

# Статистическа обработка на данните

## Статистическа обработка на данни с R

Пламен Петров и Тодор Балабанов

Център за обучение  
Институт по информационни и комуникационни технологии  
Българската академия на науките

*p.petrov@iit.bas.bg todorb@iinf.bas.bg*

5.VI.2020

## Acknowledgments

These teaching materials are funded by Velbazhd Software LLC and it is partially supported by the Bulgarian Ministry of Education and Science (contract D01–205/23.11.2018) under the National Scientific Program “Information and Communication Technologies for a Single Digital Market in Science, Education and Security (ICTinSES)”, approved by DCM # 577/17.08.2018.

# Съдържание

- 1 Описателна статистика
  - Групиране и разпръскване
- 2 Сравнителна статистика
  - Корелация и ковариация
  - Тест на Стюдънт
    - Тест на две извадки
    - Сдвоен тест на две извадки
- 3 Дисперсионен анализ
  - Сравнение при повече от две групи данни
- 4 Линейна регресия
  - Формули на линейната регресия
- 5 Заклучение
  - Дискусия

# Описателна статистика

# Централно групиране на данните и разпръскване

## Генериране на извадка от случайни числа

```
v1 <- round( rnorm(100, mean=62, sd=72) )
v2 <- v1
v2[sample(x=1:100, size=15, replace=FALSE)] <- NA
w1 <- 1 / sample(x=1:100, size=100, replace=TRUE)
```

# Най-често използвана централна статистика

## Средна стойност

```
sum(v1) / length(v1)
mean( v1 )
mean( v2 )
mean(v2, na.rm=TRUE)
```

## Претеглена средна стойност

```
weighted.mean(x=v1, w=w1)
```

# Минимум, максимум, медиана и мода

## Гранични статистики

```
min( v1 )
max( v1 )
```

## Други централни статистики

```
median( v1 )
unique(v1)[ which.max( tabulate(match(v1,unique(v1))) ) ]
```

# Статистики за разпръскване

## Дисперсия и стандартно отклонение

```
sum( (v1-mean(v1))^2 ) / (length(v1) - 1)
var( v1 )
sqrt( var(v1) )
sd( v1 )
```





# Сравнителна статистика

# Търсене на взаимни връзки

## Корелация между две променливи

```
cor(economics$psavert, economics$pce)
sum((economics$psavert-mean(economics$psavert)) *
(economics$pce-mean(economics$pce))) / ((nrow(economics)-1) *
sd(economics$psavert) * sd(economics$pce))
cor(economics[, c("pce", "psavert", "uempmed", "unemploy")])
```

# Търсене на взаимни връзки

## Ковариация между две променливи

```
cov(economics$psavert, economics$pce)
cov(economics[, c("pce", "psavert", "uempmed", "unemploy")])
identical(cov(economics$psavert, economics$pce),
cor(economics$psavert, economics$pce) * sd(economics$psavert) *
sd(economics$pce))
```

# Корелационен коефициент

$$r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y} \quad (1)$$

# Ковариационен коефициент

$$v_{xy} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) \quad (2)$$

# Сравнения на параметри

```
library(reshape2)
```

Тест на единична извадка

```
t.test(tips$tip, alternative="two.sided", mu=3.50)
```

Едностранна t-статистика

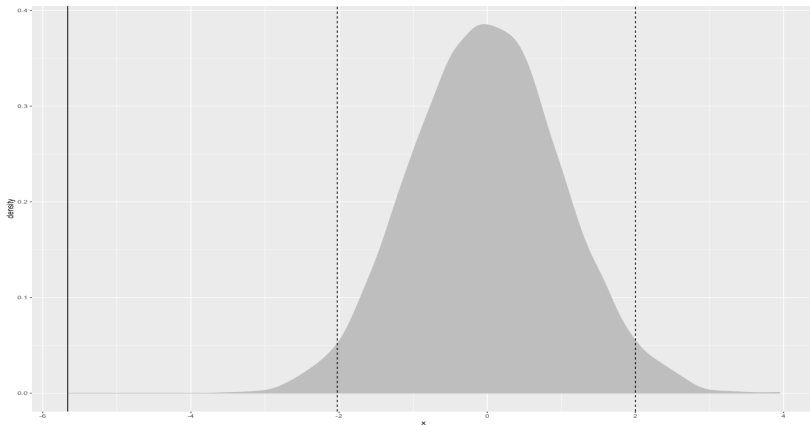
```
t.test(tips$tip, alternative="less", mu=3.50)
```

# T-статистика

$$t = \frac{\bar{X} - \mu_0}{s_{\bar{X}}/\sqrt{n}} \quad (3)$$



# Т-статистика за бакшиши



## Визуализация на t-разпределение

```
library( ggplot2 )
values <- rt(10000, df = NROW( tips ) - 1)
t <- t.test(tips$tip, alternative="two.sided", mu=3.50)
ggplot(data.frame(x=values)) + geom_density(aes(x=x),
fill="grey", color="grey") + geom_vline(xintercept=t$statistic) +
geom_vline(xintercept=mean(values) + c(-2, 2) * sd(values),
linetype=2)
```

## Сравнение на две извадки

### С прилагане на тест за нормално разпределение

```
ggplot(tips, aes(x=tip, fill=sex)) +
  geom_histogram(binwidth=1.0, alpha=0.8)
aggregate(tip~sex, data=tips, var)
shapiro.test(tips$tip[tips$sex == "Female"])
shapiro.test(tips$tip[tips$sex == "Male"])
shapiro.test( tips$tip )
ansari.test(tip~sex, tips)
t.test(tip~sex, data=tips, var.equal=TRUE)
```

## Визуализация на двете извадки

```
library( plyr )
ddply(tips, " sex", summarize, tip.mean=mean(tip), tip.sd=sd(tip),
Lower=tip.mean-2*tip.sd/sqrt(NROW(tip)),
Upper=tip.mean+2*tip.sd/sqrt(NROW(tip)))
ggplot(ddply(tips, " sex", summarize, tip.mean=mean(tip),
tip.sd=sd(tip), Lower=tip.mean-2*tip.sd/sqrt(NROW(tip)),
Upper=tip.mean+2*tip.sd/sqrt(NROW(tip))), aes(x=tip.mean,
y=sex)) + geom_point() + geom_errorbarh(aes(xmin=Lower,
xmax=Upper), height=.2)
```



## Сравнение по двойки

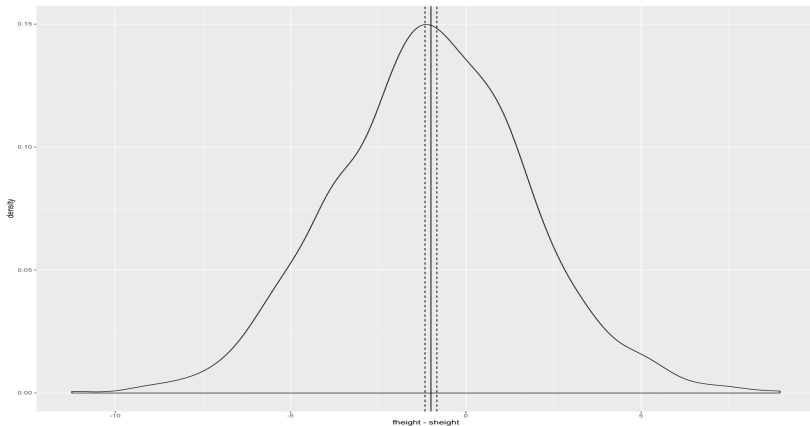
### Т-тест на сдвоени данни

```
library( UsingR )
t.test(father.son$sheight, father.son$fheight, paired=TRUE)
```

### Визуализация

```
ggplot(father.son, aes(x=fheight-sheight)) + geom_density() +
geom_vline(xintercept=mean(father.son$fheight-
father.son$sheight)) + geom_vline(xintercept =
mean(father.son$fheight-father.son$sheight) + 2*c(-1, 1 )
*sd(father.son$fheight-father.son$sheight)/sqrt(nrow(father.son)),
linetype=2)
```

# Визуализация на средните при сдвоения тест



# Дисперсионен анализ



## F-статистика за ANOVA

$$F = \frac{\sum_i n_i (\bar{Y}_i - \bar{Y})^2 / (K - 1)}{\sum_{ij} (Y_{ij} - \bar{Y}_i)^2 / (N - K)} \quad (4)$$

# Сравнение на дисперсиите

## ANOVA тест

```
library( plyr )
library( reshape2 )
aov(tip~day-1, tips)
aov(tip~day-1,tips)$coefficients
summary( aov(tip~day-1,tips) )
ddply(tips, "day", plyr::summarize, tip.mean=mean(tip),
tip.sd=sd(tip), Length=NROW(tip), tfrac=qt(p=.90,
df=Length-1), Lower=tip.mean - tfrac*tip.sd/sqrt(Length),
Upper=tip.mean + tfrac*tip.sd/sqrt(Length))
summary( lm(tip~day-1,data=tips) )
```

# Линейна регресия

# Уравнение, наклон и срез на права

$$y = ax + b + \epsilon \quad (5)$$

$$a = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad (6)$$

$$b = \bar{y} - a \quad (7)$$

# Изчисление на леинейна регресия

## Наклон и срез

```
library(ggplot2)
library(UsingR)
ggplot(father.son, aes(x=fheight, y=sheight)) + geom_point() +
geom_smooth(method="lm") + labs(x="Fathers", y="Sons")
lm(sheight~fheight, data=father.son)
summary( lm(sheight~fheight,data=father.son) )
```



# Заклучение

## Въпроси и отговори

Благодаря за вниманието!