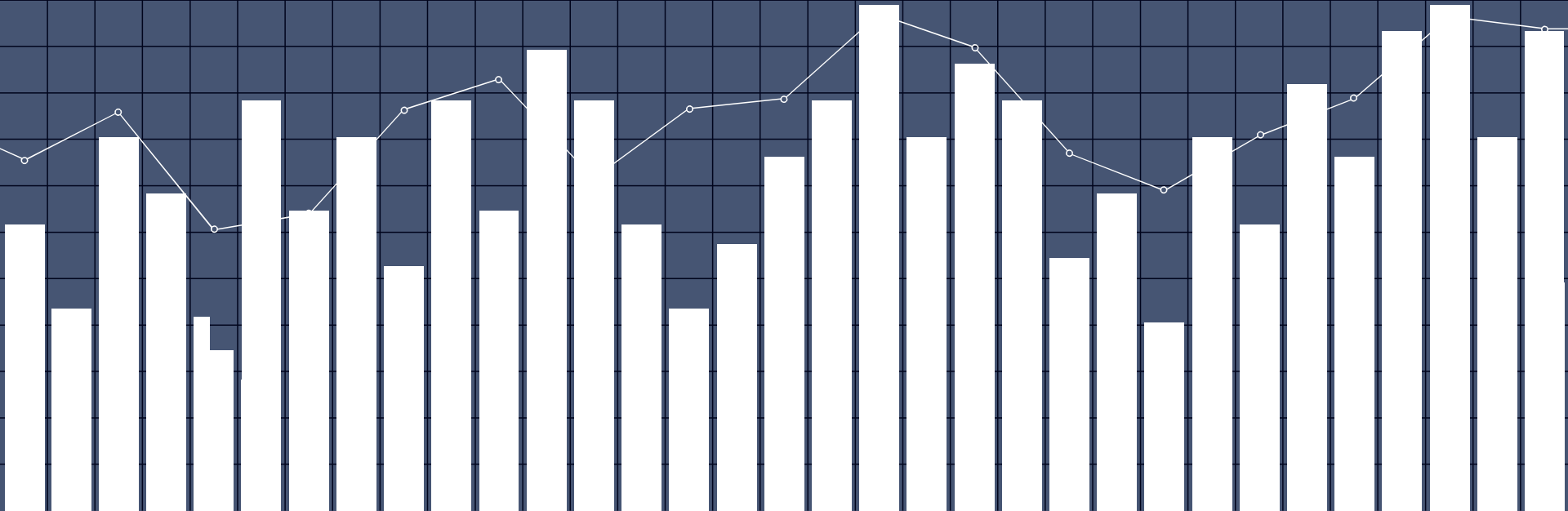
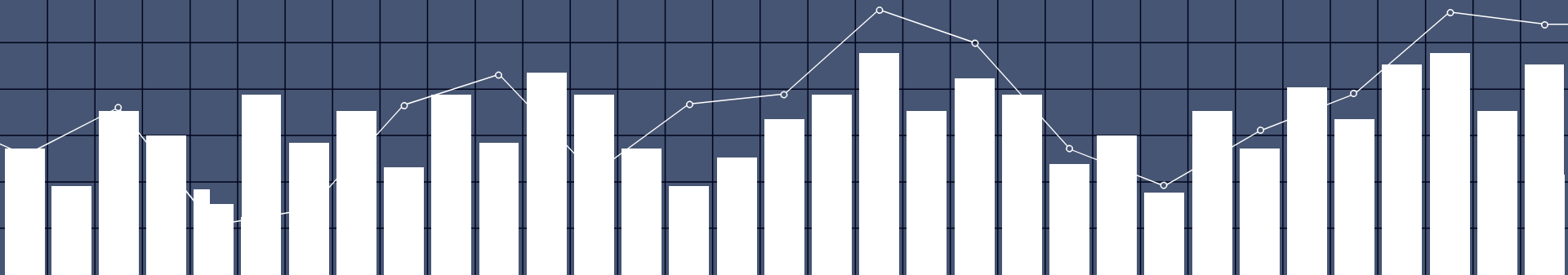
08. Srovnání skupin



Harmonogram

- 01. t-testy
- 02. (One-Way) ANOVA



Srovnání dvou průměrů (dle Conway, n.d.)

Dependent t-test - úvod

$$t = \frac{\bar{x}_D}{s_D/\sqrt{n}}$$

n is just the sample size, or the number of individuals in our sample. $\bar{x}_D$ is the mean of the difference scores, or sum of the difference scores divided by the sample size. Finally, s_D is the standard deviation of the difference scores:

$$s_D = \sqrt{\frac{\sum (x_D - \bar{x}_D)^2}{n-1}}$$

In the formula for s_D, x_D are the individual difference scores and should not be confused with $\bar{x}_D$, which is the mean of the difference scores.

Předpoklady použití:

- The sampling distribution is normally distributed. In the dependent t-test this means that the sampling distribution of the differences between scores should be normal, not the scores themselves.
- Data are measured at least at the interval level.

Srovnání dvou průměrů (dle Conway, n.d.)

Dependent t-test

```
# Bydleni_Brno <- read_excel("Bydleni_Brno.xlsx")

# In the case of our dependent t-test, we need to specify these arguments to t.test():
 ?t.test

# x: Column of Bydleni_Brno containing prices for 2015
# y: Column of Bydleni_Brno containing prices for 2016 paired: Whether we're doing a dependent (i.e.
# paired) t-test or independent t-test. In this example, it's TRUE
# Note that t.test() carries out a two-sided t-test by default

# Conduct a paired t-test using the t.test function
t.test(Bydleni_Brno$Pronajem_m2_2015,  Bydleni_Brno$Pronajem_m2_2016, paired = TRUE)
```


WHEN YOU PLAY THE GAME OF NULL HYPOTHESIS SIGNIFICANCE TESTING, YOU WIN OR YOU DIE.

Srovnání dvou průměrů (dle Conway, n.d.)

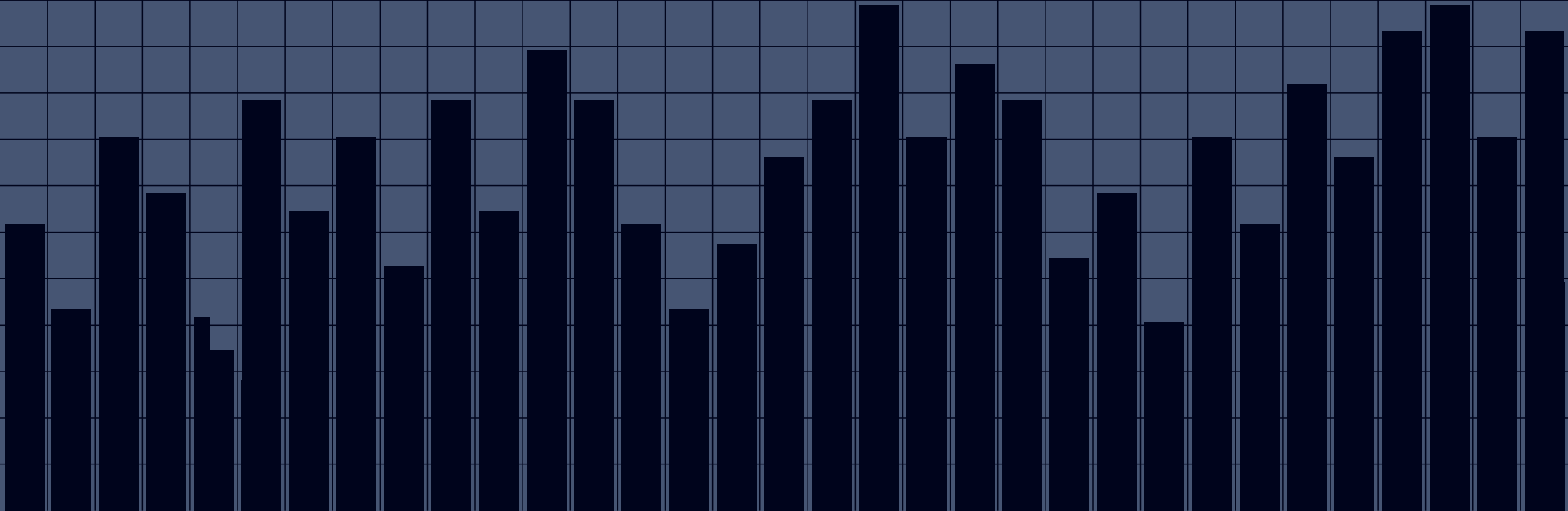
Dependent t-test – Cohenovo d – lsr

```
library("lsr")

# For cohensD(), we'll need to specify three arguments:
# x: Column of wm_t containing post-training intelligence scores
# y: Column of wm_t containing pre-training intelligence scores
# method: Version of Cohen's d to compute, which should be "paired" in this case

?cohensD()
```

Srovnání dvou průměrů (dle Conway, n.d.)

Dependent t-test – Cohenovo d – lsr – příklad

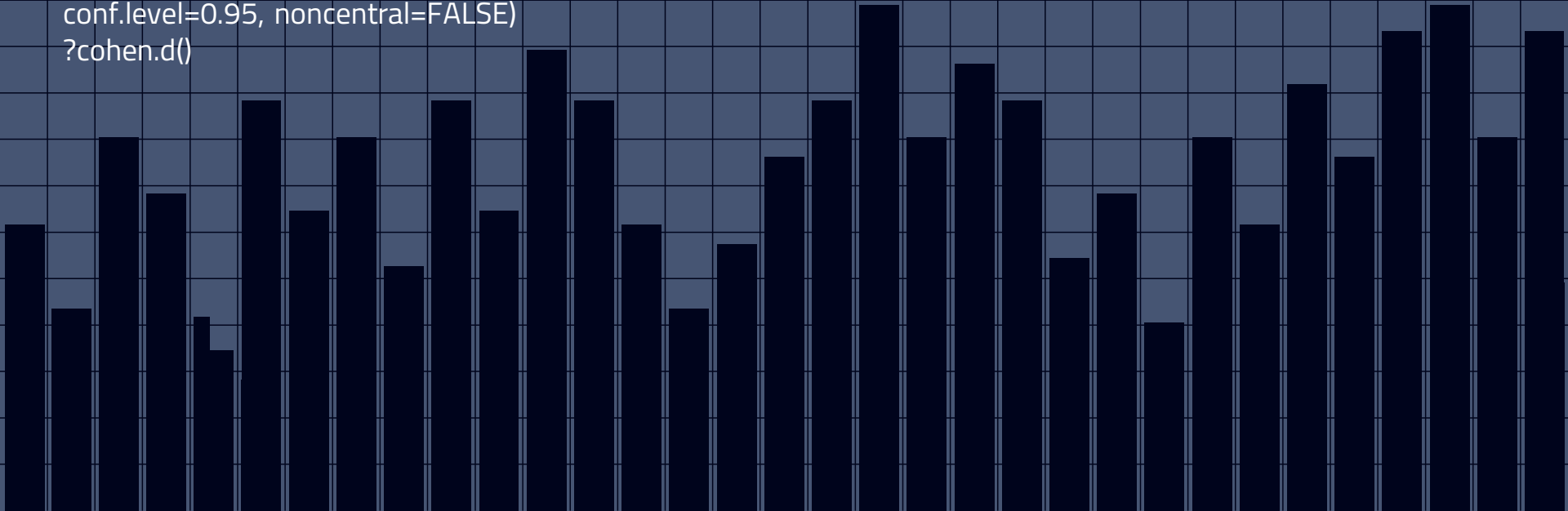
```
# Calculate Cohen's d
cohensD(Bydleni_Brno$ Pronajem_m2_2015, Bydleni_Brno$ Pronajem_m2_2016, method = "paired")
```



Srovnání dvou průměrů (dle Conway, n.d.)

Dependent t-test – Cohenovo d – effsize – argumenty

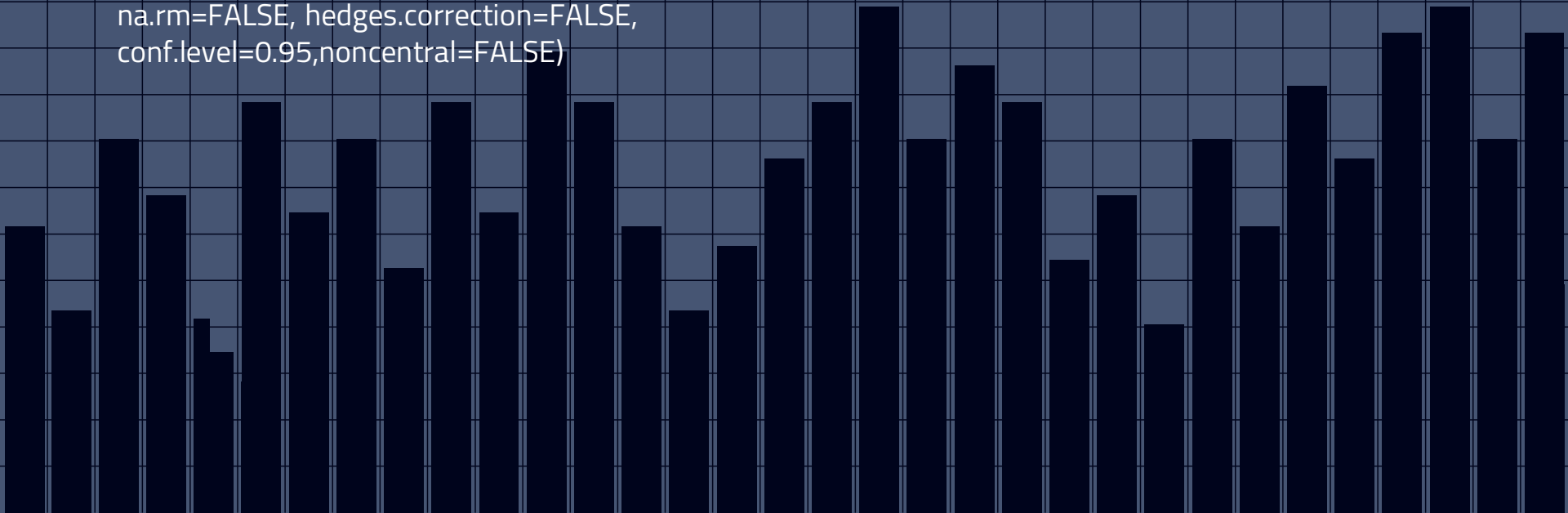
```
library("effsize")

cohen.d(x, y, pooled=TRUE, paired=TRUE,
na.rm=FALSE, hedges.correction=FALSE,
conf.level=0.95, noncentral=FALSE)
?cohen.d()
```



Srovnání dvou průměrů (dle Conway, n.d.)

Dependent t-test – Cohenovo d – effsize – příklad

```
library("effsize")

cohen.d(Bydleni_Brno$Pronajem_m2_2015, Bydleni_Brno$Pronajem_m2_2016,
      pooled=TRUE,paired=TRUE,
      na.rm=FALSE, hedges.correction=FALSE,
      conf.level=0.95,noncentral=FALSE)
```



Srovnání dvou průměrů

Cohenovo d – Guess

Guess

Display Options

- [ ] Violin Plot
- [x] BoxPlot
- [ ] Points

This app is not storing your data beyond this session.

This app will show you a graph of simulated data with a random number of observations in each of two groups and a random effect size. The effect size will be between -3 and 3, your job is to guess the size of the effect.

How much larger is group B than group A?

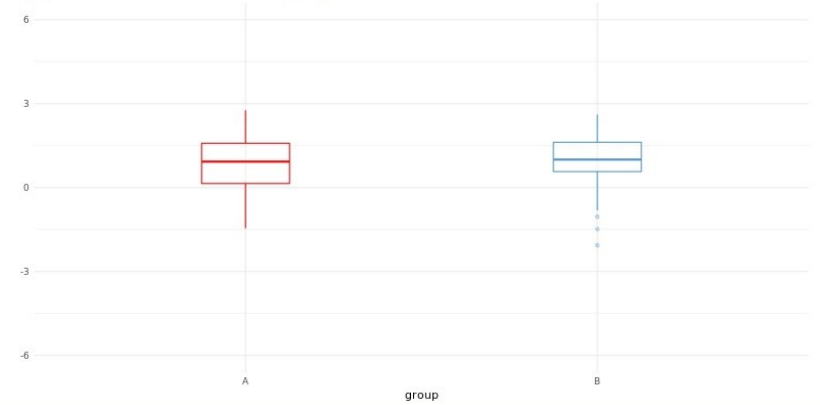
My effect size (d) guess

-3 0 3

-3 -2.4 -1.8 -1.2 -0.6 0 0.6 1.2 1.8 2.4 3

Submit Guess | Simulate a new dataset

The graph below shows 100 observations in each group.



Your guessing performance

Srovnání dvou průměrů (dle Conway, n.d.)

Independent t-test - úvod

Calculation of the observed t-value for an independent t-test is similar to the dependent t-test, but involves slightly different formulas. The t-value is now

$$t = \frac{(\bar{x_1} - \bar{x_2})}{se_p}$$

where $\bar{x_1}$ and $\bar{x_1}$ are the mean intelligence gains for group 1 and group 2, respectively. se_p is the pooled standard error, which is equivalent to

$$se_p = \sqrt{\frac{var_1}{n_1} + \frac{var_2}{n_2}}$$

Předpoklady použití:

- The sampling distribution is normally distributed.
- Data are measured at least at the interval level.
- Homogeneity of variance.
- Scores are independent (because they come from different people).

Srovnání dvou průměrů (dle Conway, n.d.)

Dependent t-test

```
# Bydleni_Brno <- read_excel("Bydleni_Brno.xlsx")

# In the case of our dependent t-test, we need to specify these arguments to t.test():
 ?t.test

# x: Column of Bydleni_Brno containing prices for 2015
# y: Column of Bydleni_Brno containing prices for 2016 paired: Whether we're doing a dependent (i.e.
# paired) t-test or independent t-test. In this example, it's TRUE
# Note that t.test() carries out a two-sided t-test by default

# Conduct a paired t-test using the t.test function
t.test(Bydleni_Brno$Pronajem_m2_2015, Bydleni_Brno$Pronajem_m2_2016, paired = TRUE)
```

Srovnání dvou průměrů

Independent t-test

```
# View the dataset
view(dfSummary(Bydleni_Brno))

# Create subsets
Sidliste <- Bydleni_Brno %>%
        select(c("Sidliste", "Pronajem_m2_2016")) %>%
        filter(Sidliste == 1, Pronajem_m2_2016 > 0)

Rodinne_Domy <- Bydleni_Brno %>%
  select(c("Sidliste", "Pronajem_m2_2016")) %>%
  filter(Sidliste == 0, Pronajem_m2_2016 > 0)

# Summary statistics
Sidliste_view <- view(dfSummary(Sidliste))
Rodinne_Domy_view <- view(dfSummary(Rodinne_Domy))

# Create a boxplot
Bydleni_Brno_Najem <- Bydleni_Brno %>%
  filter(Pronajem_m2_2016 > 0) %>%
  ggplot(aes(x = factor(Sidliste), y = Pronajem_m2_2016, fill = factor(Sidliste))) + geom_boxplot()
```

Srovnání dvou průměrů

Independent t-test - base

```
# Levene's test
library(car) # install.packages("car")

Bydleni_Brno_Najem_Levene <- Bydleni_Brno %>%
  filter(Pronajem_m2_2016 > 0)
  leveneTest(Bydleni_Brno_Najem_Levene$Pronajem_m2_2016 ~
factor(Bydleni_Brno_Najem_Levene$Sidliste))

  # Conduct an independent t-test
  t.test(Sidliste$Pronajem_m2_2016, Rodinne_Domy$Pronajem_m2_2016, var.equal = FALSE)
```

Srovnání dvou průměrů (dle Conway, n.d.)

Independent t-test - Cohen‘s d

$$t = \frac{\bar{x_1} - \bar{x_2}}{se_p}$$

where $\bar{x_1}$ and $\bar{x_1}$ are the mean intelligence gains for group 1 and group 2, respectively, and se_p is the pooled standard error.

The formula for Cohen's d for independent t-tests is

$$d = \frac{\bar{x_1} - \bar{x_2}}{sd_p}$$

where sd_p is the pooled standard deviation, which in turn is equal to

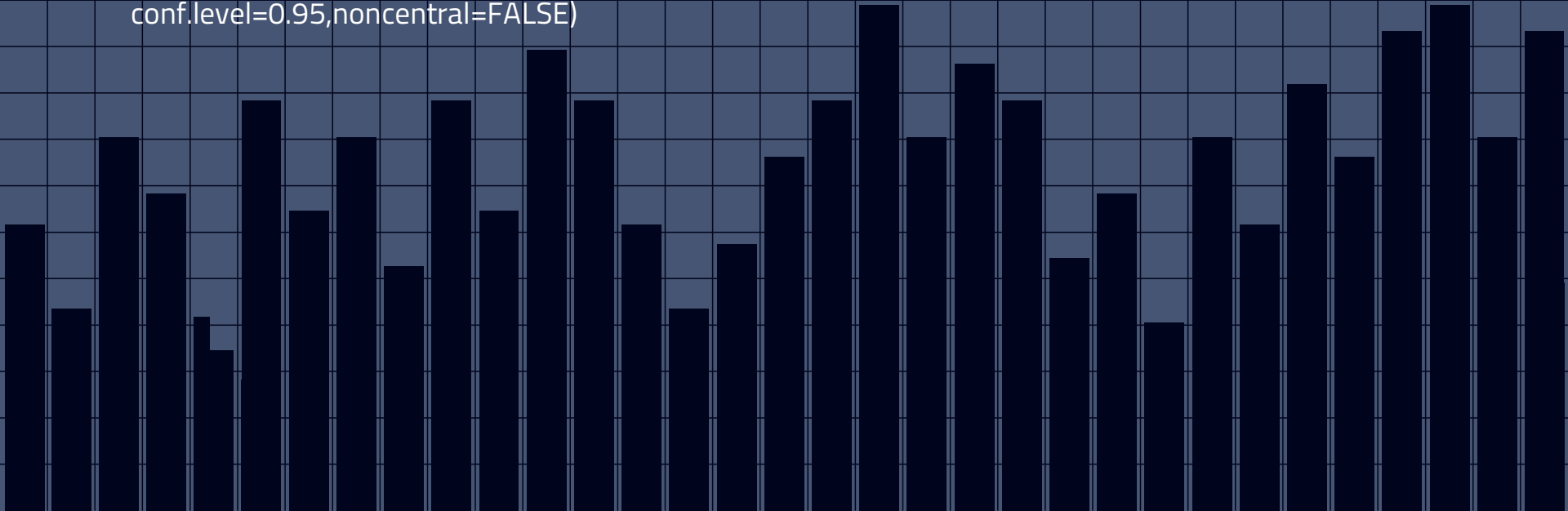
$$sd_p = \frac{sd_1 + sd_2}{2}$$

where sd_1 and sd_2 are the standard deviations of the first and second groups, respectively.

Srovnání dvou průměrů

Dependent t-test – Cohenovo d – effsize – příklad

```
# Calculate Cohen's d
cohen.d(Sidliste$Pronajem_m2_2016, Rodinne_Domy$Pronajem_m2_2016,
        pooled=FALSE,paired=FALSE,
        na.rm=FALSE, hedges.correction=F,
        conf.level=0.95,noncentral=FALSE)
```



ANOVA

Úvod

ANOVA = **AN**alysis **O**f **VA**riance

- Slouží pro **srovnání skupinových průměrů** napříč **3 a více skupinami/podmínkami**
- Dvě výchozí varianty:
  - Between design: oddělené, na sobě nezávislé skupiny (ANOVA, ANCOVA, faktoriální ANOVA atd.)
    - *Liší se jednotlivé kraje v ČR z hlediska průměrné mzdy?*
  - Within design: srovnání skupinového průměru napříč různými podmínkami (Repeated Measures ANOVA)
    - *Lišily se průměrné výdaje domácností na pohonné hmoty během posledních 5 let?*

One-Way ANOVA

Data

- **Faktor** – převažující kategorie vzdělání u obyvatel v dané brněnské čtvrti (*základní, středoškolské, vysokoškolské*)
- **Závislá proměnná** – průměrná cena za byt o rozměru 60 metrů čtverečních v roce 2016 dle městské části v Brně
- **Nulová hypotéza**:
  - Všechny skupiny mají shodný průměr (i.e. *V Brně není rozdíl v průměrné ceně bytu o rozloze 60 metrů čtverečních pro čtvrti s obyvateli s převažujícím základním, středoškolským a vysokoškolským stupněm vzdělání.*)
- **Alternativní hypotéza**:
  - *S převažujícím vyšším stupněm vzdělání v městských částech Brna se pojí vyšší průměrná cena bytu o rozloze 60 metrů čtverečních.*

One-Way ANOVA

Kód

- Bydleni_Brno$Vzdelani = factor(Bydleni_Brno$Vzdelani, order = TRUE, levels = c(0, 1, 2), labels = c("Základní", "Středoškolské", "Vysokoškolské"))
- Bydleni_Brno_60m2_2016 <- Bydleni_Brno %>%
  filter(Prodej_60m2_2016 > 0)

One-Way ANOVA

F-test a F-Ratio

- **Null hypothesis**: all groups are equal
- ANOVA provides a significance test
  - Můžeme určit kritickou hodnotu (na určité hladině významnosti) a testovat, zda ji hodnota F v našem výzkumu překračuje, tj. testovat statistickou významnost nalezených rozdílů mezi skupinami
- Test statistic is the F-test (or F-ratio)

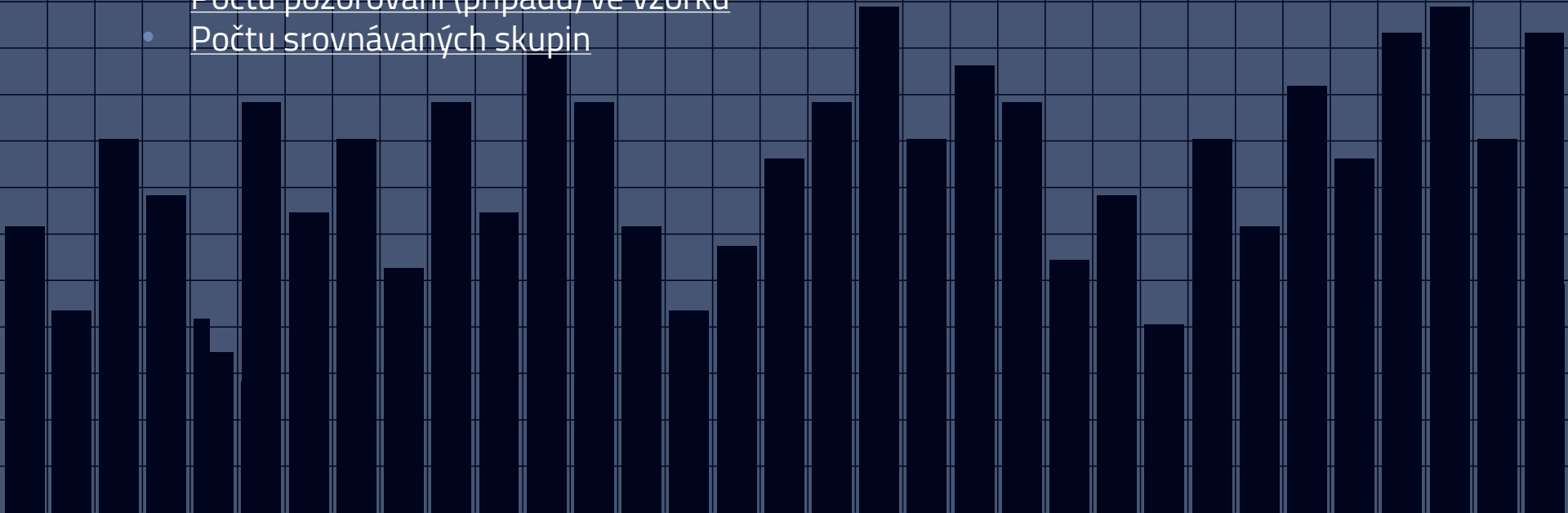
- Poměr toho, co model vysvětlit dokáže, ku tomu, co vysvětlit nedokáže
- Large F-ratio indicates significant effect
  - Čím vyšší F, tím více záleží na rozdělení lidí do jednotlivých skupin, tj. tím více se skupiny od sebe liší v závislé proměnné

One-Way ANOVA

F-test a F-Ratio

Jak získáme příslušnou p-hodnotu?

- Obdoba t-testu a "rodině" t-rozložení
- "Rodina" **F-rozložení** se odvíjí od:
  - Počtu pozorování (případů) ve vzorku
  - Počtu srovnávaných skupin



One-Way ANOVA

F-test a F-Ratio

```
# Create the vector x
x <- seq(from = 0, to = 10, length = 2000)

# Evaluate the densities
y_1 <- df(x, 3, 100)
y_2 <- df(x, 1, 1)
y_3 <- df(x, 2, 100)
y_4 <- df(x, 3, 30)
y_5 <- df(x, 3, 500)
y_6 <- df(x, 3, 50)
y_7 <- df(x, 6, 1000)

# Plot the densities
plot(x, y_1, col = 1, type = "l")
lines(x, y_2, col = 2)
lines(x, y_3, col = 3)
lines(x, y_4, col = 4)
lines(x, y_5, col = 5)
lines(x, y_6, col = 6)
lines(x, y_7, col = 7)
```

```
# Add the legend
legend("topright", title = "F distributions",
c("df = (3, 100)", "df = (1, 1)", "df = (2, 100)", "df = (3, 30)",
"df = (3, 500)", "df = (3, 50)", "df = (6, 1000)"),
col = c(1, 2, 3, 4, 5, 6, 7), lty = 1)
```

One-Way ANOVA

Summary Table

One-Way ANOVA Table

An ANOVA table is often used to record the sums of squares and to organize the rest of the calculations. *Format for the ANOVA Table:*

Source of Variation	SS	df	MS	F ratio
Between Samples	SSB	k - 1	$MSB = \frac{SSB}{k-1}$	$F = \frac{MSB}{MSW}$
Within Samples	SSW	n - k	$MSW = \frac{SSW}{n-k}$	
Total	SST = SSB+SSW	n - 1		

- The sums of squares and the degrees of freedom must check
  SS(factor) + SS(error) = SS(total) or SSB + SSW = SST
  df(factor) + df(error) = df(total) or df(between) + df(within) = df(total)

One-Way ANOVA

F-test a F-Ratio

Prozkoumání dat

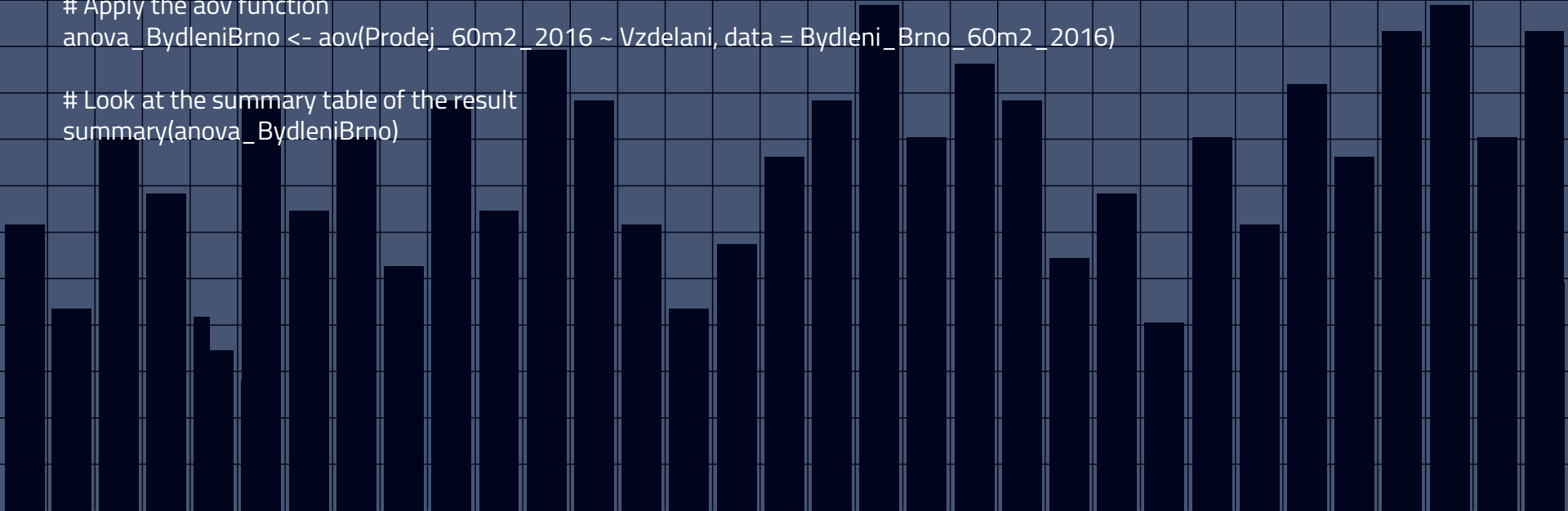
```
# Summary statistics by group
library(psych)
describeBy(Bydleni_Brno_60m2_2016$Prodej_60m2_2016, group = Bydleni_Brno_60m2_2016$Vzdelani)

# Boxplot
library(ggplot2)
bp1 = ggplot(Bydleni_Brno_60m2_2016, aes(Vzdelani, Prodej_60m2_2016))
bp1 + geom_boxplot(aes(fill=Vzdelani), alpha=I(0.5)) +
  geom_point(position="jitter", alpha=0.5) +
  geom_boxplot(outlier.size=0, alpha=0.5) +
  theme(
    axis.title.x = element_text(face="bold", color="black", size=12),
    axis.title.y = element_text(face="bold", color="black", size=12),
    plot.title = element_text(face="bold", color = "black", size=12)) +
  labs(x="Převažující kategorie vzdělání ",
       y = "Cena za byt o rozloze 60 metrů čtverečních (v Kč)",
       title= "Cena za byt o rozloze 60 metrů čtverečních (v Kč) dle převažující kategorie vzdělání") +
theme(legend.position='none')
```

One-Way ANOVA

F-test a F-Ratio

Funkce aov

```
aov(dependent_var ~ independent_var)
summary()

# Apply the aov function
anova_BydleniBrno <- aov(Prodej_60m2_2016 ~ Vzdelani, data = Bydleni_Brno_60m2_2016)

# Look at the summary table of the result
summary(anova_BydleniBrno)
```



One-Way ANOVA

F-test a F-Ratio

Velikost účinku

$$\eta^2 = \frac{SS_b}{SS_t}$$

```
library(lsr)
etaSquared(anova_wm, type = 2,
anova = FALSE)
```

$$\omega^2 = \frac{SS_b - df_b MS_w}{SS_t + MS_w}$$

```
library(sjstats)

anova_wm2 <- lm(iq ~ condition2, data =ANOVA)
r2(anova_wm2, n = NULL)
```


ONE DOES NOT SIMPLY FIND
PARTIAL OMEGA SQUARED IN SPSS.

One-Way ANOVA

Předpoklady použití

Povaha proměnných
- "Závislá" proměnná kardinální úrovně měření

Normalita rozložení závislé proměnné
- V rámci každé sledované skupiny
- Narušení nepředstavuje závažný problém, pokud jsou skupiny stejně velké + mají velikost alespoň okolo 30
- **Neparametrická** alternativa – Kruskal-Wallisův test

Homogenita rozptylu
- Sledujeme Levenův F-test, nulová hypotéza hovoří o homogenitě napříč skupinami
    - Pokud Levenův F-test vychází statisticky signifikantní:
- Sledujeme **poměr rozptylu** u skupin s největším a nejmenším rozptylem, přičemž chceme, aby byl tento **poměr menší než 3**
- Narušení by nemělo vadit, pokud jsou **skupiny stejně velké**
- Při narušení lze použít **Welchovo F**

Nezávislost pozorování

One-Way ANOVA

Předpoklady použití

library("car")

If you don't specify additional arguments, the deviation scores are calculated by comparing each score to its group median.

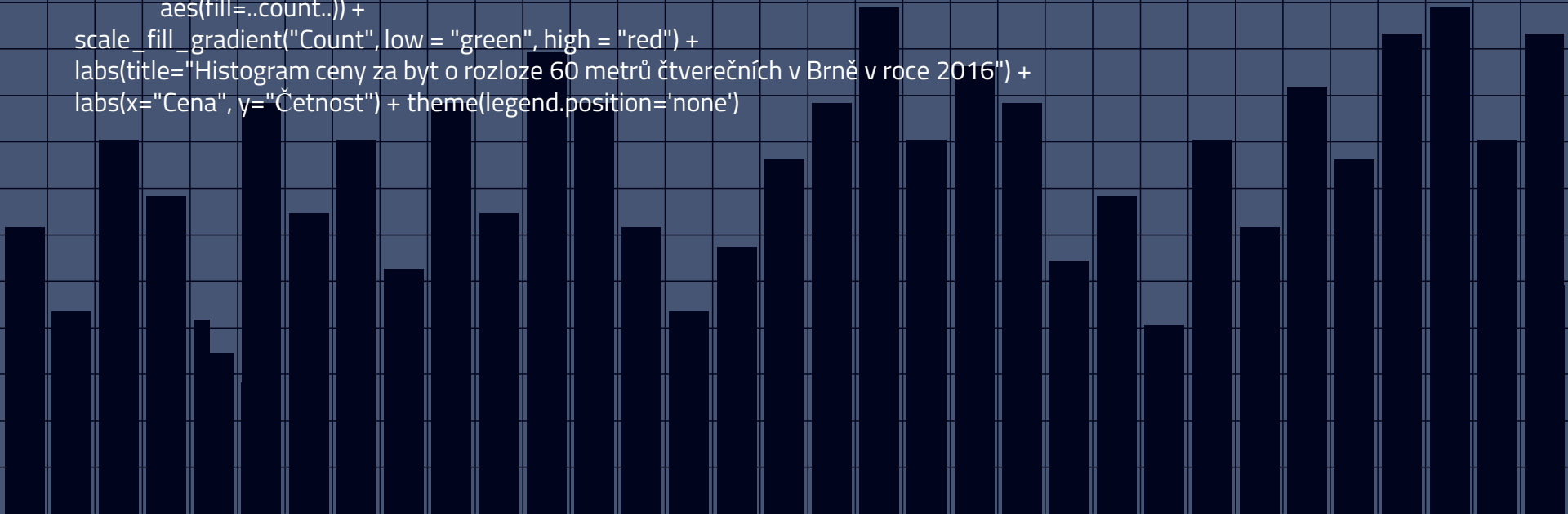
- This is the default behaviour, even though they are typically calculated by comparing each score to its group mean.
- If you want to use means and not medians, add an argument center = mean. Do this now and compare the results to the first test.

```
# Levene's test
leveneTest(Prodej_60m2_2016 ~ Vzdelani, data = Bydleni_Brno_60m2_2016)

# Levene's test with center = mean
leveneTest(Prodej_60m2_2016 ~ Vzdelani, data = Bydleni_Brno_60m2_2016, center = mean)
```

One-Way ANOVA

Předpoklady použití

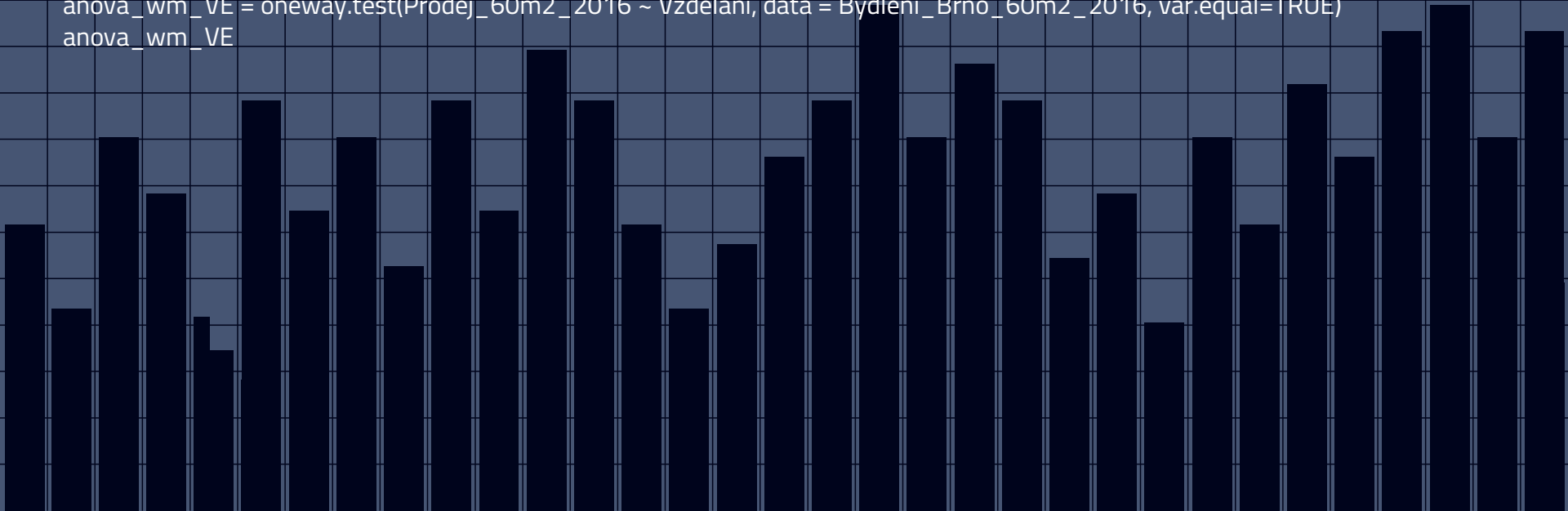
```
# Normalita rozložení
ggplot(data=Bydleni_Brno_60m2_2016, aes(Prodej_60m2_2016)) +
 geom_histogram(binwidth = 250000, col="red",
                aes(fill=..count..)) +
 scale_fill_gradient("Count", low = "green", high = "red") +
 labs(title="Histogram ceny za byt o rozloze 60 metrů čtverečních v Brně v roce 2016") +
 labs(x="Cena", y="Četnost") + theme(legend.position='none')
```



One-Way ANOVA

Welchův F-test

```
anova_wm_VNE = oneway.test(Prodej_60m2_2016 ~ Vzdelani, data = Bydleni_Brno_60m2_2016, var.equal=FALSE)
anova_wm_VNE

anova_wm_VE = oneway.test(Prodej_60m2_2016 ~ Vzdelani, data = Bydleni_Brno_60m2_2016, var.equal=TRUE)
anova_wm_VE
```



Post-Hoc Testy

Úvod

Allow for multiple pairwise comparisons without an increase in the probability of a Type I error

Používáme, pokud nemáme dopředu jasné hypotézy

- Srovnávají **vše se vším** – každou skupinu s každou (ale **neumí slučovat skupiny jako kontrasty**)

Z principu jsou oboustranné

Je jich mnoho – liší se v několika parametrech:

- **Konzervativní** (Ch. II. typu) versus **Liberální** (Ch. I. typu)
  - *Most liberal = no adjustment*
  - *Most conservative = adjust for every possible comparison that could be made*
- Ne/vhodné pro **rozdílně velké skupiny**
- Ne/vhodné pro **rozdílné skupinové rozptyly**

Post-Hoc Testy

Doporučení podle Fielda

Stejně velké skupiny a **skupinové rozptyly** (ideální situace):
- REGWQ
- Tukey

Pokud si chceme být jistí, že **P chyby I. typu** nepřekročí zvolenou hladinu:
- Bonferroni

Pokud jsou **velikosti skupin** trochu/hodně rozdílné:
- Gabriel
- Hochberg GT2

Pokud pochybujeme o **shodnosti skupinových rozptylů**:
- Games-Howell

Post-Hoc Testy

Tukey

```
# Conduct ANOVA
anova_BydleniBrno = aov(Prodej_60m2_2016 ~ Vzdelani, data = Bydleni_Brno_60m2_2016)

# View summary
summary(anova_BydleniBrno)

# Conduct Tukey procedure
tukey <- TukeyHSD(anova_BydleniBrno)

# Plot confidence intervals
plot(tukey)
```

Post-Hoc Testy

Bonferroni

The Bonferroni correction compensates for that increase by testing each individual hypothesis at a significance level of α/m, where α is the desired *overall alpha level* and m is the *number of hypotheses*.

- For example, if a trial is testing **m = 20** hypotheses with a desired α **= 0.05**, then the Bonferroni correction would test each individual hypothesis at α **= 0.05/20 = 0.0025**.

```
# Pairwise t-test
pairwise.t.test(Bydleni_Brno_60m2_2016$Prodej_60m2_2016,
Bydleni_Brno_60m2_2016$Vzdelani, p.adjust = "bonferroni")
```


10 T-TESTS AND NO CORRECTION?
HANS, BRING THE FLAMMENWERFER

Kontrasty

Úvod

Umožňují porovnat jednotlivé skupiny v jednom kroku bez nutnosti korigovat hladinu významnosti (**bez snížení síly testu**)

- Jen když máme dopředu hypotézy
- Kontrastů lze provést tolik, kolik je počet skupin – 1

Každý kontrast **srovnává 2 průměry**

- Průměr skupiny nebo průměr více skupin dohromady
- Např. „Základní" vs. „Středoškolské"

Ortogonální (nezávislé) **kontrasty**

- Skupina použitá v jednom srovnání není použitá v dalším

Neortogonální kontrasty

Kontrasty

Úvod

```
c1 = c(-1,0,1)
c2 = c(0,-1,1)
mat <- cbind(c1,c2)
contrasts(Bydleni_Brno_60m2_2016$Vzdelani) <- mat
model1 <- lm(Prodej_60m2_2016 ~ Vzdelani, data = Bydleni_Brno_60m2_2016)
summary(model1)

options(contrasts = c("contr.helmert", "contr.poly"))
contrasts(Bydleni_Brno_60m2_2016$Vzdelani) <- "contr.helmert"
model1 <- lm(Prodej_60m2_2016 ~ Vzdelani, data = Bydleni_Brno_60m2_2016)
summary(model1)
```

Zdroje

Conway, A. (n.d.) Intro to Statistics with R: Student's T-test. Dostupné online na: https://www.datacamp.com/courses/intro-to-statistics-with-r-students-t-test

Effect size (n.d.). In Wikipedia: Staženo dne 10. 10. 2016 z https://en.wikipedia.org/wiki/Effect_size

Field, A., Miles, J., & Field, Z. (2012). Discovering Statistics Using R. Sage: UK.

Navarro, D. J. (2014). Learning statistics with R: A tutorial for psychology students and other beginners. Available online: http://health.adelaide.edu.au/psychology/ccs/teaching/lsr/

Standard error (n.d.). In Wikipedia: Staženo dne 10. 10. 2016 z https://en.wikipedia.org/wiki/Standard_error

Student's t-test (n.d.). In Wikipedia: Staženo dne 10. 10. 2016 z https://en.wikipedia.org/wiki/Student%27s_t-test