

ПРИКЛАДНАЯ СТАТИСТИКА



Посиделка 4: распределение среднего

Ульянкин Ппилиф *

Чтобы забыть что-нибудь ненужное, надо сначала выучить что-нибудь ненужное.

Кот Матроскин про матстат

В этой посиделке мы начнём строить статистику, основанную на средних. Мы примем на веру, что среднее имеет нормальное распределение, посчитаем его характеристики, построим доверительный интервал и проверим свою первую гипотезу. Большинство понятий в этой лекции формализованы недостаточно чётко. Она нужна для того, чтобы на простом примере понять как строятся критерии. Все понятия мы формализуем более чётко после того как обсудим центральную предельную теорему и закон больших чисел.

1 Дядя Фёдор и конфликт

В Селе Гипотезово проживает 4 человека. У каждого из них свой рост:

| | |
|------|-----|
| Маша | 150 |
| Паша | 160 |
| Саша | 180 |
| Даша | 190 |

*https://github.com/FUlyankin/matstat_lec

Дедя Фёдор, Шарик и Матроскин проезжают через Гипотезово в Простоквашино транзитом. Ходят легенды, что в глубокой древности в Гипотезово жили великаны. Жители Гипотезово утверждают, что гены великанов передались им, поэтому в селе все очень высокие¹. Ребятам очень интересно, правда ли, что жители Гипотезово очень высокие.

- а) Посчитайте настоящий средний рост в Гипотезово по всей генеральной совокупности.
- б) Шарик гулял по деревне, встретил Сашу и Дашу. Он посчитал по ним средний рост и сказал, что это оценка среднего роста в Гипотезово. Найдите оценку Шарика. Насколько сильно эта оценка отличается от настоящего среднего?
- в) Матроскин тоже гулял по деревне. К нему в выборку попали Маша и Паша. Какую оценку он получил? Далека ли она от реального среднего?
- г) К дяде Фёдору в выборку попали Маша и Саша. Как дела обстоят с его оценкой?
- д) Подерутся ли между собой Шарик, Матроскин и дядя Фёдор? Почему результаты получились именно такими? Может ли так происходить в реальности?

Решение:

Будем обозначать настоящий средний рост, который задумала в деревне природа, буквой μ . Подсчитаем средний рост в Гипотезово по всей генеральной совокупности:

$$\mu = \frac{1}{4} \cdot (190 + 180 + 160 + 150) = 170.$$

Посчитаем такие же средние по маленьким выборкам, которые собрали исследователи:

$$\text{Фёдор: } \bar{x} = 0.5 \cdot (190 + 150) = 170$$

$$\text{Шарик: } \bar{x} = 0.5 \cdot (190 + 180) = 185$$

$$\text{Матроскин: } \bar{x} = 0.5 \cdot (150 + 160) = 155$$

Средние значения, рассчитанное по выборкам, мы используем как оценку для неизвестного μ . Каждый житель Простоквашино посчитал средний рост по двум жителям Гипотезово. У дяди Фёдора результат совпал с настоящим средним. Означает ли это, что он оценивал среднее правильнее своих коллег? **На самом деле нет. Ему просто повезло.** Из-за того, что отбор людей в выборку происходит случайно, средний рост оказывается случайной величиной. Если бы жители Простоквашино понимали это, они бы не подрались. А так, конечно, передерутся.

Происходят ли такие ситуации в реальности? Да сплошь и рядом. Каждая характеристика (среднее, доля и тп), которую мы пытаемся оценить, чтобы проверить какой-то эффект (вырастут ли продажи, если поменять дизайн бутылки) или ответить на давно мучающий нас вопрос (правда ли, что в Австралии зарплаты у девушек выше, чем у мужчин, то есть девушек дискриминируют), считается по случайной выборке из генеральной совокупности. **Любая выборочная характеристика будет случайной величиной.**

¹Если бы жители Гипотезово знали про Гальтона и его работу, у них бы не возникало таких иллюзий. Но об этом мы поговорим чуть ниже.

2 Печкин и исследование

Жизнь в Простоквашино изрядно испортилась. Почтальону Печкину надоела вся эта ругань. Чтобы раз и навсегда покончить с раздорами, он сел на велосипед и поехал в Гипотезово. Там он измерил рост всех четверых жителей, а после стал фантазировать, что могло бы получиться в качестве среднего, если бы он опросил только двух каких-то жителей.

- Найдите распределение среднего, посчитанного по двум наблюдениям.
- Постройте гистограмму для распределения среднего. Столбики стройте с шагом 5, верхнюю границу включайте в столбик. Отметьте на картинке рост, который получил Шарик, дядя Фёдор и Матроскин. Какая из оценок ближе всего к центру распределения?
- Какова вероятность оказаться в хвостах распределения? Какова вероятность оказаться в его центре?

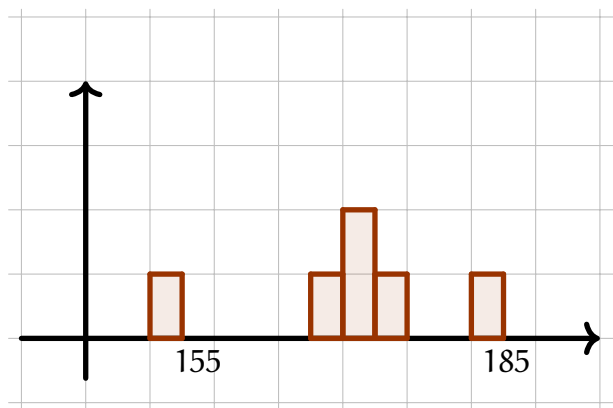
Решение:

Вспомним из комбинаторики сочетания и посчитаем, сколько разных значений может принять среднее. В случайную выборку мы всегда берём двоих. Это число сочетаний из 4 по 2: $C_4^2 = \frac{4!}{2!2!} = 6$.

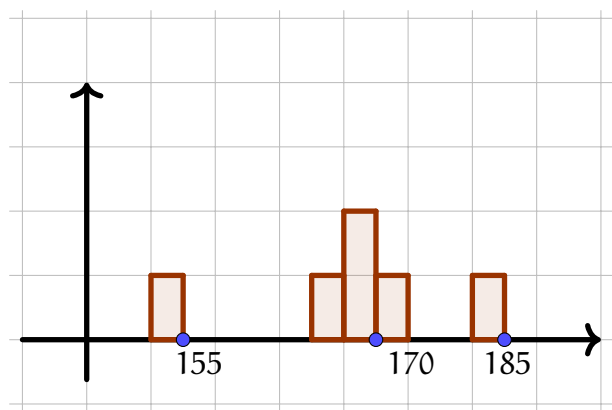
Давайте выпишем все значения, которые может принять средний рост:

| | |
|-------------|-------------------------------|
| Маша и Паша | $0.5 \cdot (150 + 160) = 155$ |
| Маша и Саша | $0.5 \cdot (150 + 190) = 170$ |
| Маша и Даша | $0.5 \cdot (150 + 180) = 165$ |
| Паша и Саша | $0.5 \cdot (160 + 190) = 175$ |
| Паша и Даша | $0.5 \cdot (160 + 180) = 170$ |
| Саша и Даша | $0.5 \cdot (190 + 180) = 185$ |

Мы берём двух человек из выборки случайно. При исследовании мы можем получить любое из этих значений. Давайте нарисуем гистограмму для распределения выборочного среднего.



Отметим на гистограмме точки Фёдора, Шарика и Матроскина:



Что мы видим? Мы видим, что Шарик и Матроскин своими оценками попали в хвосты распределения. То есть им очень не повезло. Вероятность попасть в хвосты равна $\frac{2}{6} = \frac{1}{3}$. Оценка дяди Фёдора находится в центре распределения, и из-за этого является адекватной.

3 Дядя Фёдор и распределение среднего

Построив распределение для среднего значения роста в Гипотезово, Печкин очень сильно удивился. Оказалось, что это случайная величина. Печкин решил узнать у своего друга по переписке, Роналда Фишера, как правильно делать выводы, когда ты видишь только **часть генеральной совокупности, то есть выборку**.

Фишер объяснил Печкину, что \bar{x} при большом числе наблюдений, имеет нормальное распределение. Когда мы хотим сделать выводы о среднем, нам нужно работать сразу со всем распределением.

- а) Найдите математическое ожидание и дисперсию случайной величины \bar{x}
- б) Кто такой Роналд Фишер? Хороших ли друзей заводит себе Печкин?

Решение:

Будем предполагать, что наша выборка пришла к нам из какой-то генеральной совокупности с математическим ожиданием μ и дисперсией σ^2 . Будем предполагать, что все наблюдения независимы друг от друга²

$$X_1, X_2, \dots, X_n \sim \text{iid} (\mu, \sigma^2).$$

Обратите внимание, что про конкретный вид распределения никаких предположений не сделано. По этой выборке мы посчитали среднее. Найдём его математическое ожидание. Вспом-

²Из-за конечности нашей выборки это неправда, но об этом мы ещё поговорим.

ним, что константа выносится за знак математического ожидания, а **математическое ожидание суммы всегда разваливается в сумму математических ожиданий**

$$\begin{aligned}\mathbb{E}(\bar{x}) &= \mathbb{E}\left(\frac{X_1 + X_2 + \dots + X_n}{n}\right) = \\ &= \frac{1}{n} \cdot \mathbb{E}(X_1 + X_2 + \dots + X_n) = \frac{1}{n} \cdot (\mathbb{E}(X_1) + \mathbb{E}(X_2) + \dots + \mathbb{E}(X_n)) = \\ &= \frac{1}{n} \cdot (\mu + \mu + \dots + \mu) = \frac{1}{n} \cdot n \cdot \mu = \mu.\end{aligned}$$

При поиске дисперсии мы активно пользуемся тем, что наши **наблюдения независимы**. Только для такой ситуации дисперсия суммы равна сумме дисперсий

$$\begin{aligned}\text{Var}(\bar{x}) &= \text{Var}\left(\frac{X_1 + X_2 + \dots + X_n}{n}\right) = \\ &= \frac{1}{n^2} \cdot \text{Var}(X_1 + X_2 + \dots + X_n) = \frac{1}{n^2} \cdot (\text{Var}(X_1) + \text{Var}(X_2) + \dots + \text{Var}(X_n)) = \\ &= \frac{1}{n^2} \cdot (\sigma^2 + \sigma^2 + \dots + \sigma^2) = \frac{1}{n^2} \cdot n \cdot \sigma^2 = \frac{\sigma^2}{n}.\end{aligned}$$

Получается, что наше среднее имеет асимптотически нормальное распределение с параметрами μ и $\frac{\sigma^2}{n}$:

$$\bar{x} \overset{asy}{\sim} \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right).$$

Слова асимптотически нормальное означают, что при $n \rightarrow \infty$ распределение среднего сходится к нормальному. На самом деле тут мы пользуемся **центральной предельной теоремой**. О ней мы подробнее поговорим в следующих посиделках. Пока что придётся принять асимптотическую нормальность средних на веру.

Отталкиваясь от нормального распределения, мы можем построить для нашего неизвестного параметра μ доверительный интервал. Он будет хорошо работать при очень большом числе наблюдений n , то есть **будет асимптотическим**.

Фишер — очень неплохое знакомство для Печкина. Он считается отцом современной частотной статистики. Именно он и его ученики в течение первой половины XX века придумали аппарат для проверки гипотез и метод максимального правдоподобия.

4 Дядя Фёдор и доверительный интервал

Печкин рассказал дяде Фёдору, Шарик и Матроскину про то, что выводы по точечным оценкам делать некорректно. Нужно понять насколько сильно среднее может измениться, если мы пересоберём выборку. То есть для оценки нужно построить доверительный интервал.

- а) Вспомните правило 3-х сигм для нормального распределения. Отталкиваясь от него постройте доверительный интервал для μ .
- б) Найдите стандартное отклонение для Шарика, Матроскина и Фёдора по формуле

$$\hat{\sigma} = \sqrt{\frac{1}{n-1} \sum (x_i - \bar{x})^2}.$$

- в) Постройте для каждого из парней доверительный интервал по правилу трёх сигм. Обратите внимание, что стандартное отклонение, которое мы посчитали в первом пункте — стандартное отклонение для роста. Нам нужно скорректировать его на число наблюдений, чтобы получить стандартное отклонение для среднего, то есть надо построить интервал

$$\left(\bar{x} - 3 \cdot \frac{\hat{\sigma}}{\sqrt{n}}; \quad \bar{x} + 3 \cdot \frac{\hat{\sigma}}{\sqrt{n}} \right)$$

- г) Лежит ли настоящий средний рост во всех трёх доверительных интервалах? Что это означает? Насколько широкими вышли интервалы?

Решение:

Итак, по ЦПТ $\bar{x} \overset{asy}{\sim} \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$. Правило трёх сигм говорит нам, что

$$\mathbb{P}\left(\mu - 3 \cdot \sqrt{\frac{\sigma^2}{n}} \leq \bar{x} \leq \mu + 3 \cdot \sqrt{\frac{\sigma^2}{n}}\right) \approx 0.997$$

Такой интервал для случайной величины \bar{x} называется **предиктивным интервалом**. Его границы фиксированы, а в центре находится случайная величина. Давайте разрешим неравенство относительно μ . Тогда мы получим, что:

$$\mathbb{P}\left(\bar{x} - 3 \cdot \sqrt{\frac{\sigma^2}{n}} \leq \mu \leq \bar{x} + 3 \cdot \sqrt{\frac{\sigma^2}{n}}\right) \approx 0.997$$

Таким образом μ оказалась по центру. Если мы оценим по нашей выборке σ , тогда мы сможем получить диапазон в границах которого лежит неизвестный параметр μ с вероятностью 0.997:

$$\mathbb{P} \left(\bar{x} - 3 \cdot \sqrt{\frac{\hat{\sigma}^2}{n}} \leq \mu \leq \bar{x} + 3 \cdot \sqrt{\frac{\hat{\sigma}^2}{n}} \right) \approx 0.997.$$

Такой интервал называется **доверительным интервалом**. Его границы — случайные величины, а в середине стоит неизвестная константа, которую доверительный интервал покрывает с вероятностью 0.997.

Если мы возьмём вместо 3 другую засечку, например 1.96, мы получим доверительный интервал с другой вероятностью

$$\mathbb{P} \left(\bar{x} - 1.96 \cdot \sqrt{\frac{\hat{\sigma}^2}{n}} \leq \mu \leq \bar{x} + 1.96 \cdot \sqrt{\frac{\hat{\sigma}^2}{n}} \right) \approx 0.95.$$

Теперь давайте найдём стандартные отклонения для Шарика, Матроскина и Фёдора:

$$\text{Фёдор:} \quad \sqrt{\frac{1}{2-1} \cdot [(190-170)^2 + (150-170)^2]} \approx 28.3$$

$$\text{Шарик:} \quad \sqrt{\frac{1}{2-1} \cdot [(190-185)^2 + (180-185)^2]} = 7.1$$

$$\text{Матроскин:} \quad \sqrt{\frac{1}{2-1} \cdot [(150-155)^2 + (160-155)^2]} = 7.1$$

А затем построим для каждого из них доверительные интервалы:

$$\text{Фёдор:} \quad \left(170 - 3 \cdot \frac{28.3}{\sqrt{2}}; \quad 170 + 3 \cdot \frac{28.3}{\sqrt{2}} \right) = (110; \quad 230)$$

$$\text{Шарик:} \quad \left(185 - 3 \cdot \frac{7.1}{\sqrt{2}}; \quad 185 + 3 \cdot \frac{7.1}{\sqrt{2}} \right) = (169.9; \quad 200)$$

$$\text{Матроскин:} \quad \left(155 - 3 \cdot \frac{7.1}{\sqrt{2}}; \quad 155 + 3 \cdot \frac{7.1}{\sqrt{2}} \right) = (140; \quad 170.1)$$

Все три доверительных интервала покрывают 170. Для всех трёх ситуаций доверительные интервалы получились довольно широкими. Это означает, что по двум наблюдениям оценка среднего получилась очень неточной. Чтобы повысить её точность, нужно собрать ещё наблюдений. Тогда доверительные интервалы станут уже, а наши выводы достовернее.

На практике обычно строят точечную оценку, понимают какое у этой точечной оценки распределение, а после на основе распределения прикидывают насколько оценка получилась точной с помощью доверительного интервала.

5 Мир в Простоквашино

Печкин приехал на велосипеде из Гипотезово в Простоквашино и принёс его жителям новое знание. Матроскин, Шарик и дядя Фёдор были поражены этим знанием. Все склоки и ссоры закончились. Жители Простоквашино помирились. Прошла неделя. Как-то вечером ребята пили чай да призадумались: а можно ли по собранным наблюдениям как-то проверить гипотезу о том, что граждане Гипотезово не очень высокие. Например, что их рост, в среднем, равен 160.

Посреди ночи все собрались на кухне у Печкина. Мудрый почтальон набросал на бумаге следующие мысли:

- 1) Мы знаем, что \bar{x} — случайная величина, которая имеет нормальное распределение.
- 2) Значит расстояние $\bar{x} - 160$ — это тоже случайная величина с нормальным распределением.
- 3) Если наша гипотеза верна, $\bar{x} - 160$ должно быть близко к нулю. Это случайная величина. Её распределение должно концентрироваться около нуля.
- 4) Значит мы можем построить для расстояния $\bar{x} - 160$ доверительный интервал. Если окажется, что ноль оказался внутри доверительного интервала, мы не можем отвергнуть гипотезу. Если он оказалось за пределами интервала, мы отвергаем гипотезу.
- 5) При этом, если мы будем пользоваться правилом 3-х сигм, при отвержении гипотезы, мы ошибёмся с вероятностью 1%, так как наш доверительный интервал будет покрывать истинное значение с вероятностью 99% (есть ещё другая ошибка, **ошибка 2 рода**, зря согласиться с гипотезой).

Проверьте гипотезу о том, что μ , так обычно обозначают то среднее значение, которое задумала природа, равно 160. Будем использовать выборку дяди Фёдора. Используя её же, проверим гипотезу о том, что $\mu = 100$.

Решение:

Гипотеза $H_0 : \mu = 160$. Альтернативная гипотеза: $H_a : \mu \neq 160$. Оценкой для μ будет $\bar{x} = 170$. Наблюдаемое расстояние составит $\bar{x} - \mu = 170 - 160 = 10$.

Стандартное отклонение для среднего мы уже искали. Оно оказалось равно $\frac{28.3}{\sqrt{2}} \approx 20$. Доверительный интервал составит $(10 - 3 \cdot 20; 10 + 3 \cdot 20) = (-50; 70)$. Ноль входит в этот интервал. Гипотеза о том, что $\mu = 160$ не отвергается. Гипотеза не противоречит нашим данным.

Пришёл черёд второй гипотезы, $H_0 : \mu = 100$. Альтернативная гипотеза $H_a : \mu \neq 100$. Наблюдаемое расстояние составит $\bar{x} - \mu = 170 - 100 = 70$. Доверительный интервал для него составит $(-70 - 3 \cdot 20; -70 + 3 \cdot 20) = (-130; -10)$. Ноль не входит в этот интервал. Значит гипотеза о том, что $\mu = 100$ отвергается. Эта гипотеза противоречит собранным данным.

Обратите внимание, что если мы захотим протестировать гипотезу $H_0 : \mu = 150$ или $H_0 : \mu = 165$,

они тоже не будут отвергаться, так как тест каждой гипотезы делается против конкретной альтернативы. Если мы хотим в ходе тестирования получать не такие размытые результаты, нам нужно собрать больше наблюдений, тогда доверительные интервалы станут уже, мы сможем улавливать более мелкие изменения, и наши выводы будут точнее.

Мы, в задачке выше смотрели войдёт ли ноль в доверительный интервал для расстояния между предполагаемым нами μ_0 и просчитанным по выборке \bar{x} . То есть строили

$$\left((\bar{x} - \mu_0) - 3 \cdot \frac{\hat{\sigma}}{\sqrt{n}}; (\bar{x} - \mu_0) + 3 \cdot \frac{\hat{\sigma}}{\sqrt{n}} \right)$$

На практике часто считают вот такую штуку:

$$z = \frac{\bar{x} - \mu_0}{\frac{\hat{\sigma}}{\sqrt{n}}}.$$

Её обычно называют **z-статистикой**. При большом n и отсутствии выбросов она имеет нормальное распределение. Эту статистику можно рассчитать по данным, а после сравнить с 3 и -3 . Если наблюдаемое значение попало между ними, гипотеза не отвергается. То есть мы делаем всё ровно то же самое, что с доверительным интервалом, просто немного переписали.

Если мы постоянно будем искать разные средние и проверять гипотезы про них, при сравнении z-статистики с 3, мы будем зря отвергать гипотезу H_0 в 0.3% случаев. Если мы согласны совершать такую ошибку чаще, например в 5% случаев, мы можем воспользоваться засечкой 1.96 и строить 95% доверительный интервал. Другими словами, для этого надо посчитать z-статистику и сравнить её с 1.96.

Выбирая разные цифры для сравнения (их ещё называют **критическими значениями**), мы будем допускать разные ошибки. Обойтись без ошибок не получится, так как мы всегда будем работать с какой-то конечной выборкой.