RUDN university

5100

# Программирование в среде R

Шевцов Василий Викторович,
директор ДИТ РУДН, shevtsov_vv@rudn.university

# Анализ номинативных данных

# Определения

Нулевая гипотеза — принимаемое по умолчанию предположение о том, что не существует связи между двумя наблюдаемыми событиями, феноменами. Так, нулевая гипотеза считается верной до того момента, пока нельзя доказать обратное.

Часто в качестве нулевой гипотезы выступают предположения об отсутствии взаимосвязи или корреляции между исследуемыми переменными, об отсутствии различий (однородности) в распределениях (параметрах распределений) в двух и/или более выборках. Для обозначения нулевой гипотезы часто используют символ $H_0$.

Причём крайним значением невозможного (маловероятного) считается от 0.01 до 0.05 или менее

# Определения

Уровень значимости — процент появления ошибок первого рода (отклонение верной нулевой гипотезы).

- первый уровень — 5% или 0.05, т. е. вероятность ошибиться 5 к 100 или 1 к 20.
- второй уровень — 1% или 0.01, т. е. вероятность 1 к 100.
- третий уровень — 0.1% или 0.001, вероятность 1 к 1000.

df <- read.csv("C:\\Users\\Администратор\\Downloads\\grants.csv")

| npersons | years_in_uni | oldest_age | field | RFCD.Code.1 | midpoint | status |
|---|---|---|---|---|---|---|
| 2 | < 5 | 66 | bio | 270799 | 24999.5 | 1 |
| 1 | < 5 | 51 | bio | 270106 | 24999.5 | 0 |
| 1 | < 5 | 36 | bio | 270708 | 24999.5 | 0 |
| 1 | 5-10 | 46 | bio | 270603 | 24999.5 | 0 |
| 4 | > 10 | 46 | physics | 240402 | 24999.5 | 1 |
| 2 | < 5 | 46 | bio | 270603 | 24999.5 | 0 |
| 1 | < 5 | 36 | chem | 250103 | 24999.5 | 1 |
| 2 | 5-10 | 45 | bio | 270603 | 24999.5 | 1 |

# table

table uses the cross-classifying factors to build a contingency table of the counts at each combination of factor levels.

Usage
table(...,
exclude = if (useNA == "no") c(NA, NaN),
useNA = c("no", "ifany", "always"),
dnn = list.names(...), deparse.level = 1)

```
> df <- read.csv("C:\\Users\\Администратор\\Downloads\\grants.csv")
> df$status <- factor(df$status,labels=c("not funded","funded"))
> t1 <- table(df$status)
> t1

not funded     funded
       747        673
```

```
> dim(t1)
[1] 2
> dim(t2)
[1] 2 5
```

```
> t2 <- table(df$status,df$field)
> t2

             beh_cog bio chem physics soc
  not funded     100 473   60      70  44
  funded          65 432   66      78  32
```

```
> t2 <- table(status=df$status,field=df$field)
> t2
            field
status       beh_cog bio chem physics soc
  not funded     100 473   60      70  44
  funded          65 432   66      78  32
```

# prop.table

```
> prop.table(t2)
            field
status          beh_cog        bio       chem    physics        soc
  not funded 0.07042254 0.33309859 0.04225352 0.04929577 0.03098592
  funded     0.04577465 0.30422535 0.04647887 0.05492958 0.02253521
> prop.table(t2,1)
            field
status          beh_cog        bio       chem    physics        soc
  not funded 0.13386881 0.63319946 0.08032129 0.09370817 0.05890228
  funded     0.09658247 0.64190193 0.09806835 0.11589896 0.04754829
> prop.table(t2,2)
            field
status          beh_cog       bio       chem   physics       soc
  not funded 0.6060606 0.5226519 0.4761905 0.4729730 0.5789474
  funded     0.3939394 0.4773481 0.5238095 0.5270270 0.4210526
```

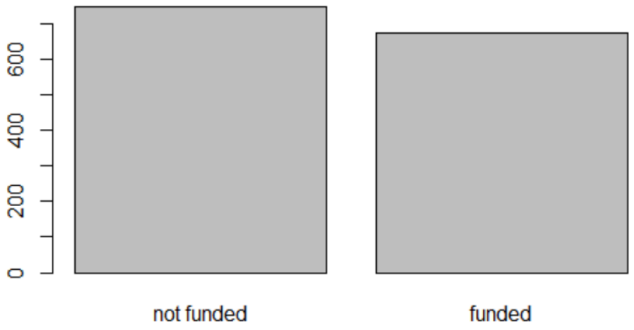
# table

```
> t3 <- table(years <- df$years_in_uni,field=df$field,status=df$status)
> t3
, , status = not funded

      field
       beh_cog bio chem physics soc
  < 5       57 198   31      20  22
  > 10      29 144   28      47  16
  5-10      14 131    1       3   6

, , status = funded

      field
       beh_cog bio chem physics soc
  < 5       27 180   41      22  14
  > 10      30 155   19      54  15
  5-10       8  97    6       2   3
```

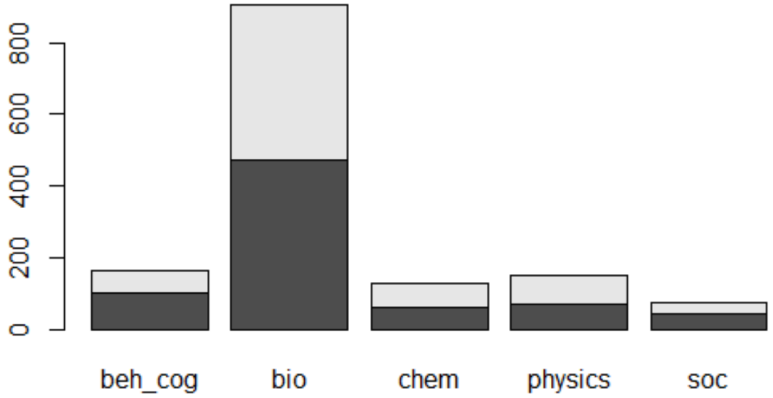
```
> dim(t3)
[1] 3 5 2
```

# barplot

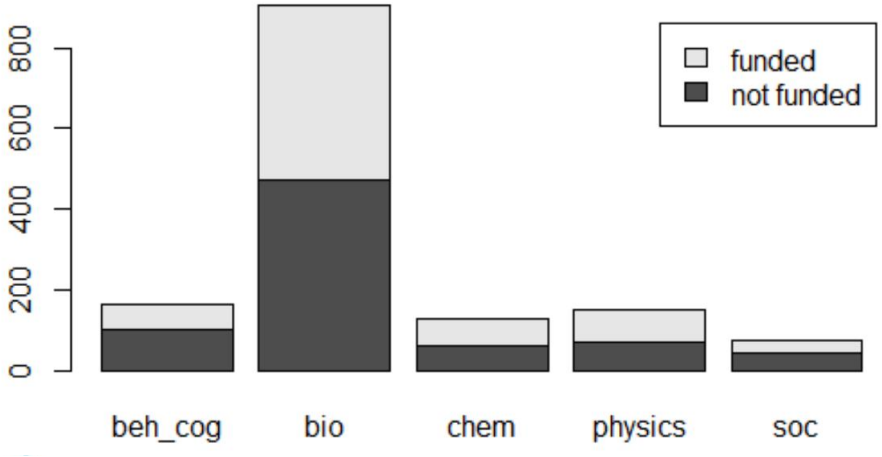
```
barplot(t1)
```



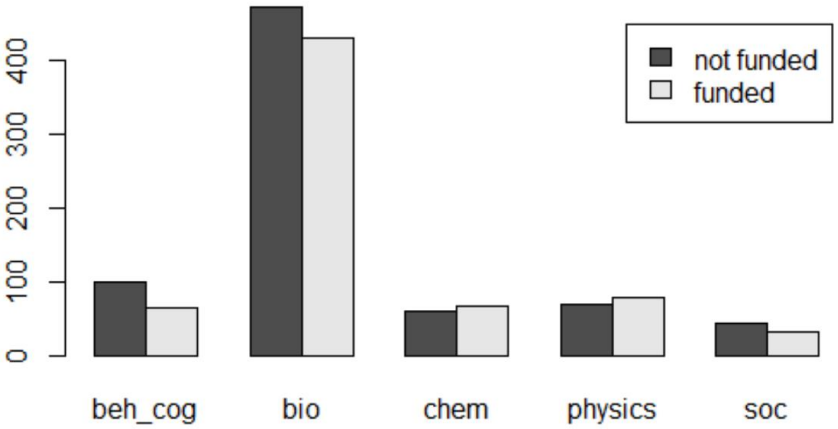

```
barplot(t2)
```




```
barplot(t2,legend.text = TRUE)
```

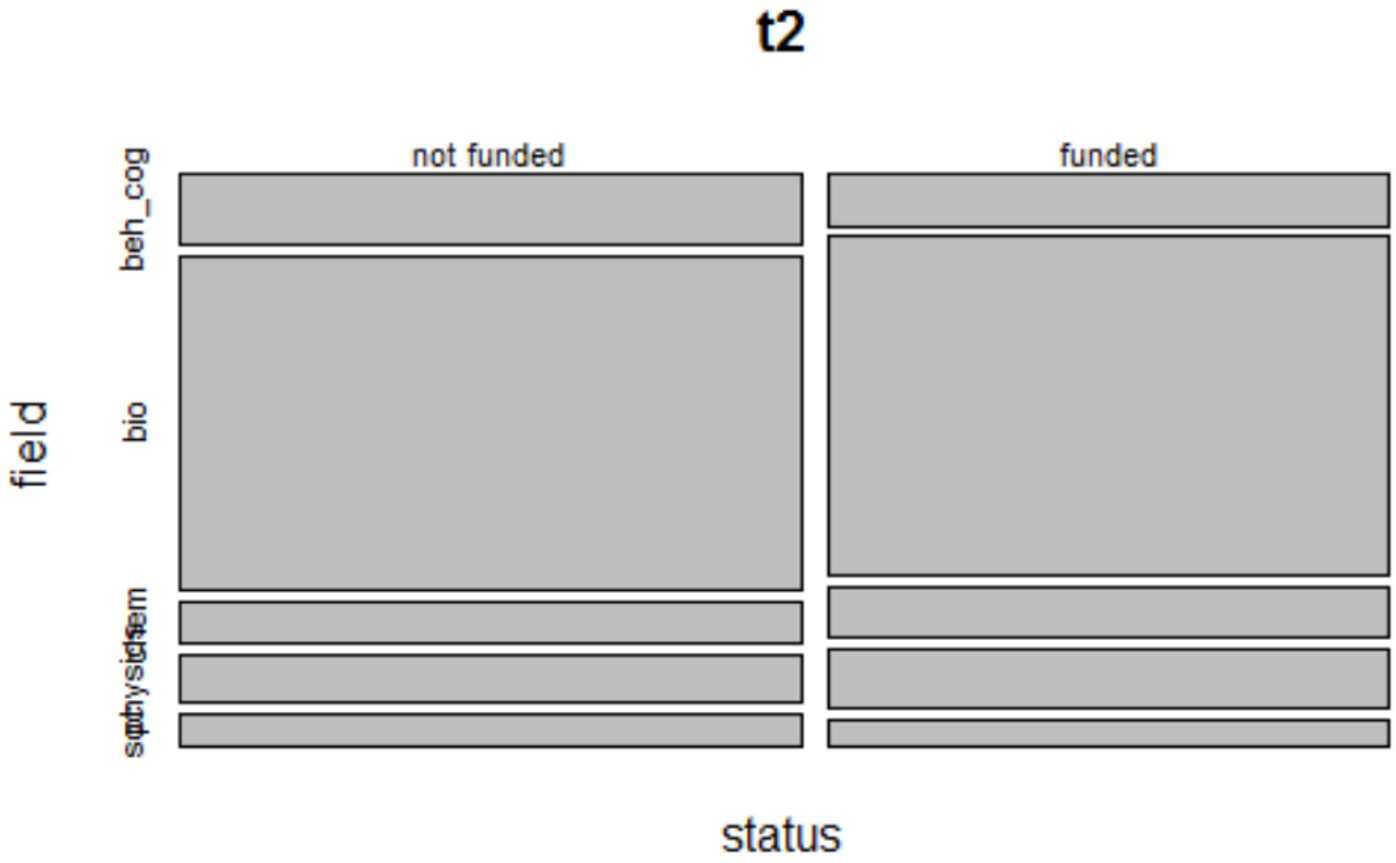



```
barplot(t2,legend.text = TRUE,beside = TRUE)
```



# mosaicplot

```
mosaicplot(t2)
```



# binomial test

Биномиальное распределение в теории вероятностей — распределение количества «успехов» в последовательности из n независимых случайных экспериментов, таких, что вероятность «успеха» в каждом из них постоянна и равна p.

```
> binom.test(x=5,n=20,p=0.5)

        Exact binomial test

data:  5 and 20
number of successes = 5, number of trials = 20, p-value = 0.04139
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.08657147 0.49104587
sample estimates:
probability of success
                  0.25
```

<0.05 нулевая гипотеза отвергается, принимается альтернативная
>0.05 нулевой гипотезой нельзя пренебречь

# binomial test

```
> t1

not funded     funded
       747        673
> binom.test(t1)

        Exact binomial test

data:  t1
number of successes = 747, number of trials = 1420, p-value = 0.05268
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.4997023 0.5523023
sample estimates:
probability of success
             0.5260563
```

# Chi-Square

```
> chisq.test(t1)

        Chi-squared test for given probabilities

data:  t1
X-squared = 3.8563, df = 1, p-value = 0.04956
```

```
> ch <- chisq.test(t1)
> str(ch)
List of 9
 $ statistic: Named num 3.86
  ..- attr(*, "names")= chr "X-squared"
 $ parameter: Named num 1
  ..- attr(*, "names")= chr "df"
 $ p.value  : num 0.0496
 $ method   : chr "Chi-squared test for given probabilities"
 $ data.name: chr "t1"
 $ observed : 'table' int [1:2(1d)] 747 673
  ..- attr(*, "dimnames")=List of 1
  .. ..$ : chr [1:2] "not funded" "funded"
 $ expected : Named num [1:2] 710 710
  ..- attr(*, "names")= chr [1:2] "not funded" "funded"
 $ residuals: 'table' num [1:2(1d)] 1.39 -1.39
  ..- attr(*, "dimnames")=List of 1
  .. ..$ : chr [1:2] "not funded" "funded"
 $ stdres   : 'table' num [1:2(1d)] 1.96 -1.96
  ..- attr(*, "dimnames")=List of 1
  .. ..$ : chr [1:2] "not funded" "funded"
 - attr(*, "class")= chr "htest"
> ch$expected
not funded     funded
       710        710
> ch$observed

not funded     funded
       747        673
```

# Определения

Количество степеней свободы — это количество значений в итоговом вычислении статистики, способных варьироваться. Иными словами, количество степеней свободы показывает размерность вектора из случайных величин, количество «свободных» величин, необходимых для того, чтобы полностью определить вектор.

```
> t2
             field
status        beh_cog bio chem physics soc
  not funded      100 473   60       70  44
  funded           65 432   66       78  32
> chisq.test(t2)

        Pearson's Chi-squared test

data:  t2
X-squared = 8.0601, df = 4, p-value = 0.0894
```

# Fisher's Exact Test

Точный тест Фишера — тест статистической значимости, используемый в анализе таблиц сопряжённости для выборок маленьких размеров. Назван именем своего изобретателя Р. Фишера. Относится к точным тестам значимости, поскольку не использует приближения большой выборки (асимптотики при размере выборки стремящемся к бесконечности).

```
> fisher.test(t2)

        Fisher's Exact Test for Count Data

data:  t2
p-value = 0.08921
alternative hypothesis: two.sided
```

# Сравнение двух групп

# iris

чашелистик лепесток

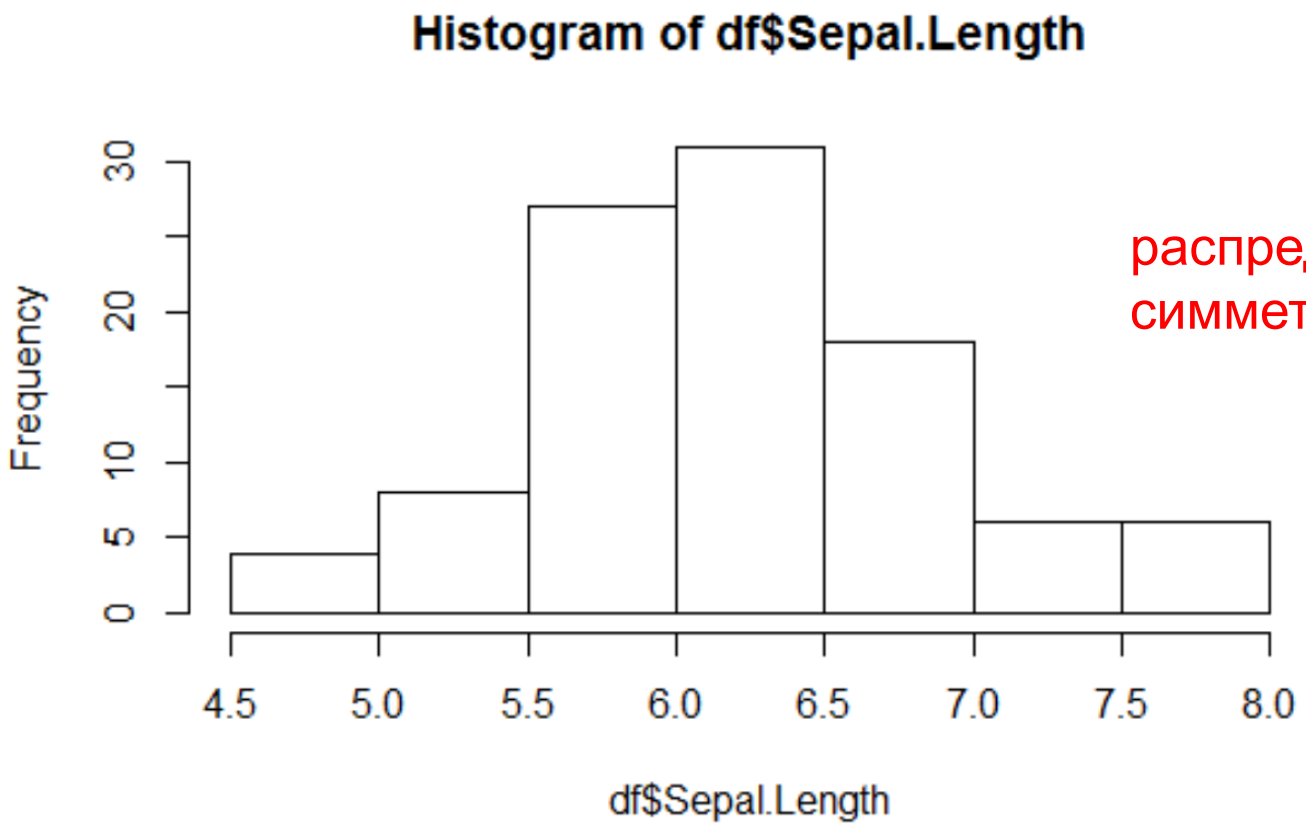
| Sepal.Length | Sepal.Width | Petal.Length | Petal.Width | Species |
|---|---|---|---|---|
| 5.1 | 3.5 | 1.4 | 0.2 | setosa |
| 4.9 | 3.0 | 1.4 | 0.2 | setosa |
| 4.7 | 3.2 | 1.3 | 0.2 | setosa |
| 4.6 | 3.1 | 1.5 | 0.2 | setosa |
| 5.0 | 3.6 | 1.4 | 0.2 | setosa |
| 5.4 | 3.9 | 1.7 | 0.4 | setosa |
| 4.6 | 3.4 | 1.4 | 0.3 | setosa |
| 5.0 | 3.4 | 1.5 | 0.2 | setosa |
| 4.4 | 2.9 | 1.4 | 0.2 | setosa |
| 4.9 | 3.1 | 1.5 | 0.1 | setosa |
| 5.4 | 3.7 | 1.5 | 0.2 | setosa |

# Выборка двух групп

```
> df <- subset(df,Species!="setosa")
> df <- iris
> str(df)
'data.frame':   150 obs. of  5 variables:
 $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
 $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
 $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
 $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
 $ Species     : Factor w/ 3 levels "setosa","versicolor",..: 1 1 1 1 1 1 1 1 1 1 ...
> df <- subset(df,Species!="setosa")
> str(df)
'data.frame':   100 obs. of  5 variables:
 $ Sepal.Length: num  7 6.4 6.9 5.5 6.5 5.7 6.3 4.9 6.6 5.2 ...
 $ Sepal.Width : num  3.2 3.2 3.1 2.3 2.8 2.8 3.3 2.4 2.9 2.7 ...
 $ Petal.Length: num  4.7 4.5 4.9 4 4.6 4.5 4.7 3.3 4.6 3.9 ...
 $ Petal.Width : num  1.4 1.5 1.5 1.3 1.5 1.3 1.6 1 1.3 1.4 ...
 $ Species     : Factor w/ 3 levels "setosa","versicolor",..: 2 2 2 2 2 2 2 2 2 2 ...
> table(df$Species)

    setosa versicolor  virginica
         0         50         50
```

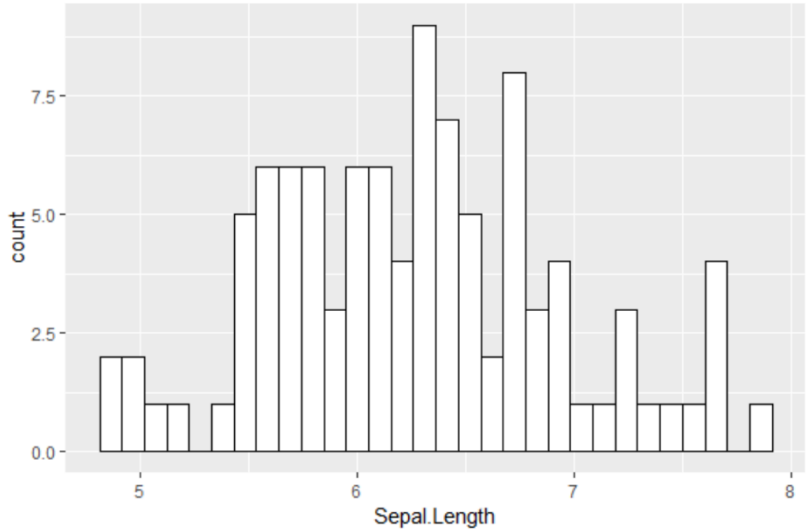
# Распределение длины чашелистика

```
hist(df$Sepal.Length)
```

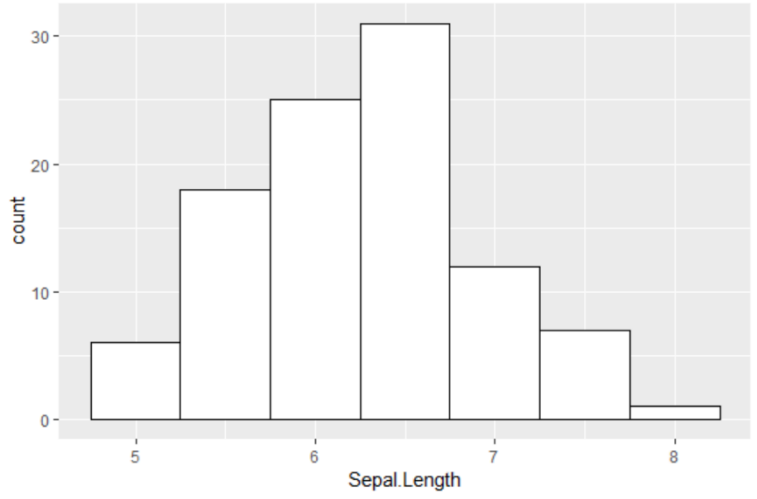



распределение - симметричное

```
> ggplot(df, aes(x=Sepal.Length))+
+   geom_histogram(fill="white",col="black")
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

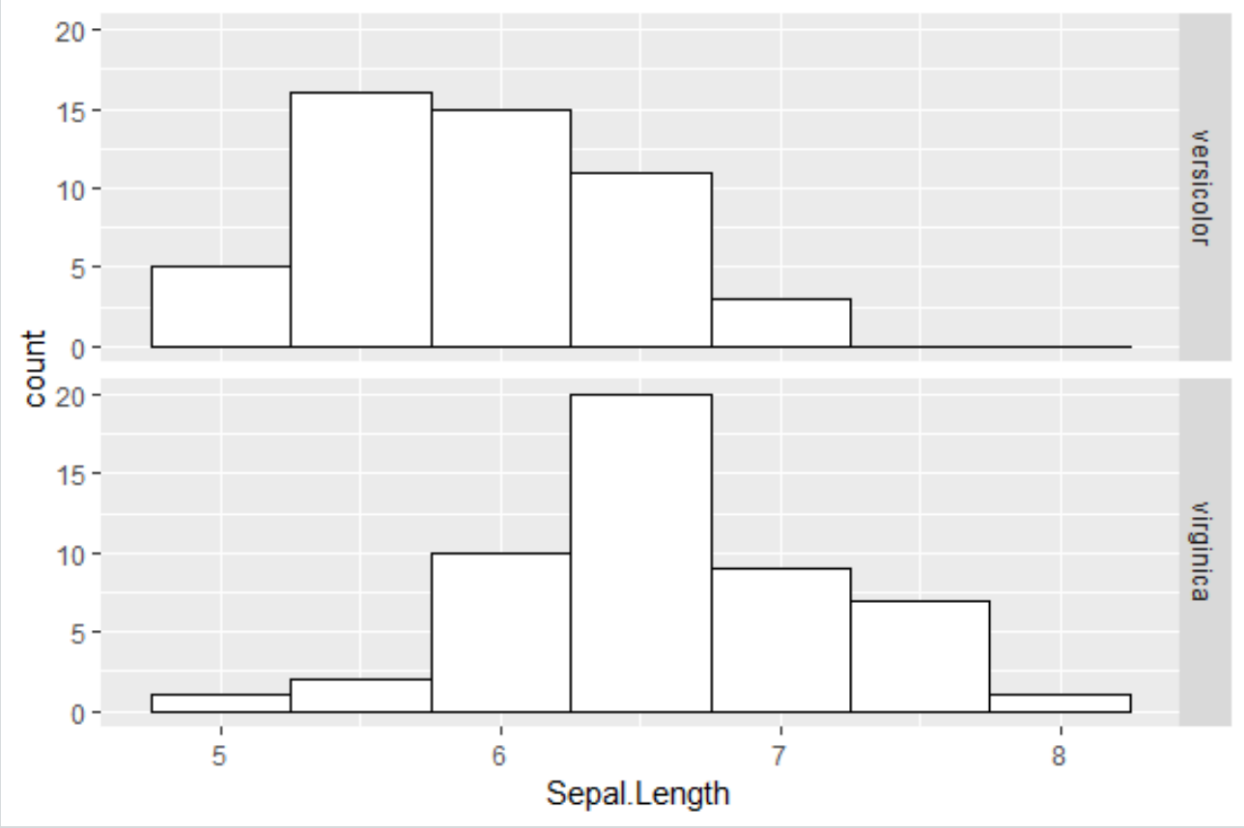



```
> ggplot(df, aes(x=Sepal.Length))+
+   geom_histogram(fill="white",col="black",binwidth = 0.5)
```



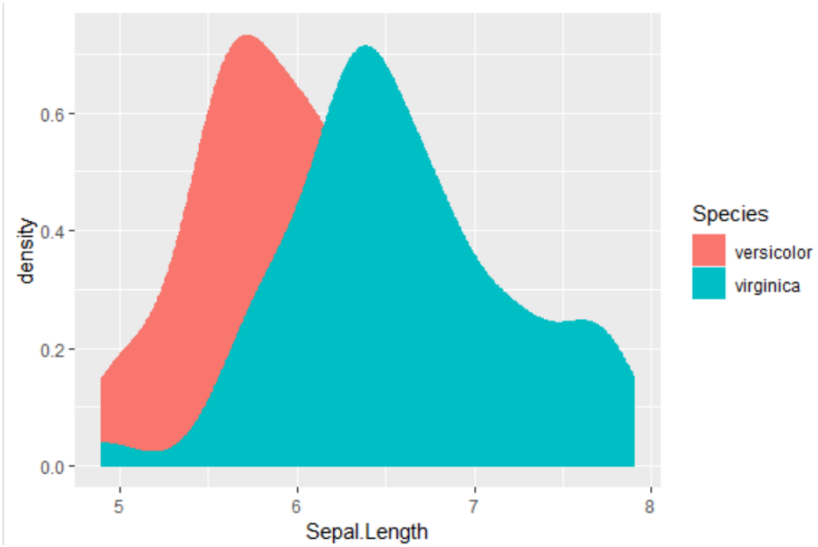
# Гистограмма, распределение по группам

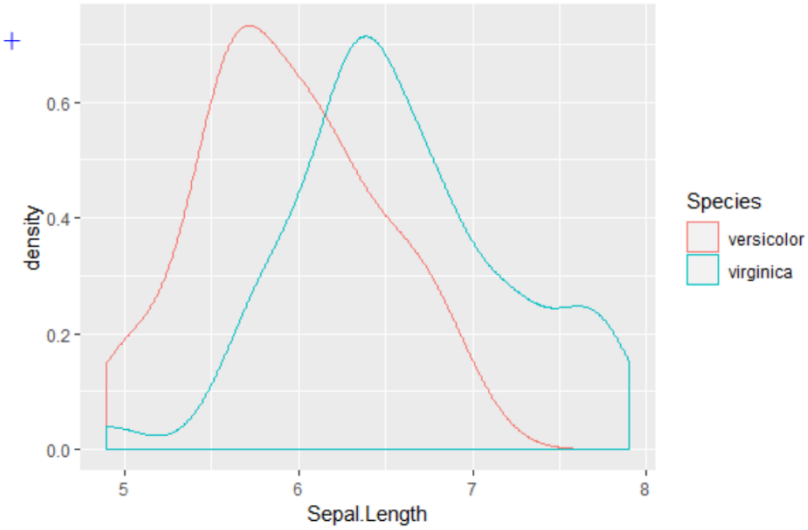
```
> ggplot(df, aes(x=Sepal.Length))+
+   geom_histogram(fill="white",col="black",binwidth = 0.5)+
+   facet_grid(Species~.)
```

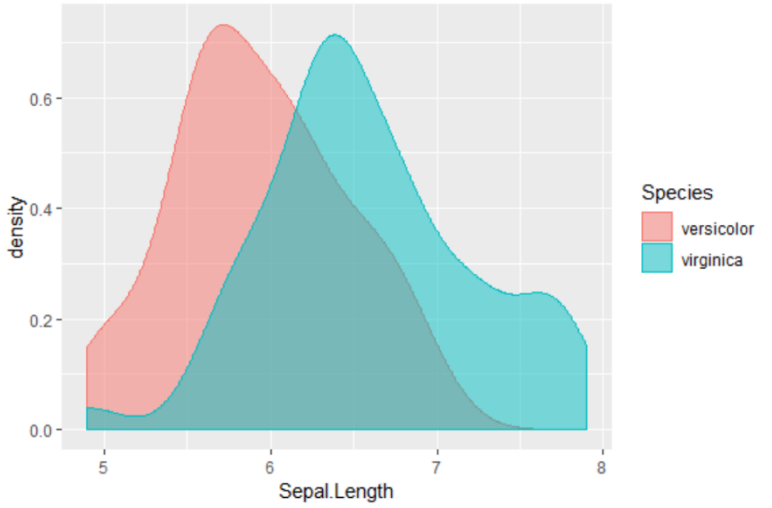


# Плотность распределения, по группам

```
> ggplot(df, aes(x=Sepal.Length,col=Species))+
+   geom_density()
```

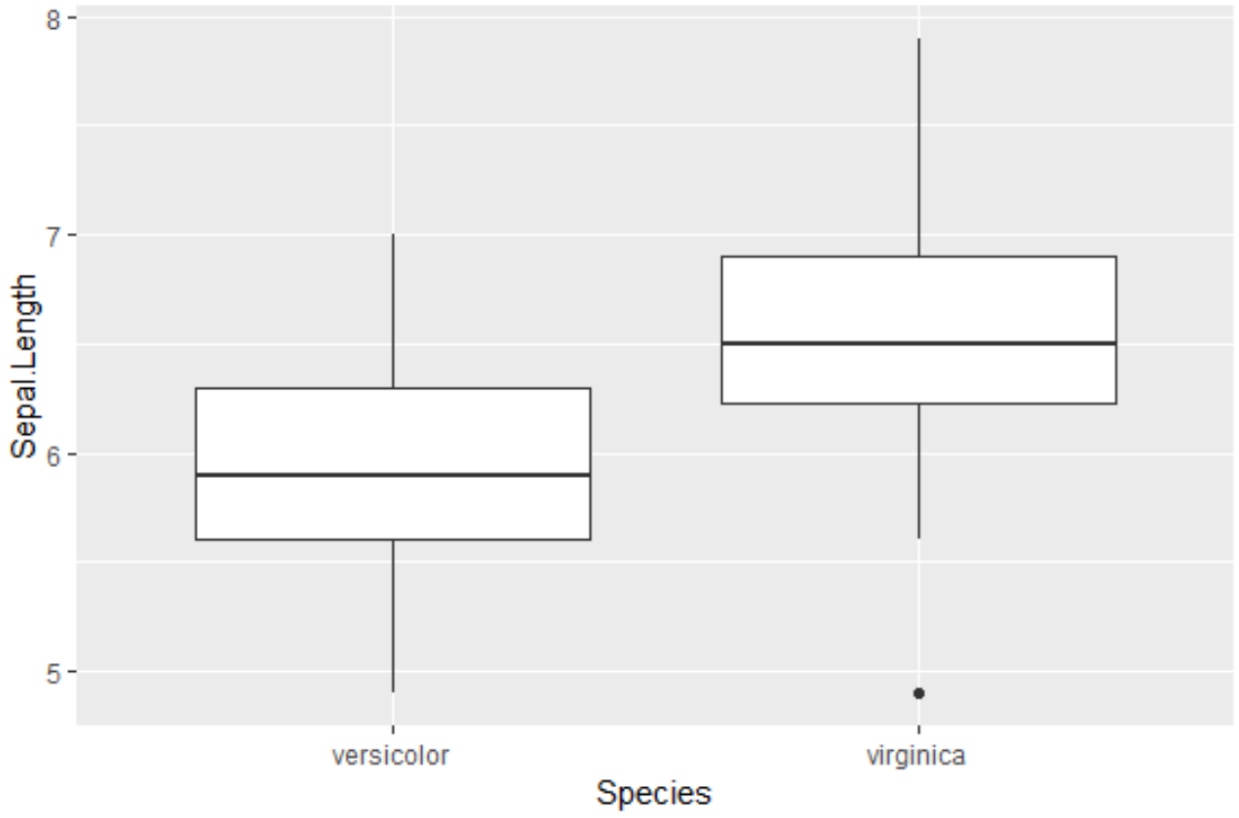


# Ящик с усами

```
> ggplot(df,aes(Species,Sepal.Length))+
+    geom_boxplot()
```



# Определение статистической значимости

# Определения

В статистике величину (значение) переменной называют статистически значимой, если мала вероятность случайного возникновения этой или ещё более крайних величин. Здесь под крайностью понимается степень отклонения тестовой статистики от нуль-гипотезы.

Разница называется статистически значимой, если появление имеющихся данных (или ещё более крайних данных) было бы маловероятно, если предположить, что эта разница отсутствует; это выражение не означает, что данная разница должна быть велика, важна, или значима в общем смысле этого слова.

Этот критерий позволяют исследователю оценить вероятность того, что результаты могли появиться чисто случайно.

# Определения. Нормальное распределение

Стандартным нормальным распределением называется нормальное распределение с математическим ожиданием μ = 0 и стандартным отклонением σ = 1.


RUDN university

5100

# Определения. Гомогенность дисперсий

Вторым важным условием применимости классического дисперсионного анализа является однородность (также "гомоскедастичность") групповых дисперсий (англ. homogeneity of variance, или homoscedasticity of variance).

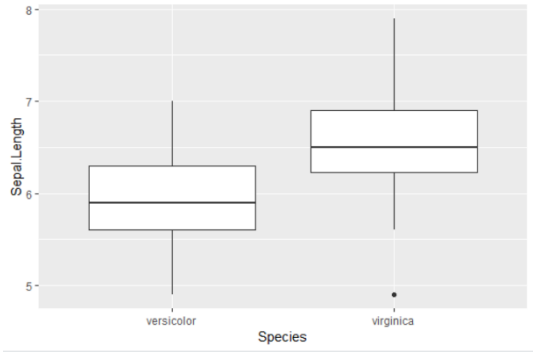
Речь здесь идет о том, что помимо нормального распределения в каждой группе, значения зависимой переменной должны также иметь одинаковую степень разброса. Необходимость выполнения этого условия определяется способом вычисления внутри- и межгрупповых дисперсий, применяемым в классическом дисперсионном анализе: при значительно различающихся групповых дисперсиях используемые формулы просто не будут работать корректно.

# Проверка на нормальное распределение

```
> shapiro.test(x=df$Sepal.Length)

        Shapiro-Wilk normality test

data:  df$Sepal.Length
W = 0.98054, p-value = 0.1464
```

уровень значимости >0.05, нулевая гипотеза об отсутствии нормального распределения не подтверждается

```
> shapiro.test(x=df$Sepal.Length[df$Species=="versicolor"])

        Shapiro-Wilk normality test

data:  df$Sepal.Length[df$Species == "versicolor"]
W = 0.97784, p-value = 0.4647
```



# Проверка на гомогенность дисперсий

```
> bartlett.test(Sepal.Length~Species,df)

        Bartlett test of homogeneity of variances

data:  Sepal.Length by Species
Bartlett's K-squared = 2.0949, df = 1, p-value = 0.1478
```

уровень значимости >0.05, нулевая гипотеза об отсутствии гомогенности дисперсий не подтверждается

# Student's t-Test

t-критерий Стьюдента — общее название для класса методов статистической проверки гипотез (статистических критериев), основанных на распределении Стьюдента. Наиболее частые случаи применения t-критерия связаны с проверкой равенства средних значений в двух выборках.

```
> t.test(Sepal.Length~Species,df)

        Welch Two Sample t-test

data:  Sepal.Length by Species
t = -5.6292, df = 94.025, p-value = 1.866e-07
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.8819731 -0.4220269
sample estimates:
mean in group versicolor  mean in group virginica
                   5.936                    6.588
```

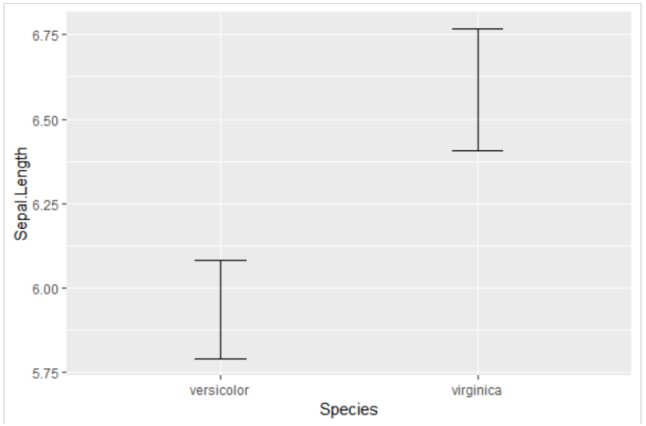
доверительный интервал с указанием разницы между средними групп

уровень значимости <0.05, нулевая гипотеза о равенстве средних в группах не подтверждается

# Графическое представление

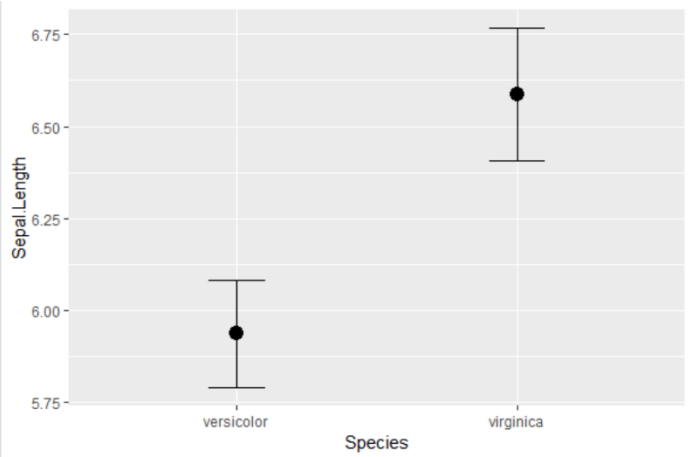
```
> ggplot(df,aes(Species,Sepal.Length))+
+   stat_summary(fun.data = mean_cl_normal,geom="errorbar",width=0.2)
```

вывод доверительных интервалов

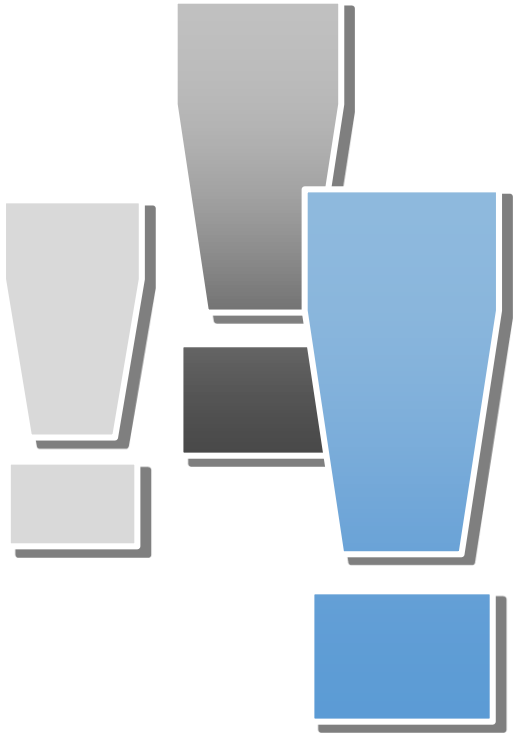



```
> ggplot(df,aes(Species,Sepal.Length))+
+   stat_summary(fun.data = mean_cl_normal,geom="errorbar",width=0.2)+
+   stat_summary(fun.y = mean,geom="point",size=4)
```

вывод доверительных интервалов
и среднего значения



# Спасибо за внимание!



Шевцов Василий Викторович

shevtsov_vv@rudn.university
+7(903)144-53-57