ПРИКЛАДНАЯ СТАТИСТИКА

Деревни Гипотезово да Простоквашино

«Если выпадет решка, то отпустим конечно. И в небо летит монетка с одинаковыми сторонами.»

Олег о неудачном дизайне АБ-теста

Упражнение 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

Дедя Фёдор, Шарик и Матроскин проезжают через Гипотезово в Простоквашино транзитом. Каждый из них заинтересовался ростом местных жителей и решил по небольшой подвыборке из двух человек посчитать средний рост всех жителей Гипотезово.

а) Посчитайте настоящий средний рост в Гипотезово по всей генеральной совокупности.

б) Шарик посчитал средние по Саше и Даше и сказал, что это оценка среднего роста в Гипотезово. Сколько у него получилось? Насколько сильно эта оценка отличается от настоящего среднего?

в) К Матроскину в выборку затесались 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 ребята:

$$\begin{aligned} &\text{Фёдор:} && \bar{x} = 0.5 \cdot (190 + 150) = 170 \\ &\text{Шарик:} && \bar{x} = 0.5 \cdot (190 + 180) = 185 \\ &\text{Матроскин:} && \bar{x} = 0.5 \cdot (150 + 160) = 155 \end{aligned}$$

Средне значения, расчитанное по выборке, мы используем как оценку для неизвестного μ.

Видим, что значения у выборочных средних получились очень разными. При этом расстояние от среднего Шарика до настоящего равно 15, от среднего Матроскина тоже 15. От среднего Фёдора до настоящего нулевое.

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 с распределением, которое мы построим в следующей задачке. 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.** Из-за этого нам нужно придумать какой-то способ работать с такими характеристиками, чтобы несмотря на случайность выборки, почти не ошибаться.

Упражнение 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 опросил всех четверых жителей, а после стал фантазировать что могло бы получиться в качестве

среднего, если бы он опросил только двух каких-то жителей.

а) Является ли средний рост случайной величиной? Сколько значений принимает эта случайная величина (сколько вариантов опросить местных жителей есть у Печкина)?

б) Найдите все возможные значения среднего роста в Гипотезово. Постройте гистограмму для этого среднего значения. Как и в прошлый раз, 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?

в) Какова вероятность оказаться в хвостах распределения? Какова вероятность оказаться в его центре?

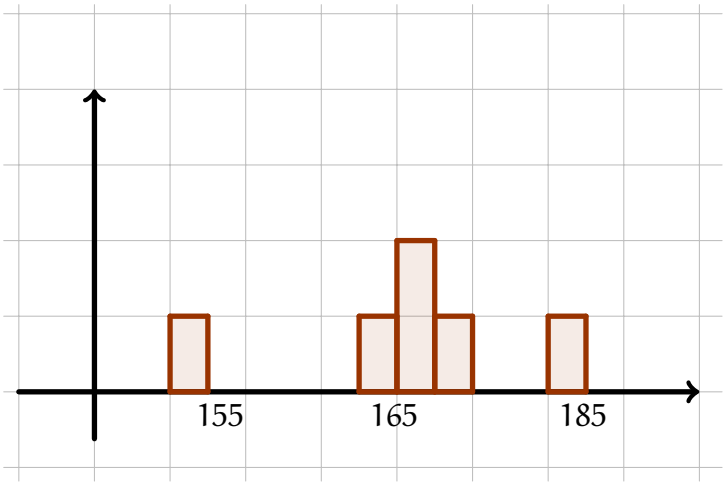
Решение:

Да. Средний рост — случайная величина. Вспомним из комбинаторики сочетания и посчитаем, сколько разных значений она может принимать. Это число сочетаний из 4 по 2: $C_4^2 = \frac{4!}{2!2!} = 6$.

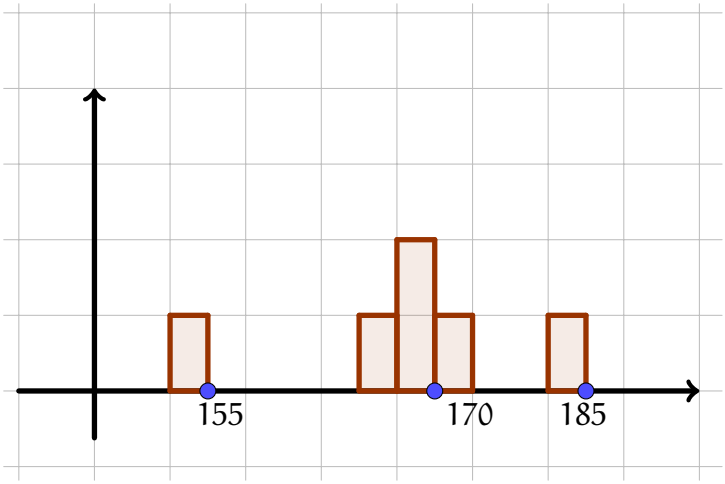
Давайте выпишем все значения, которые может принять средний рост:

$$\begin{aligned}
&\text{Маша и Паша} \quad 0.5 \cdot (150 + 160) = 155 \\
&\text{Маша и Саша} \quad 0.5 \cdot (150 + 190) = 170 \\
&\text{Маша и Даша} \quad 0.5 \cdot (150 + 180) = 165 \\
&\text{Паша и Саша} \quad 0.5 \cdot (160 + 190) = 175 \\
&\text{Паша и Даша} \quad 0.5 \cdot (160 + 180) = 170 \\
&\text{Саша и Даша} \quad 0.5 \cdot (190 + 180) = 185
\end{aligned}$$

Каждое из этих значений — случайная величина принимает равновероятно, так как мы берём двух человек из генеральной совокупности абсолютно случайно. Давайте нарисуем гистограмму для этого распределений.



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.

Упражнение 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, вошедших в него, 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:

Пусть $X_1, X_2, \ldots, X_n \sim iid\ (\mu, \sigma^2)$. Распределение мы не знаем, но знаем дисперсию и математическое ожидание. По этой выборке мы посчитали среднее. Найдём его математическое ожидание:

$$\mathbb{E}(\bar{x}) = \mathbb{E}\left(\frac{X_1 + X_2 + \ldots + X_n}{n}\right) =$$
$$= \frac{1}{n} \cdot \mathbb{E}(X_1 + X_2 + \ldots + X_n) = \frac{1}{n} \cdot (\mathbb{E}(X_1) + \mathbb{E}(X_2) + \ldots + \mathbb{E}(X_n)) =$$
$$= \frac{1}{n} \cdot (\mu + \mu + \ldots + \mu) = \frac{1}{n} \cdot n \cdot \mu = \mu.$$

Теперь дисперсия:

$$\mathrm{Var}(\bar{x}) = \mathrm{Var}\left(\frac{X_1 + X_2 + ... + X_n}{n}\right) =$$
$$= \frac{1}{n^2} \cdot \mathrm{Var}(X_1 + X_2 + ... + X_n) = \frac{1}{n^2} \cdot (\mathrm{Var}(X_1) + \mathrm{Var}(X_2) + ... + \mathrm{Var}(X_n)) =$$
$$= \frac{1}{n^2} \cdot (\sigma^2 + \sigma^2 + ... + \sigma^2) = \frac{1}{n^2} \cdot n \cdot \sigma^2 = \frac{\sigma^2}{n}.$$

При поиске дисперсии мы активно пользуемся тем, что наши **наблюдения независимы.** Только для такой ситуации дисперсия суммы равна сумме дисперсий. В свою очередь, математическое ожидание суммы равно сумме математических ожиданий всегда.

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

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

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 20 века придумали аппарат для проверки гипотез и метод максимального правдоподобия.

Упражнение 4 (дядя Фёдор и 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 озна-

чает? Насколько широкими вышли интервалы?

Решение:

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

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

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:

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

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:

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

Такой интервал называется **доверительным интервалом.** Его границы — случайные величины, а в середине стоит неизвестная константа, которую доверительный интервал накрывает с вероятностью 0.997. Дальше мы будем подробнее говорить про доверительные интервалы.

Если мы возьмём не 3 для нормального распределения, а другую засечку, например 1.96, мы получи доверительный интервал с другой вероятностью

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

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

$$\begin{aligned}
&\text{Фёдор:} && \sqrt{\frac{1}{2-1} \cdot \left[(190-170)^2 + (150-170)^2\right]} \approx 28.3 \\
&\text{Шарик:} && \sqrt{\frac{1}{2-1} \cdot \left[(190-185)^2 + (180-185)^2\right]} = 7.1 \\
&\text{Матроскин:} && \sqrt{\frac{1}{2-1} \cdot \left[(150-155)^2 + (160-155)^2\right]} = 7.1
\end{aligned}$$

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 накрывают 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 так делают очень часто: строят точечную оценку по какой-то выборке, понимают какое у этой точечной оценки распределение, а после на основе распределения прикидывают насколько оценка получилась точной с помощью доверительного интервала.

Упражнение 5 (в котором в Простоквашино наступает мир)

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 средний рост в Гипотезово 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 накрывать 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 $(10 - 3 \cdot 20; \quad 10 + 3 \cdot 20) = (-50; \quad 70)$. 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 $(-70 - 3 \cdot 20; \quad -70 + 3 \cdot 20) = (-130; \quad -10)$. 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 $H_0 = 150$ или $H_0 = 165$, 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

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

На практике часто считают вот такую штуку:

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

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. **Мы с вами поверхностно познакомились с проверкой гипотез. На будущих неделях мы обсудим то, как это делается в мельчайших деталях.**