Занятие пятое. Критерии согласия или как перейти к параметрической модели

О критериях вообще и согласия в частности

Пусть у нас есть выборка $X_1, ..., X_n \sim F$. Критерием для проверки гипотезы $H_0 : F \in \mathcal{A}_0$ против альтернативы $H_1 : F \in \mathcal{A}_1$ мы называем правило, которое по выборке выдает число 0 (принять H_0 гипотезу) или число 1 (принять гипотезу H_1). Здесь $\mathcal{A}_0$, $\mathcal{A}_1$ — некоторые множества распределений.

Легко понять, что любой критерий задается критическим множеством D— таким множеством выборок, при попадании в которое критерий выдает 1.

Для критерия существует два вида ошибок: отвергнуть H_0, когда она верна или принять H_0, когда на самом деле верна H_1. Ошибка первого типа называется *ошибкой первого рода*, второго — *ошибка второго рода*. В рамках курса математической статистики вы рассматривали критерии с заданным уровнем значимости α, то есть те, у которых вероятности всех возможных ошибок 1 рода равны (меньше или равны) α, т.е.

$$\mathbf{P}_F((X_1, ..., X_n) \in D) = \alpha, \ \forall F \in \mathcal{A}_0.$$

Более удобным являются следующий подход — рассмотрим семейство критических множеств D_c с некоторым параметром, $D_{c_2} \subseteq D_{c_1}$, $c_1 \leq c_2$. Наиболее распространенным вариантом таких множеств являются $D_c = \{T(X_1, ..., X_n) \geq c\}$, где T — некоторая статистика. Найдем максимальное такое c, что D_c содержит нашу реализацию $x_1, ..., x_n$. Тогда $\sup_{F \in \mathcal{A}_0} \mathbf{P}_F((X_1, ..., X_n) \in D_c)$ называют *фактическим уровнем значимости* (p-value). Это минимальный уровень значимости, при котором гипотеза H_0 при нашей выборке отвергается. Маленький фактический уровень значимости свидетельствует о том, что гипотеза крайне маловероятна и отвергается практически наверняка, большой — что данный критерий не отвергает нашу гипотезу.

Вопрос 1. Рассмотрим критерий $\overline{X} > c$ для проверки гипотезы $H_0 : \mu = 0$ с альтернативой $H_1 : \mu > 0$ для $X_i \sim \mathcal{N}(\mu, 1)$. Статистика $\overline{x}$ приняла значение 3. Чему равен фактически уровень значимости?

Рассматривая критерии вида $\{T(X_1, ..., X_n) \geq c\}$, где T — статистика с непрерывной функцией распределения, и считая их фактический уровнь значимости, мы можем заметить, что при выполнении гипотезы он распределен равномерно на [0,1]. Это простое следствие того, что $F(X) \sim R[0, 1]$ при $X \sim F$.
Таким образом, если у нас в доступе есть большое число выборок, мы можем по каждой найти уровень значимости и посмотреть на распределение этих уровней. При выполнении гипотезы оно должно быть близко к равномерному.

В рамках сегодняшнего занятия мы будем рассматривать критерии согласия и в качестве H_0 будем рассматривать $F \in F_\theta$, т.е. принадлежность F какому-то параметрическому семейству. Альтернативой H_1 мы будем считать все остальные распределения. В англоязычной среде такие тесты называют goodness of fit.
Для построения критерия уровня α достаточно найти некоторое свойство, которые бы выполнялось для всех распределений нашего класса достаточно вероятно (с вероятностью не менее $1 - \alpha$).
При этом сколько-то удовлетворительно мажорировать вероятность ошибки второго рода не удается, поскольку вне нашего параметрического семейства есть сколь угодно похожие на наши распределения. Но по-крайней мере, можно искать критерий, от которого мы ожидаем, что при альтернативе он чаще попадает в критическое множество.

Пример 1. Теорема Пирсона утверждает, что если вектор $\vec{X} = (X_1, ..., X_k)$ принимается значения $(1, 0, ..., 0)$, ... , $(0, 0, ..., 1)$ с вероятностями $p_1, ..., p_k$, то при n-кратном розыгрыше н.о.р. таких векторов, частоты $(\nu_1, ..., \nu_k) = \vec{X}_1 + ... + \vec{X}_n$ появления различных исходов удовлетворяют соотношению

$$T = \sum_{i=1}^{k} \frac{(\nu_i - np_i)^2}{np_i} \xrightarrow{d} Y \sim \chi^2_{k-1},$$

где χ^2_{k-1} — распределение хи-квадрат с $k - 1$ степенью свободы.
Пусть мы хотим проверить гипотезу о том, что вероятности — заданные $p_1, ..., p_k$. Пусть y_p — квантиль

уровня p (то есть $F^{-1}(p)$, где F — функция распределения) для χ^2_{k-1}. Тогда любое из критических множеств $\{T < y_\alpha\}$, $\{T > y_{1-\alpha}\}$, $\{T < y_{\alpha/2}\} \cup \{T > y_{1-\alpha/2}\}$ задает критерий с уровнем значимости $1-\alpha$. При этом при альтернативе мы ожидаем, что статистика T будет принимать большие значения, поскольку ν_i/n будут стремиться не к p_i, а к каким-то другим вероятностям. Следовательно, критерий с первым критическим множеством будет крайне неудачным. Он действительно редко (с вероятностью α) будет попадать в критическое множество при верной гипотезе, но еще реже будет попадать туда при альтернативе. Из этих соображений понятно, что наиболее разумный критерий имеет второй вид. Это и есть критерий хи-квадрат, который мы изучали в рамках математической статистики.

Можно наложить и более формальные требования на критерии. Распространенными являются
1) Несмещенность, т.е. $\mathbf{P}_F((X_1, ..., X_n) \in D) \geq \alpha$ при $F \in \mathcal{A}_2$. Иначе говоря, при альтернативе мы попадаем в критическое множество не реже, чем при гипотезе.
2) Состоятельность, $\mathbf{P}_F((X_1, ..., X_n) \notin D) \to 0$, $F \in \mathcal{A}_2$. Иначе говоря, вероятности ошибок второго рода стремятся к 0 при каждой альтернативе с ростом n.

В данном случае (при конечном числе значений $\vec{X}$) критерий хи-квадрат не будет, вообще говоря, несмещенным, но будет состоятельным.

Сделаем еще одно важное замечание. Большинство указанных нами теорем основываются на асимптотических утверждениях. Тем не менее, они работают и при небольших выборках. В этом случае предельное распределение заменяется на явно подсчитанных или моделированных методом Монте-Карло квантилях распределения тестовых статистик.

О простой гипотезе и сложной альтернативе

Рассмотрим сперва проверку гипотезы для случая, когда гипотеза H_0 простая — $F = F_0$. С двумя критериями такого типа вы уже знакомы:
1) Критерий хи-квадрат.
Разобьем область значений нашей величины на непересекающиеся диапазоны $\Delta_1, ..., \Delta_k$, подсчитаем $p_i = \mathbf{P}(X \in \Delta_i)$, $X \sim F_0$. Тогда если H_0 верна, то ν_i — количества X_1,...,X_n, попавших в Δ_i, удовлетворяют теореме Пирсона

$$T = \sum_{i=1}^{k} \frac{(\nu_i - np_i)^2}{np_i} \xrightarrow{d} Y \sim \chi^2_{k-1}.$$

Критерий χ^2 предлагает взять критическое множество $D = \{T > y_{1-\alpha}\}$.
Этот критерий не будет несмещенным (он является асимптотическим и в целом не гарантирует никаких особенных свойств при конечных n) и не будет состоятельным (поскольку любое другое распределение с теми же вероятностями попадания в Δ_i неотличимо от нашего), но достаточно разумен.
Для дискретных данных с конечным числом значений критерий состоятелен.
2) Критерий Колмогорова.
Этот критерий функционирует только для непрерывных F_0. В силу теоремы Колмогорова в этом случае

$$\sqrt{n} D_n = \sqrt{n} \sup_x |\hat{F}_n - F_0| \xrightarrow{d} K \sim K(x),$$

где $K(x)$ — распределение Колмогорова, $\hat{F}_n$ — ЭФР. Критерий Колмогорова предлагает использовать критическое множество $\{\sqrt{n} D_n > k_{1-\alpha}\}$, где $k_{1-\alpha}$ — квантиль распределения Колмогорова.
Этот критерий также асимптотический, а потому, вообще говоря, смещенный, но состоятельный
Вопрос 2. Доказать состоятельность критерия Колмогорова.
Вопрос 3. Рассмотрим выборку из $\mathcal{N}(0,1)$, будем брать из нее подвыборки без возвращения и считать на их основе статистику Колмогорова. Должен ли фактический уровень значимости иметь равномерное распределение?
Кроме этих критериев, полезны также
3) Критерии omega-квадрат. Подобно критерию Колмогорова они работают лишь в непрерывном случае,

опираются на $|\hat{F}_n(x) - F(x)|$, но пользуются соотношениями

$$\omega_i^2 = n \int_{\mathbb{R}} (\hat{F}_n(x) - F(x))^2 g_i(F(x)) dF(x) \xrightarrow{d} W_i \sim W_i(x), i = 1, 2$$

где $g_1(x) = 1$, $g_2(x) = 1/(x(1-x))$, а $W_1(x)$, $W_2(x)$ — некоторые распределения. Соответственно, первый критерий (он называется Крамера-Мизеса) предлагает критическое множество $\{\omega_1 > w_{1-\alpha,1}\}$, второй (Андерсона-Дарлинга) — $\{\omega_2 > w_{1-\alpha,2}\}$, где $w_{1-\alpha,i}$ — квантили W_i, $i = 1, 2$. Зачем нужны эти критерии и чем они отличаются от критерия Колмогорова?

Идейно разница между ними такова:
Критерий Колмогорова улавливает наибольшее отклонение между ЭФР и ф.р. Неважно, как долго встречалось это отклонение, возникло ли оно на узком диапазоне или было при большом количестве x, критична здесь только величина максимального перепада.
Критерий Крамера-Мизеса лучше реагирует на продолжительные по времени отклонения.
Критерий Андерсона-Дарлинга фокусируется на отклонении при тех значениях, которые редки для предполагаемой ф.р. На практике критерий Андерсона-Дарлинга выглядит заметно сильнее обоих конкурентов.
В R критерий Колмогорова-Смирнова ks.test есть в стандартном пакете stats, аналогично критерий хи-квадрат задается chisq.test. Критерии Андерсона-Дарлинга и Крамера-Мизеса есть в goftest и заданы там функциями ad.test и cvm.test.
Упражнение 1. Проведите тестирование всех 4 методов для проверки 1) нормальности $\mathcal{N}(0, 1)$ 2) равномерности $R[-1.7, 1.7]$ для выборок: а) стандартной нормальной б) $R[-1.7, 1.7]$ в) из распределения Лапласа г) $\mathcal{N}(0.2, 1)$, д) $\mathcal{N}(0, 1)$ при значениях больших 1 по модулю и $R[-1, 1]$ иначе; размеров 25, 50, 100
Для тестирования сгенерируйте по 100 выборок для каждого распределения и подсчитайте для них фактические уровни значимости p-value. Постройте график для э.ф.р. p-value для каждого метода. Для правильного распределения эта э.ф.р. должна быть близка к равномерной ф.р., а для неправильного по возможности быть сильно выше.

Нетрудно заметить, что проверку гипотезы $F = F_0$ можно свести к проверке гипотезы равномерности для непрерывных F_0, просто применив F_0 к элементам выборки. Это позволяет осуществлять и визуальную проверку, близость к равномерному распределению вполне успешно идентифицируется графически.
Такого рода механизм предлагает так называемый quantile-quantile plot. Для его построения по оси абсцисс откладываются квантили распределения F_0, а по оси ординат — упорядоченная выборка. В случае выполнения гипотезы график должен быть близок к прямой y=x.

О сложной гипотезе и сложной альтернативе

Эта глава требуется только при сдаче на сложном уровне
Более частая ситуация заключается в том, что у нас есть гипотеза о принадлежности к параметрическому семейству, например, что выборка нормальная, но с неизвестными параметрами.
Можно, конечно, оценить неизвестные параметры состоятельными оценками, но эта процедура может сместить распределение статистик критерия.

1) Критерий хи-квадрат в этом случае удается модернизировать, если вместо неизвестных параметров подставить оценки ОМП для них.
Более конкретно, рассмотрим вероятности $\mathbf{P}_\theta(\Delta_i)$. Подсчитаем количества попаданий ν_i в Δ_i. Найдем ОМП для θ по функции правдоподобия

$$L(x_1, ..., x_n; \theta) = \prod_{i=1}^{n} \mathbf{P}_\theta(\Delta_i)^{\nu_i}.$$

Подставив в $\mathbf{P}_\theta(\Delta_i)$ полученную оценку для θ, можно найти новую статистику хи-квадрат. Теорема Пирсона утверждает, что полученная статистика будет иметь распределение χ^2_{k-l-1}, где l — размерность параметра θ, k — число Δ_i.
2) Критерий отношения правдоподобий — более общий метод, позволяющий работать со сложными па-

раметрическими моделями. Критерий хи-квадрат, в действительности, является апрроксимацией этого критерия.
Можно представлять этот критерий как некоторое обобщение критерия Неймана-Пирсона. Пусть $\theta = (\theta_1, ..., \theta_r)$ и основная гипотеза

$$H_0 : (\theta_{q+1}, ...\theta_r) = (\theta_{0,q}, ..., \theta_{0,r}),$$

то есть часть параметров фиксирована, а остальные произвольны. Тогда найдем статистику отношения правдоподобия

$$T = 2 \ln \frac{L(x_1, ..., x_n, \hat{\theta})}{L(x_1, ..., x_n, \hat{\theta}_0)},$$

где $\hat{\theta}_0$ — ОМП при H_0, $\hat{\theta}$ — ОМП в общей параметрической модели. Оказывается, при H_0 в так называемых сильно регулярных моделях справедливо соотношение $T \xrightarrow{d} Y \sim \chi^2_{r-q}$, $n \to \infty$, откуда $T > y_{1-\alpha}$ задает асимптотический критерий уровня α. Не будем останавливаться подробно на формулировке условий сильной регулярности, отметим лишь ключевые: правдоподобие должно достаточно гладко зависить от параметра и область изменения параметра является открытым множеством
В частности, критерий работает для выборок из дискретного распределения, где параметричность модели уже не требуется.
Вопрос 4. Как будет выглядеть критерий отношения правдоподобий для дискретных выборок с k возможными значениями?
3) Критерий Колмогорова-Смирнова также применим к сложной гипотезе при подстановке состоятельных оценок (например, ОМП). Однако в этом случае предельное распределение уже не будет Колмогоровским, а будет своим для каждого класса распределений. Это замечание крайне важно и зачастую игнорируется малоопытными исследователями.
Так для нормальных распределений при этом получится так называемое распределение Лиллиефорса. В общем случае можно определить критическое множество с помощью метода Монте-Карло.
4) Критерии Андерсона-Дарлинга и Крамера-Мизеса будут верны и в случае параметрических семейств, но опять-таки предельное распределение станет иным и будет зависеть от распределения выборки.

О некоторых параметрических семействам

Для некоторых семейств существуют довольно мощные специализированные критерии. Большая подборка таких тестов находится в пакете PoweR
1) Для нормальных распределений в R есть удобный пакет nortest. Наиболее удачными методами являются
a) Метод Шапиро-Уилка, задающийся статистикой

$$\frac{\left(\sum_{i=1}^n a_i X_{(i)}\right)^2}{\sum_{i=1}^n (X_i - \overline{X})^2},$$

где a_i — некоторые константы. В R он задан функцией shapiro.test, входящим в стандартные пакеты. Этот тест показывает наиболее хорошие результаты даже при небольших выборках.
b) Хорошее качество проверки дает тест Андерсона-Дарлинга ad.test, который можно найти в nortest
c) Качественным критерием является критерий Jarque, Beta, представленный в moments функцией jarque.test. Этот критерий основан на коэффициентах асимметрии Sk и эксцесса C и использует статистику

$$JB = n\left(\frac{Sk^2}{6} + \frac{(K-3)^2}{24}\right),\ Sk^2 = \frac{\hat{\mu}_3}{\hat{\mu}_2^3},\ K = \frac{\hat{\mu}_4}{\hat{\mu}_2^4},$$

где $\hat{\mu}_i = \overline{(X - \overline{X})^i}$ — i-ый центрированный выборочный момент.
2) Для экспоненциальных распределений большой спектр критериев предлагает пакет exptest. В част-

ности, критерий Гини (Gini) базируется на статистике

$$G = \frac{\sum_{j=1}^{n} X_{(j)}(2j - n - 1)}{n(n-1)\overline{X}},$$

которая при нормировке $\sqrt{12(n-1)}(G - 0.5)$ имеет асимптотическое нормальное распределение. Также можно найти критерий Шапиро-Уилка для экспоненциального случая и ряд других критериев. Критерий Андерсона-Дарлинга неплохо проявляет себя и в этом случае.
Упражнение 2. Проверить нормальные и экспоненциальные выборки размеров а) 20, б) 50, в) 100 на экспоненциальность и нормальность.

Как проверять много гипотез?

Предположим, что у нас есть цепочка гипотез $H_{0,i}$ против $H_{1,i}$, $i \leq k$. Если мы хотим получить итоговую вероятность ошибки I рода не больше α, то как организовать процесс?
Конечно, если статистики критериев независимы, то мы можем просто проверять каждую из гипотез на уровне значимости $1 - \sqrt[k]{1-\alpha}$.
В общем случае возможны несколько методов:
1) Метод Бонферрони.
Каждую из гипотез проверять на уровне $1 - \alpha/k$. Этот метод даст вероятность не более α того, что мы отвергнем хотя бы одну верную гипотезу.
Соответственно, фактический уровень значимости проверки всех наших гипотез мы оцениваем суммой фактических уровней значимости каждой.
2) Метод Беньямини-Хучберга.
Если у нас нет необходимости не ошибаться, то мы можем наблюдать за долей ошибочно отвергнутых гипотез $H_{i,0}$. Предположим, что m_0 из гипотез H_0 верны, а остальные $m - m_0$ — нет. Пусть N — число отвергнутых гипотез H_0, N_1 — число ошибочно отвергнутых H_0. Тогда N_1/N называют FDP (False Discovery Proportion), где в случае $N = 0$ FDP=0.
Назовем FDR (False Discovery Risk) $\mathbf{E}(N_1/N)$ — среднее число ошибочно отвергнутых H_0.
Тогда разумно рассматривать систему критериев, таких что при любом $m_0 \leq m$ $FDR \leq \alpha$, т.е. среднее число отвергнутых гипотез не больше α.
Метод Беньямини-Хучберга строит такую процедуру отвержения/принятия. Упорядочим фактические уровни значимости имеющихся критериев $p_{(1)} \leq p_{(2)} ... \leq p_{(k)}$. Положим $l_i = i\alpha/(kC_k)$, $C_k = \sum_{i=1}^{k} i^{-1}$ в случае зависимых критериев и $C_k = 1$ иначе. Тогда положим $R = \max\{i : p_{(i)} < l_i\}$, $P = p_{(R)}$. Метод предлагает отвергать те из $H_{0,i}$, для которых $p_i < P$.

Пример 2. Пусть p-value 10 критериев приняли значения 0.00017, 0.00448, 0.00671, 0.00907, 0.01220, 0.33626, 0.39341, 0.53882, 0.58125, 0.98617.
Мы хотим проверить эту совокупность на уровне значимости 0.05. Тогда в случае метода Бонферрони мы отвергнем те, у которых p-value меньше 0.005, то есть первые 2.
В случае метода Беньямини-Хучберга в общем случае мы считаем $C_k = 2.92$, $l_i = i\alpha/29.2$, сравнивая $p_{(i)}$ и l_i, убеждаемся, что гипотеза отвергается лишь в первом случае.
В случае независимости в том методе $l_i = i\alpha/k$ и гипотеза отвергнется в первых 5 ситуациях.

Вопрос 5. После операции у многих людей появляются неприятные ощущения (nausea), которые можно снять некоторыми лекарствами. Известно, что плацебо помогает в 55% случаев. Проверить эффективность лекарств на уровне 0.05, пользуясь а) методом Бонферрони, б) методом Беньямини-Хучберга

	Number of Patients	Incidence of Nausea
Chlorpromazine	75	26
Dimenhydrinate	85	52
Pentobarbital (100 mg)	67	35
Pentobarbital (150 mg)	85	37

Ответы на вопросы

1. Фактический уровень значимости равен $\mathbf{P}_{H_0}(\overline{X} > 3) = 1 - \Phi(3) \approx 0.0013$.

2. Предположим, что верна альтернатива, то есть $F = F_1 \neq F_0$. Тогда $\hat{F}_n$ сходится к F_1, а значит $\sup|\hat{F}_n(x) - F_0(x)$ сходится к $d = \sup|F_1(x) - F_0(x)| > 0$. Значит $\sqrt{n}D_n \to \infty$, $\mathbf{P}_{F_1}(\sqrt{n}D_n > c) \to 0$ при всех c.

3. Наиболее очевидным контрпримером является случай, когда наша подвыборка имеет тот же размер, что вся выборка. Тогда набор p-value будет состоять из одного и того же числа. В случае, если подвыборки будут меньше размера, этот эффект будет не так заметен, но все же p-value будут зависимы через то, что они считаются по одной и той же выборки. Тем не менее, если исходная выборка будет большого размера, а подвыборки не столь большого, то мы добьемся нужного эффекта.

4. Для k возможных значений мы имеем параметрическую гипотезу $H_0 : p_1 = p_1^0, ..., p_k = p_k^0$, где $p_1, ..., p_{k-1}$ — параметры модели, $p_k = 1 - ... - p_{k-1}$, p_1^0,...,p_k^0 — известные значения. Тогда при гипотезе H_0 ОМП для параметров это p_1^0,...,p_{k-1}^0, поскольку наша модель фактически не имеет неизвестных параметров. В общем случае правдоподобие имеет вид

$$L(x_1, ..., x_n) = p_1^{N_1}...p_{k-1}^{N_{k-1}}(1 - p_1 - ... - p_{k-1})^{N_k},$$

где N_i — число i-х значений в нашей выборке. Логарифмируя и дифференцируя правдоподобие, мы получим уравнения

$$\frac{N_i}{p_i} - \frac{N_k}{1 - p_1 - ... - p_{k-1}} = 0,$$

т.е. $N_i = cp_i$, $i = 1, ..., k$ при некотором c. Тогда $N = N_1 + ... + N_k = c(p_1 + ... + p_k) = c$, то есть оценки будут иметь вид $\hat{p}_i = N_i/N$. Нетрудно убедиться, что это действительно максимум. Тогда

$$T = 2\ln\frac{L(x_1, ..., x_n, \hat{p}_1, ..., \hat{p}_k)}{L(x_1, ..., x_n, p_1^0, ..., p_k^0)} = 2\sum_{i=1}^{k} \ln\left(\frac{N_i}{Np_i^0}\right)^{N_i} = 2\sum_{i=1}^{k} N_i \ln\left(\frac{N_i}{Np_i^0}\right).$$

Критерий имеет вид $T > c$, где $c = y_{1-\alpha}$, $y_{1-\alpha}$ — $1 - \alpha$-квантиль распределения χ^2_{k-1}.

5. Если лекарства эффективны, то для бернуллиевских выборок $X_i \sim Bern(p)$, представленных в таблице, верно $H_1 : p > 0.55$, а если нет, то $H_0 : p = p_0 = 0.55$. Возьмем в качестве критерия $\overline{X} > c$, при гипотезе в силу ЦПТ $\mathbf{P}(\overline{X} > c)$ сходится к $\Phi\left(\sqrt{n}\frac{c-p_0}{\sqrt{p_0(1-p_0)}}\right)$. Подсчет p-value дает нам 0.03, 0.99, 0.88, 0.39. Следовательно, в методе Бонферрони мы сравниваем их с $\alpha/4 = 0.0125$ и не отвергаем. В методе Беньямини-Хучберга мы получим $C_k \approx 2$, $l_i = i\alpha/8$, что меньше любого из заданных p-value. Увы, ни про одно из лекарств мы не можем уверенно утверждать, что оно работает.