ЛЕКЦИЯ 5

Доверительные интервалы и проверка гипотез

В этой лекции мы обсуждаем, как строить интервальные оценки и как на основе этого проверять первые простые гипотезы.

5.1 Интервальные оценки

Метод максимального правдоподобия является одним из способов построения точечных оценок, которые по данной выборке выдают число (по определению, оценка или статистика – любая измеримая функция от выборки). Сама оценка, несмотря на то, что она обладает разными полезными свойствами, не позволяет ничего утверждать о своей устойчивости. Сильно ли изменится оценка, если мы получим ещё раз новую выборку из той же вероятностной модели? С другой стороны, не совсем ясно, близка ли оценка к настоящему значению параметра.

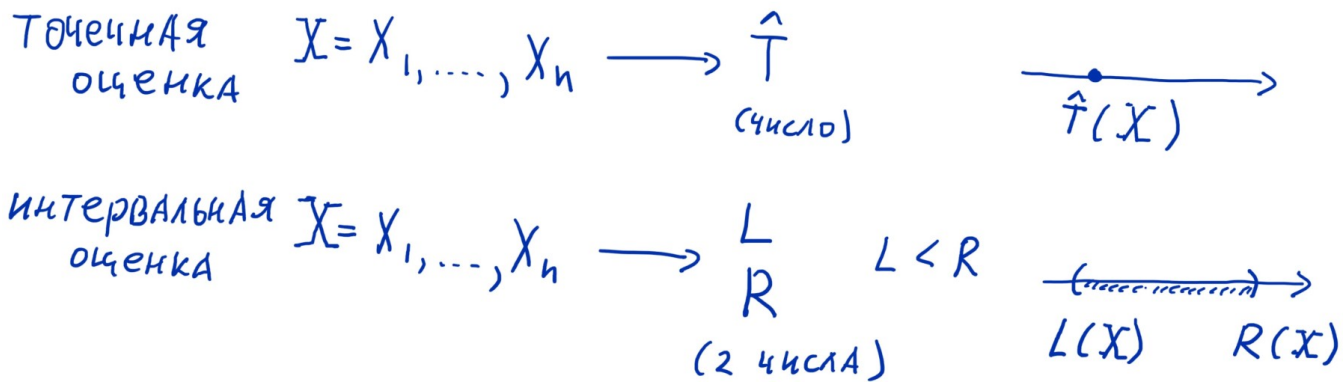
По этой причине нам нужны интервальные оценки параметра. Их фундаментальное различие от обычных оценок состоит в том, что теперь по выборке нужно возвращать интервал, причём обладающий заданным уровнем доверия, то есть, истинное значение параметра должно лежать в интервале с заданной вероятностью.

Определение 5.1. *Пусть $X_1, .., X_n$ – выборка из вероятностной модели, имеющей параметр θ. Доверительным интервалом уровня $1-\alpha \in [0,1]$ называется интервал (L, R) со случайными концами такой, что*

1. *L, R – статистики (функции от выборки) и $L < R$ с вероятностью 1;*

2. *Вероятность*

$$\mathbb{P}\left(L < \theta < R\right) = 1 - \alpha.$$



Для того, чтобы построить интервальную оценку, нужно придумать такие две статистики L, R, которые бы удовлетворяли определению.

Пример 5.1. *Если выборка $X_1, .., X_n$ состоит из независимых и одинаково распределённых по $\mathcal{N}(\mu, \sigma^2)$ случайных величин, то можно заметить, что их выборочное среднее имеет нормальное распределение и ещё*

$$\mathbb{P}\left(z_{\alpha/2} < \frac{\frac{1}{n}\sum_{i=1}^n X_i - \mu}{\sigma/\sqrt{n}} < z_{1-\alpha/2}\right) = 1 - \alpha.$$

Поэтому можно предложить для среднего μ при известной дисперсии интервальную оценку

$$L = \frac{1}{n}\sum_{i=1}^n X_i - \frac{\sigma}{\sqrt{n}} z_{1-\alpha/2}, \;\; R = \frac{1}{n}\sum_{i=1}^n X_i + \frac{\sigma}{\sqrt{n}} z_{1-\alpha/2}.$$

5.2 Асимптотические доверительные интервалы

Для некоторых случайных величин, скажем, для гауссовских или гамма, известны распределения сумм. Но для большинства случаев точной компактной формулы нет. Кроме того, даже в нормальном случае, чтобы воспользоваться идеей выше нужна дополнительно дисперсия, которой изначально не дано. Тут к нам в первую очередь приходит на помощь центральная предельная теорема.

Теорема 5.1. *Пусть $X_1, .., X_n, ...$ – последовательность независимых одинаково распределённых случайных величин с конечной дисперсией σ^2 и матожиданием μ. Тогда*

$$\frac{\frac{1}{n}\sum_{i=1}^n X_i - \mu}{\sigma/\sqrt{n}} \to^d \mathcal{N}(0, 1)$$

А Лемма Слуцкого позволяет получить вариант с выборочной дисперсией внизу:

Теорема 5.2. *Пусть $X_1, .., X_n, ...$ – последовательность независимых одинаково распределённых случайных величин с конечной дисперсией σ^2 и матожиданием μ, а $\hat{\sigma}_n^2$ – выборочная дисперсия, посчитанная по выборке $X_1, .., X_n$. Тогда*

$$\frac{\frac{1}{n}\sum_{i=1}^n X_i - \mu}{\hat{\sigma}/\sqrt{n}} \to^d \mathcal{N}(0, 1).$$

Такая ЦПТ уже позволяет приближённо строить доверительные интервалы; они называются асимптотическими и формально лишь при росте числа наблюдений n сходятся к гауссовским. Тем не менее, скорость сходимости в ЦПТ достаточно хорошо изучена и для многих распределений она наступает достаточно быстро; часто хватает уже сотен и тысяч наблюдений, чтобы получить качественный интервал.

Пример 5.2. *Если выборка $X_1, .., X_n$ состоит из независимых и одинаково распределённых по $\mathcal{N}(\mu, \sigma^2)$ случайных величин, с помощью ЦПТ мы можем получить*

$$\mathbb{P}\left(z_{\alpha/2} < \frac{\frac{1}{n}\sum_{i=1}^{n} X_i - \mu}{\hat{\sigma}/\sqrt{n}} < z_{1-\alpha/2}\right) \approx 1 - \alpha.$$

Поэтому можно предложить для среднего μ даже при неизвестной дисперсии интервальную оценку

$$L = \frac{1}{n}\sum_{i=1}^{n} X_i - \frac{\hat{\sigma}}{\sqrt{n}} z_{1-\alpha/2},\ R = \frac{1}{n}\sum_{i=1}^{n} X_i + \frac{\hat{\sigma}}{\sqrt{n}} z_{1-\alpha/2}.$$

При $n \to \infty$ статистики L, R будут сходиться к матожиданию μ, это следствие Закона Больших Чисел(ЗБЧ).

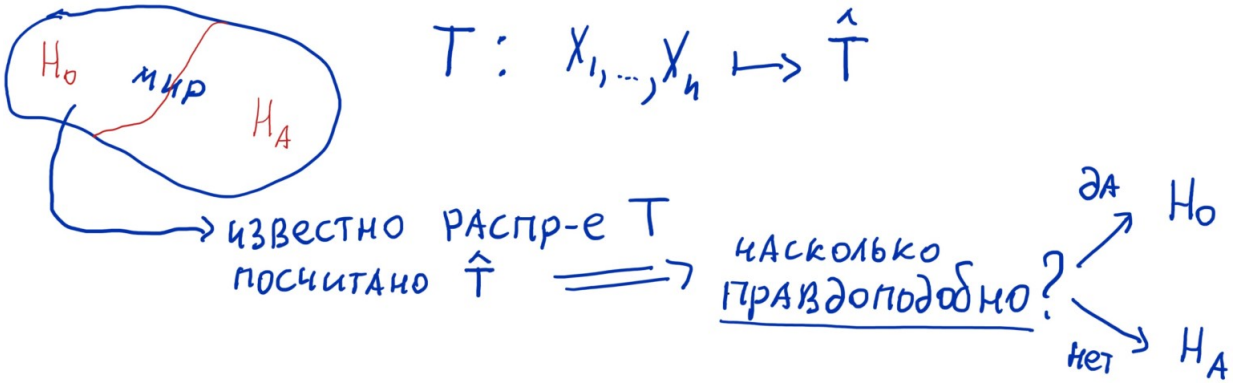
5.3 Проверка гипотез и доверительные интервалы

Проверка гипотез – совершенно отдельная задача, как может показаться на первый взгляд, но на самом деле вся идея статистической проверки гипотез очень похожа на идею доверительных интервалов. С точки зрения математики, аппарат проверки гипотез состоит из нескольких компонент:

1. Два взаимоисключающих утверждения (гипотеза H_0 и альтернатива H_A) о вероятностной модели данных. Либо верна гипотеза H_0, либо альтернатива H_A, третьего не дано;

2. Тестовая статистика (функция от выборки) T, для которой известно распределение при верной гипотезе;

3. Процедура принятия однозначного решения в пользу H_0 или H_A на основе посчитанного по данным значения статистики T и заданного уровня значимости (он же ошибка первого рода) α;



Из математической статистики мы примерно представляем первые два пункта, давайте детальнее рассмотрим третий. Допустим, мы посчитали значение статистики T, мы знаем её распределение при верной гипотезе. Значит, мы можем проверить, насколько полученное значение T вероятно. Если оно попадает в область высокой вероятности, то мы принимаем решение в пользу гипотезы, а иначе – в пользу альтернативы, при этом вероятность ошибки точно контролируется уровнем значимости. Лучше всего это можно проиллюстрировать на примере.

Пример 5.3. *Алексей ходил n дней за кофе и записывал свои наблюдения как $X_1, .., X_n$ в предположении независимости походов за кофе. Каждый день он писал, хороший или плохой кофе ему выпал сегодня. Мерой качества он считает вероятность хорошего кофе p. Алексей хотел бы проверить, удовлетворяет ли кофе его требованиям ($p > 0.7$).*

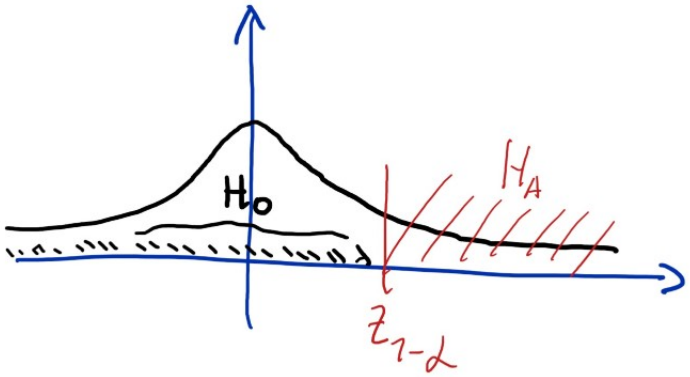
Сначала формируем гипотезу и альтернативу, которые в нашем случае можно записать как

$$H_0 : p = 0.7, H_A : p > 0.7.$$

Если $p < 0.7$, то Алексей уже не отличает, ему одинаково плохо, потому что его как исследователя прежде всего интересует альтернатива. Для таких гипотезы и альтернативы на основе центральной предельной теоремы можем построить z-тест (символом z_α обозначают обычно гауссовские квантили). Заметим, что статистика

$$T = \frac{\frac{1}{n}\sum_{i=1}^{n} X_i - \mu}{\hat{\sigma}/\sqrt{n}}$$

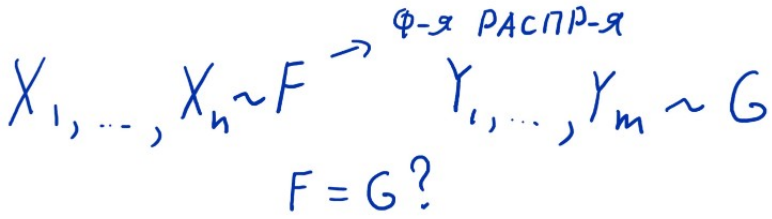
при условии, что μ – настоящее матожидание X, имеет нормальное распределение. Следовательно, если гипотеза верна, то $\mu = 0.7$, остальное можем вычислить. Согласно распределению, T лежит в интервале $(-\infty, z_{1-\alpha})$ примерно (при больших n это точнее) с вероятностью $1 - \alpha$.



Скажем, что мы отвергаем гипотезу, если T выходит за рамки интервала. Мы же можем ошибиться? Если гипотеза всё же верна, то при отвержении мы совершим ошибку, а интервал построен так, что её вероятность как раз равна α, то есть, мы полностью это контролируем.

То, что мы только что проделали, называют ещё (асимптотическим) z-тестом для проверки гипотезы о долях. На самом деле, z-тестом принято называть все критерии, в которых статистика при верной гипотезе имеет нормальное распределение.

5.4 Гипотеза об однородности и больше критериев



Часто на практике нам даётся две выборки и требуется проверить, из одного ли они распределения. Такая задача называется проверкой однородности и для неё есть достаточно развитый аппарат статистических критериев. Мы вернёмся к этому вопросу позже ещё подробнее, пока давайте посмотрим, как эту задачу можно решить уже известными нам средствами.

Итак, пусть даны две несвязанных выборки $X_1, .., X_n$ и $Y_1, .., Y_m$, независимые в совокупности внутри и между собой и каждая из распределений F и G соответственно. Если мы сразу попробовали бы поставить гипотезу и альтернативу, вероятно, они бы выглядели так:

$$H_0 : \forall x \ F(x) = G(x), \ H_A : \exists x \ F(x) \neq G(x).$$

Оказывается, мы можем проверять такие гипотезы, например, с помощью тестов Хи-квадрат (критерий Согласия Пирсона, для дискретных распределений) или Колмогорова-Смирнова (для непрерывных распределений). Однако для такой задачи эти критерии оказываются не очень состоятельными при выборках меньше тысяч наблюдений в силу сложности альтернативы, говорящей, что распределения просто разные.

Вместо этого можно попробовать упростить альтернативу и выиграть в мощности. Кроме того, в практических задачах нас часто интересует не распределение в целом, а, например, сдвиг одного распределения относительно другого или изменение масштаба. Первое направление приводит нас к проверке гипотезы о сдвиге, которую можно формализовать с помощью средних, о второе – к гипотезе о разности масштабов, которая формализуется исследованием дисперсии.

Пример 5.4. *Алексей несколько месяцев чередовал две кофейни КофеТочка и ТочкаКофе, записывая качество кофе в каждой, совершая одно посещение за день. В результате он получил две выборки X_i и Y_i размеров m и n, соответствующие каждой кофейне. Его теперь интересует, как он серьёзно (с-т-а-т-и-с-т-и-ч-е-с-к-и) может проверить, где кофе был лучше, чтобы рекомендовать коллегам.*

Такой вопрос можно сформулировать, например, в виде

$$H_0 : p_1 = p_2, \quad H_A : p_1 \neq p_2,$$

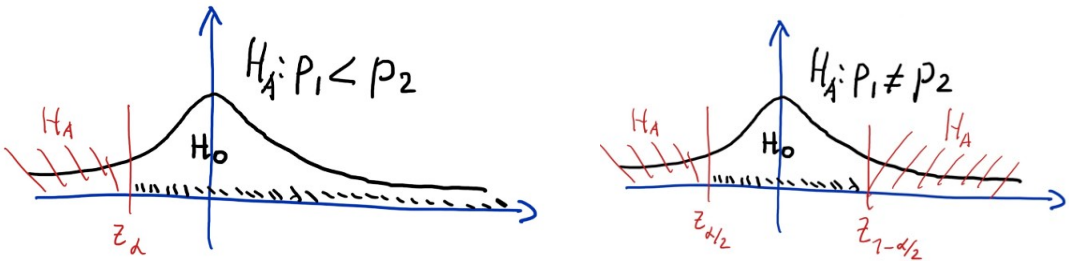
а можно по аналогии с предыдущим примером задать вопрос "Лучше ли кофе в ТочкаКофе?"и записать

$$H_0 : p_1 = p_2, \quad H_A : p_1 < p_2.$$

Статистику можем задать аналогично:

$$T = \frac{\overline{X} - \overline{Y}}{\sqrt{\hat{\sigma}_X^2/n + \hat{\sigma}_Y^2/m}}$$

и она тоже будет асимптотически нормальной. Доверительные области для разных альтернатив будут устроены по разному: для общего неравенства доверительная область критерия – интервал $(z_{\alpha/2}, z_{1-\alpha/2})$, в случае второй альтернативы – полуинтервал $(z_\alpha, +\infty)$. Это можно просто запомнить, если подумать, в пользу какого из утверждений говорят очень маленькие или очень большие значения статистики.



С другой стороны, нас может интересовать разброс в двух группах и это приводит к критериям для проверки масштаба и, в частности, к дисперсионному анализу (Analysis of Variance, ANOVA).

Пример 5.5. *Где кофе более непредсказуемый? Как попробовать ответить на такой вопрос? Можно подумать со стороны дисперсионного анализа. Сформулируем гипотезу*

$$H_0 : \sigma_X^2 = \sigma_Y^2, H_A : \sigma_X^2 > \sigma_Y^2.$$

Для проверки такой гипотезы можно использовать статистику

$$T = \frac{\hat{\sigma}_X^2}{\hat{\sigma}_Y^2}$$

которая при верной гипотезе асимптотически имеет распределение Фишера $F(n-1, m-1)$ *(отсюда название критерия) с* $n-1$ *(по* X*) и* $m-1$ *(по* Y*) степенями свободы. В пользу альтернативы говорят очень большие значения статистики, поэтому критическая область – полуинтервал* $(F_{1-\alpha}, +\infty)$.

Гипотеза об однородности – одна из самых популярных на практике, для её проверки есть и другие критерии. Мы вернёмся к этому более подробно через несколько недель.

5.5 Гипотеза о независимости

Формально, для выборки пар $(X_1, Y_1), .., (X_n, Y_n)$ гипотеза о независимости формулируется как

$$H_0 : \forall x, y \quad F_{XY}(x, y) = F_X(x) F_Y(y),$$

однако против общей альтернативы такую гипотезу в общем случае проверять достаточно сложно. Здесь мы начнём с одного критерия для случайных величин с конечным числом значений, а для непрерывного случая будет предложена другая идея на следующей неделе.

Критерий Пирсона придуман для случая, когда обе выборки принимают значения на конечном множестве. Это могут быть возрастные группы, цвет волос, уровень образования, район проживания... любые признаки, которые принято называть категориальными и которые часто не имеют числовой шкалы.

Пример 5.6. *Среди младших исследователей университета Математики провели опрос. Их спросили, сколько они в среднем за день выпивают чашек кофе и сколько статей было опубликовано за последний год. Результаты приведены в таблице ниже:*

чашки|статьи	*0*	*1*	*2*	*3*
не пьёт	*19*	*21*	*9*	*7*
1	*3*	*9*	*3*	*6*
2	*4*	*10*	*15*	*6*
3	*6*	*16*	*13*	*11*
4	*1*	*2*	*0*	*2*

Верно ли, что кофе помогает публиковать больше статей?

Критерий Пирсона строится на основании следующей идеи: сравниваются эмпирические распределения по одной из переменных для каждого фиксированного значения другой переменной (по сути, условные распределения). Если они *похожи*, то мы склонны думать, что величины независимы. Остаётся понять меру похожести и попробовать на её основе построить статистический критерий.

Обозначим за n_{ij} число наблюдений, попавших в ячейку таблицы (i, j); если бы два признака были бы независимы, то вероятность $p_{ij} = p_{i\cdot}p_{\cdot j}$, где точкой обозначена сумма по соответствующей размерности. Статистика Пирсона задаётся как

$$T = n\sum_{i=1}^{m}\sum_{j=1}^{k}\frac{\left(n_{ij} - \frac{n_{i\cdot}n_{\cdot j}}{n}\right)^2}{n_{i\cdot}n_{\cdot j}}$$

и если присмотреться, то можно увидеть в числителе разность частот. Такая статистика имеет распределение $\chi^2((m-1)(k-1))$. При верной гипотезе статистика скорее принимает малые значения, поэтому критическая область – полуинтервал $(\chi^2_{1-\alpha}, +\infty)$. В примере выше можно вычислить значение статистики (оно равно 577.566), которое далеко в критической области. Посчитанный $p - value = 0.0414$.