma}{\sqrt{n}}; \bar{x} + z\frac{\sigma}{\sqrt{n}}\right]$, где $\bar{x}$ – выборочное среднее, и такое значение z, для которого $\Phi(z) = 1 - \frac{1}{2}\alpha$.

Взаимосвязь между α и z проиллюстрировано в следующем рисунке:



Пример 5.5.1.
Известно, что погрешность измерения при измерении концентрации (количество частиц на миллион (ppm)) нитрат-ионов в воде определенным методом распределено по нормальному закону распределения со средним значением 0 и стандартным отклонением 0,05.
(а)Если измерения концентрации в 10-ти образцах дали $\sum x = 11.37$частиц на миллион(ppm), определите симметричный доверительный интервал 99,5% для концентрации нитрата-иона в образце, определив границы интервалов с точностью до 2-х знаков после запятой.
(б) Сколько измерений потребуется, чтобы уменьшить ширину этого интервала максимум до 0,03 ppm?
Решение:
(а) Измеренное значение X, концентрации нитрат-ионов равно $\mu + Y$, где Y - погрешность измерения.Таким образом,
$M(X) = M(\mu + Y) = \mu + M(Y) = \mu + 0 = \mu$
$Д(X) = Д(\mu + Y) = Д(Y)$
Это означает, что $X \sim N(\mu; 0.05^2)$. Для 99.5%-ного доверительного интервала:
$1 - \alpha = 0.995$, тогда $\alpha = 0.005$и $\frac{1}{2}\alpha = 0.0025$. Поэтому:
$$\Phi(z) = 1 - \frac{1}{2}\alpha = 1 - 0.0025 = 0.9975$$
Из таблицы нормального распределения z=2.807
Тогда 99.5% доверительный интервал для μ:
$$\left[\bar{x} - 2.807\frac{\sigma}{\sqrt{n}}; \bar{x} + 2.807\frac{\sigma}{\sqrt{n}}\right] = \left[\frac{11.37}{10} - 2.807\frac{0.05}{\sqrt{10}}; \frac{11.37}{10} + 2.807\frac{0.05}{\sqrt{10}}\right] =$$
$= [1.092 \ldots; 1.181 \ldots] = [1.09; 1.18]$, с точностью до двух цифр после запятой.
(б) Ширина 99.5%-ного доверительного интервала для выборки размера n равна $2 \cdot 2.807\frac{0.05}{\sqrt{n}}$. Эта ширина должна быть не более 0.03. Значит:
$2 \cdot 2.807\frac{0.05}{\sqrt{n}} \le 0.03$. Упрощаем: $\sqrt{n} \ge \frac{2 \cdot 2.807 \cdot 0.05}{0.03}$, или $n \ge 87.5$
Поскольку n должно быть целым числом, требуется 88 или более измерений, чтобы обеспечить доверительный интервал шириной не более 0,03 ppm.

5.6.Доверительный интервал для большой выборки

В рассмотренных выше примерах предполагалось, что выборки взяты из нормально распределенной генеральной совокупности и что дисперсия этого распределения известна. На практике вы не всегда будете уверены, что генеральная совокупность распределена по нормальному закону распределения, и вы можете иметь или не иметь точную информацию о ее отклонениях. Однако метод вычисления доверительного интервала, который был описан, все еще может применяться при условии большой выборки.
Сначала рассмотрим распределение генеральной совокупности, из которого берется выборка. Хотя это распределение может быть и не нормальным законом распределения, центральная предельная теорема

говорит, что среднее значение выборки, $\bar{X}$, распределена приблизительно по нормальному закону распределения при условии, что выборка большая.
Далее рассмотрим дисперсию генеральной совокупности. Несмещенной оценкой дисперсии является:
$s^2 = \frac{1}{n-1}\sum(x-\bar{x})^2 = \frac{1}{n-1}\left(\sum x^2 - \frac{(\sum x)^2}{n}\right)$
Эта оценка будет варьироваться от выборки к выборке, но для больших выборок вариация настолько мала, что s можно рассматривать как постоянную, и ее можно использовать вместо σв формуле.
Эмпирическое правило заключается в том, что «большой» означает размер выборки 30 или более.

Для достаточно большой выборки $(\boldsymbol{n \geq 30})$ любой генеральной совокупности, $\mathbf{100(1-\alpha)}$%-ый доверительный интервал для среднего значения генеральной совокупности (μ) задается следующим образом:
$\left[\boldsymbol{\bar{x} - z\frac{s}{\sqrt{n}}; \bar{x} + z\frac{s}{\sqrt{n}}}\right]$***, где $\boldsymbol{\bar{x}}$ – выборочное среднее, и такое значение z, для которого $\Phi(\boldsymbol{z}) = \mathbf{1} - \frac{\mathbf{1}}{\mathbf{2}}\boldsymbol{\alpha}$ и***
$\boldsymbol{s^2 = \frac{1}{n-1}\left(\sum x^2 - \frac{(\sum x)^2}{n}\right)}$.

Пример 5.6.1.
1 января в определенном здании было установлено 100 новых лампочек Eternity вместе с устройством, которое регистрирует продолжительность использования каждой лампочки. К 1 марта все 100 лампочек вышли из строя. Данные для зарегистрированных сроков службы лампочек, t (в часах), суммированы как $\sum t = 10500$и $\sum t^2 = 1712500$. Предполагая, что установленные лампочки представляют собой случайную выборку лампочек Eternity, получите симметричный 99% доверительный интервал для среднего срока службы лампочек Eternity, с точностью до одного часа.
Решение:
В этом примере не известно ни распределение генеральной совокупности по сроку службы лампочек, ни дисперсия. Однако размер выборки намного больше 30, поэтому можно использовать формулы доверительного интервала. Для начала необходимо вычислить $\bar{t}$ и s^2.

$$\bar{t} = \frac{\sum t}{n} = \frac{10500}{100} = 105$$

$$s^2 = \frac{1}{n-1}\left(\sum t^2 - \frac{(\sum t)^2}{n}\right) = \frac{1}{100-1}\left(1712500 - \frac{10500^2}{100}\right) = 6161.6 \ldots$$

Поэтому: s=78.49…
Для 99%-го доверительного интервала, $1-\alpha = 0.99$, то есть $\alpha = 0.01$
и $\frac{1}{2}\alpha = 0.005$. Таким образом, $\Phi(z) = 1 - \frac{1}{2}\alpha = 1 - 0.005 = 0.995$
Из таблицы нормального распределения: z=2.576
Итак, 99%-ый интервал для среднего срока службы лампочек Eternity, составляет:
$\left[105 - 2.576\frac{78.49\ldots}{\sqrt{100}}; 105 + 2.576\frac{78.49\ldots}{\sqrt{100}}\right] = [85; 125]$, с точностью до часа.

5.7.Доверительный интервал для доли генеральной совокупности.
Многие статистические исследования связаны с определением **доли** генеральной совокупности, имеющего определенный признак. Предположим, вы являетесь производителем устройства, предназначенного для левшей. Для того чтобы оценить потенциальный рынок, Вас заинтересует доля левшей в общей численности населения. Было бы невозможно спросить всех, и поэтому вам пришлось бы изучить выборку. Предположим, что вам удалось получить информацию от случайной выборки из 500 человек, и вы обнаружили, что 60 из них были левшами. Представляется разумным подсчитать, что доля левшей в общей численности населения составляет $\frac{60}{500} = 0.12$ или 12%. Однако вы должны быть уверены, что этот метод дает вам несмещенную оценку p.
Рассмотрим более общую ситуацию, когда опрашивается случайная выборка из n человек. При условии, что n намного меньше, чем численность населения, распределение случайной величины X - количество левшей в выборке размера n, будет B (n, p), то есть распределена по биномиальному закону распределения, так как:
- есть фиксированное количество испытаний (n человек опросили);
- каждое испытание имеет два возможных исхода (левша или правша);

- исходы являются несовместными (при условии, что никто не является одновременно и левшой, и правшой);
- вероятность того, что человек левша постоянна;
- испытания являются независимыми.

Пусть P_s – случайная величина «доля левшей в выборке». Тогда $P_s = \frac{X}{n}$. Математическое ожидание P_s:
$M(P_s) = M\left(\frac{X}{n}\right) = \frac{1}{n}M(X) = \frac{1}{n} \cdot np = p$
Таким образом, доля в выборке дает несмещенную точечную оценку доли генеральной совокупности. Однако было бы более полезно найти доверительный интервал для p, поскольку это дает представление о точности оценки. Чтобы сделать это, вам нужно рассмотреть распределение P_s. Для большой выборки X будет распределен приблизительно по нормальному закону распределения, и поэтому P_s также будет распределен приблизительно по нормальному закону распределения. Чтобы рассчитать доверительный интервал, вам также нужна дисперсия P_s. Ее можно найти следующим образом.
$$Д(P_s) = Д\left(\frac{X}{n}\right) = \frac{1}{n^2} \cdot Д(X) = \frac{1}{n^2} \cdot npq = \frac{pq}{n}$$

Если X ~ B (n, p), то доля выборки $\boldsymbol{P_s}$***, где*** $\boldsymbol{P_s = \frac{X}{n}}$***, распределена приблизительно по нормальному закону распределения, то есть*** $\boldsymbol{N\left(p; \frac{pq}{n}\right)}$***, когда n достаточно велико, чтобы выполнялись неравенства np> 5 и nq>5. На практике это означает, что если в n испытаниях будет более 5 успехов и 5 неудач.***

Ранее мы показали, что выборочное среднее распределено по нормальному закону распределения, то есть $\bar{X} \sim N\left(\mu; \frac{\sigma^2}{n}\right)$, и доверительный интервал для математического ожидания генеральной совокупности μ задается как:
$\left[\bar{x} - z\frac{\sigma}{\sqrt{n}}; \bar{x} + z\frac{\sigma}{\sqrt{n}}\right]$. По аналогии, доверительный интервал для p находится путем замены p_sвместо $\bar{x}$, где $p_s = \frac{x}{n}$изамены $\sqrt{\frac{pq}{n}}$ вместо $\frac{\sigma}{\sqrt{n}}$. Получим приблизительный доверительный интервал:
$\left[p_s - z\sqrt{\frac{pq}{n}}; p_s + z\sqrt{\frac{pq}{n}}\right]$, где значение z находится как и раньше.
Этот доверительный интервал выражается через p и q, которые неизвестны (в противном случае доверительный интервал не потребовался бы!). Однако, опять же, аналогично нахождению доверительного интервала для математического ожидания, эти неизвестные величины можно заменить их оценками из выборки при условии, что выборка велика. В результате получим следующий интервал:
$\left[p_s - z\sqrt{\frac{p_s q_s}{n}}; p_s + z\sqrt{\frac{p_s q_s}{n}}\right]$,
где $q_s = 1 - p_s = 1 - \frac{x}{n}$.
Строго говоря, поскольку в этом расчете дискретное распределение было заменено на непрерывное, следует применить коррекцию непрерывности. Однако она обычно игнорируется при расчете доверительного интервала для пропорции. Кроме того, строго говоря, следует использовать несмещенную оценку дисперсии генеральной совокупности, рассчитываемую как $\frac{n}{n-1}\left(\frac{p_s q_s}{n}\right)$, а не $\frac{p_s q_s}{n}$. Выражение $\frac{n}{n-1}$ представляет собой поправку, аналогичную той, что содержится в формуле для s^2. Когда n велико, эта корректировка не имеет большого значения, и поэтому мы ее проигнорируем.

Для достаточно большого размера выборки n из генеральной совокупности, где p – доля членов этой генеральной совокупности с определенным признаком, приблизительный $\boldsymbol{100(1-\alpha)}$***%-ый доверительный интервал для p:***
$\boldsymbol{\left[p_s - z\sqrt{\frac{p_s q_s}{n}}; p_s + z\sqrt{\frac{p_s q_s}{n}}\right]}$***. Где*** p_s ***– выборочная доля с этим признаком,*** $\boldsymbol{p_s = 1 - q_s}$***, а значение zтакое, что*** $\Phi(\boldsymbol{z}) = \boldsymbol{1 - \frac{1}{2}\alpha}$***.***
Этот доверительный интервал является приблизительным, потому что:
- ***дискретное распределение приближается непрерывным;***

- ***не применяется коррекция непрерывности;***
- ***дисперсия генеральной совокупности оценивается по выборке, а используемая оценка является смещенной;***
- ***случайная величина P_s – только приблизительно распределена по нормальному закону распределения.***

Эту формулу теперь можно использовать для расчета доверительного интервала для доли левшей в генеральной совокупности. В нашей выборке: ы n = 500, p_s = 0,12 и
q_s = 0,88.Таким образом, 95%-ный доверительный интервал доли левшей:
$\left[0.12 - 1.96\sqrt{\frac{0.12\cdot 0.88}{500}}; 0.12 + 1.96\sqrt{\frac{0.12\cdot 0.88}{500}}\right] = [0.092; 0.148]$, с точностью до трех цифр после запятой.
Тот же самый ответ будет получен, если применить коррекцию непрерывности и использовать несмещенную оценку дисперсии.
Интересно отметить, что ответ не зависит от размера генеральной совокупности, а только от размера выборки. Это будет верно во всех случаях, когда размер выборки намного меньше, чем размер генеральной совокупности, так что значение *p* фактически постоянно. Этот факт можно использовать для расчета размера выборки, необходимого для определения доверительного интервала заданной ширины, как показано в следующем примере.
Пример 5.7.1.
С целью оценки доли избирателей страны, которые проголосуют «да» на предстоящем референдуме, должен быть проведен опрос общественного мнения. Во время пробного опроса была опрошена случайная выборка из 100 человек; 42 сказали, что они будут голосовать "да". Оцените размер случайной выборки, необходимый для получения 99% доверительного интервала доли с шириной 0,02.
Решение:
Для 99%-ного доверительного интервала значение z=2.576. Тогда ширина этого доверительного интервала: $2\cdot 2.576\sqrt{\frac{p_s q_s}{n}}$, где p_s и q_s используются для оценки p и q. Согласно пробного опроса p_s = 0,42 и, следовательно,
q_s = 0,58. Таким образом, требуемое значение n, удовлетворяет уравнению:
$2\cdot 2.576\sqrt{\frac{0.42\cdot 0.58}{n}} = 0.02$. Решая которое получим, n = 16165, с точностью до целого числа. Этот ответ будет только приблизительным по вышеуказанным 4-м причинам.
Интересно отметить, что, если этот пример относится к определенной стране с электоратом около 40 миллионов, то требуемая выборка составляет всего около 0,04% электората. Проблема заключается в получении случайной выборки. На практике опросы общественного мнения не опираются на случайные выборки, а используют сложные методы, предназначенные для обеспечения репрезентативных выборок.

Задачи.

№1. Выборочное среднее случайной выборки из n наблюдений над случайной величиной X, где $X \sim N(\mu, \sigma^2)$, обозначается как $\bar{X}$. В каждой из следующих случаев выразите значение $b - a$ через n и σ.

а) $P(\bar{X} < a) = 0{,}03$, $P(\bar{X} > b) = 0{,}07$

б) $P(\bar{X} < a) = 0{,}04$, $P(\bar{X} > b) = 0{,}06$

в) $P(\bar{X} < a) = 0{,}05$, $P(\bar{X} > b) = 0{,}05$

Что подтверждают эти результаты о ширине симметричного доверительного интервала для μ?

Ответ №1: а) $\frac{3{,}357\sigma}{\sqrt{n}}$ б) $\frac{3{,}306\sigma}{\sqrt{n}}$ в) $\frac{3{,}29\sigma}{\sqrt{n}}$. Они подтверждают, что симметричный доверительный интервал имеет наименьшую ширину.

№2. Масса шоколадных батончиков, сходящих с производственной линии, распределена по нормальному закону распределения со средним значением μ и стандартным отклонением 1,5 грамма. Масса, в граммах, 9-ти случайно выбранных батончиков равна:

251,79 251,13 251,34 248,66 249,74 249,91 250,45 249,32 251,36.

Вычислите 95%-ный симметричный доверительный интервал для среднего значения генеральной совокупности, μ.

Ответ №2: $[249{,}43; 251{,}39]$

№3. Продолжительность времени, x часов, в течение которых автомобили были припаркованы на автостоянке в центре города, была измерена для случайной выборки из 200 автомобилей, и вычислили, что $\sum x = 358{,}2$ и $\sum x^2 = 773{,}18$. Обозначив среднее значение и дисперсию времени стоянки генеральной совокупности через μ и σ^2 соответственно, вычислите:

а) несмещенные оценки μ и σ^2

б) симметричный 90%-ный доверительный интервал для μ.

Также было установлено, что 39 из 200 автомобилей оставались на стоянке более 2,5 часов. Рассчитайте симметричный 90% - ный доверительный интервал для доли генеральной совокупности автомобилей, остающихся дольше 2,5 часов, с указанием границ интервалов с точностью до трех цифр после запятой. Менеджер автостоянки считает, что только 10% всех автомобилей остаются дольше 2,5 часов. Это подтверждается вашими расчетами?

Ответ №3: а) 1,791 часов, 0,6615 часов^2 б) $[1{,}696; 1{,}886]$

$[0{,}149; 0{,}241]$. Нет, доверительный интервал указывает на среднее значение, превышающее 10%.

№4. Объем краски, разлитый в литровую банку, был измерен для 100 случайно выбранных банок, и результаты, x литров, сведены в следующие суммы: $\sum x = 104{,}0$ и $\sum x^2 = 110{,}06$.

а) Рассчитайте несмещенные оценки среднего и дисперсии объемов краски, разлитых в банки.

б) Вычислите симметричный 90% доверительный интервал для среднего объема краски, определяя границы интервала с соответствующей степенью точности.

в) Оцените наименьший размер выборки, необходимый для получения симметричного 90% - ного доверительного интервала шириной не более 0,02 литра.

Ответ №4: а) 1,04 литра, 0,0192 литра^2 б) $[1{,}02; 1{,}06]$ в) 520

№5. Попечительский совет благотворительного фонда желает внести изменения в устав фонда. Для того, чтобы это произошло, по крайней мере две трети членов должны проголосовать за изменение. Перед голосованием секретарь консультируется со случайной выборкой из 60 членов и выясняет, что 75% из них проголосуют за изменение. Вычислите симметричный 95% доверительный интервал для совокупности всех членов, которые будут голосовать за изменение. Ближе к тому времени, когда должно быть проведено голосование, секретарь консультируется со случайной выборкой из n членов и обнаруживает, что 75% из них проголосуют за изменение. Используя эту цифру, он вычисляет симметричный 99% доверительный интервал для совокупности всех членов, которые проголосуют за изменение. Этот интервал не включает в себя значение две трети. Найдите наименьшее возможное значение n.

Ответ №5: $[0{,}640; 0{,}860]$, $n = 180$

№6. Определенная марка бензина использовалась в 80-ти случайно выбранных автомобилях той же модели и возраста. Расход бензина, x миль на галлон, был получен для каждого автомобиля. Результаты

обобщены в следующих суммах: $\sum x = 1896$ и $\sum x^2 = 45959$. Рассчитайте приблизительный симметричный 98% доверительный интервал среднего расхода бензина для всех автомобилей этой модели и возраста. Приведите причину, по которой интервал является приблизительным.
Ответ №6: $[22{,}8; 24{,}6]$. нормальное распределение выборочного среднего является приближенным (центральная предельная теорема) и используется оценка дисперсии (то есть приблизительное значение дисперсии)

№7. Случайная величина X распределена по нормальному закону распределения со средним значением μ и дисперсией σ^2. Симметричный 98% интервал для μ, основанный на случайной выборке из 25 наблюдений над X равен $[12{,}4; 12{,}6]$. Найдите:
а) значение σ.
б) размер случайной выборки, которая даст симметричный 95% доверительный интервал для μ ширины, максимально близкий к 0,8.
Ответ №7: а) 3,439 (0,215?) б) 284

№8. Случайная выборка из 200 рыб была поймана из пруда, содержащего большое количество рыб. Каждая рыба была помечена и возвращена в пруд. На следующий день было поймана 400 рыб, и 22 из них оказались помеченными. Их тоже вернули в пруд.
а) Получите симметричный 90% доверительный интервал для доли меченых рыб в пруду.
б) Если предположить, что количество меченых рыб все-таки 200, то что можно сказать о численности популяции рыб в этом пруду?
Ответ №8: а) $[0{,}0362; 0{,}0738]$
б) От 2710 до 5520 на основе доверительного интервала. 3640 на основе p_s.

№9. Составитель кроссвордов классифицирует головоломку как "легкую", если 60% или более людей, пытающихся решить головоломку, могут правильно завершить ее в течение 20 минут. Она классифицируется как "трудная", если менее 30% людей могут выполнить ее правильно в течение 20 минут. Все остальные головоломки классифицируются как "средние". Конкретная головоломка была дана 150 участникам конкурса, и 74 правильно выполнили ее в течение 20 минут. Компилятор хочет быть на 90% уверенным в правильной классификации головоломки.
а) Как следует классифицировать данную головоломку?
б) Может ли она быть на 95% уверена, что правильно классифицировала данную головоломку?
Ответ №9: а) 90% доверительный интервал для p составляет $[0{,}426; 0{,}561]$, поэтому данная головоломка классифицируется как «средняя»
б) 95% доверительный интервал составляет $[0{,}413; 0{,}573]$, поэтому ответ «да, может быть уверена, что правильно классифицировала головоломку»

№10. Телезритель замечает, что многие телепрограммы начинаются на несколько минут позже объявленного времени. Для случайной выборки из 120 программ несмещенные оценки среднего и дисперсии времени задержки составляют 1,75 минуты и 2,46 минуты^2 соответственно, с точностью до 3-х значащих цифр. Используйте эти оценки для расчета симметричного 99% - ного доверительного интервала для среднего времени задержки всех телепрограмм (генеральной совокупности телепрограмм). Зритель рассматривает нормальное распределение $N(1{,}75; 2{,}46)$ как модель задержки в минутах произвольно выбранной телепрограммы. Покажите, что эта модель не согласуется с наблюдением зрителя о том, что телевизионные программы почти никогда не начинаются раньше объявленного времени. Влияет ли тот факт, что распределение времени задержки может быть не быть нормально распределенной, на достоверность расчета доверительного интервала для среднего времени задержки? Объясните причину своего ответа.
Ответ №10: $[1{,}38; 2{,}12]$; $P(T < 0) = 0{,}132$, не достаточно малая вероятность (13,2%). Выборка достаточно велика для применения центральной предельной теоремы, поэтому интервал справедлив для наиболее разумного распределения.