RStudio CheatSheet by Adéla Vrtková, Martina Litschmannová, Katedra aplikované matematiky, FEI, VŠB-TUO

via cheatography.com (aktualizováno: 16. března 2023)

Nápověda, Pracovní adresář, Knihovny

```
?boxplot
help(boxplot)
```
- **nápověda** pro funkci *boxplot*

```
getwd()
```
- vypsání současného **pracovního adresáře**

```
setwd("C:/Users/RStudio")
setwd("C:\\Users\\RStudio")
```
- nastavení **pracovního adresáře**

```
install.packages("packageZ") # na daném PC pouze jednou
library(packageZ) # vždy při novém otevření RStudia
```
- stažení a instalace **knihovny** *packageZ*, aktivace knihovny pro použití

```
packageZ::functionF(x)
```
- **zavolání funkce** *functionF* z knihovny *packageZ*

Rozdělení pravděpodobnosti - Předpony

r-	generování hodnot z určitého rozdělení
d-	hustota pravděpodobnosti $f(x)$ nebo pravděpodobnostní funkce $P(X = x)$
p-	$P(X \leq x)$
q-	kvantilová funkce $F^{-1}(x)$

Rozdělení pravděpodobnosti - Diskrétní

-binom	Binomické rozdělení
-hyper	Hypergeometrické rozdělení R vyžaduje jiné nastavení parametrů
-nbinom	Negativně binomické rozdělení Definice v R - počet neúspěšných pokusů
-geom	Geometrické rozdělení Definice v R - počet neúspěchů před prvním úspěchem
-pois	Poissonovo rozdělení

Rozdělení pravděpodobnosti - Spojité

-unif	Rovnoměrné rozdělení
-exp	Exponenciální rozdělení
-norm	Normální rozdělení R vyžaduje jiné nastavení parametrů
-weibull	Weibullovo rozdělení
-lnorm	Logaritmicko-normální rozdělení
-t	Studentovo rozdělení
-chisq	χ^2 rozdělení
-f	Fisherovo-Snedecorovo rozdělení

Import dat

```
data = read.csv2("C:/Users/RStudio/dataset.csv")
```
- import dat v **csv formátu** z konkrétní složky a uložení jako *data*

```
data = read.csv2("http://am-nas.vsb.cz/DATA/dataset.csv")
```
- import dat v **csv formátu** z internetu a uložení jako *data*

```
data = read_excel("C:/MojeSoubory/dataset.xlsx", sheet = "List1")
```
- import dat v **xlsx formátu** pomocí knihovny readxl z konkrétní složky a uložení jako *data*

Práce s datovým souborem

```
data = as.data.frame(data)
```
- uložení dat jako objekt typu *data.frame* (je-li jiného typu)

```
data.S = stack(data.tab)
```
- **převod dat** z tabulky do st. datového formátu; všechny sloupce jsou **sloučeny do jednoho** (*values*) a k němu je **přidán nový sloupec** (*ind*, typu factor) určující, ze kterého původního sloupce data pocházejí

```
data.tab = unstack(data.S)
```
- **převod dat** ze st. datového formátu do tabulky, vyžaduje konkrétní strukturu, viz R-Help

```
data.S.omit = na.omit(data.S)
```
- vynechání řádků, ve kterých se vyskytují **chybějící hodnoty** (NA)
- další užitečné funkce (knihovna tidyr): pivot_longer(), pivot_wider()

Práce s datovým souborem pomocí *dplyr*

filter	**vybere řádky** na základě daných podmínek
select	**vybere sloupce** podle jejich názvu nebo čísla
arrange	**seřadí řádky** podle zvolené proměnné
group_by	**seskupí** hodnoty do skupin podle zvolené proměnné
summarise	generuje **souhrnné charakteristiky** různých proměnných
mutate	**přidá** novou proměnnou nebo **transformuje** existující

%>%	pomocí **operátoru *pipe*** (Ctrl+Shift+M) lze řetězit funkce
	`data %>% group_by(skupina) %>% summarise(m=max(hodnoty))`
	`data %>% filter(skupina=="A") %>% arrange(hodnoty)`

EDA pro kvalitativní proměnnou *skupina* v souboru *mydata*

```
mydata$skupina = as.factor(mydata$skupina)
```
- předefinování proměnné na typ *factor*

```
table(mydata$skupina)
```
- **tabulka** absolutních četností

```
prop.table(table(mydata$skupina))
```
- **tabulka** relativních četností

```
barplot(table(mydata$skupina))
```
- **sloupcový** graf (základní R)

```
ggplot(tabulka, aes(x = skupina, y = abs.cet))+
    geom_bar(stat = "identity")+
    geom_text(aes(label = abs.cet), vjust = -0.5)
```
- **sloupcový** graf (knihovna ggplot2); vstupem je tabulka abs. četností, která je uložena jako typ *data.frame*; sloupec *skupina* - identifikace kategorie, sloupec *abs.cet* - odpovídající abs. četnosti

(více: rstudio.com/resources/cheatsheets/)

EDA pro kvantitativní proměnnou *hodnoty* v souboru *mydata*

summary(mydata$hodnoty)	vybrané míry polohy
length(mydata$hodnoty)	rozsah (pozor s NA v datech)
min(mydata$hodnoty)	minimum
mean(mydata$hodnoty)	aritmetický průměr
quantile(mydata$hodnoty,probs=0.3)	30% kvantil
max(mydata$hodnoty)	maximum
sd(mydata$hodnoty)	směrodatná odchylka
var(mydata$hodnoty)	rozptyl
moments::skewness(mydata$hodnoty)	šikmost (moments)
moments::kurtosis(mydata$hodnoty)-3	stand. špičatost (moments)
mydata %>% summarise(min = min(hodnoty), prumer = mean(hodnoty), max = max(hodnoty))	využití knihovny dplyr pro výpočet souhrnných charakteristik
boxplot(mydata$hodnoty)	krabicový graf (základní R)
hist(mydata$hodnoty)	histogram (základní R)
hist(mydata$hodnoty,freq=FALSE) lines(density(mydata$hodnoty))	histogram a empirická hustota pravděpodobnosti
qqnorm(mydata$hodnoty) qqline(mydata$hodnoty)	QQ-graf (základní R)

EDA pro kvantitativní proměnnou *hodnoty* v souboru *mydata* - ggplot2

```
ggplot(mydata, aes(x = "", y = hodnoty))+
    geom_boxplot()
```
- **krabicový** graf kvantitativní proměnné *hodnoty* v souboru *mydata*

```
ggplot(mydata, aes(x = hodnoty))+
    geom_histogram(bins = 20)
```
- **histogram** kvantitativní proměnné *hodnoty* v souboru *mydata*

```
ggplot(mydata, aes(sample = hodnoty))+
    stat_qq()+
    stat_qq_line()
```
- **QQ-graf** kvantitativní proměnné *hodnoty* v souboru *mydata*

```
+labs(x = "popisek x", y = "popisek y", title = "nazev") # popisky
+theme_bw() # změna barevného schématu
+theme_classic() # jiná změna barevného schématu
```
- některé další volitelné parametry

(více: rstudio.com/resources/cheatsheets/)

RStudio CheatSheet by Adéla Vrtková, Martina Litschmannová, Katedra aplikované matematiky, FEI, VŠB-TUO

via cheatography.com (aktualizováno: 16. března 2023)

EDA pro kvantitativní proměnnou *hodnoty* dle kvalitativní proměnné *skupina* v souboru *mydata*

```
tapply(mydata$hodnoty, mydata$skupina, mean)
```

- průměr pro *hodnoty* (numeric) podle proměnné *skupina* (factor)

```
tapply(mydata$hodnoty, mydata$skupina, quantile, probs=0.5)
```

- 50% kvantil pro *hodnoty* (numeric) podle proměnné *skupina* (factor)

```
mydata %>%
  group_by(skupina) %>%
  summarise(prumer = mean(hodnoty),
            med = quantile(hodnoty, probs = 0.5)
```

- využití knihovny dplyr pro výpočet souhrnných charakteristik pro *hodnoty* (numeric) podle proměnné *skupina* (factor)

EDA pro kvantitativní proměnnou *hodnoty* dle kvalitativní proměnné *skupina* v souboru *mydata* - ggplot2

```
ggplot(mydata, aes(x = skupina, y = hodnoty))+
    geom_boxplot()
```

- **vícenásobný krabicový** graf kvantitativní proměnné *hodnoty* dle kvalitativní proměnné *skupina* v souboru *mydata*

```
ggplot(mydata, aes(x = hodnoty))+
    geom_histogram(bins = 20)+
    facet_wrap("skupina")
```

- **sada histogramů** kvantitativní proměnné *hodnoty* dle kvalitativní proměnné *skupina* v souboru *mydata*

```
ggplot(mydata, aes(sample = hodnoty))+
    stat_qq()+
    stat_qq_line()+
    facet_wrap("skupina", scales = "free")
```

- **sada QQ-grafů** kvantitativní proměnné *hodnoty* dle kvalitativní proměnné *skupina* v souboru *mydata*

```
+labs(x = "popisek x", y = "popisek y", title = "nazev") # popisky
+theme_bw() # změna barevného schématu
+theme_classic() # jiná změna barevného schématu
```

- některé další volitelné parametry

(více: rstudio.com/resources/cheatsheets/)

Identifikace a odstranění odlehlých pozorování

```
mydata %>%
  identify_outliers(hodnoty)
```

- **identifikace odlehlých pozorování** u kvantitativní proměnné *hodnoty* ze souboru *mydata* (vyžaduje knihovny dplyr a rstatix)

```
mydata %>%
  group_by(skupina) %>%
  identify_outliers(hodnoty)
```

- **identifikace odlehlých pozorování** u kvantitativní proměnné *hodnoty* dle kvalitativní proměnné *skupina* ze souboru *mydata* (vyžaduje knihovny dplyr a rstatix)

```
outliers = mydata %>%
            identify_outliers(hodnoty)
mydata = mydata %>%
            mutate(hodnoty_out =
                    ifelse(ID %in% outliers$ID, NA, hodnoty))
```

- **odstranění odlehlých pozorování** u kvantitativní proměnné *hodnoty* ze souboru *mydata* (vyžaduje knihovny dplyr a rstatix a proměnnou s unikátním identifikátorem statistické jednotky *ID*)

```
outliers = mydata %>%
            group_by(skupina) %>%
            identify_outliers(hodnoty)
mydata = mydata %>%
            mutate(hodnoty_out =
                    ifelse(ID %in% outliers$ID, NA, hodnoty))
```

- **odstranění odlehlých pozorování** u kvantitativní proměnné *hodnoty* dle kvalitativní proměnné *skupina* ze souboru *mydata* (vyžaduje knihovny dplyr a rstatix a proměnnou s unikátním identifikátorem statistické jednotky *ID*)

Statistická indukce pro jeden náhodný výběr

```
shapiro.test(mydata$hodnoty)
```

- Shapirův-Wilkův test

```
symmetry.test(mydata$hodnoty, boot = FALSE) # lawstat knihovna
```

- Test symetrie

```
varTest(mydata$hodnoty, sigma.squared=400, alternative="two.sided",
        conf.level=0.95) # EnvStats knihovna
```

- 95% oboustranný int. odhad **rozptylu** a oboustranný jednovýběrový test o rozptylu (odpovídá $H_0 : \sigma^2 = 400, H_A : \sigma^2 \neq 400$)
- 95% oboustranný int. odhad **sm. odchylky** a oboustranný jednovýběrový test o sm.odchylce (odpovídá $H_0 : \sigma = \sqrt{400}, H_A : \sigma \neq \sqrt{400}$)

```
t.test(mydata$hodnoty, mu=5, alternative="two.sided", conf.level=0.95)
```

- 95% oboustranný int. odhad **střední hodnoty** a oboustranný jednovýběrový Studentův t-test (odpovídá $H_0 : \mu = 5, H_A : \mu \neq 5$)

```
wilcox.test(mydata$hodnoty, mu=8, alternative="two.sided",
            conf.level=0.95, conf.int=T)
```

- 95% oboustranný int. odhad **mediánu** (pro symetrická data) a oboustranný jednovýběrový Wilcoxonův test (odpovídá $H_0 : x_{0,5} = 8$, $H_A : x_{0,5} \neq 8$)

```
SIGN.test(mydata$hodnoty, md=8, alternative="two.sided",
          conf.level=0.95) # BSDA knihovna
```

- 95% oboustranný int. odhad **mediánu** a oboustranný jednovýběrový znaménkový test (odpovídá $H_0 : x_{0,5} = 8, H_A : x_{0,5} \neq 8$)

```
binom.test(16, 100, 0.18, alternative="two.sided",conf.level=0.95)
```

- 95% oboustranný int. odhad **parametru binomického rozdělení** a oboustranný jednovýběrový test o parametru binomického rozdělení (Clopperova-Pearsonova metoda) (odpovídá $p = \frac{16}{100}, H_0 : \pi = 0.18, H_A : \pi \neq 0.18$)
- **parametr *alternative*** může nabývat hodnot {"two.sided","less","greater"}

Statistická indukce pro dva nezávislé náhodné výběry

```
tapply(mydata$hodnoty,mydata$skupina,shapiro.test)
```

- Shapirův-Wilkův test proměnné *hodnoty* (numeric) pro každou variantu proměnné *skupina* (factor)

```
var.test(dataA,dataB,ratio=1,alternative="two.sided",conf.level=0.95)
```

- 95% oboustranný int. odhad **poměru rozptylů** a test poměru rozptylů (odpovídá $H_0 : \frac{\sigma_A^2}{\sigma_B^2} = 1, H_A : \frac{\sigma_A^2}{\sigma_B^2} \neq 1$)

```
t.test(dataA,dataB,alternative="two.sided",var.equal=T,conf.level=0.95)
```

- 95% oboustranný int. odhad **rozdílu středních hodnot** a dvouvýběrový t-test pro homoskedasticitní data (odpovídá $H_0 : \mu_A - \mu_B = 0, H_A : \mu_A - \mu_B \neq 0$)

```
t.test(dataA,dataB,alternative="two.sided",var.equal=F,conf.level=0.95)
```

- 95% oboustranný int. odhad **rozdílu středních hodnot** a Aspinové-Welchův test pro heteroskedasticitní data (odpovídá $H_0 : \mu_A - \mu_B = 0, H_A : \mu_A - \mu_B \neq 0$)

```
wilcox.test(dataA,dataB, alternative="two.sided",
            conf.level=0.95, conf.int=T)
```

- 95% oboustranný int. odhad **rozdílu mediánů** (pro data s podobným tvarem rozdělení) a Mannův-Whitneyho test (odpovídá $H_0 : x_{0,5}^A - x_{0,5}^B = 0,\ H_A : x_{0,5}^A - x_{0,5}^B \neq 0$)

```
prop.test(c(10,30),c(100,255), alternative="two.sided",conf.level=0.95)
```

- 95% oboustranný int. odhad **rozdílu parametrů binomického rozdělení** a Pearsonův χ^2 test shody parametrů dvou binomických rozdělení s Yatesovou korekcí (odpovídá $p_A = \frac{10}{100}, p_B = \frac{30}{255}, H_0 : \pi_A - \pi_B = 0, H_A : \pi_A - \pi_B \neq 0$)
- **parametr *alternative*** může nabývat hodnot {"two.sided","less","greater"}

RStudio CheatSheet by Adéla Vrtková, Martina Litschmannová, Katedra aplikované matematiky, FEI, VŠB-TUO
via cheatography.com (aktualizováno: 16. března 2023)

Statistická indukce pro tři a více nezávislých náhodných výběrů

```
tapply(mydata$hodnoty,mydata$skupina,shapiro.test)
```

- Shapirův-Wilkův test proměnné *hodnoty* (numeric) pro každou variantu proměnné *skupina* (factor)

```
bartlett.test(mydata$hodnoty~mydata$skupina)
```

- **Bartlettův** test homoskedasticity

Při kopírování je nutné symbol ~ v R-skriptu ručně přepsat na klávesnici.

```
leveneTest(mydata$hodnoty~mydata$skupina) # car knihovna
```

- **Leveneho** test homoskedasticity

Při kopírování je nutné symbol ~ v R-skriptu ručně přepsat na klávesnici.

```
vysledky = aov(mydata$hodnoty~mydata$skupina)
summary(vysledky)
```

- uložení výsledku testu **ANOVA** a vypsání tabulky ANOVA

Při kopírování je nutné symbol ~ v R-skriptu ručně přepsat na klávesnici.

```
TukeyHSD(vysledky)
```

- Tukeyho **post-hoc** analýza po ANOVA

```
plot(TukeyHSD(vysledky))
```

- grafická prezentace Tukeyho post-hoc analýzy

```
kruskal.test(mydata$hodnoty~mydata$skupina)
```

- **Kruskallův-Wallisův** test

Při kopírování je nutné symbol ~ v R-skriptu ručně přepsat na klávesnici.

```
dunnTest(hodnoty~skupina, data = mydata,
         method = "bonferroni") # FSA knihovna
```

- Dunnové **post-hoc** analýza s Bonferroniho korekcí po Kruskalovu-Wallisovu testu

Při kopírování je nutné symbol ~ v R-skriptu ručně přepsat na klávesnici.

Kontingenční tabulky

```
tab = table(data$factor1, data$factor2)
```

- **kontingenční tabulka** pro *factor1* a *factor2* (obojí typu factor)

```
tab = matrix(c(12,45,23,54), ncol=2, byrow=T)
```

- ruční **sestavení kontingenční tabulky** pomocí funkce *matrix*

$\begin{bmatrix} 12 & 45 \\ 23 & 54 \end{bmatrix}$ (může být doplněno použitím *rownames* a *colnames*)

```
mosaicplot(tab)
```

- **mozaikový** graf v základním R

```
ggplot(data)+
    geom_mosaic(aes(x = product(factor2, factor1), fill = factor2)
```

- **mozaikový** graf s knihovnou ggmosaic

```
cramersV(tab) # lsr knihovna
```

- **Cramérovo** V (míra kontingence)

```
chisq.test(tab)
chisq.test(tab)$expected
chisq.test(tab)$p.value
```

- χ^2 **test nezávislosti** v kont. tabulce, očekávané četnosti a p-hodnota

```
epi.2by2(tab) # epiR knihovna
```

- bodové a int. odhady relativního rizika a poměru šancí; je nutné pohlídat strukturu tabulky; (1. řádek - "exponovaná" skupina, 1. sloupec - výskyt jevu)

Testy dobré shody

```
obs_cet = c(979, 1002, 1015, 980, 1040, 984) # pozorované četnosti
exp_pst = c(1/6, 1/6, 1/6, 1/6, 1/6, 1/6) # oček. pravděpodobnosti
chisq.test(obs_cet, p = exp_pst, rescale.p = T)
```

- χ^2 **test dobré shody** (nelze použít pro neúplně specifikovaný test)