T-test и тест Манна-Уитни (Вилкоксона). Сравнение по мощности и по устойчивости к выбросам.

Редкокош Кирилл

23.10.2021

Рассмотрим случайную величину ξ, определенную на некотором вероятностном пространстве $(\Omega, \mathfrak{A}, \mathsf{P})$.

T-test

Одновыборочный t-test

Одновыборочный t-test применяется для проверки нулевой гипотезы $H_0 : \mathsf{E}\xi = a = a_0$ о равенстве математического ожидания $\mathsf{E}\xi$ некоторому известному значению a_0.

Статистика критерия будет иметь следующий вид:

– $\mathsf{D}\xi = \sigma^2 < \infty$: $t = z = \sqrt{n}\frac{(\overline{x}-a)}{\sigma} \xrightarrow[n\to\infty]{} N(0,1)$
Если $\xi \sim N(a, \sigma^2)$, то $t = z \sim N(0,1)$.

В этом случае разбиение будет иметь следующий вид:
- $H_1 : \mathsf{E}\xi \neq a_0$ $A^{(\alpha)}_{\text{крит}} = \mathbb{R} \setminus \left(z_{a/2}, z_{1-a/2}\right)$
- $H_1 : \mathsf{E}\xi > a_0$ $A^{(\alpha)}_{\text{крит}} = (z_{1-a}, \infty)$
- $H_1 : \mathsf{E}\xi < a_0$ $A^{(\alpha)}_{\text{крит}} = (-\infty, z_a)$

– $\mathsf{D}\xi$ неизвестна:

$t = \sqrt{n-1}\,\frac{\bar{x}-a_0}{s} = \sqrt{n}\,\frac{\bar{x}-a_0}{\tilde{s}} \xrightarrow[n\to\infty]{} N(0,1)$
Если $\xi \sim N(a, \sigma^2)$, то $t \sim t(n-1)$. В этом случае разбиение будет иметь следующий вид:
- $H_1 : \mathsf{E}\xi \neq a_0$ $A^{(\alpha)}_{\text{крит}} = \mathbb{R} \setminus \left(qnt_{t(n-1)}(a/2), qnt_{t(n-1)}(1-a/2)\right)$
- $H_1 : \mathsf{E}\xi > a_0$ $A^{(\alpha)}_{\text{крит}} = (qnt_{t(n-1)}(1-a), \infty)$
- $H_1 : \mathsf{E}\xi < a_0$ $A^{(\alpha)}_{\text{крит}} = (-\infty, qnt_{t(n-1)}a)$

Двухвыборочный t-test для независимых выборок с равной дисперсией

Если дисперсия известна, то статистика критерия будет выглядеть следующим образом:

$$t = \frac{\bar{x}-\bar{y}}{\sigma\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}} \xrightarrow[n_1,n_2\to\infty]{} N(0,1).$$

Если данные нормальные, то $t \sim N(0,1)$

Если дисперсия неизвестна:

$$t = \frac{\bar{x} - \bar{y}}{\tilde{s}_{1,2}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \xrightarrow[n_1, n_2 \to \infty]{} N(0,1).$$

Если данные нормальные, то $t \sim t(n_1 + n_2 - 2)$

Двухвыборочный t-test для независимых выборок с различной дисперсией

Если дисперсия известна, то статистика критерия будет выглядеть следующим образом:

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \xrightarrow[n_1, n_2 \to \infty]{} N(0,1).$$

Если данные нормальные, то $t \sim N(0,1)$

Если дисперсия неизвестна:

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \xrightarrow[n_1, n_2 \to \infty]{} N(0,1).$$

Условия применимости

– Наблюдения в выборке независимы друг от друга (для двухвыборочного критерия- в обеих выборках);

– Данные имеют распределение близкое к нормальному или объем выборки достаточно велик.

Mann–Whitney U test

Используется для проверки нулевой гипотезы: $P(\xi_1 > \xi_2) = \frac{1}{2}$.

Для построения статистики критерия U необходимо составить единый ранжированный ряд из обеих сопоставляемых выборок, расставив их элементы по степени нарастания признака и приписав меньшему значению меньший ранг (при наличии повторяющихся элементов в выборке использовать средний ранг). Разделить единый ранжированный ряд на два, состоящих соответственно из единиц первой и второй выборок. Подсчитать отдельно сумму рангов, пришедшихся на долю элементов первой выборки W_1, и отдельно — на долю элементов второй выборки W_2, статистика критерия будет иметь следующий вид:

$U := \max(n_1 n_2 + \frac{n_1(n_1+1)}{2} - W_1, n_1 n_2 + \frac{n_2(n_2+1)}{2} - W_2)$.

Если верна H_0, то $\mathsf{E}U = \frac{n_1 n_2}{2}$, $\mathsf{D}U = n_1 n_2 \frac{n_1+n_2+1}{12}$ и $\frac{U - \mathsf{E}U}{\sqrt{\mathsf{D}U}} \xrightarrow[n_1, n_2 \to \infty]{} N(0,1)$

Критерий состоятельный против альтернативы $H_1 : P(\xi_1 > \xi_2) \neq P(\xi_1 < \xi_2)$. Если формы распределений одинаковы, то эта альтернатива обозначает сдвиг. Для симметричных распределений это условие обозначает равенство медиан (а для нормального — математических ожиданий).

Так как критерий является ранговым, то он устойчив к выбросам, хоть и за счет небольшой потери мощности при сравнении с t-test.

##Условия применимости

– Наблюдения в выборке независимы друг от друга;

– Наблюдения в выборке, по крайней мере, порядковые;

– В выборочных данных не должно быть совпадающих значений (все числа — разные) или таких совпадений должно быть очень мало.

Моделирование

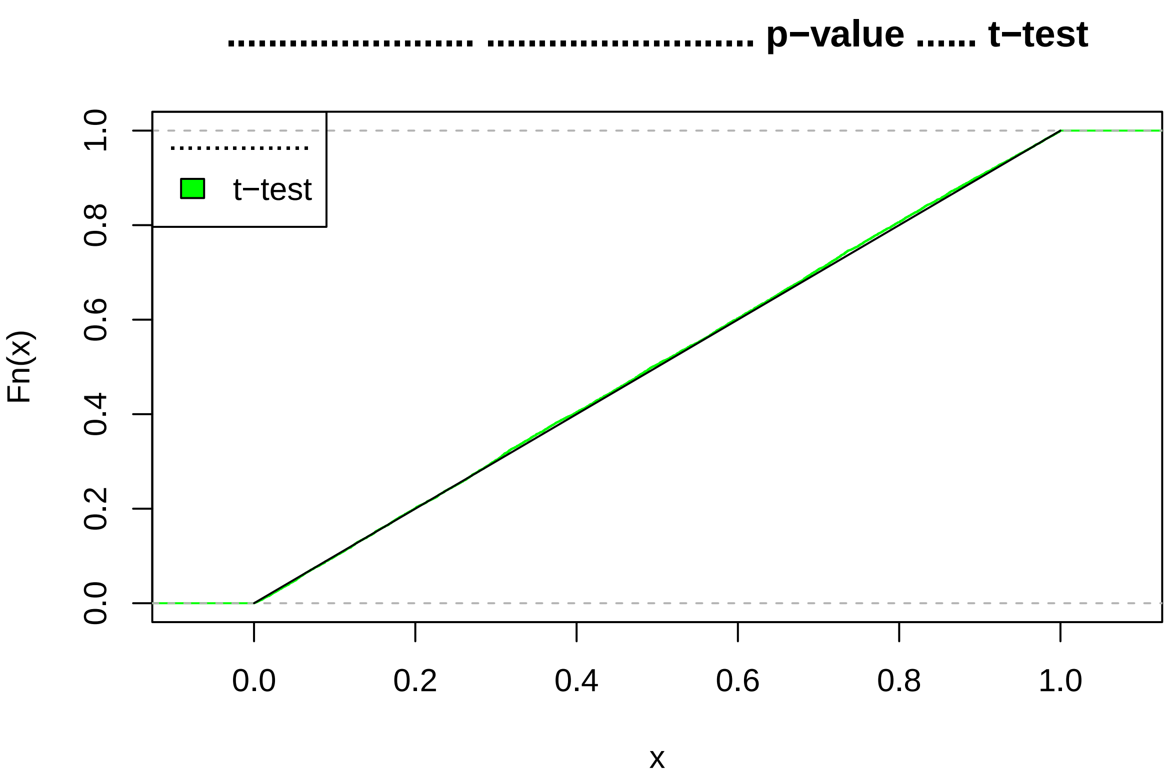
Рассмотрим p-value как случайную величину:

$\alpha_1 = P_{H_0}(H_0 \text{ отвергается}) = \alpha \Leftrightarrow P_{H_0}(\alpha > p) = \alpha \Leftrightarrow P_{H_0}(p < \alpha) = \alpha$

Соответственно, если верна H_0, то p-value равномерно распределено на [0,1].

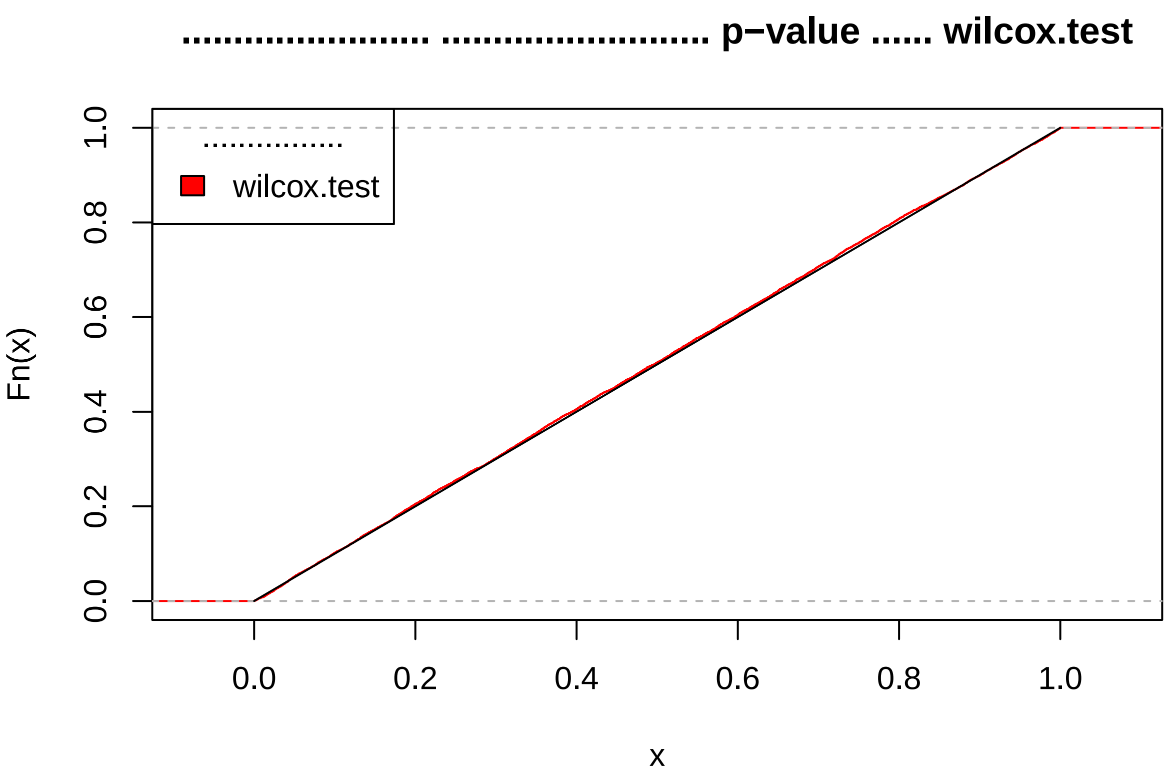
Смоделируем 10000 выбор (по 1000 индивидов) из стандартного нормального распределения. Подсчитываем выборки p-value (отдельно для t-test и wilcox test) и строим функции распределения.

```
set.seed(1)
size <- 10000
individuals <- 1000
samp <- matrix(rnorm(individuals * size, 0, 1), ncol = size)
pvalT <- apply(samp, 2, function(x)
t.test(x, mu = 0)$p.value)
plot(ecdf(pvalT), col='green', main ="Эмперическое распределение p-value для t-test")
lines(c(0,1), c(0,1))
legend("topleft", c("t-test"), title = "Критерий", fill=c("green"))
```




```
pvalW <- apply(samp, 2, function(x)
wilcox.test(x, mu = 0)$p.value)
plot(ecdf(pvalW), col='red', main ="Эмперическое распределение p-value для wilcox.test")
```

```
lines(c(0,1), c(0,1))
legend("topleft", c("wilcox.test"), title = "Критерий", fill=c("red"))
```

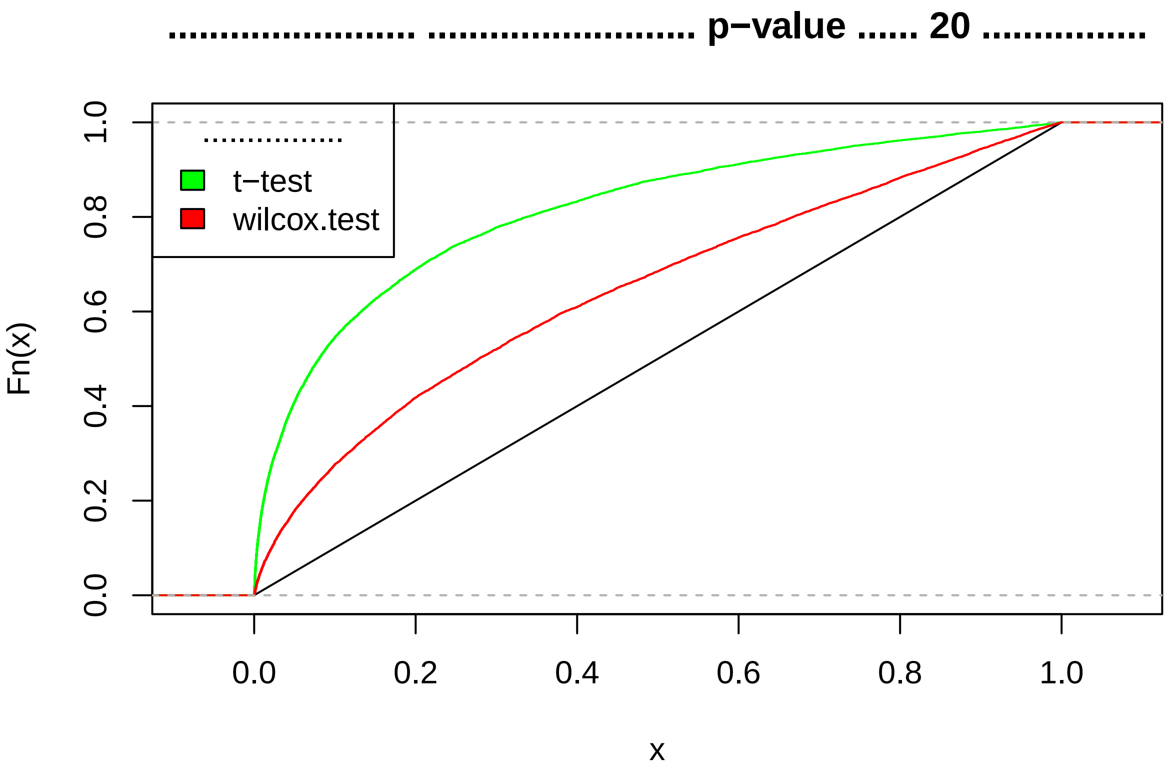


Эмпирические функции распределения совпадают с прямой y=x (на [0,1]), значит оба критерия точные (так как p-value равномерно распределено на [0,1]).

Устойчивость к выбросам

Добавим 20 выбросов (небольших) с одной из сторон (в данном случае справа) и посмотрим на поведения эмпирической функции распределения p-value.

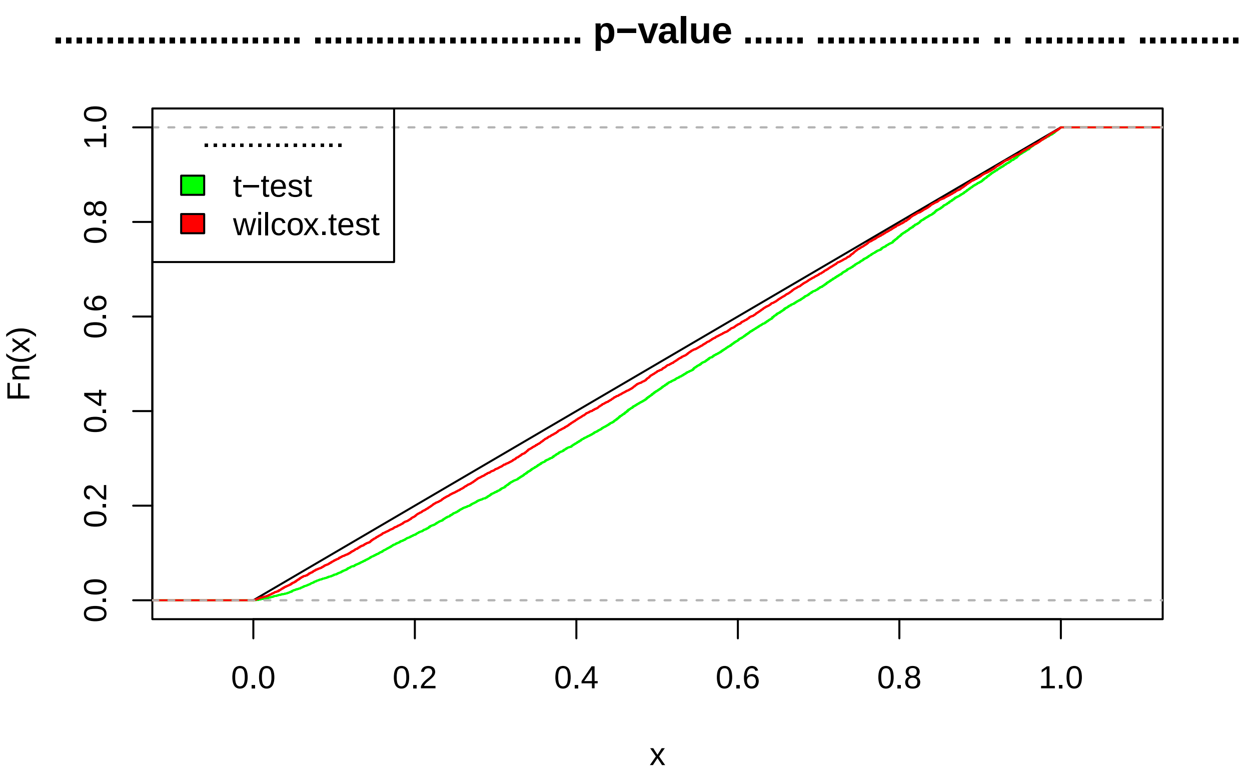
```
amount_of_outliers <- 20
right <- matrix(rnorm(amount_of_outliers * size, 3, 1), ncol = size)
sampRight <- matrix(data=rbind(samp, right), ncol = size)
t_r <- apply(sampRight, 2, function(x)
t.test(x, mu = 0)$p.value)
plot(ecdf(t_r), col='green', main ="Эмперическое распределение p-value при 20 выбросах")
lines(c(0,1), c(0,1))
w_r <- apply(sampRight, 2, function(x)
wilcox.test(x, mu = 0)$p.value)
lines(ecdf(w_r), col='red')
legend("topleft", c("t-test", "wilcox.test"), title = "Критерий", fill=c("green", "red"))
```



Оба критерия стали радикальными, а значит неприменимыми. Стоит заметить, что wilcox test ближе к прямой y=x (на [0,1]), соответственно он является более устойчивым к выбросам такого типа (но все еще неприменимым при таком количестве выбросов).

Добавим ещё 20 выбросов (небольших) с другой стороны (слева) и посмотрим на поведения эмпирической функции распределения p-value.

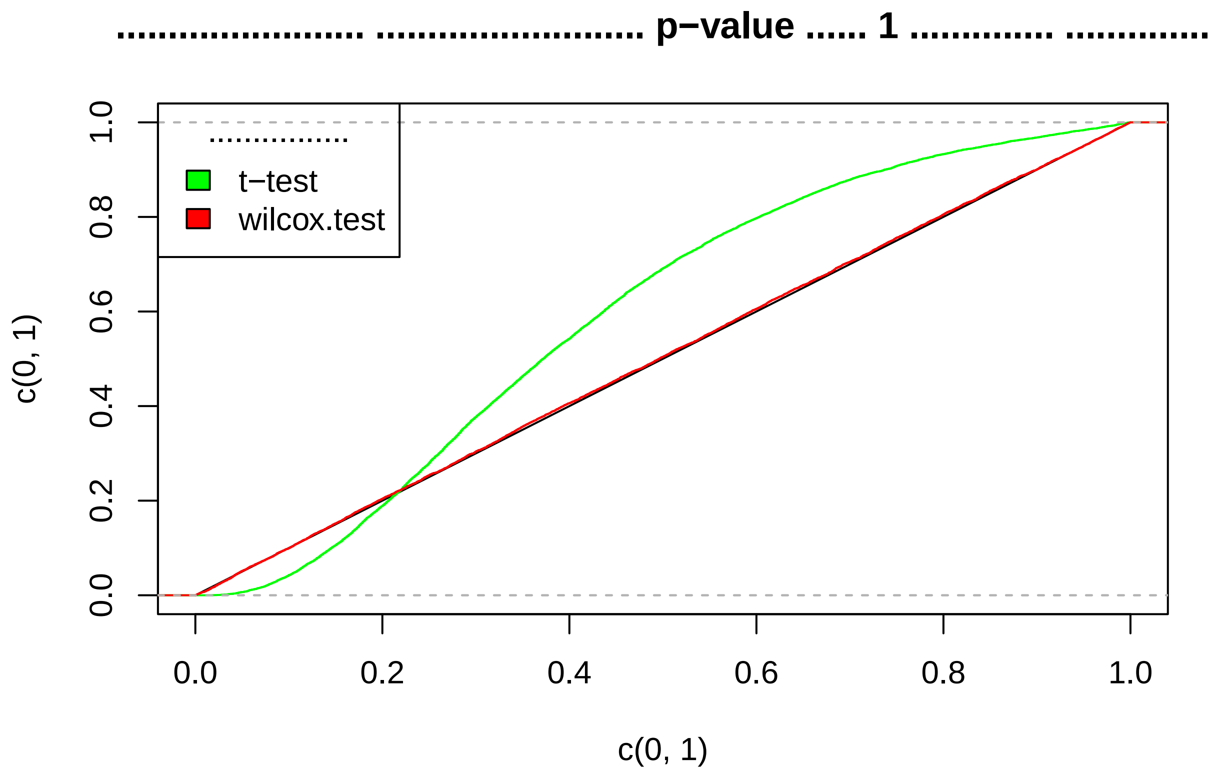
```
left <- matrix(rnorm(amount_of_outliers * size, -3, 1), ncol = size)
sampRightAndLeft <- matrix(data=rbind(sampRight, left), ncol = size)
t_rl <- apply(sampRightAndLeft, 2, function(x)
t.test(x, mu = 0)$p.value)
plot(ecdf(t_rl), col='green', main ="Эмперическое распределение p-value при выбросах с обеих сторон")
lines(c(0,1), c(0,1))
w_rl <- apply(sampRightAndLeft, 2, function(x)
wilcox.test(x, mu = 0)$p.value)
lines(ecdf(w_rl), col='red')
legend("topleft", c("t-test", "wilcox.test"), title = "Критерий", fill=c("green", "red"))
```



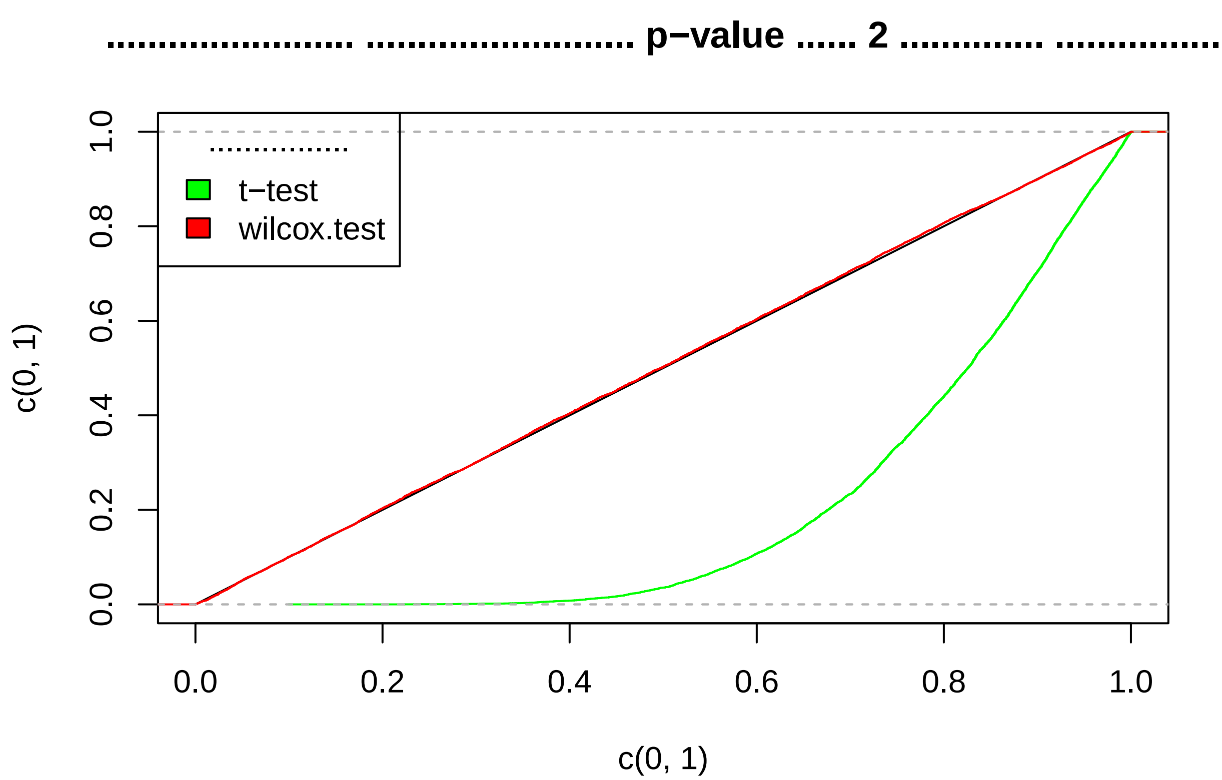
Оба критерия стали консервативными (в силу увеличения дисперсии распределения при таком же математическом ожидании), wilcox test ближе к прямой y=x (на [0,1]).

Рассмотрим исходную модель с добавлением одного выброса, но очень отдаленного. А также двух отдальенных выбросов с обеих сторон.

```
outlier1 <- matrix(rnorm(size, -65, 1), ncol = size)
sampOut1 <- matrix(data=rbind(samp, outlier1), ncol = size)
t_out1 <- apply(sampOut1, 2, function(x)
t.test(x, mu = 0)$p.value)
plot(c(0,1), c(0,1), type="l", main ="Эмперическое распределение p-value при 1 сильном выбросе")
lines(ecdf(t_out1), col='green')
w_out1 <- apply(sampOut1, 2, function(x)
wilcox.test(x, mu = 0)$p.value)
lines(ecdf(w_out1), col='red')
legend("topleft", c("t-test", "wilcox.test"), title = "Критерий", fill=c("green", "red"))
```




```
outlier2 <- matrix(rnorm(size, 65, 1), ncol = size)
sampOut2 <- matrix(data=rbind(sampOut1, outlier2), ncol = size)
t_out2 <- apply(sampOut2, 2, function(x)
t.test(x, mu = 0)$p.value)
plot(c(0,1), c(0,1), type="l", main ="Эмперическое распределение p-value при 2 сильных выбросах")
lines(ecdf(t_out2), col='green')
w_out2 <- apply(sampOut2, 2, function(x)
wilcox.test(x, mu = 0)$p.value)
lines(ecdf(w_out2), col='red')
legend("topleft", c("t-test", "wilcox.test"), title = "Критерий", fill=c("green", "red"))
```



Случай с 1 выбросом: wilcox test почти неотличим от прямой y=x (на [0,1]), а значит он устойчив к таким выбросам. В то же время t-test для стандартных уровней значимости ($\alpha < 0.2$) является консервативным, после чего становится радикальным (чем более отдаленный выброс мы рассматриваем тем дальше находится точка пересечения с прямой y=x).

Случай с 2 выбросами: wilcox test почти неотличим от прямой y=x (на [0,1]), а значит он устойчив к таким выбросам. В то же время t-test является консервативным.

Мощность

Рассматривая p-value как случайную величину:

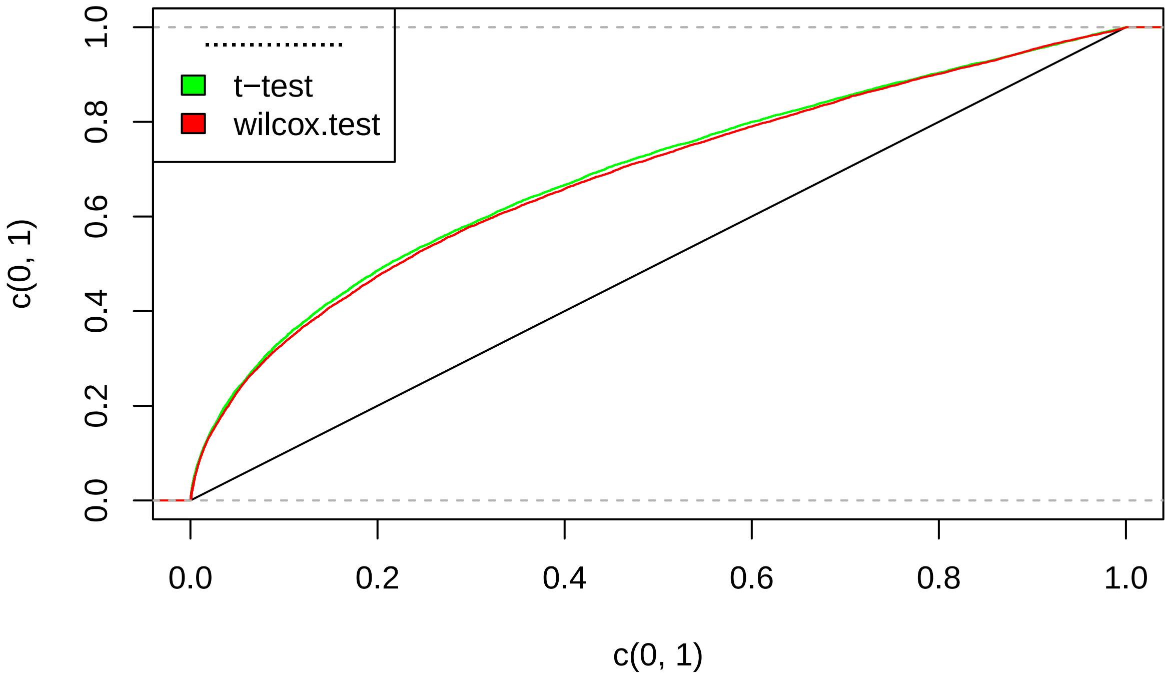
$\alpha_{II} = P_{H_1}(H_0 \text{ не отвергается}) \Leftrightarrow \alpha_{II} = P_{H_1}(\alpha < p) \Leftrightarrow \alpha_{II} = 1 - P_{H_1}(p < \alpha) \Leftrightarrow P_{H_1}(p < \alpha) = \beta.$

Таким образом, если верна альтернативная гипотеза, то через эмпирическую функцию распределения p-value можно определить мощность критерия против альтернативы H_1. Соответственно, для сравнение критериев по мощности против альтернативы H_1 необходимо, что бы H_1 была одинакова для обоих тестов.

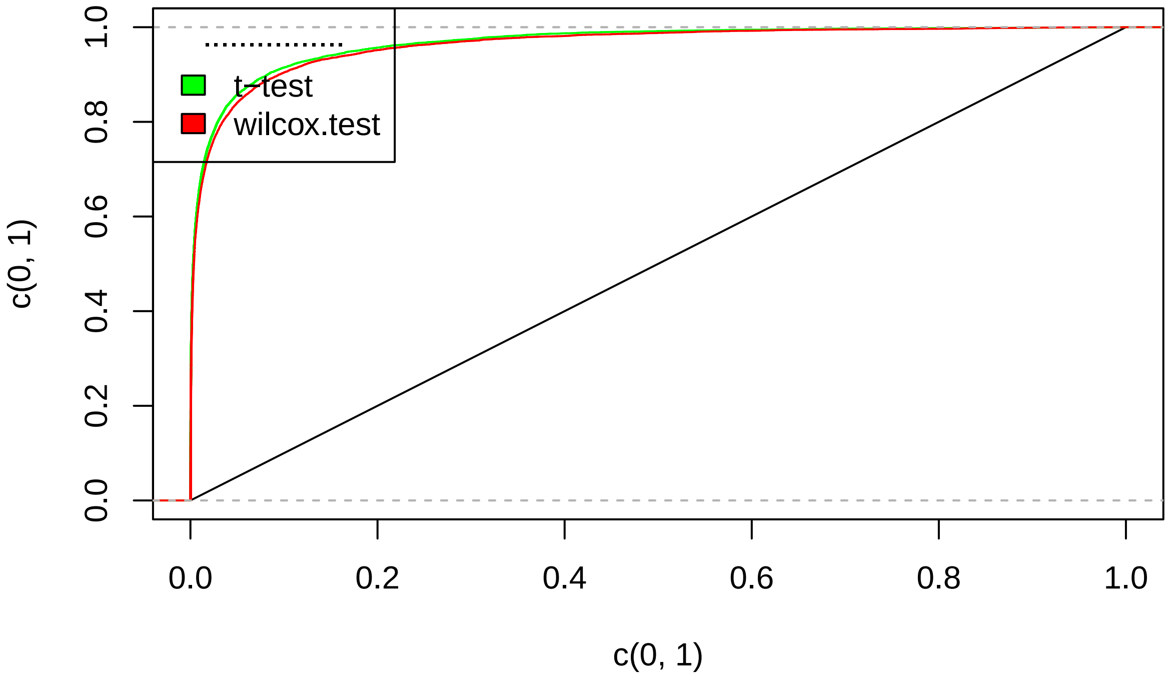
Оценим мощность этих двух критериев, для этого два раза смоделируем 150 выборок (по 10000 индивидов) из нормального распределения (1. С математическим ожиданием 0.1 и дисперсией 1; 2. С математическим ожиданием 0.25 и дисперсией 1). Подсчитываем выборки p-value (отдельно для t-test и wilcox test) и строим функции распределения. Первый случай – мощность против альтернативной гипотезы, что математическое ожидание равно 0.1, второй – мощность против альтернативной гипотезы, что математическое ожидание равно 0.25.

```
samp1 <- matrix(rnorm(10000 * 150, 0.1, 1), ncol = 10000)
samp2 <- matrix(rnorm(10000 * 150, 0.25, 1), ncol = 10000)
pval1_t <- apply(samp1, 2, function(x)
t.test(x, mu = 0)$p.value)
plot(c(0,1), c(0,1), type="l")
```

```
lines(ecdf(pval1_t), col='green')
pval1_w <- apply(samp1, 2, function(x)
wilcox.test(x, mu = 0)$p.value)
lines(ecdf(pval1_w), col='red')
legend("topleft", c("t-test", "wilcox.test"), title = "Критерий", fill=c("green", "red"))
```



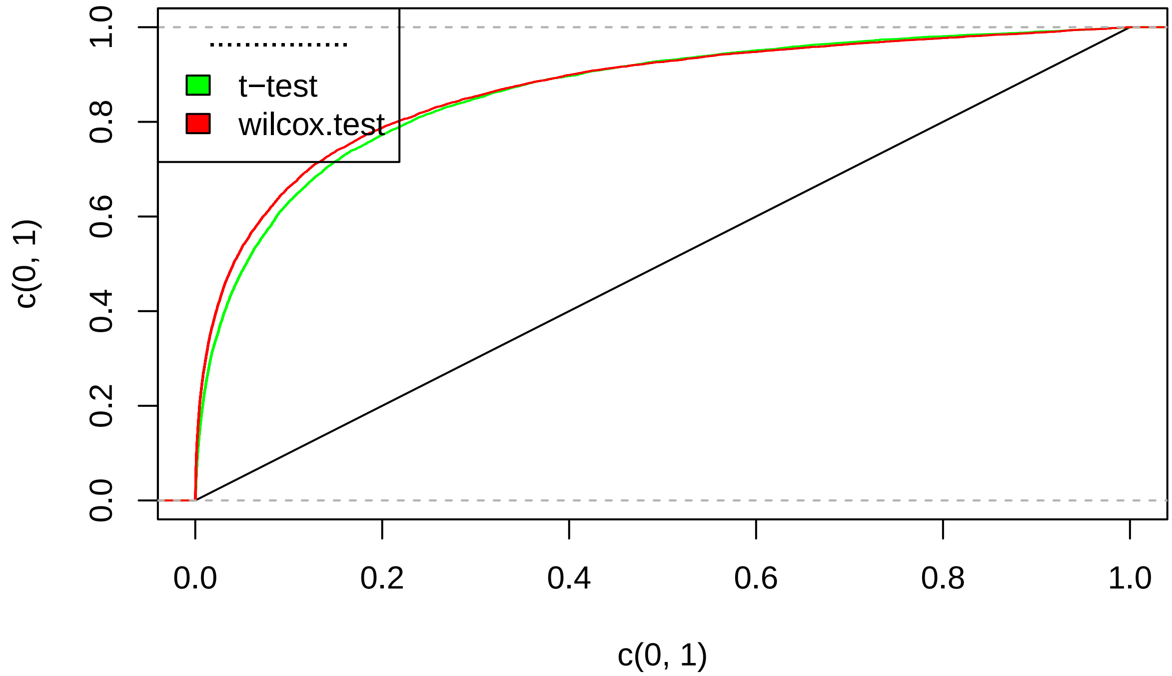

```
pval2_t <- apply(samp2, 2, function(x)
t.test(x, mu = 0)$p.value)
plot(c(0,1), c(0,1),  type="l")
lines(ecdf(pval2_t), col='green')
pval2_w <- apply(samp2, 2, function(x)
wilcox.test(x, mu = 0)$p.value)
lines(ecdf(pval2_w), col='red')
legend("topleft", c("t-test", "wilcox.test"), title = "Критерий", fill=c("green", "red"))
```



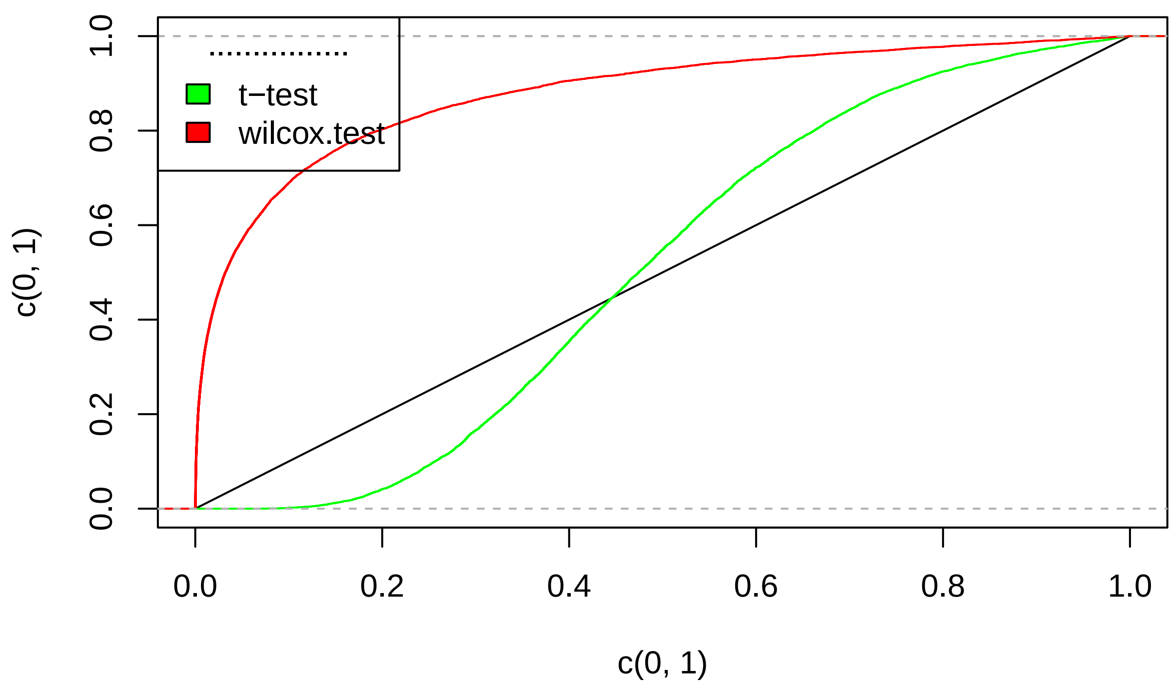
Мощность против обеих альтернатив t-test больше мощности wilcox test (против соответствующих альтернатив), а также заметим, что оба критерия являются более мощными против альтернативной гипотезы, что математическое ожидание равно 0.25, так как альтернатива находится дальше от нулевой гипотезы.

Сравним мощности t-test и wilcox test в случае с выбросами с обеих сторон (так как они применимы-не радикальны) против альтернативы, что математическое ожидание равно 0.07.

```
pvalrl_t <- apply(sampRightAndLeft, 2, function(x)
t.test(x, mu = 0.07)$p.value)
plot(c(0,1), c(0,1), type="l")
lines(ecdf(pvalrl_t), col='green')
pvalrl_w <- apply(sampRightAndLeft, 2, function(x)
wilcox.test(x, mu = 0.07)$p.value)
lines(ecdf(pvalrl_w), col='red')
legend("topleft", c("t-test", "wilcox.test"), title = "Критерий", fill=c("green", "red"))
```



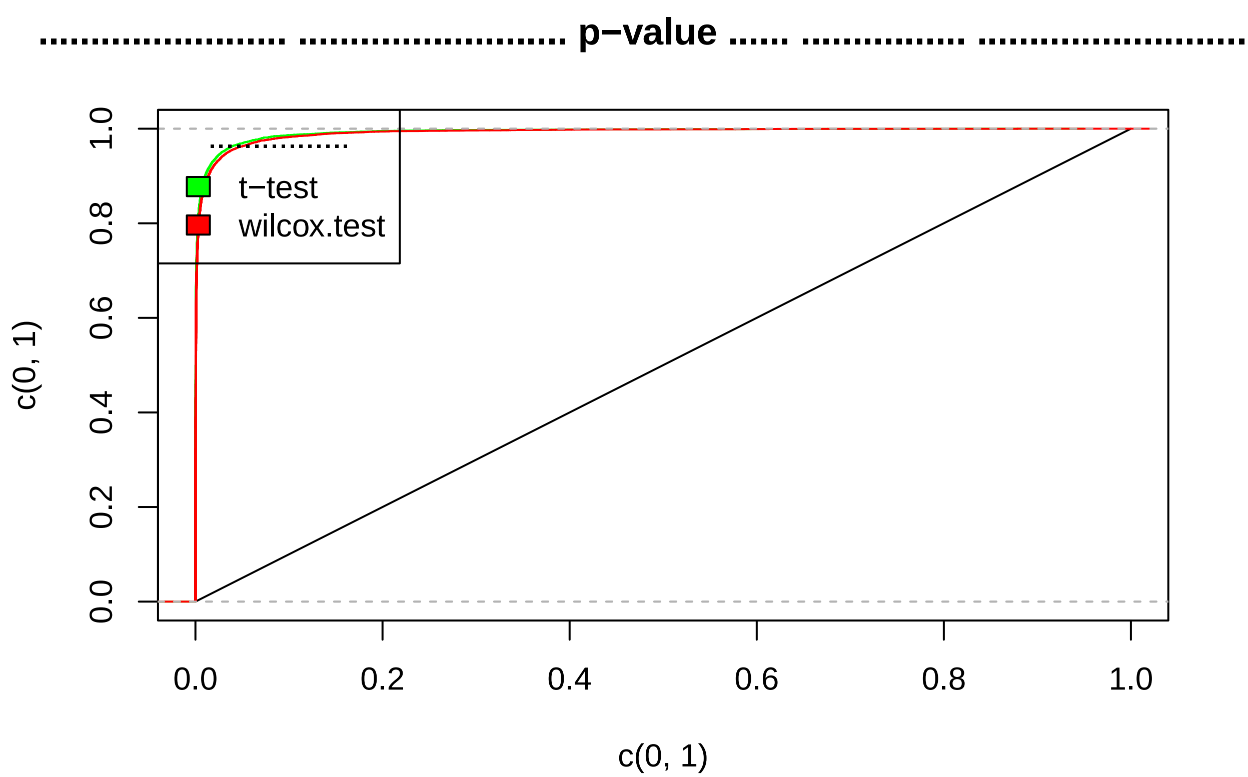

```
pvalOut2_t <- apply(sampOut2, 2, function(x)
t.test(x, mu = 0.07)$p.value)
plot(c(0,1), c(0,1), type="l")
lines(ecdf(pvalOut2_t), col='green')
pvalOut2_w <- apply(sampOut2, 2, function(x)
wilcox.test(x, mu = 0.07)$p.value)
lines(ecdf(pvalOut2_w), col='red')
legend("topleft", c("t-test", "wilcox.test"), title = "Критерий", fill=c("green", "red"))
```



В обоих случаях wilcox test является более мощным против гипотезы, что математическое ожидание равно 0.07.

Проверим состоятельность критериев, для этого рассмотрим выборки из $N(0.1, 1)$, но в 10 увеличим число выборок.

```
samp3 <- matrix(rnorm(10000 * 1500, 0.1, 1), ncol = 10000)
pval3_t <- apply(samp3, 2, function(x)
t.test(x, mu = 0)$p.value)
plot(c(0,1), c(0,1), type="l", main ="Эмперическое распределение p-value для проверки состоятельности")
lines(ecdf(pval3_t), col='green')
pval3_w <- apply(samp3, 2, function(x)
wilcox.test(x, mu = 0)$p.value)
lines(ecdf(pval3_w), col='red')
legend("topleft", c("t-test", "wilcox.test"), title = "Критерий", fill=c("green", "red"))
```



При росте числа выборок мощность (против альтернативной гипотезы, что математическое ожидание равно 0.1) стремится к 1, значит оба критерия состоятельные (против альтернативной гипотезы, что математическое ожидание равно 0.1).

Выводы

Было продемонстрировано, что оба критерия являются точными.

Преимущества wilcox test:

- Более устойчив к выбросам, нечувствителен к единичным даже очень сильным выбросам.

Преимущества t-test:

- Является более мощным (при одинаковых альтернативных гипотезах).

Оба критерия состоятельные против альтернатив об отличном математическом ожидании.

Применение к реальным данным

```
library(GGally)
library(ggpubr)
```

Были выбраны данные, содержащие GRE.Score и TOEFL.Score студентов из различных вузов– https://www.kaggle.com/mohansacharya/graduate-admissions.

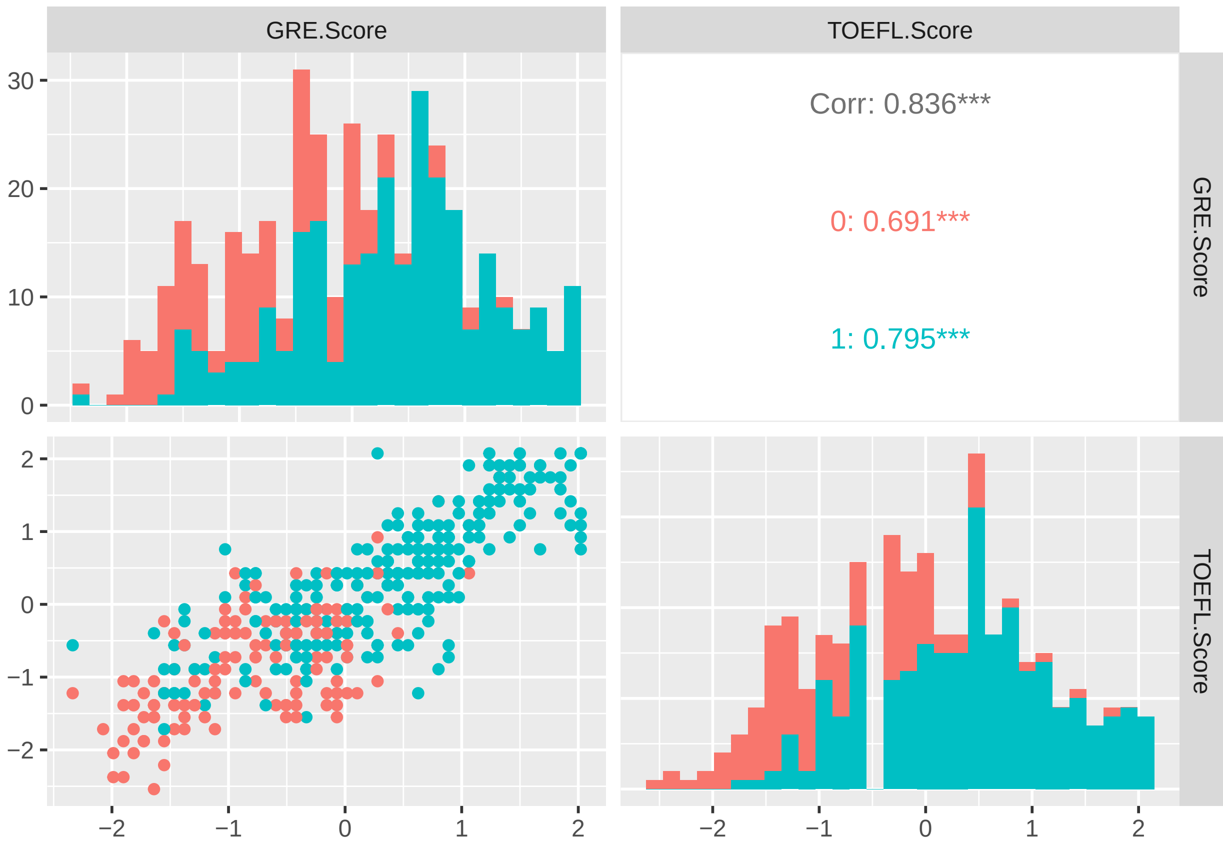
Значения GRE.Score и TOEFL.Score были стандартизованы, для того, чтобы они находились на одном интервале.

Группы индивидов выбирались следующим образом– в нулевую группу попали те студенты, которые обучались в университете с рейтингом ниже 3, а в первую- не меньше 3.

```
df <- read.csv("Admission_Predict.csv", header = TRUE, as.is = FALSE)
head(df)
```

```
##   Serial.No. GRE.Score TOEFL.Score University.Rating SOP LOR CGPA Research Chance.of.Admit
## 1          1       337         118                 4 4.5 4.5 9.65        1            0.92
## 2          2       324         107                 4 4.0 4.5 8.87        1            0.76
## 3          3       316         104                 3 3.0 3.5 8.00        1            0.72
## 4          4       322         110                 3 3.5 2.5 8.67        1            0.80
## 5          5       314         103                 2 2.0 3.0 8.21        0            0.65
## 6          6       330         115                 5 4.5 3.0 9.34        1            0.90
```

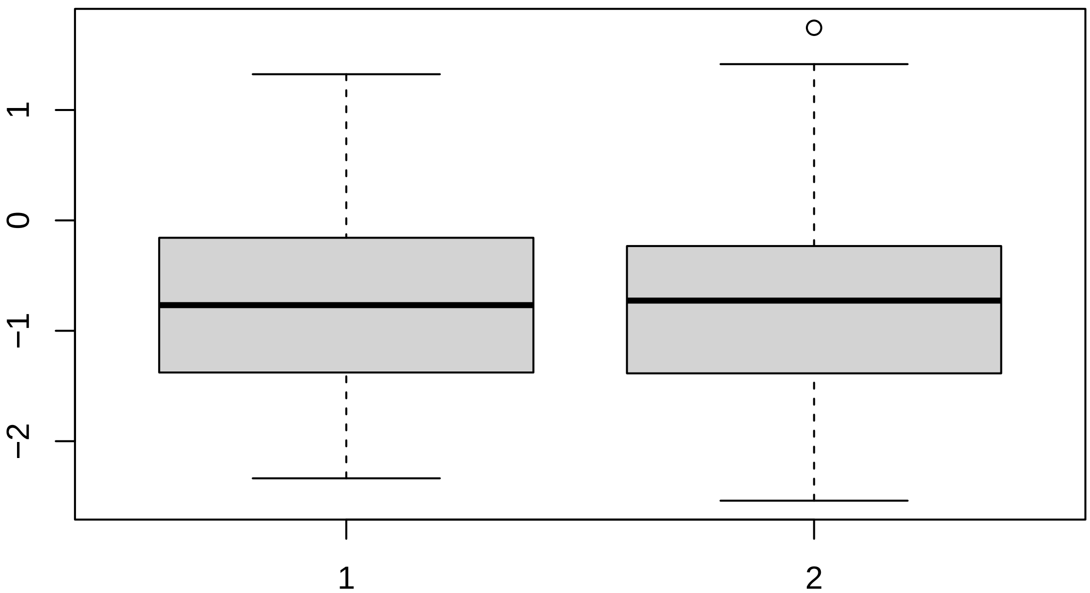
```
df$GRE.Score <- (df$GRE.Score-mean(df$GRE.Score))/sd(df$GRE.Score)
df$TOEFL.Score <- (df$TOEFL.Score-mean(df$TOEFL.Score))/sd(df$TOEFL.Score)
df$col[df$University.Rating > 2] <- '1'
df$col[df$University.Rating < 3] <- '0'
ggpairs(df, columns=c(2, 3), ggplot2::aes(colour=col), diag = list(continuous = "barDiag"))
```



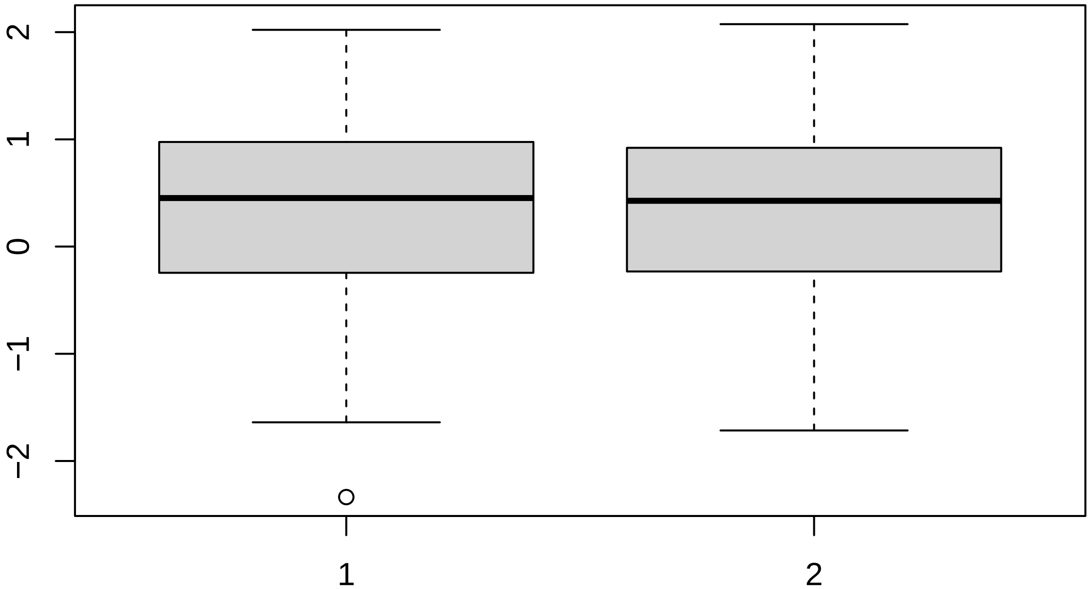
Выборка состоит из 400 наблюдений, можно говорить об асимптотическом схождении к стандартному нормальному распределению статистик критериев t-test и wilcox.test.

Посмотрим на boxplot данных.

```
boxplot(subset(df, col == "0")$GRE.Score, subset(df, col == "0")$TOEFL.Score)
```



```
boxplot(subset(df, col == "1")$GRE.Score, subset(df, col == "1")$TOEFL.Score)
```



Медианы выборок слабо отличаются, как и разброс данных, а также наблюдаются единичные потенциальные outliers.

Проверим гипотезу о равенстве средних при помощи t-test и гипотезу о равном сдвиге (так как формы распределения одинаковы) при помощи wilcox.test.

```
t.test(subset(df, col == "0")$GRE.Score, subset(df, col == "0")$TOEFL.Score)
```

```
##
##  Welch Two Sample t-test
##
## data:  subset(df, col == "0")$GRE.Score and subset(df, col == "0")$TOEFL.Score
## t = 0.16914, df = 263.82, p-value = 0.8658
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.1695944  0.2014690
## sample estimates:
##  mean of x  mean of y
## -0.7676287 -0.7835660
```

```
wilcox.test(subset(df, col == "0")$GRE.Score, subset(df, col == "0")$TOEFL.Score)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  subset(df, col == "0")$GRE.Score and subset(df, col == "0")$TOEFL.Score
## W = 8886, p-value = 0.9479
## alternative hypothesis: true location shift is not equal to 0
```

```
t.test(subset(df, col == "1")$GRE.Score, subset(df, col == "1")$TOEFL.Score)
```

```
##
##  Welch Two Sample t-test
##
## data:  subset(df, col == "1")$GRE.Score and subset(df, col == "1")$TOEFL.Score
## t = -0.10547, df = 531.73, p-value = 0.916
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.1558022  0.1399245
## sample estimates:
## mean of x mean of y
## 0.3823768 0.3903157
```

```
wilcox.test(subset(df, col == "1")$GRE.Score, subset(df, col == "1")$TOEFL.Score)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  subset(df, col == "1")$GRE.Score and subset(df, col == "1")$TOEFL.Score
## W = 36173, p-value = 0.767
## alternative hypothesis: true location shift is not equal to 0
```

Все 4 рассматриваемых гипотезы не отвергаются, так как p-value больше стандартного уровня значимости.