Проверка гипотез

Высшая школа цифровой культуры
Университет ИТМО
dc@itmo.ru

Содержание

1 Проверка статистических гипотез

1.1 А что такое гипотеза?

Здравствуйте, уважаемые слушатели. Вот мы и подошли к последней лекции нашего курса – лекции, освещающей задачу проверки гипотез, решаемую методами математической статистикой.

При построении вероятностных моделей случайных экспериментов мы то и дело выдвигаем какие-то предположения о генеральной совокупности: о вероятности того или иного события, о типе распределения, о параметрах этого распределения и так далее. Все эти предположения логично даже с точки зрения русского языка назвать гипотезами. Почему гипотезы? Да потому, что верного ответа на момент наблюдения у нас нет, а вот гипотезы (или предположения), которые хочется подтвердить или опровергнуть – есть.

Например, мы можем измерять объем продаж какого-то товара и хотим проверить гипотезу, заключающуюся в том, что продажи в этом году, в среднем, больше, чем в прошлом году. Просто сравнивать средние на больше-меньше нельзя, так как результат мог быть случайным: могло случайно получиться больше. Так что вопрос глубже, ведь гипотеза не о сравнении средних двух выборок (то есть не о сравнении выборочных средних), а о сравнении математических ожиданий двух генеральных совокупностей, распределение и параметры которых нам неизвестны.

Можно привести и другой пример. Скажем, пусть вопрос «а кто сегодня готовит ужин?» решается случайно, с помощью монетки. Конечно очень важно, чтобы монетка была правильной (то есть давала справедливое решение) и выпадала как орлом, так и решкой с вероятностью, очень близкой к 0.5. Как это проверить? Например, можно бросить монетку 100 раз и посчитать, сколько раз выпал орел, а сколько – решка, и сравнить эти числа. Но неужели мы когда-нибудь получим ровно 50 орлов и ровно 50 решек?

Скорее всего нет. Именно поэтому, наверное, если числа приблизительно равны, то мы склонны сделать вывод в пользу того, что монетка правильная. Но и тут есть подводные камни. Ведь не исключено, что такой результат в серии экспериментов мог получиться и при использовании неправильной монетки (скажем, с вероятностью выпадения орла равной 0.3).

Наоборот, если выпало далекое от половины число гербов, то мы, скорее всего, склонны сказать, что монетка неправильная. Но и тут нет уверенности, что это так: ведь может быть всякое, эксперимент-то случаен. Так что и в такой ситуации монета вполне может оказаться правильной.

Все дело в том, что по выборке конечного объема, как правило, безошибочных выводов о распределении сделать нельзя, поэтому возможность выбрать неверную гипотезу – то, с чем обязательно нужно считаться.

Вообще, говоря на вероятностном языке, последний описанный эксперимент (с монеткой) – это серия испытаний Бернулли с (неизвестной) вероятностью успеха p в каждом испытании. Тогда гипотеза, которую мы проверяем, может быть записана так: $p = 0.5$ (ну или, например, $p \in (0.45, 0.55)$, чтобы позволить какой-никакой логичный брак).

Итак, давайте дадим определение гипотезы. Пусть дана выборка $X = (X_1, X_2, ..., X_n)$ из некоторого распределения $\mathcal{P}$. Если не оговорено противное, то мы будем считать, что все наблюдения (все случайные величины, входящие в выборку), имеют одно и то же распределение. Оказывается, в ряде случаев это предположение нуждается в проверке! То же самое касается и независимости.

Определение 1.1.1 *Гипотезой* H *называется произвольное предположение о распределении генеральной совокупности* ξ.

Если же учесть только что сказанное замечание, гипотезой H *правильнее назвать произвольное предположение о распределении наблюдений.*

Давайте сразу же здесь введем понятия простой и сложной гипотез.

Определение 1.1.2 *Гипотеза* H *называется простой, если она указывает только на одно распределение* $\mathsf{H} = \{\mathcal{P} = \mathcal{P}_1\}$. *Иначе гипотеза* H *называется сложной* $\mathsf{H} = \{\mathsf{P} \in \mathbb{P}\}$, *где* $\mathbb{P}$ *– некоторое подмножество множества всех распределений.*

Часто, если гипотез всего две, то одну из них называют основной, а другую – альтернативой. Тут же важно отметить, что альтернатива, конечно, не пересекается с основной гипотезой, но вовсе нее обязана быть дополнением основной до всего семейства распределений.

Давайте приведем типичные примеры (или типичные постановки задач проверки гипотез), чтобы проиллюстрировать сказанное.

Пример 1.1.1 *Осуществляется выбор из нескольких простых гипотез:* $\mathsf{H}_1 = \{\mathcal{P} = \mathcal{P}_1\}$, $\mathsf{H}_2 = \{\mathcal{P} = \mathcal{P}_2\}$, ..., $\mathsf{H}_k = \{\mathcal{P} = \mathcal{P}_k\}$. *Скажем,*

$$\mathsf{H}_1 = \{\mathcal{P} = \mathsf{B}_{1/2}\}, \ \mathsf{H}_2 = \{\mathcal{P} = \mathsf{U}_{0,1}\}, \ \mathsf{H}_3 = \{\mathcal{P} = \mathsf{N}_{2,4}\}.$$

Еще один классический пример.

Пример 1.1.2 *Имеется простая основная гипотеза и сложная альтернатива:* $\mathsf{H}_1 = \{\mathcal{P} = \mathcal{P}_1\}$ *и* $\mathsf{H}_2 = \{\mathcal{P} \in \mathbb{P}\}$, *где* $\mathbb{P}$ *– какое-то подмножество семейства всех распределений, не содержащее* $\mathcal{P}_1$. *Например,* $\mathsf{H}_1 = \{\mathcal{P} = \mathsf{B}_{0.5}\}$ *и* $\mathsf{H}_2 = \{\mathcal{P} = \mathsf{B}_p, \ p < 0.5\}$.

Могут, конечно, как и основная гипотеза, так и альтернатива быть сложными, например

Пример 1.1.3 $\mathsf{H}_1 = \{\mathcal{P} = \mathsf{N}_{a,\sigma^2},\ a \in \mathbb{R}, \sigma > 0\}$ *и* $\mathsf{H}_2 = \{$*гипотеза* H_1 *неверна*$\}$.

Часто встречается и так называемая гипотеза однородности.

Пример 1.1.4 *Пусть дана выборка* $X_{11}, X_{12}, ..., X_{1n_1}$ *из распределения* $\mathcal{P}_1$, $X_{21}, X_{22}, ..., X_{2n_2}$ *из распределения* $\mathcal{P}_2$ *и так далее* $X_{k1}, X_{k2}, ..., X_{kn_k}$ *из распределения* $\mathcal{P}_k$. *Проверяется сложная гипотеза* $\mathsf{H}_1 = \{\mathcal{P}_1 = \mathcal{P}_2 = ... = \mathcal{P}_k\}$ *против сложной альтернативы* $\mathsf{H}_2 = \{$*гипотеза* H_1 *неверна*$\}$.

В последнем примере ставится важная задача – задача проверки того, берутся ли выборки из одного и того же распределения, или нет. Эта задача очень важна в реальной жизни, ведь очень важно понимать, какое именно распределение мы изучаем по конкретной выборке, чтобы не получилось казусов вроде: по случайно взятой выборке из роста детей в детском саду спрогнозировать размер одежды игроков баскетбольной команды. Ну и так далее.

Хорошо, а как же понять, какая гипотеза лучше? Конечно, с помощью выборки и так называемого критерия.

1.2 Критерий и его ошибки

Определение 1.2.1 *Пусть имеются гипотезы* H_1, H_2, ..., H_k *и выборка* $X = (X_1, ..., X_n)$. *Тогда критерий* $\delta = \delta(X_1, X_2, ..., X_n)$ *– это отображение*

$$\delta : \mathbb{R}^n \to \{\mathsf{H}_1, \mathsf{H}_2, ..., \mathsf{H}_n\}.$$

Итак, критерий – это функция от выборки, возвращающая одну из возможных гипотез. Бывают еще и так называемые рандомизированные критерии, которые принимают каждую гипотезу с некоторой вероятностью, но о таких критериях мы говорить не будем.

Конечно, как мы уже обсуждали ранее, критерий не всегда выдает верную гипотезу, потому что по выборке, по конечному числу наблюдений значений генеральной совокупности, обычно не удается получить о ней полной информации. Именно поэтому возникают так называемые ошибки $1, 2$ и так далее родов. Мы, как правило, будем рассматривать две гипотезы: основную и альтернативу, поэтому определение несколько упростим.

Определение 1.2.2 *Говорят, что произошла ошибка* 1 *рода, если гипотеза* H_1 *отвергнута критерием, в то время как она верна. Говорят, что произошла ошибка* 2 *рода, если гипотеза* H_2 *отвергнута критерием, в то время как она верна.*

Естественно, наряду с ошибками критерия, возникают и вероятности ошибок 1 и 2 родов. Давайте их определим. Итак, вероятность ошибки первого рода – это

$$\alpha_1 = \mathsf{P}_{\mathsf{H}_1}(\delta(X) \neq \mathsf{H}_1) = \mathsf{P}_{\mathsf{H}_1}(\delta(X) = \mathsf{H}_2),$$

а вероятность ошибки второго рода – это

$$\alpha_2 = \mathsf{P}_{\mathsf{H}_2}(\delta(X) \neq \mathsf{H}_2) = \mathsf{P}_{\mathsf{H}_2}(\delta(X) = \mathsf{H}_1).$$

Отметим, что говоря, например, что гипотеза H_1 верна, и вычисляя $\mathsf{P}_{\mathsf{H}_1}(\cdot)$, мы предполагаем, что распределение выборки именно такое, как и предполагает гипотеза H_1, и вычисляем вероятность в соответствии с этим распределением. Аналогично с H_2. Давайте приведем такой простой пример.

Пример 1.2.1 *Пусть наудачу взятое изделие некоторого производства оказывается бракованным с вероятностью p. Контроль продукции тоже допускает ошибки: он бракует годное изделие с вероятностью γ и пропускает бракованное с вероятностью ε.*

Для наудачу взятого изделия логично ввести две гипотезы: $\mathsf{H}_1 = \{$изделие годное$\}$ и $\mathsf{H}_2 = \{$изделие бракованное$\}$. Давайте найдем вероятности ошибок первого и второго родов.

$$\alpha_1 = \mathsf{P}_{\mathsf{H}_1}(\delta = \mathsf{H}_2) = \mathsf{P}_{\textit{изделие годное}}(\textit{контроль забраковал изделие}) = \gamma,$$

$$\alpha_2 = \mathsf{P}_{\mathsf{H}_2}(\delta = \mathsf{H}_1) = \mathsf{P}_{\textit{изделие бракованное}}(\textit{контроль пропустил изделие}) = \varepsilon.$$

Итак, на данный момент неплохо бы сделать следующие выводы. Во-первых, статистический критерий не говорит точно: верна или нет проверяемая гипотеза. Он лишь решает, противоречат или не противоречат выборочные данные выдвинутой гипотезе. Именно поэтому, разумнее говорить «отвергаем» и «не отвергаем» гипотезу, вместо «отвергаем» и «принимаем» (хотя последняя терминология тоже очень популярна). Во-вторых, если есть одна основная гипотеза, а остальное – это нежелательные от нее отклонения, то вывод «данные противоречат гипотезе» куда весомее, чем вывод «данные не противоречат гипотезе». Ну и в-третьих, нам никогда неизвестно, какая гипотеза реально верна, а потому нужно считаться с вероятностями ошибок критерия. Про последнее поговорим совсем чуть-чуть подробнее.

1.3 Уровень значимости и мощность

Давайте остановимся подробно только на случае, когда рассматривается две простые гипотезы о распределении наблюдений:

$$\mathsf{H}_1 = \{\mathcal{P} = \mathcal{P}_1\}, \quad \mathsf{H}_2 = \{\mathcal{P} = \mathcal{P}_2\}.$$

Любой критерий δ в этом случае на выборке $X = (X_1, X_2, ..., X_n)$ принимает не более двух значений (ведь, напомним, критерий δ – это функция из $\mathbb{R}^n$ в $\{\mathsf{H}_1, \mathsf{H}_2\}$). Ну а тогда можно считать, что все $\mathbb{R}^n$ распадается на две части S и $\mathbb{R}^n \setminus S$, то есть

$$\mathbb{R}^n = S \cup (\mathbb{R}^n \setminus S),$$

и критерий δ в общем виде принимает вид

$$\delta(X) = \begin{cases} \mathsf{H}_1, & \text{если } X \in \mathbb{R}^n \setminus S \\ \mathsf{H}_2, & \text{если } X \in S \end{cases}.$$

Область S часто называют критической областью.

Определение 1.3.1 *Уровнем значимости критерия δ называют вероятность ошибки первого рода α_1:*

$$\alpha_1 = \alpha_1(\delta) = \mathsf{P}_{\mathsf{H}_1}(\delta \neq \mathsf{H}_1) = \mathsf{P}_{\mathsf{H}_1}(\delta = \mathsf{H}_2) = \mathsf{P}_{\mathsf{H}_1}(X \in S).$$

Итак, уровень значимости критерия – это вероятность того, что выборка попадает в критическую область в условиях того, что верна гипотеза H_1.

Вместе с уровнем значимости критерия часто рассматривают и такую характеристику, как мощность.

Определение 1.3.2 *Мощностью критерия δ называют $1 - \alpha_2$, то есть*

$$1 - \alpha_2 = 1 - \alpha_2(\delta) = 1 - \mathsf{P}_{\mathsf{H}_2}(\delta \neq \mathsf{H}_2) = \mathsf{P}_{\mathsf{H}_2}(\delta = \mathsf{H}_2) = \mathsf{P}_{\mathsf{H}_2}(X \in S).$$

Итак, мощность критерия – это вероятность того, что выборка попадет в критическую область при условиях того, что верна гипотеза H_2.

Важно отметить, что вероятности ошибок первого и второго рода вычисляются при разных предположениях о распределении (верна гипотеза H_1 или гипотеза H_2), а потому никаких фиксированных соотношений вроде $\alpha_1 = 1-\alpha_2$ и иже с ними просто напросто не существует.

Ясно, что хочется: чтобы уровень значимости был как можно меньше, ведь уровень значимости – это вероятность ошибки первого рода, а ошибки надо искоренять. Однако не менее важно увеличивать и мощность: ведь чем больше мощность, тем меньше ошибка второго рода. К сожалению, на практике следить за тем и другим параметрами одновременно не представляется возможным из-за отсутствия явной функциональной зависимости между мощностью и уровнем значимости. На практике классически выбирается достаточно маленький уровень значимости, а уж мощность оказывается такой, какой оказывается, мы ее не выбираем.

Оказывается, что дело обстоит даже вот как. Уменьшение уровня значимости как правило ведет к уменьшению и мощности, то есть одновременно уменьшить уровень значимости и увеличить мощность мы, как правило, не можем. Ведь уменьшая ошибку первого рода, то есть уменьшая вероятность $\mathsf{P}_{\mathsf{H}_1}(X \in S)$, мы будем уменьшать критическую область S, тем самым уменьшая и мощность, равную $\mathsf{P}_{\mathsf{H}_2}(X \in S)$, так что между мощностью и уровнем значимости нужно сохранять некий компромисс. Покажем это и на таком шуточном примере.

Пусть у нас есть выборка объема 1 из нормального распределения $\mathsf{N}_{a,1}$. Рассмотрим две гипотезы: $\mathsf{H}_1: \ a = 0$ и $\mathsf{H}_2: \ a = 1$ и критерий

$$\delta(X_1) = \begin{cases} \mathsf{H}_1, & X_1 \le d \\ \mathsf{H}_2, & X_1 > d \end{cases}.$$

Что такое ошибка первого рода? Это вероятность не отвергнуть гипотезу H_2 в случае, когда верна гипотеза H_1, то есть $\alpha = \mathsf{P}_{\mathsf{H}_1}(\delta = H_2) = \mathsf{P}_{\mathsf{H}_1}(X_1 > d)$. Что такое ошибка второго рода? Это вероятность не отвергнуть гипотезу H_1 в случае, когда верна гипотеза H_2, то есть $\mathsf{P}_{\mathsf{H}_2}(\delta = \mathsf{H}_1) = \mathsf{P}_{\mathsf{H}_2}(X_1 \le d)$.

На рисунке четко видно, что уменьшая вероятность ошибки первого рода (α), автоматически возрастает вероятность ошибки второго рода, тем самым уменьшается мощность критерия, и наоборот.

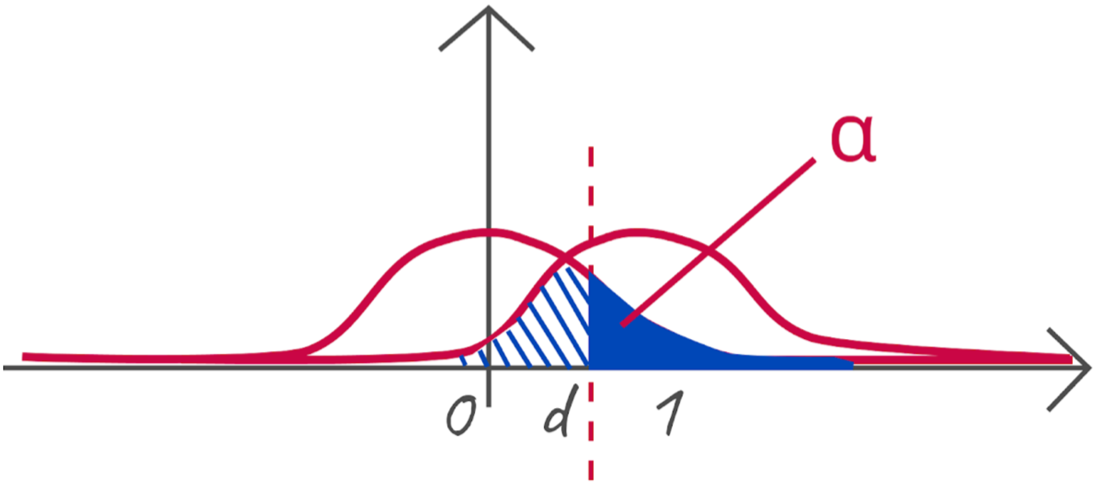

Рис. 1: Ошибки первого и второго рода

Еще раз отметим, что ошибку первого рода мы фиксируем, а ошибку второго – нет. А это значит, что может быть вовсе не все равно, что рассматривать в качестве первой гипотезы, а что – в качестве альтернативы к ней. Скажем, путь мы собираемся лететь на самолете, и у нас есть две гипотезы: $\mathsf{H}_1 = \{$самолет исправный$\}$ и $\mathsf{H}_2 = \{$самолет неисправный$\}$. При проверке гипотезы возникает две ошибки: отвергнуть гипотезу, что самолет исправный, если самолет исправный (ошибка первого рода), и отвергнуть гипотезу, что он неисправный, если он неисправный (ошибка второго рода). При таком выборе гипотез мы контролируем ошибку события, которое, в худшем случае, принесет нам потерю денег за сданный билет, и совсем не контролируем вероятность события, которое отвечает за возможную авиакатастрофу.

В такой ситуации, гипотезы стоит сформулировать иначе: $\mathsf{H}_1 = \{$самолет неисправный$\}$ и $\mathsf{H}_2 = \{$самолет исправный$\}$. Теперь мы будем контролировать вероятность события «отвергнуть гипотезу, что самолет неисправный, если самолет неисправный», сделаем ее, конечно, известной и очень маленькой, и перестанем бояться возможной (с неизвестной вероятностью происходящей) авиакатастрофы.

2 Критерии согласия

2.1 Понятие критерия согласия

Проверка гипотез – тема, заслуживающая не одной лекции, а целого отдельного курса. Мы остановимся только на так называемых критериях согласия и некоторых их обобщениях.

Определение 2.1.1 *Критериями согласия обычно называют критерии, предназначенные для проверки простой гипотезы* $\mathsf{H}_1 = \{\mathcal{P} = \mathcal{P}_1\}$ *при сложной альтернативе* $\mathsf{H}_2 = \{$*гипотеза* H_1 *неверна*$\}$.

Мы будем работать по следующему принципу. Пусть задана некоторая функция, показывающая отклонение эмпирического распределения (построенного по выборке $X = (X_1, X_2, ..., X_n)$) от теоретического, распределение которой существенно разнится от того верна основная гипотеза или нет. Логично, что в зависимости от значения этой «функции» отклонения, можно не отвергать или отвергать основную гипотезу. Давайте попробуем сформулировать некоторый алгоритм.
1. Пусть удается задать некоторую функцию $\rho(X)$ такую, что:

- если гипотеза H_1 верна, то

$$\rho(X) \xrightarrow[n\to+\infty]{\mathsf{d}} Y \sim \mathcal{G},$$

  где $\mathcal{G}$ – некоторое непрерывное распределение;

- если гипотеза H_1 неверна, то

$$|\rho(X)| \xrightarrow[n\to+\infty]{\mathsf{P}} \infty.$$

2. Тогда для случайной величины Y, имеющей распределение $\mathcal{G}$ найдем постоянную C из условия $\varepsilon = \mathsf{P}(|Y| \geq C)$. Построим критерий δ следующим

образом:

$$\delta(X) = \begin{cases} \mathsf{H}_1, & \text{если } |\rho(X)| < C \\ \mathsf{H}_2, & \text{если } |\rho(X)| \geq C \end{cases}$$

Итак, построенный критерий работает вот по какому правилу. Если на данной выборке функция отклонения по абсолютному значению оказывается велика, то выбор свидетельствует в пользу альтернативы, и наоборот.

Можно показать, что при увеличении n уровень значимости нашего критерия стремится к ε, а мощность – к единице. Последнее понятие нуждается в уточнении, ведь теперь альтернатива состоит не из одного распределения, а из какого-то множества. В нашем случае это означает, что мощность для каждого конкретного распределения из альтернативы стремится к единице.

Давайте теперь приведем конкретные примеры критериев и их применений.

2.2 Критерий Колмогорова

Одна из самых важных задач – по выборке $X = (X_1, X_2, ..., X_n)$ определить распределение генеральной совокупности ξ. Критерий Колмогорова и позволяет это сделать в случае, когда функция распределения генеральной совокупности непрерывна.

Пусть $X = (X_1, X_2, ..., X_n)$ – выборка из некоторого распределения $\mathcal{P}$. Проверяется гипотеза $\mathsf{H}_1 = \{\mathcal{P} = \mathcal{P}_1\}$ против сложной альтернативы $\mathsf{H}_2 = \{\mathcal{P} \neq \mathcal{P}_1\}$. Если предполагаемое распределение $\mathcal{P}_1$ имеет непрерывную функцию распределения F_1, то удобно пользоваться так называемым критерием Колмогорова. Рассмотрим функцию отклонения

$$\rho(X) = \sqrt{n} \sup_t |F_n^*(t) - F_1(t)|,$$

где F_n^*, как обычно, эмпирическая функция распределения, построенная по выборке X.

Покажем, что введенная функция удовлетворяет описанным выше условиям, но для начала напомним теорему Колмогорова.

Теорема 2.2.1 (Колмогорова) *Пусть $X_1, ..., X_n$ – выборка из генеральной совокупности ξ с непрерывной функцией распределения F_ξ. Тогда для эмпирической функция распределения $F_n^*(t)$ выполняется*

$$Y_n = \sqrt{n} \cdot \sup_{t \in \mathbb{R}} |F_n^*(t) - F_\xi(t)| \xrightarrow[n \to +\infty]{\mathsf{d}} Y,$$

где случайная величина Y имеет распределение Колмогорова с функцией рас-

пределения

$$F_Y(t) = \begin{cases} \sum\limits_{i=-\infty}^{+\infty} (-1)^i e^{-2i^2t^2}, & t \geq 0 \\ 0, & t < 0 \end{cases}.$$

Итак, проведем проверку по пунктам. Пусть $\rho(X) = \sqrt{n} \sup\limits_t |F_n^*(t) - F_1(t)|$, тогда

- Если H_1 верна, то все X_i имеют распределение $\mathcal{P}_1$, а тогда, по теореме Колмогорова, $\rho(X)$ по распределению сходится к случайной величине, имеющей распределение Колмогорова.

- Если H_1 неверна, то все случайные величины X_i имеют некоторое распределение $\mathcal{P}_2$, отличное от $\mathcal{P}_1$. Значит, согласно теореме Гливенко-Кантелли,
$$F_n^* \xrightarrow[n\to+\infty]{\mathsf{P}} F_2$$
для каждого $y \in \mathbb{R}$. Но так как $\mathcal{P}_1 \neq \mathcal{P}_2$, то найдется t_0, что $F_1(t_0) \neq F_2(t_0)$, а значит
$$\sup_{t\in\mathbb{R}} |F_n^*(t) - F_1(t)| \geq |F_n^*(t_0) - F_1(t_0)| \xrightarrow[n\to+\infty]{\mathsf{P}} |F_2(t_0) - F_1(t_0)| > 0.$$
Но тогда
$$\rho(X) = \sqrt{n} \sup_{t\in\mathbb{R}} |F_n^*(t) - F_1(t)| \xrightarrow[n\to+\infty]{\mathsf{P}} \infty.$$

Как же осуществляется проверка гипотез? Пусть случайная величина Y имеет распределение с функцией распределения Колмогорова. Это распределение табулировано (таблицу можно найти в дополнительных материалах), так что по заданному $\varepsilon > 0$ легко найти такое C, что $\varepsilon = \mathsf{P}(Y \geq C)$, а тогда Критерий Колмогорова выглядит так:

$$\delta(X) = \begin{cases} \mathsf{H}_1, & \rho(X) < C \\ \mathsf{H}_2, & \rho(X) \geq C \end{cases}.$$

Пример применения критерия на конкретных данных с пошаговым алгоритмам приведен в опросе к данному фрагменту.

Кроме критерия Колмогорова, часто используют критерий χ^2 Пирсона как для проверки параметрической, так и непараметрической гипотез. Для ознакомлением с этим критерием мы отправляем заинтересованного слушателя к дополнительным материалам.

2.3 Проверка гипотезы однородности

Как мы не раз отмечали, часто, особенно при сборе данных, полезно понимать: берутся ли данные из одного и того же распределения, или нет. Скажем, подчиняется ли рост членов баскетбольной команды и рост детей в детском саду одному и тому же распределению? Вероятно, нет, скажете вы. Но ответ на поставленный вопрос не всегда так очевиден.

Итак, пусть даны две выборки $X = (X_1, X_2, ..., X_n)$ и $Y = (Y_1, Y_2, ..., Y_m)$ из неизвестных распределений $\mathcal{P}$ и $\mathcal{G}$, соответственно. Будем проверять сложную гипотезу $\mathsf{H}_1 = \{\mathcal{P} = \mathcal{G}\}$ против альтернативы $\mathsf{H}_2 = \{\mathsf{H}_1 \text{ неверна}\}$. Предположим также, что распределения $\mathcal{P}$ и $\mathcal{G}$ имеют непрерывные функции распределения.

Оказывается, что, по аналогии с критерием Колмогорова, резонно рассмотреть функцию отклонения вида

$$\rho(X, Y) = \sqrt{\frac{mn}{m+n}} \sup_{t \in \mathbb{R}} |F_n^*(t) - G_m^*(t)|,$$

где функция $F_n^*(t)$ – эмпирическая функция распределения, построенная по выборке X, а $G_m^*(t)$ – эмпирическая функция распределения, построенная по выборке Y.

Оказывается справедливой следующая теорема

Теорема 2.3.1 *Если гипотеза* H_1 *верна, то*

$$\rho(X, Y) \xrightarrow[m,n \to +\infty]{\mathsf{d}} Y,$$

где случайная величина Y *имеет распределение Колмогорова.*

Аналогично предыдущему пункту, про данному ε найдем такое C, что $\varepsilon = \mathsf{P}(Y \geq C)$, тогда критерий Колмогорова-Смирнова имеет вид:

$$\delta(X, Y) = \begin{cases} \mathsf{H}_1, & \rho(X, Y) < C \\ \mathsf{H}_2, & \rho(X, Y) \geq C \end{cases}.$$

Пример применения критерия на конкретных данных с пошаговым алгоритмам приведен в опросе к данному фрагменту.

2.4 Гипотезы о нормальном распределении

Одно из самых важных распределений – нормальное распределение. Давайте рассмотрим некоторые гипотезы (и, конечно, методы их проверки), касающиеся его параметров.

2.4.1 Совпадение средних двух нормальных выборок с равными дисперсиями

Итак, пусть есть две независимые выборки $X = (X_1, X_2, ..., X_n)$ из распределения $\mathsf{N}_{a_1,\sigma^2}$ и $Y = (Y_1, Y_2, ..., Y_m)$ из распределения $\mathsf{N}_{a_2,\sigma^2}$, причем дисперсия σ^2 одинакова для обоих распределений, но, вообще говоря, неизвестна. Особенно часто возникает необходимость проверить равенство средних двух независимых выборок из нормальных распределений в медицине для выяснения наличия или отсутствия действия препарата. Ясно, что данная задача – частный случай задачи проверки однородности двух выборок. Оказывается, что в случае, если дисперсии различны, то задача решается лишь в частных случаях, о которых мы говорить не будем.

Итак, мы будем проверять гипотезу $\mathsf{H}_1 = \{a_1 = a_2\}$, используя случайную величину

$$t_{n+m-2} = \sqrt{\frac{nm}{n+m}} \cdot \frac{\left(\overline{X} - a_1\right) - \left(\overline{Y} - a_2\right)}{\sqrt{\frac{(n-1)S_0^2(X)+(m-1)S_0^2(Y)}{n+m-2}}},$$

где $S_0^2(X)$ – несмещенная выборочная дисперсия, пстроенная по выборке X, а $S_0^2(Y)$ – по выборке Y.

Оказывается, справедлива следующая теорема.

Теорема 2.4.1 *Случайная величина t_{n+m-2} имеет распределению Стьюдента T_{n+m-2} с $(n + m - 2)$ степенями свободы.*

Если гипотеза H_1 верна, то резонно рассмотреть функцию отклонения

$$\rho(X, Y) = \sqrt{\frac{nm}{n+m}} \cdot \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{(n-1)S_0^2(X)+(m-1)S_0^2(Y)}{n+m-2}}},$$

которая, по только что сформулированной теореме имеет распределение Стьюдента T_{n+m-2}, а значит критерий имеет смысл строить следующим образом.

По заданному ε, найдем квантиль $C = \tau_{1-\varepsilon/2}$ – $(1 - \varepsilon/2)$ квантиль распределения Стьюдента T_{n+m-2}. Для такой квантили

$$\mathsf{P}(|t_{n+m-2}| > C) = \varepsilon$$

и критерий Стьюдента имеет такой же вид, как и критерии согласия:

$$\delta(X, Y) = \begin{cases} \mathsf{H}_1, & |\rho(X, Y)| < C \\ \mathsf{H}_2, & |\rho(X, Y)| \geq C. \end{cases}$$

Пример применения критерия на конкретных данных с пошаговым алгоритмам приведен в опросе к данному фрагменту.

2.4.2 Совпадение дисперсий двух нормальных выборок

Естественно, даже если известно, что выборки получены из нормального распределения, зачастую совершенно неизвестно: одинаковы у них дисперсии или нет, и непонятно: можно ли применять только что описанный критерий Стьюдента, или нет. Для проверки этого факта часто применяют так называемый критерий Фишера.

Итак, пусть есть две независимые выборки $X = (X_1, X_2, ..., X_n)$ из распределения $\mathsf{N}_{a_1,\sigma_1^2}$ и $Y = (Y_1, Y_2, ..., Y_m)$ из распределения $\mathsf{N}_{a_2,\sigma_2^2}$. Проверяется гипотеза $\mathsf{H}_1 = \{\sigma_1^2 = \sigma_2^2\}$.

Зададим функцию отклонения следующим образом:

$$\rho(X, Y) = \frac{S_0^2(X)}{S_0^2(Y)}$$

Справедлива следующая теорема.

Теорема 2.4.2 *Если гипотеза H_1 верна, то $\rho(X, Y)$ имеет распределение Фишера $\mathsf{F}_{n-1,m-1}$ с $(n-1, m-1)$ степенями свободы.*

Справку о распределении Фишера можно найти в дополнительных материалах.

Построим критерий Фишера. Пусть $f_{\varepsilon/2}$ и $f_{1-\varepsilon/2}$ – соответсвующие квантили распределения Фишера $F_{n-1,m-1}$. Тогда критерий таков:

$$\delta(X, Y) = \begin{cases} \mathsf{H}_1, & f_{\varepsilon/2} \leq \rho(X, Y) \leq f_{1-\varepsilon/2}, \\ \mathsf{H}_2, & \text{иначе} \end{cases}.$$

Пример применения критерия на конкретных данных с пошаговым алгоритмам приведен в опросе к данному фрагменту.

2.4.3 Гипотеза о среднем нормальной совокупности с известной дисперсией

Пусть имеется выборка из нормального распределения N_{a,σ^2} с известной дисперсией σ^2. Проверяется простая гипотеза $\mathsf{H}_1 = \{a = a_0\}$ против сложной альтернативы $\mathsf{H}_2 = \{a \neq a_0\}$.

Рассмотрим функцию отклонения

$$\rho(X) = \sqrt{n}\frac{\overline{X} - a_0}{\sigma}.$$

Ясно, что если гипотеза H_1 верна, то $\rho(X)$ имеет стандартное нормальное распределение. Пусть $\varepsilon > 0$, тогда выберем $C = \tau_{1-\varepsilon/2}$ – квантиль уровня

$(1-\varepsilon/2)$ стандартного нормального распределения, а тогда

$$\varepsilon = \mathsf{P}_{\mathsf{H}_1}(|\rho(X)| \geq C)$$

и критерий, как и все критерии согласия, выглядит так:

$$\delta(X) = \begin{cases} \mathsf{H}_1, & |\rho(X)| < C \\ \mathsf{H}_2, & |\rho(X)| \geq C. \end{cases}$$

Пример применения критерия на конкретных данных с пошаговым алгоритмам приведен в опросе к данному фрагменту.

2.4.4 Гипотеза о среднем нормальной совокупности с неизвестной дисперсией

Пусть теперь имеется выборка из нормального распределения N_{a,σ^2} с неизвестной дисперсией σ^2. Проверяется простая гипотеза $\mathsf{H}_1 = \{a = a_0\}$ против сложной альтернативы $\mathsf{H}_2 = \{a \neq a_0\}$.

Рассмотрим функцию отклонения

$$\rho(X) = \sqrt{n}\frac{\overline{X} - a_0}{S_0(X)}.$$

Ясно, что если гипотеза H_1 верна, то $\rho(X)$ имеет распределение Стьюдента T_{n-1}. Пусть $\varepsilon > 0$, тогда выберем $C = \tau_{1-\varepsilon/2}$ – квантиль уровня $(1-\varepsilon/2)$ распределения T_{n-1}, тогда

$$\varepsilon = \mathsf{P}_{\mathsf{H}_1}(|\rho(X)| \geq C)$$

и критерий, как и все критерии согласия, выглядит так:

$$\delta(X) = \begin{cases} \mathsf{H}_1, & |\rho(X)| < C \\ \mathsf{H}_2, & |\rho(X)| \geq C. \end{cases}$$

Пример применения критерия на конкретных данных с пошаговым алгоритмам приведен в опросе к данному фрагменту.

2.5 Резюме

Итак, в этой лекции мы рассмотрели лишь некоторые подходы к проверке гипотез, и на этом наш курс подошел к концу. Можно еще много говорить и про многомерные выборки (которых в нашем курсе мы не коснулись), и про сравнение критериев, и про сравнение оценок: аппарат статистики очень богат и обширен. Все это, конечно, невозможно запихнуть в семестровый курс. Но мы надеемся, что полученных в этом курсе базовых знаний, умений и навыков хватит для освоения и тех специфических методов, которые используются в конкретно вашей предметной области. Удачи и до новых встреч!