# 1 Проверка гипотез

## 1.1 Общие сведения

### 1.1.1 Примеры гипотез

Пусть $H_0$ — это гипотеза (**h**ypothesis), т.е. некоторое предположение о случайной величине $\xi$, которое мы хотим проверить (модель — это предположение, которое считается верным без проверки). Она называется нулевой (null hypothesis), потому что позднее появится альтернативная к ней.

Важно, что гипотеза — предположение о неизвестном законе распределения $\xi$, а не о выборке.

Например, гипотеза о том, что мат.ож. давления до и после приема лекарств одинаково. Или гипотеза о том, что распределение ошибки прибора нормальное, или о том, что распределение генератора псевдослучайных чисел равномерное на [0,1], или что . Или о том, что зависимости между временем на смотрением анимэ и успехами в учебе нет. Тут прослеживается такая особенность: в гипотезе обычно предполагается, что эффекта нет, так как решение принимается, если гипотеза отвергается.

**Задание 1**: Напишите, какую гипотезу вам было бы интересно проверить и какие данные для этого нужно было бы собрать?

### 1.1.2 Критерий

Для проверки гипотезы применяется критерий. Критерий — это правило, по которому гипотеза либо отвергается, либо не отвергается. Про построение такого правила — позже.

Правило строится на основе выборки. Разумное решение: нам нужно построить правило, которое, как уже говорилось, показывает, насколько выборка отличается от тех предположений, которые постулируются в гипотезе. Если отличается сильно, то гипотеза отвергается.

Измерять отличие удобно в числах. Поэтому вводится статистика критерия (функция от выборки, на основе которой строится критерий) $t = t(x_1, \ldots, x_n)$, которая выборке сопоставляет число.

Например, гипотеза про то, что $H_0 : \mathsf{E}\xi = 0$ (например, условия тренировки не влияют на результат (средняя разница в результатах нулевая)). В этом случае, 0 — ожидаемое значение (expected), выборочное среднее $\bar{x}$ — наблюдаемое значение (observed). Их разница как раз измеряет отличие. Однако, просто по разнице не сказать, отличие большое или нет. И вообще, числа могут получиться случайно, на их основе теорию не построить. Поэтому нужен статистический подход, который мы опишем ниже.

На основе результатов критерия принимают решения. Например, если лекарство показало эффективность (гипотезу о том, что оно неэффективно, отвергли), то его запускают в производство.

### 1.1.3 Нет безошибочных решений

Проблема: случайно может произойти что угодно, т.е. безошибочных решений практически не бывает. Приходится задавать максимальный уровень вероятности ошибки, на который можно согласиться при принятии решения.

Задаем маленький уровень значимости (significance level) $0 < \alpha < 1$ и соглашаемся, что с вероятностью $\alpha$ будем принимать неправильное решение.

Что такое маленький? Это зависит от критичности ошибки при принятии решения. Например, принять решение о полете и полететь на неисправном самолете или принять решение взять зонт и зря носить зонт в сумке весь день.

**Задание 2**: Придумайте ситуацию (гипотезу), когда она верна, а вы ошибочно считаете, что на не верна и поэтому принимаете неверное решение. В одном случае на вероятность ошибочного решение больше 0.001 вы бы точно не согласились. Вторая ситуация — когда согласились бы и на 0.2, но, пожалуй, не больше.

### 1.1.4 Статистический критерий

Чтобы построить теорию и отвечать на вопрос, маленькое или большое значение статистики критерия, отвергать гипотезу или нет, нужно перейти на теоретический язык. Т.е., нужно рассматривать абстрактную выборку, ‘до эксперимента’, где $x_i$ — одинаково распределенные независимые случайные величины с тем же распределением, что и у $\xi$.

Таким образом, и статистика критерия $t = t(x_1, \ldots, x_n)$ — тоже случайная величина. Если верна $H_0$, то $t$ имеет некоторое распределение и принимает некоторый диапазон значений.

Например, для модели $\xi \sim N(a, \sigma^2)$ и гипотезы $H_0 : \mathsf{E}\xi = 0$, статистика критерия $t = \sqrt{n}\bar{x}/\sigma$ имеет распределение $N(0, 1)$. Видим, что возможны любые значения. Однако, если мы допускаем некоторую вероятность ошибочно отвергнуть верную нулевую гипотезу, то критерий можем построить. Обычно главное — контролировать эту вероятность.

Разбиваем значения статистики критерия на две части, доверительную и критическую область так, что вероятность для статистика критерия попасть в критическую область равна $\alpha$.

**Задание 3**. Нарисуйте плотность распределения статистики критерия $t = \sqrt{n}\bar{x}/\sigma$ при условии, что верна нулевая гипотезы, и разбейте область значений на две части, доверительную и критическую. Сделайте это разумным образом, чтобы было разумно значения из критической области считать не соответствующими справедливости нулевой гипотезы.

**Формальное определение** Назовем критерием разбиение области значений статистики критерия на две части, $A_\alpha^{(\text{крит})}$ и $A_\alpha^{(\text{дов})}$, такие что вероятность ошибки первого рода $\alpha_I = \mathsf{P}_{H_0}(t \in A_\alpha^{(\text{крит})}) = \alpha$.

После того как критерий построен, пользуемся им уже в режиме 'после эксперимента', когда выборка и значение статистики критерия — числа. Если число $t$ попадает в критическую область, то гипотеза отвергается. Иначе — не отвергается (но нельзя говорить, что принимается, это обсудим позднее).

Итак, важно (!): разбиение на доверительную и критическую область строится на теор. языке, для абстрактной выборки. А используется это разбиение уже для конкретной выборки, чисел.

Допустимо строить разбиение так, чтобы выполнялось $\alpha_I \le \alpha$ (тогда критерий называется консервативным).

Часто удается построить только асимптотический критерий, когда $\alpha_I \to \alpha$ при $n \to \infty$. В этом случае критерий можно применять при достаточно (для критерия) большом объеме выборки, где допустимый объем выборки зависит от скорости сходимости.

Ниже более подробно.

## 1.2 Схема построение критерия на основе статистики критерия

1. Строим статистику критерия $t$ так, что:

   - Статистика критерия $t$ должна измерять то, насколько выборка соответствует гипотезе. В этом случае мы получаем значение статистики критерия для «идеального соответствия».

     Например, если гипотеза про математическое ожидание $H_0 : \mathsf{E}\xi = a_0$, то $t = \bar{x} - a_0$ подходит под это требование. Если гипотеза про дисперсию $H_0 : \mathsf{D}\xi = \sigma_0^2$, то соответствие правильнее измерять отношением и поэтому подошло бы $t = s^2/\sigma_0^2$.

     **Пример.** Пусть $H_0 : \mathsf{E}\xi = a_0$; тогда $t = \bar{x} - a_0$ и «идеальное значение» $t = 0$.

- Распределение $t$ при верной $H_0$ должно быть известно хотя бы асимптотически. Из-за этого часто преобразовывают меры несоответствия, приведенные выше. Для $H_0 : \mathsf{E}\xi = a_0$ в модели $\xi \sim N(a, \sigma^2)$ с известной дисперсией $\sigma^2$ удобно использовать статистику критерия $t = \sqrt{n}(\bar{\mathbf{x}} - a_0)/\sigma \sim N(0, 1)$.

  Для $H_0 : \mathsf{D}\xi = a_0$ в модели $\xi \sim N(a, \sigma^2)$ известно распределение статистики критерия $t = ns^2/\sigma_0^2 \sim \chi^2_{n-1}$.

2. Строим разбиение области значений статистики критерия $t$ так, что:
   - $\mathsf{P}(t \in A_\alpha^{\text{крит}}) = \alpha$.
   - Если альтернативная гипотеза $H_1$ (см. про нее в след.разделе) не конкретизирована, то $A_\alpha^{(\text{крит})}$ следует выбрать так, чтобы она располагалась как можно дальше от идеального значения.

     **Пример.** Обозначения: pdf (probability distribution function) — это плотность, а cdf (cumulative distribution function) — это функция распределения.

     В случае $t \sim \mathrm{N}(0, 1)$ при идеальном значении 0, разумно определить $A_\alpha^{(\text{крит})}$ «на хвостах» графика плотности $\mathrm{pdf}_{\mathrm{N}(0,1)}$ симметрично по обе стороны от 0 так, что для $A_\alpha^{(\text{крит})} = (-\infty, -t_\alpha) \cup (t_\alpha, \infty)$

     $$\alpha/2 = \int_{-\infty}^{-t_\alpha} \mathrm{pdf}_{\mathrm{N}(0,1)}(y)\,\mathrm{d}y = \int_{t_\alpha}^{+\infty} \mathrm{pdf}_{\mathrm{N}(0,1)}(y)\,\mathrm{d}y.$$

     Иными словами,

     $$\alpha/2 = 1 - \mathrm{cdf}_{\mathrm{N}(0,1)}(t_1) \implies t_1 = \mathrm{cdf}^{-1}_{\mathrm{N}(0,1)}(1 - \alpha/2)$$

     и аналогично для $t_0$. Границы доверительной области часто называют критическими значениями.
   - На будущее: если $H_1$ известна, то $A_\alpha^{(\text{крит})}$ выбирается так, чтобы максимизировать мощность критерия против альтернативы $H_1$, определения будут позже.

### 1.2.1 Пример с числами

Общая схема всех примеров будет как написано ниже.

- Модель/предположения: (необязательно, но если есть, то это нужно проверять/обсуждать до использования критерия)
- Гипотеза: $H_0$ : ...
- Статистика критерия $t = ...$:
- Ее распределение при условии, что верная $H_0$: ... — выписано распределение. Если распределение асимптотическое, то при применении критерия нужно обращать внимание на объем выборки.
- Разбиение значений статистики критерия на доверительную и критическую области.
- Дана выборка, дан уровень значимости.
- Задание: проверить гипотезу, сказать, отвергается она или нет.

Как решать:

1. Теор.часть, выборка абстрактная, уровень значимости $\alpha$ тоже произвольный. По виду статистики критерия вы понимаете, какое значение соответствует 'идеальному' соответствию данных гипотезе. Рисуете график плотности статистики критерия и разбиваете значения на доверит. и крит. части, чтобы вероятность попасть в крит. область была равна $\alpha$. В крит.область включаете значения, наиболее далекие от 'идеального'.

2. Практическая часть, выборка состоит из чисел, уровень значимости — конкретное число. Подставляете в формулу статистики критерия числа, получаете число, обозначим $t_0$. Затем считаем, чему равны критические значения (граница(ы) между критической и доверительной областями). Эти числа выражаются через обратную функцию распределения (это квантили). Значения можно вычислить в R или Python, см. ниже приложение. Рисуем снова график плотности статистики критерия, отмечаем там найденные числа, показываем, где критическая область, где доверительная. На основе того, куда попало $t_0$, делаем вывод, отвергается или нет нулевая гипотеза.

Ниже я буду давать три варианта, в зависимости от остатка деления дня рождения на 3.

**Задание 4**: Провести эту схему (записать то, что выше, с самого начала, со слова Модель) для уже разобранного выше критерия со статистикой критерия $t = \sqrt{n}(\bar{\mathbf{x}} - a_0)/\sigma$. Там дисперсия $\sigma^2$ предполагается известной. Пусть она равна 1.44. Гипотеза: $H_0 : \mathsf{E}\xi = a_0$, где (0) $a_0 = -1$, (1) 0.5, (2) 1. Примените критерий для выборки $(0, 2, 1, -1, -2)$ и уровня значимости (0) $\alpha = 0.05$, (1) 0.1, (2) 0.2.

**Задание 5**: Провести эту схему для следующей постановки задачи.

Модель: $\xi$ имеет распределение Бернулли с неизвестным параметром, вероятностью успеха $p$. Напомню, что это означает, что она принимает значения 0 и 1, 1 (успех) с вероятностью $p$.

Гипотеза: $H_0 : p = p_0$

Статистика критерия:

$$t = \sqrt{n}\frac{\hat{p} - p_0}{\sqrt{p_0(1 - p_0)}},$$

где $\hat{p} = \bar{x}$, что логично, так как сумма значений выборки — это в точности число успехов.

Ее распределение при условии, что $H_0$ верна: $t \xrightarrow{\mathrm{d}} \mathrm{N}(0, 1)$ (т.е. это асимптотический критерий).

Дана выборка в виде: число успешных собеседований 45, неуспешных — 55. Проверить гипотезу (0) $H_0 : p = 0.45$, (1) 0.5, (2) 0.4, уровень значимости (0) $\alpha = 0.2$, (1) 0.05, (2) 0.1.

Задание: проверить гипотезу, сказать, отвергается она или нет.

**Задание 6**: Провести эту схему для следующей постановки задачи.

Модель: нет (но предполагается, что $\xi$ принимает конечное число значений).

Гипотеза:

$$H_0 : \mathcal{P}_\xi = \mathcal{P}_0, \text{ где } \mathcal{P}_0 : \begin{pmatrix} x_1^* & \dots & x_k^* \\ p_1 & \dots & p_k \end{pmatrix}.$$

Статистика критерия:

$$T = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i}.$$

Здесь такие обозначения: выборка $\mathbf{x}$ сгруппирована, т.е. каждому $x_i^*$ сопоставляем *наблюдаемую* абсолютную частоту $n_i$ (сколько раз оно встретилось в выборке); $np_i$ — *ожидаемая* абсолютная частота.

Ее распределение при условии, что $H_0$ верна: $T \xrightarrow{\mathrm{d}} \chi^2(k - 1)$. (т.е. это асимптотический критерий). По поводу свойств и вида плотности распределения хи-квадрат отсылаем к википедии https://ru.wikipedia.org/wiki/%D0%A0%D0%B0%D1%81%D0%BF%D1%80%D0%B5%D0%B4%D0%B5%D0%BB%D0%B5%D0%BD%D0%B8%D0%B5_%D1%85%D0%B8-%D0%BA%D0%B2%D0%B0%D0%B4%D1%80%D0%B0%D1%82.

Дана выборка в виде: число успешных собеседований 45, неуспешных — 55. Проверить гипотезу, что это распределение Бернулли с $p = 0.05$. Проверить для значений $\alpha$ от 0.01 до 0.99 с шагом 0.01.

Задание: проверить гипотезу, сказать, отвергается она или нет. Когда надоест перебирать уровни значимости, найдите такое пороговое значение, называемое $p$-значение, что при меньших уровнях значимости гипотеза не отвергается, а при бОльших — отвергается.

## 1.3 Понятие вероятностного уровня $p$-value.

**Определение.** *$p$-value* — это такой значение, что при значениях уровня значимости $\alpha$, больших $p$-value, $H_0$ отвергается (по причине попадания $t$ в $A_\alpha^{\text{крит}}$), а при меньших — не отвергается.

$p$-value — не вероятность, это пороговое значение. Неформально его можно интерпретировать как меру согласованности $H_0$ и выборки. Например, при больших значениях $p$-value практически при всех разумных уровнях значимости гипотеза не отвергается. При близких к нулю значениях $p$-value, наоборот, гипотеза будет отвергаться.

$p$-value — максимальное значение уровня значимости, при котором гипотеза не отвергается (значение статистики критерия попадает в доверит. область). Или, что эквивалентно, минимальное значение уровня значимости, при котором гипотеза отвергается.

Если критическая область определяется через превышение статистики критерия некоторого значения $t_0$, то есть еще определение $p$-value как вероятности того, что при повторных экспериментах статистика критерия будет больше, чем значение в текущем эксперименте. Это определение написано и в wikipedia, но оно не универсальное. Тем не менее, лучше его знать.

**Задание 7** В заданиях 4, 5 и 6 найти p-value и сформулировать ответ в виде: при таких-то уровнях значимости гипотеза отвергается, при таких-то — не отвергается.

## 1.4 Приложение. Вычисление функции распределения и обратной к ней

https://rdrr.io/snippets/

По этому адресу можно делать вычисления он-лайн, вставив туда нужную часть кода

```
###normal distribution N(a, sd^2)
a <- 0
sd <- 1
x <- 2

#cumulative distribution function (cdf)
cdf <- pnorm(x, mean = a, sd = sd) print(cdf)

#inverse to this cdf
x <- qnorm(cdf, mean = a, sd = sd)
print(x)

###chi-square distribution chi2(m), where m is degree of freedom
x <- 240
m <- 200

#cumulative distribution function (cdf) of chi2(m)
cdf <- pchisq(x, df = m) print(cdf)

#inverse to this cdf
x <- qchisq(cdf, df = m)
print(x)
```

Просто онлайн калькуляторы:
https://planetcalc.ru/4986/, https://www.statdistributions.com/normal/ (для ф.р. нужен left tail) — нормальное распределение,
https://www.statdistributions.com/chisquare/ (для ф.р. нужен left tail) — распределение хи-квадрат.