Семинар 9: Введение в АБ-тесты

Задача 1 (которая сеет в наших головах раздор и сомнение)

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

Маша	140
Паша	150
Саша	200
Даша	190

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

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

б) Шарик посчитал средние по Саше и Даше и сказал, что это оценка среднего роста в АБтестово. 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:

Давайте вспоминать семинар по матстату, который мы с вами не так давно решали. Подсчитаем средний рост в АБтестово по всей генеральной совокупности:

$$\frac{1}{4} \cdot (190 + 200 + 150 + 140) = 170.$$

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

$$\begin{aligned} &\text{Фёдор:} && 0.5 \cdot (200 + 140) = 170 \\ &\text{Шарик:} && 0.5 \cdot (200 + 190) = 195 \\ &\text{Матроскин:} && 0.5 \cdot (140 + 150) = 145 \end{aligned}$$

Видим, что значения у выборочных средних получились очень разными. При этом расстояние от среднего Шарика до настоящего равно 25, от среднего Матроскина 35. От среднего Фёдора до настоящего нулевое.

Каждый житель Простоквашино посчитал средний рост по двум жителям АБтестово. 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 следущей 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), 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.

Задача 2 (в которой происходит 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 АБтестово. Там он опросил всех четверых жителей села, 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 АБтестово. 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?

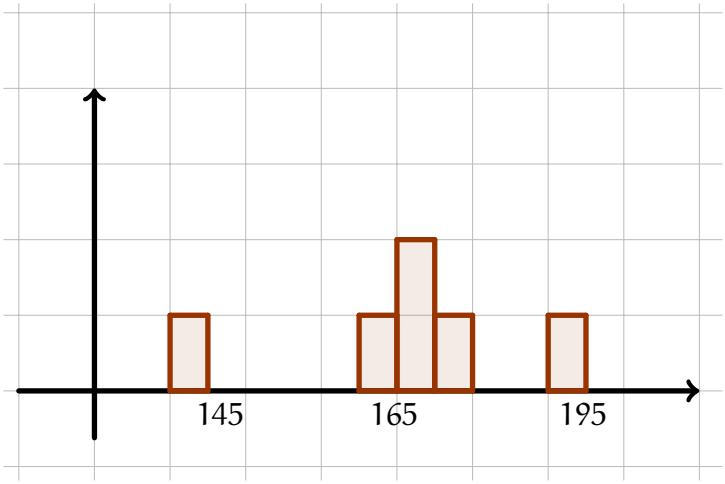
Решение:

Да. Средний рост — случайная величина. Если вы немного помните со школы комбинаторику, вы можете посчитать сколько значений она принимает. Это число сочетаний из 4 по 2: $C_4^2 = \frac{4!}{2!2!} = 6$. Если не помните, вы можете в явном виде перечислить все пары жителей. Но комбинаторику я бы на вашем месте повторил. Это полезно — уметь считать количество разных комбинаций.

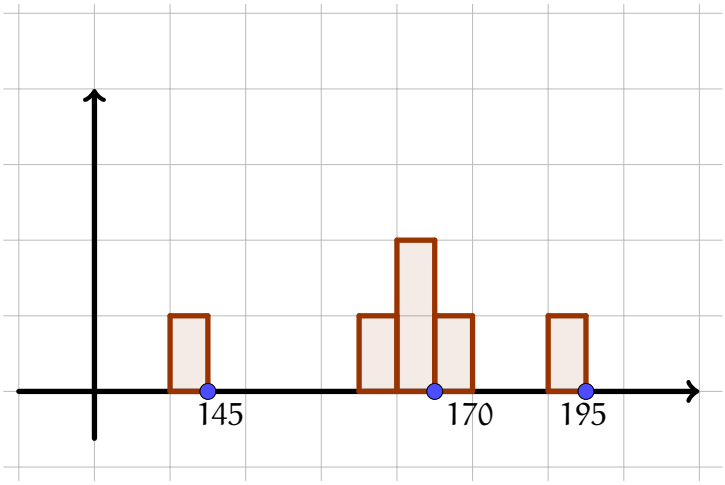
Давайте выпишем все значения, которые может принять средний рост:

$$
\begin{array}{ll}
\text{Маша и Паша} & 0.5 \cdot (140 + 150) = 145 \\
\text{Маша и Саша} & 0.5 \cdot (140 + 200) = 170 \\
\text{Маша и Даша} & 0.5 \cdot (140 + 190) = 165 \\
\text{Паша и Саша} & 0.5 \cdot (150 + 200) = 175 \\
\text{Паша и Даша} & 0.5 \cdot (150 + 190) = 170 \\
\text{Саша и Даша} & 0.5 \cdot (200 + 190) = 195
\end{array}
$$

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.

Задача 3 (в которой Печкин находит решение проблемы)

Построив распределение для среднего значения роста в АБтестово Печкин очень силь но 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.

Фишер объяснил Печкину, что $\bar{x}$ имеет нормальное распределение. Когда мы хотим сделать выводы о среднем, нам нужно работать сразу со всем распределением. Например, с помощью правила двух сигм для него можно построить доверительный интервал, то есть интервал, в котором с вероятностью 95% лежит истиное значение среднего.

а) Найдите стандартное отклонение для Шарика, Матроскина и Фёдора по формуле

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

б) Постройте для каждого из парней доверительный интервал по правилу двух 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

$$\left(\bar{x} - 2 \cdot \frac{\hat{\sigma}}{\sqrt{n}};\quad \bar{x} + 2 \cdot \frac{\hat{\sigma}}{\sqrt{n}}\right)$$

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

г) Кто такой Роналд Фишер? Хороших ли друзей заводит себе Печкин?

Найдём стандартные отклонения:

$$\text{Фёдор:}\quad \sqrt{\frac{1}{2-1}\cdot\left[(200-170)^2+(140-170)^2\right]} \approx 42$$

$$\text{Шарик:}\quad \sqrt{\frac{1}{2-1}\cdot\left[(200-195)^2+(190-195)^2\right]} = 25$$

$$\text{Матроскин:}\quad \sqrt{\frac{1}{2-1}\cdot\left[(140-145)^2+(150-145)^2\right]} = 25$$

Пришло время построить доверительные интервалы:

$$\text{Фёдор:}\quad \left(170 - 2\cdot\frac{42}{\sqrt{2}};\quad 170 + 2\cdot\frac{42}{\sqrt{2}}\right) = (110.6;\quad 229.4)$$

$$\text{Шарик:}\quad \left(195 - 2\cdot\frac{25}{\sqrt{2}};\quad 195 + 2\cdot\frac{25}{\sqrt{2}}\right) = (159.64;\quad 230.35)$$

$$\text{Матроскин:}\quad \left(145 - 2\cdot\frac{25}{\sqrt{2}};\quad 145 + 2\cdot\frac{25}{\sqrt{2}}\right) = (109.64;\quad 180.35)$$

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.

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.

Задача 4 (в которой в Простоквашино наступает мир)

Печкин приехал на велосипеде из АБтестово в Простоквашино и принёс его жителям новое знание. Матроскин, Шарик и дядя Фёдор были поражены этим знанием. Все склоки и соры 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 АБтестово равен 160?

Посреди ночи простоквашинская братва завалилась к Печкину и стала мучать его вопросами. Мудрый почтальон набросал следущие мысли:

1. Мы знаем, что $\bar{x}$ — случайная величина, которая имеет нормальное распределение.
2. Значит расстояние $\bar{x} - 160$ — это тоже случайная величина.
3. Если наша гипотеза верна, $\bar{x} - 160 = 0$ и распределение концентрируется вокруг нуля.
4. Значит мы можем построить для расстояния $\bar{x} - 160$ доверительный интервал. Если окажется, что наблюдаемое нами расстояние оказалось внутри доверительного интервала, мы не можем отвергнуть гипотезу. Если оно оказалось за пределами интервала, мы отвергаем гипотезу.
5. При этом, если мы будем пользоваться правилом 3-х сигм, мы ошибёмся с вероятностью 5%, так как наш доверительный интервал будет накрывать истиное значение с вероятностью 95%.

Проверьте гипотезу о том, что $\mu = 160$ по этому алгоритму, используя выборку дяди Фёдора. Используя её же проверьте гипотезу о том, что $\mu = 120$. В данном случае буквой μ мы обозначили настоящее среднее.

Решение:

Гипотеза $H_0 : \mu = 160$. Альтернативная гипотеза: $H_1 : \mu \neq 160$. Оценкой для μ будет $\bar{x} = 160$. Наблюдаемое расстояние составит $\bar{x} - \mu = 170 - 160 = 10$.

Стандартное отклонение для среднего мы уже искали. Оно оказалось равно $\frac{42}{\sqrt{2}} \approx 29.7$. Доверительный интервал составит $(10 - 2 \cdot 29.7; \quad 10 + 2 \cdot 29.7) = (-50; \quad 70)$. Ноль входит в этот интервал. Гипотеза о том, что $\mu = 170$ не отвергается.

Пришёл черёд второй гипотезы, $H_0 : \mu = 100$. Альтернативная гипотеза $H_1 : \mu \neq 100$. Наблюдаемое расстояние составит $\bar{x} - \mu = 170 - 100 = -70$. Доверительный интервал для него

составит $(-70-2\cdot29.7;\quad -70+2\cdot29.7) = (-130;\quad -10)$. Ноль не входит в этот интервал. Значит гипотез о том, что $\mu = 100$ отвергается.

Обратите внимание, что если мы захотим протестировать гипотезу $H_0 = 160$ 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, а наши выводы точнее.

Задача 5 (в которой дядя Фёдор помогает людям)

Дядя Фёдор настолько был в восторге от проведённого исследования, что написал статью об этом на habr.ru. Теперь ему пишут со всех концов мира. Например, вчера дяде Фёдору пришло три письма:

- Аристарх, Пантелей и Иван исследуют рост людей. Они сделали три выборки. Аристарх занимается баскетболом, поэтому он опросил своих друзей. Пантелей измеряет рост людей у остановки, где люди ждут автобус. Иван залезает в дома к молодым девушкам и измеряет их рост, пока они спят. Что такое репрезентативность выборки? Чья выборка будет репрезентативной? Почему?
- Хипстер Сергей пишет, что он опросил в Москве и Питере по 100 человек. Каждому он задавал вопрос: ”Кофе любишь?”В Москве ”Да”сказали 50 человек, в Питере 55 человек. Можно ли исходя из этого сделать вывод, что в Питере кофе любят больше? Как правильно узнать, где кофе любят больше?
- Знахарка Акулина пишет, что смешала в тазике ”доктор Мом”с соком редьки. Этот настой она дала простудившейся внучке. Внучка выздоровела. Означает ли это, что лекарство работает? Как правильно проверить работоспособность лекарства?

Помогите дяде Фёдору ответить на эти вопросы.

Решение:

- Аристарх — балбес! Он собирает данные по людям из баскетбольной команды, в которую специально отбирают высоких людей. Все его оценки среднего роста окажутся смещёнными. Его выборка нерепрезентативна. Аналогичные проблемы у Ивана. Он собирает данные только по молодым девушкам, а после собирается говорить о среднем росте всех людей. Его выборка тоже смещена. Самая адекватная стратегия у Пантелея. В людях, которые проходят мимо метро есть элемент случайности и, скорее всего, выборка будет репрезентативной. Но это неточно. Возможно, у района, где собираются данные есть какие-то скрытые особенности, которые приведут к некорректным результатам.
- Конечно же нельзя. Мы с вами поняли выше, что каждая выборочная характеристика — случайная величина. В данном случае отклонение в 5 человек может быть случайным, а разница в долях любящих кофе незначимой. Для того, чтобы правильно узнать где кофе любят больше, нужно проверить гипотезу о том, что $p_m = p_{spb}$ с помощью техники похо-

жей на то, что мы делали с вами выше.

- У Акулины есть только одно наблюдение. Чтобы по-нормульному выяснить работает ли лекарство, нужно взять две группы больных одной и той же болезнью внучек и поделить их на две части. Одной части давать плацебо, второй знахарское средство. После нужно проверить гипотезу о том, что самочувствие тех, кто принимал лекарству лучше, чем тех, кто не принимал.

Ещё задачи

В этом разделе находится ещё пара задач, которые можно порешать руками. Возможно, что-то похожее будет на самостоятельной работе.

Задача 6 (про экзамены)

Ежегодно более 200000 людей по всему миру сдают стандартизированный экзамен GMAT при поступлении на программы MBA. Средний результат составляет 525 баллов, стандартное отклонение — 100 баллов.

Сто студентов закончили специальные подготовительные курсы и сдали экзамен. Средний полученный ими балл — 541.4. Проверьте гипотезу о неэффективности программы.

Решение:

Ну, поехали. Гипотеза $H_0 : \mu = 525$, то есть программа не даёт никакого улучшения в знаниях. Альтернативная гипотеза $H_1 : \mu \neq 525$ — улучшения есть. Зафиксируем вероятность ошибки первого рода на уровне 5% и построим 95% доверительный интервал для разности $541.4 - 525$ по формуле:

$$\bar{x} - \mu \pm 2 \cdot \frac{\hat{\sigma}}{\sqrt{n}}.$$

Получим, что разность с вероятностью 0.95 лежит в промежутке $(-3.6; 36.4)$. Доверительный интервал накрывает ноль, значит гипотеза о том, что программа плохо работает не отвергается.

Часто ту же процедуру проделывают немного иначе. Вместо того, чтобы выписывать доверительный интервал и смотреть что он накрывает, считают значение t—статистики:

$$\frac{\bar{x} - \mu}{\frac{\hat{\sigma}}{\sqrt{n}}}$$

и сравнивают его с критическим значением. В случае нормального распределения и правила двух сигм, это 2. На самом деле число два взято довольно грубо. Настоящим критическим

значением для 95% доверительного интервала является 1.96.

Давайте найдём наблюдаемое значение t—статистики и сравним его с критическим значением, как взрослые.

$$t_{\text{набл.}} = \frac{541.4 - 525}{100/10} = 1.63$$

Получаем значение, которое меньше 2. Наша t—статистика попала в доверительный интервал и гипотеза о том, что курсы не повлияли на результат не отвергается. Если использовать в качестве критического значения уточнённые 1.96, получаем то же самое.

Задача 7 (проблемы с монеткой)

Олег подбрасывает монетку и орёт: "ОРЁЛ-РЕШКА-ОРЁЛ-РЕШКА!". Ещё он недавно посмотрел фильм Кристофера Нолана "Тёмный рыцарь". Там ему очень понравился Харви Дент. Потому что у него тоже была монетка, которую тот подбрасывал. ЛСП стало интересно: а правильная ли у него монетка. Действительно ли она выпадает орлом с вероятностью $\frac{1}{2}$?

1. Олег подбросил монетку трижды и получил комбинацию: OPP. Найдите долю выпадения орла. Дальше будем обозначать эту долю как $\hat{p}$.
2. На теории вероятностей вы докажете, что стандартное отклонение для доли считается по формуле $\sqrt{\frac{\hat{p}\cdot(1-\hat{p})}{n}}$. Найдите среднее отклонение доли.
3. Можно показать, что $\hat{p}$ имеет нормальное распределение. Постройте для вашей оценки доли 95% доверительный интервал по формулам:

   $$\hat{p} \pm 1.96\sqrt{\frac{\hat{p}\,(1-\hat{p})}{n}}$$

   Найдите его ширину. Лежит ли $\frac{1}{2}$ в этом интервале?
4. Олег подбросил монетку ещё два раза и получил OPPOP. Найдите доверительный интервал для этой ситуации. Найдите его ширину. Стал ли он уже? Почему это произошло?

Решение:

Все вычисления, связанные с этой задачкой удобно сделать в python. Там они есть :)

Доля орлов, $\hat{p} = \frac{1}{3}$. Стандартное отклонение, $\sqrt{\frac{1/3 \cdot 2/3}{3}} \approx 0.27$. Доверительный интервал:

$$\left(\frac{1}{3} - 1/96 \cdot 0.27; \quad \frac{1}{3} + 1/96 \cdot 0.27\right) = (-0.2; \quad 0.87).$$

Он захватывает $\frac{1}{2}$. Это означает, что гипотеза о том, что монетка правильная не отвергается.

Так как доля не может быть меньше нуля, левую часть интервала можно загрубить и считать нулевой. При таком раскладе получаем ширину интервала $0.87 - 0 = 0.87$. Если мы увеличим число наблюдений до 5 и построим такой же интервал, его ширина станет 0.82. Оценка стала точнее.

Задача 8 (про Вальда)

Во время Второй Мировой войны американские военные собрали статистику попаданий пуль в фюзеляж самолёта. По самолётам, вернувшимся из полёта на базу, была составлена карта повреждений среднестатистического самолёта. С этими данными военные обратились к статистику Абрахаму Вальду с вопросом, в каких местах следует увеличить броню самолёта. Что посоветовал Абрахам Вальд и почему? Как это связано с репрезентативностью?

Решение:

Выборка повреждений самолётов в наших руках нерепрезентативна. Самолёты с критическими повреждениями не долетают до аэродрома. Значит нужно укреплять те части самолётов, где пробоин нет. Именно такой совет дал Абрахам Вальд.