Прикладная статистика в R
Лекция 3. Проверка статистических гипотез. Критерии согласия. Критерии однородности данных.

Елена Михайловна Парилина

д. ф.-м. н., проф.

2021

Модуль 2. Проверка статистических гипотез.

- Общий принцип проверки статистических гипотез.
- Критерии согласия с заданным распределением.
- Критерии о значении параметров нормального распределения:
    - Одновыборочный критерий.
    - Двухвыборочный критерий Стьюдента.
    - Критерий Фишера.
    - Критерий Бартлетта.
    - Критерий Шапиро-Уилкса.
- Критерий однородности данных (Колмогорова-Смирнова).
- Непараметрические критерии: критерий Вилкоксона, ранговый критерий Краскела-Уоллиса.
- Проверка статистических гипотез в R.

Общий принцип проверки статистических гипотез

Гипотеза

Статистической гипотезой называется любое предположение о законе распределения генеральной совокупности.

В дальнейшем гипотезу будем обозначать буквой H. Выдвинутую изначально гипотезу, подлежащую статистической проверке, называют нулевой (основной) и обозначают через H_0. Гипотезу, альтернативную к нулевой, называют альтернативной (конкурирующей) гипотезой и обычно обозначают через H_1.

Простая vs. сложная гипотеза

Гипотеза называется *простой*, если в ней единственным образом определяется закон распределения генеральной совокупности. В противном случае гипотеза называется *сложной*.

Общий принцип проверки статистических гипотез

Например, гипотеза типа: «генеральная совокупность подчиняется нормальному закону распределения с параметрами $(0, 1)$» является простой, а гипотеза: «распределение генеральной совокупности нормальное» является сложной, поскольку определяет вид распределения генеральной совокупности с точностью до параметров.

Критерий

Статистический критерий — это метод статистического сопоставления высказанной нулевой гипотезы с имеющимися выборочными данными $X_{[n]} = \{x_1, \ldots, x_n\}$, сопровождаемый количественной оценкой достоверности получаемого вывода.

Статистика критерия

Статистика критерия $\gamma(X_{[n]})$ — функция от выборочных данных, на основании численного значения которой принимается решение об отклонении или принятии нулевой гипотезы.

Общий принцип проверки статистических гипотез

Если значение статистики $\gamma(X_{[n]})$ попадает в *критическую область*, то нулевая гипотеза H_0 отвергается (соответственно, принимается альтернативная гипотеза H_1). Если значение статистики $\gamma(X_{[n]})$ не попадает в *критическую область*, то нет оснований отвергнуть нулевую гипотезу.

Как правило, статистику $\gamma(X_{[n]})$ выбирают таким образом, чтобы ее распределение при справедливости нулевой гипотезы H_0 и при справедливости альтернативной гипотезы H_1 как можно более сильно различалось.

При проверке статистических гипотез возможны ошибочные выводы двух типов:

- отклонение нулевой гипотезы H_0, когда на самом деле она верна — ошибка первого рода,
- принятие нулевой гипотезы H_0, если на самом деле она неверна — ошибка второго рода.

Общий принцип проверки статистических гипотез

- Вероятность ошибки первого рода будем обозначать через α, вероятность ошибки второго рода будем обозначать через β.
- Мощность критерия $\mu = 1 - \beta$ представляет собой вероятность отклонения нулевой гипотезы H_0, когда верна альтернативная гипотеза H_1.
- Вероятность ошибки первого рода будем также называть уровнем значимости статистического критерия.

Последовательность проверки любой статистической гипотезы следующая:

1. Выдвигается нулевая гипотеза H_0 и альтернативная гипотеза H_1. Задается уровень значимости критерия α. Обычно α выбирается равным 0.001; 0.01 или 0.05.
2. Выбирается статистика критерия $\gamma(X_{[n]})$ так, что при условии справедливости гипотезы H_0 статистика $\gamma(X_{[n]})$ подчиняется некоторому известному закону распределения вероятностей.

Общий принцип проверки статистических гипотез

3. Определяется критическая область. В качестве критической области для гипотезы H_0 выбирается такая область возможных значений статистики $\gamma(X_{[n]})$, попадание в которую при условии справедливости гипотезы H_0 выглядит маловероятным по сравнению с возможностью попадания статистики $\gamma(X_{[n]})$ в указанную область при условии справедливости гипотезы H_1.

   Критическая область может состоять из одного интервала, как правило, следующего вида: $(-\infty, z_\alpha)$ или $(z_{1-\alpha}, \infty)$, где z_α и $z_{1-\alpha}$ — квантили уровней α и $1-\alpha$ закона распределения, которому должна подчиняться (возможно, асимптотически) статистика $\gamma(X_{[n]})$.

   Критическая область может состоять из двух интервалов: $(-\infty, z_{\frac{\alpha}{2}})$ и $(z_{1-\frac{\alpha}{2}}, \infty)$, где $z_{\frac{\alpha}{2}}$ и $z_{1-\frac{\alpha}{2}}$ — квантили уровней $\frac{\alpha}{2}$ и $1-\frac{\alpha}{2}$ соответствующего закона распределения, которому должна подчиняться (возможно, асимптотически) статистика $\gamma(X_{[n]})$.

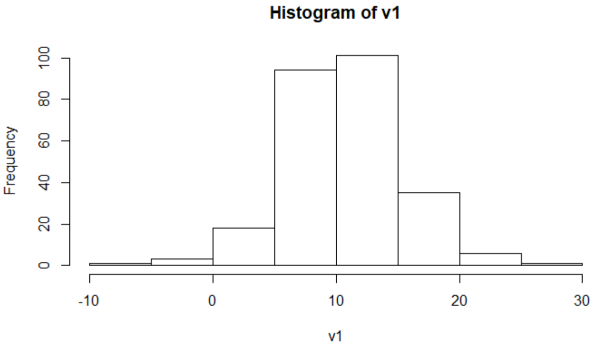
Общий принцип проверки статистических гипотез

4. Делается вывод о принятии или отклонении нулевой гипотезы. Попадание численного значения статистики $\gamma(X_{[n]})$ в критическую область говорит о противоречии имеющихся выборочных данных и нулевой гипотезы, поэтому в этом случае гипотеза H_0 отклоняется (с вероятностью ошибки α), и принимается альтернативная гипотеза H_1. Если численное значение статистики $\gamma(X_{[n]})$ не попадает в критическую область, то нет оснований отвергнуть нулевую гипотезу.
5. Возможен альтернативный подход. Найдем вероятность, которая называется p-значением (p-value). Это вероятность при справедливости нулевой гипотезы получить более экстремальные значения. Если p-value меньше заданного уровня значимости (обычно 0.05), то нулевая гипотеза отвергается в пользу альтернативной. Если p-value больше заданного уровня значимости, то нулевая гипотеза не отвергается.
6. В программах R, SAS, STATA реализован второй подход.

Критерии согласия. Нормальное распределение

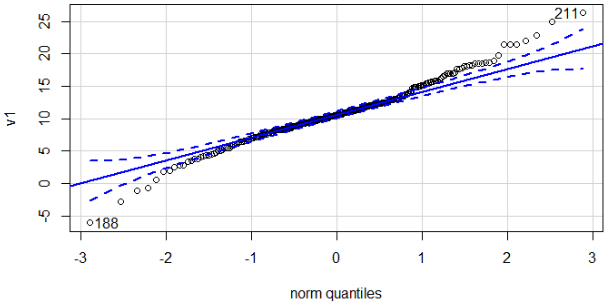
Проверка нормального распределения данных

Сначала посмотрим на графики:



Проверка нормального распределения данных

Посмотрим на график qqplot (сравним эмпирические и теоретические квантили):



Проверка нормального распределения данных

Нулевая гипотеза H_0: выборка извлечена из нормальной генеральной совокупности.
Альтернативная гипотеза H_1: распределение отлично от нормального.

Критерий Колмогорова или Lilliefors test

$$D^+ = \max_{i=1,\dots,n}\left[\frac{i}{n} - F_0\left(\frac{x_{(i)} - \bar{x}}{s}\right)\right], \quad D^- = \max_{i=1,\dots,n}\left[F_0\left(\frac{x_{(i)} - \bar{x}}{s}\right) - \frac{i-1}{n}\right],$$

$$D = \max\{D^+, D^+\}.$$

Статистика критерия: $D\left(\sqrt{n} - 0.01 + \frac{0.85}{\sqrt{n}}\right)$.

Функция в R

```
install.packages("nortest")
library(nortest)
lillie.test(v1)
```

Критерий Колмогорова или Lilliefors test

Результат:

```
> lillie.test(v1)

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  v1
D = 0.072214, p-value = 0.002328
```

Гипотезу о нормальном распределении данных не принимаем, поскольку $p - value = 0.002328$, что меньше уровня значимости критерия 0.05.

Пример

Исследования компании, производящей рубероидную кровельную плитку в Бостоне и Вермонте, показали, что основным фактором, влияющим на оценку качества продукции, является ее вес. На последнем этапе плитка пакуется, а затем размещается на деревянных стеллажах. После заполнения стеллажа регистрируется его вес. Имеются данные о весе (в фунтах) 368 стеллажей, заполненных плитками, произведенными в бостонском отделении компании, и 330 стеллажей, загруженных в Вермонте.

Задача: проверить нормальность данных. Далее проверить однородность данных двух заводов.

```
> PALLET <- read_excel("PALLET.xls")
> View(PALLET)
> p1<-PALLET$Бостон
> p2<-PALLET$Вермонт
> hist(p1)
> lillie.test(p1)

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  p1
D = 0.043246, p-value = 0.09502

> hist(p2)
> lillie.test(p2)

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  p2
D = 0.042088, p-value = 0.166
```

Обе выборки извлечены из нормальной генеральной совокупности, так как $p - value > 0.05$ в обоих случаях.

Проверка нормального распределения данных

Критерий Андресона-Дарлина или Anderson–Darling test

$$A = -n - \frac{1}{n}\sum_{i=1}^{n}(2i-1)\left[\ln F_0\left(\frac{x_{(i)}-\bar{x}}{s}\right) + \ln(1 - F_0\left(\frac{x_{(n-i+1)}-\bar{x}}{s}\right)\right]$$

Статистика критерия: $A\left(1 + \frac{0.75}{n} + \frac{2.25}{n^2}\right)$.

Функции в R

```
library(nortest)
ad.test(p1)
ad.test(p2)
```

```
> ad.test(p1)

        Anderson-Darling normality test

data:  p1
A = 0.84132, p-value = 0.03001

> ad.test(p2)

        Anderson-Darling normality test

data:  p2
A = 0.78753, p-value = 0.04072
```

Мы не принимаем гипотезу о нормальности данных в двух случаях в пользу альтернативной, так как $p - value < 0.05$ в обоих случаях.

Критерий Крамера-фон-Мизеса или Cramer–von Mises test

$$\omega^2 = \frac{1}{12n} + \sum_{i=1}^{n} \left(F_0 \left(\frac{x_{(i)} - \bar{x}}{s} \right) - \frac{2i-1}{2n} \right)^2.$$

Статистика критерия: $\omega^2 \left(1 + \frac{0.5}{n}\right)$.

cvm.test

```
 library(nortest)
cvm.test(x)
cvm.test(rnorm(100, mean = 5, sd = 3))
cvm.test(runif(100, min = 2, max = 4))
```

```
> cvm.test(p1)

        Cramer-von Mises normality test

data:  p1
W = 0.10321, p-value = 0.1012

> cvm.test(p2)

        Cramer-von Mises normality test

data:  p2
W = 0.098797, p-value = 0.1163

> cvm.test(rnorm(100, mean = 5, sd = 3))

        Cramer-von Mises normality test

data:  rnorm(100, mean = 5, sd = 3)
W = 0.024767, p-value = 0.9112

> cvm.test(runif(100, min = 2, max = 4))

        Cramer-von Mises normality test

data:  runif(100, min = 2, max = 4)
W = 0.27126, p-value = 0.000696
```

Мы принимаем гипотезы о нормальности всех данных, так как $p - value > 0.05$ во всех случаях.

Критерий Шапиро-Уилка или Shapiro–Wilk test

Статистика критерия:

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2},$$

где $x_{(i)}$ — элементы вариационного ряда, $\overline{x}$ — выборочное среднее.
Коэффициенты a_i вычислены и берутся из таблиц (или рассчитываются в программе).
Распределение статистики критерия не носит специального названия, а вычисляется методом Монте-Карло в программе.

Функции в R

```
shapiro.test(p1)
shapiro.test(p2)
```

```
> shapiro.test(p1)

        Shapiro-Wilk normality test

data:  p1
W = 0.98195, p-value = 0.0001452

> shapiro.test(p2)

        Shapiro-Wilk normality test

data:  p2
W = 0.99083, p-value = 0.03768
```

Мы отвергаем гипотезы о нормальности всех данных, так как $p - value < 0.05$ во всех случаях.

Функция uniNorm

Функция uniNorm в R

```
 library(MVN)
uniNorm(data, type = "CVM" , desc = TRUE)
```

- Функция проверяет нормальность выборки следующими критериями Shapiro-Wilk, Cramer-von Mises, Lilliefors (Kolmogorov-Smirnov), Shapiro-Francia и Anderson-Darling.
- Аргументы функции: `data` — данные; `type` задает один из тестов: `SW` Shapiro-Wilk, `CVM` Cramer-von Mises, `Lillie` Lilliefors (Kolmogorov-Smirnov), `SF` Shapiro-Francia, `AD` Anderson-Darling; если `desc` = TRUE, то функция отображает описательную статистику, включая mean, standard deviation, median, minimum, maximum, 25th and 75th percentiles, skewness and kurtosis.

Критерий Харке–Бера или Jarque-Bera test

Это статистический тест, проверяющий ошибки наблюдений на нормальность посредством сверки их третьего момента (асимметрия) и четвёртого момента (эксцесс) с моментами нормального распределения, у которого $S = 0$, $K = 3$.

В тесте Харке–Бера проверяется нулевая гипотеза
$H_0: S = 0,\ K = 3,$
$H_1: S \neq 0,\ K \neq 3,$
где S — коэффициент асимметрии (Skewness), K — коэффициент эксцесса (Kurtosis).

Статистика критерия:

$$JB = n\left(\frac{S^2}{6} + \frac{(K-3)^2}{24}\right),$$

где $S = \frac{\sum e_i^{\mathbf{3}}}{n\hat{\sigma}^{\mathbf{3}}_{ML}}$, $K = \frac{\sum e_i^{\mathbf{4}}}{n\hat{\sigma}^{\mathbf{4}}_{ML}}$ и e_i — остатки модели, n — количество наблюдений, $\hat{\sigma}^2_{ML} = \frac{\sum e_i^{\mathbf{2}}}{n}$, ML — обозначение метода максимального правдоподобия (Maximal Likelihood).

Критерий Харке–Бера или Jarque-Bera test

Данная статистика имеет распределение хи-квадрат с двумя степенями свободы, поскольку коэффициенты S и K асимптотически нормальны.

Чем ближе распределение ошибок к нормальному, тем меньше статистика Харке—Бера отличается от нуля. При достаточно большом значении статистики p-value будет мало, и тогда будет основание отвергнуть нулевую гипотезу (статистика попала в «хвост» распределения).

Тест Харке—Бера является асимптотическим тестом, то есть применим к большим выборкам.

JB test

```
library(tseries)
jarque.bera.test(p1)
p2 <- p2[!is.na(p2)]
jarque.bera.test(p2)
```

```
> jarque.bera.test(p1)

        Jarque Bera Test

data:  p1
X-squared = 25.313, df = 2, p-value = 3.187e-06

> jarque.bera.test(p2)
Error in jarque.bera.test(p2) : NAs in x
> p2 <- p2[!is.na(p2)]
> jarque.bera.test(p2)

        Jarque Bera Test

data:  p2
X-squared = 4.9771, df = 2, p-value = 0.08303
```

Критерий Пирсона

Статистика критерия:

$$\sum_{i=1}^{r} \frac{\left(n_i - np_i^{(0)}(\hat{\theta})\right)^2}{np_i^{(0)}(\hat{\theta})}.$$

pearson.test

```
library(nortest)
pearson.test(p1, adjust = TRUE)
```

- `n.classes` — количество интервалов разбиения. По умолчанию используется формула Moore (1986).
- `adjust` логическая переменная. Если `TRUE` (default), то гипотеза сложная (мы не знаем параметров распределения, т.е. не знаем мат.ож. и дисперсии нормального распределения), если FALSE, то гипотеза простая (мы знаем параметры распределения).

Критерий Пирсона для нормального распределения

```
> library(nortest)
> pearson.test(p1, adjust = TRUE)

        Pearson chi-square normality test

data:  p1
P = 24.533, p-value = 0.1765

> pearson.test(p2, adjust = TRUE)

        Pearson chi-square normality test

data:  p2
P = 32.982, p-value = 0.01678
```

Для первой выборки гипотеза о нормальности подтверждается, для второй — отвергается при уровне значимости 0.05.

Критерии однородности данных (по распределению)

Двувыборочный критерий Колмогорова-Смирнова

Проверяется нулевая гипотеза
$H_0 : F_1 = F_2$ (выборки однородны),
$H_1 : F_1 \neq F_2$ (выборки неоднородны).

$$D^+ = \max_{r=1,\dots,m}\left[\frac{r}{m} - F_n(y_{(r)})\right] = \max_{s=1,\dots,n}\left[G_m(x_{(s)}) - \frac{s-1}{n}\right],$$

$$D^- = \max_{r=1,\dots,m}\left[F_n(y_{(r)}) - \frac{r-1}{m}\right] = \max_{s=1,\dots,n}\left[\frac{s}{n} - G_m(x_{(s)})\right],$$

$$D = \max\{D^+, D^-\}.$$

Статистика критерия: $D\sqrt{\frac{mn}{m+n}}$.

ks.test

```
ks.test(p1,p2, alternative = c("two.sided" , "less" , "greater"),
    exact = NULL)
```

Двувыборочный критерий Колмогорова-Смирнова

```
> ks.test(p1,p2,exact = NULL)

        Two-sample Kolmogorov-Smirnov test

data:  p1 and p2
D = 1, p-value < 2.2e-16
alternative hypothesis: two-sided
```

```
> ks.test(p1,p2,alternative="g",exact = NULL)

        Two-sample Kolmogorov-Smirnov test

data:  p1 and p2
D^+ = 1, p-value < 2.2e-16
alternative hypothesis: the CDF of x lies above that of y
```

```
> ks.test(p1,p2,alternative="l",exact = NULL)

        Two-sample Kolmogorov-Smirnov test

data:  p1 and p2
D^- = 2.6888e-17, p-value = 1
alternative hypothesis: the CDF of x lies below that of y
```

Критерий Вилкоксона или Wilcoxon rank sum test / Mann–Whitney U-test

Проверяется нулевая гипотеза
$H_0 : F_1 = F_2$ (выборки однородны),
$H_1 : F_1 \neq F_2$ (выборки неоднородны).

Статистика критерия рассчитывается на основе вычисления рангов элементов двух выборок с последующим сравнением этих ранжирований.

Оба критерия однородности распределений (Колмогорова-Смирнова и Вилкоксона) предполагают непрерывность распределений исследуемых величин. Т.е. предположительно в выборках не должно быть одинаковых значений, хотя такие могут встречаться и у непрерывных распределений при округлении данных при моделировании выборок.

Критерий Вилкоксона или Wilcoxon rank sum test / Mann–Whitney U-test

wilcox.test

```
wilcox.test(p1, p2,
   alternative = c("two.sided" , "less" , "greater"),
   mu = 0, paired = FALSE, exact = NULL, correct = TRUE,
   conf.int = FALSE, conf.level = 0.95, ...)
```

```
> wilcox.test(p1, p2)

        Wilcoxon rank sum test with continuity correction

data:  p1 and p2
W = 0, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
```

Согласно результатам, распределения выборок отличаются (так как *p*-value гораздо меньше 0.05), т.е. продукция, изготавливаемая, на разных заводах, отличается.

Итоги

Что мы узнали на Лекции 3?

- Общий принцип работы любого статистического критерия.
- Что такое критерии согласия.
- Как проверить нормальность данных при помощи нескольких критериев.
- Как проверить однородность данных двух выборок (т.е. что данные неразличимы по распределению).

Что мы узнаем на Лекции 4?

Мы узнаем,

- как проверять гипотезы о параметрах распределения данных в R.
- как проверять гипотезы об однородности данных, если имеется две и более выборок.

Спасибо за внимание и до встречи на Лекции 4!