Pracovní adresář

- `getwd()` # výpis pracovního adresáře
- `setwd("C:/Moje/Pracovni")` # nastavení pracovního adresáře
- `setwd("C:\\Moje\\Pracovni")` # nastavení pracovního adresáře

Nápověda

- `?funkce` # nápověda pro funkci `funkce`
- `help(funkce)` # nápověda pro funkci `funkce`

Instalování a načtení nového balíčku

- `install.packages("novy")` # instalace balíčku jménem `novy`
- `library(novy)` # načte balíček jménem `novy`

Importování datového souboru

- `data = read.csv2(file="C:/Martina/STA1/DATA/aku.csv`
- `data = read.csv2(file="http://am-nas.vsb.cz/lit40/DATA/aku.csv")`
- `data = readWorkbook("C:/Martina/STA1/DATA/aku.xlsx",sheet=1, startRow=4,colNames=TRUE,cols=2:9)` # balíček openxlsx

Práce s datovým souborem

- `data.s<-stack(data.tab)` # převod z tabulky `data.tab` do standardního datového formátu a uložení do dat. rámce `data.s`

- `data.s.bezNA<-na.omit(data.s)` # odstranění chybějících hodnot (NA) ze `data.s` a uložení do dat. rámce `data.s.bezNA`

- `data.tab<-unstack(data.s)` # převod ze standardního datového formátu `data.s` (v prvním sloupci musí být hodnoty proměnné (numeric), ve druhém sloupci id (factor)) do tabulky a uložení do `data.tab`

- `data.mat=data.matrix(data.data.frame)` # Do proměnné `data.mat` (datový typ matrix) se uloží tabulka zapsaná v proměnné `data.data.frame` (datový typ data.frame)

Obecnější způsob převodu dat z tabulky do standardního dat. formátu

- ```
  data.s=reshape(data.tab,        # dat. rámec, který bude převáděn do std. dat. formátu
                 direction="long",
                 varying=c("A","B"),
                 v.names="hodnoty",
                 times=c("sk.A","sk.B"),
                 timevar="skupina")
  ```
  ```

`direction` - parameter určující „long“ nebo „wide“ formát (nutno nastavit na `"long"`)
`varying` - názvy proměnných, které mají být zařazeny do sloupce hodnot (values)
`v.names` - pojmenování sloupce hodnot (values names)
`times` - varianty id (jaké id bude přiřazeno hodnotám ze sloupců uvedených v parametru varying)
`timevar` - pojmenování proměnné, v níž je uvedeno id

Způsob převodu párových dat z tabulky do standardního dat. formátu

- ```
  data.s=reshape(data.tab,        # dat. rámec, který bude převáděn do std. dat. formátu
                 direction="long",
                 varying=list(c("pred.A","před.B"),c("po.A","po.B"))
                 v.names=c("hodnoty.pred","hodnoty.po")
                 times=c("sk.A","sk.B"),
                 timevar="skupina")
  ```

Rozdělení pravděpodobnosti

Pro každé rozdělení pravděpodobnosti nabízí R několik funkcí, které mají speciální prefixy (první znaky názvu):

- `r` # generování náhodných výběrů
- `d` # hustota pravděpodobnosti $f(x)$, pravděpodobnostní funkce $P(X = x)$
- `p` # kumulativní pravděpodobnostní funkce $P(X \leq x)$
- `q` # kvantilová funkce $F^{-1}(x)$ (Pozor! Distribuční funkce je v R definována jako $F(x) = P(X \leq x)$.)

Diskrétní rozdělení pravděpodobnosti
- `-binom` # binomické rozdělení
- `-hyper` # hypergeometrické rozdělení
- `-nbinom` # negativně binomické rozdělení (Pozor! V R je definováno jako počet neúspěchů před k-tým úspěchem.)
- `-geom` # geometrické rozdělení (Pozor! V R je definováno jako počet neúspěchů před prvním úspěchem.)
- `-pois` # Poissonovo rozdělení

Spojitá rozdělení pravděpodobnosti
- `-unif` # rovnoměrné rozdělení
- `-norm` # normální rozdělení
- `-exp` # exponenciální rozdělení
- `-weibull` # Weibullovo rozdělení
- `-lnorm` # logaritmicko-normální rozdělení
- `-t` # Studentovo rozdělení
- `-chisq` # Chí-kvadrát rozdělení
- `-f` # Fisherovo-Snedecorovo rozdělení
  ```

Explorační analýza

Kategoriální proměnná (proměnná `barva`, která je uložena v souboru `my.data`)

- `summary(my.data$barva)` # tabulka četností proměnné `barva`

- `barplot(summary(my.data$barva))` # sloupcový graf
- `pie(summary(my.data$barva))` # výsečový graf

Kvantitativní proměnná (proměnná `vyska`, která je uložena v souboru `my.data`)

- `summary(my.data$vyska)` # míry polohy (minimum, dolní kvartil, medián, průměr, horní kvartil, maximum) proměnné `vyska`

- `length(my.data$vyska)` # rozsah
- `min(my.data$vyska)` # minimum
- `mean(my.data$vyska)` # průměr
- `quantile(my.data$vyska,0.03)` # 3% kvantil
- `max(my.data$vyska)` # maximum
- `sd(my.data$vyska)` # směrodatná odchylka
- `100*sd(my.data$vyska)/mean(my.data$vyska)` # variační koeficient (%)
- `skewness(my.data$vyska)` # šikmost (balíček moments)
- `kurtosis(my.data$vyska)` # špičatost (balíček moments – Pozor! Nejde o normovanou špičatost. Pro výpočet normované špičatosti je nutno odečíst 3.)

- `boxplot(my.data$vyska)` # krabicový graf
- `hist(my.data$vyska)` # histogram
- `qqnorm(my.data$vyska)` # QQ-graf pro posouzení normality
- `plot(density(my.data$vyska))` # jádrový odhad hustoty pravděpodobnosti
- `hist(my.data$vyska,freq=FALSE,density=-1)`
  `lines(density(my.data$vyska))` # histogram + odhad hustoty

Kvantitativní proměnná dle tříd (pro každou statistickou jednotku jsou v souboru `my.data` uvedeny atributy `vyska(numeric)` a `skupina(factor)` – data ve standardním datovém formátu)

Poznámka: Doporučujeme kategoriální proměnnou (factor) před vlastní analýzou překódovat na číselné hodnoty (tj. každé variantě proměnné přiřadit číselný kód – 1, 2, 3, …).

- `tapply(my.data$vyska,my.data$skupina,mean)` # průměr proměnné `vyska` pro každou variantu proměnné `skupina` (směrodatnou odchylku, min, max, šikmost a špičatost určíte obdobně)

- `tapply(my.data$vyska,my.data$skupina,quantile,probs=0.25)` # dolní kvartil proměnné `vyska` pro každou variantu proměnné `skupina` (obdobně určíte libovolné kvantily)

- `boxplot(split(my.data$vyska,my.data$skupina))` # vícenásobný krabicový graf

Statistická indukce pro jednu proměnnou

Ověření normality

- `shapiro.test(my.data$vyska)` # Shapirův-Wilkův test aplikovaný na proměnnou `vyska`

Test o rozptylu

- POZOR! V prostředí R není implementována funkce pro test a intervalový odhad směrodatné odchylky. Nutno řešit „ručním“ výpočtem.

Test o střední hodnotě

- `t.test(my.data$vyska,mu=5,alternative="two.sided",conf.level=0.95)` # jednovýběrový oboustranný t-test aplikovaný na proměnnou `vyska` ($H_0: \mu = 5, H_A: \mu \neq 5$) + oboustranný 95% intervalový odhad střední hodnoty, parametr `alternative` může nabývat hodnot {„two.sided“, “less“, “greater“}

Testy o mediánu

- `wilcox.test(my.data$vyska,mu=5,alternative="two.sided",conf.level=0.95,conf.int=TRUE)` # jednovýběrový oboustranný Wilcoxonův test aplikovaný na proměnnou `vyska` ($H_0: x_{0.5} = 5, H_A: x_{0.5} \neq 5$) + oboustranný 95% intervalový odhad mediánu, parametr `alternative` může nabývat hodnot {"two.sided", "less", "greater"}
- `signmedian.test(my.data$vyska,mu=5,alternative="two.sided",conf.level=0.95,conf.int=TRUE)` # jednovýběrový oboustranný znaménkový test aplikovaný na proměnnou `vyska` ($H_0: x_{0.5} = 5, H_A: x_{0.5} \neq 5$) + oboustranný 95% intervalový odhad mediánu, parametr `alternative` může nabývat hodnot {"two.sided", "less", "greater"}

Test o parametru binomického rozdělení

- `binom.test(20,120,0.18,alternative="two.sided",conf.level=0.95)` # jednovýběrový oboustranný Clopperův-Pearsonův test parametru binomického rozdělení ($H_0: \pi = 0.18, H_A: \pi \neq 0.18$) + 95% Clopperův-Pearsonův intervalový odhad parametru binomického rozdělení pro $p = \frac{20}{200}$, parametr `alternative` může nabývat hodnot {"two.sided", "less", "greater"}

Statistická indukce pro dvě nezávislé proměnné (standardní datový formát)

Kvantitativní proměnná dle tříd (pro každou statistickou jednotku jsou v souboru `my.data` uvedeny atributy `vyska(numeric)` a `skupina(factor)` – data ve standardním datovém formátu, atribut `skupina` nabývá dvou variant)

Ověření normality

- `tapply(my.data$vyska,my.data$skupina,shapiro.test)` # test normality proměnné `vyska` pro každou variantu proměnné `skupina`

Test o shodě rozptylů

- `var.test(my.data$vyska~my.data$skupina,ratio=1,alternative="two.sided",conf.level=0.95)` # test poměru rozptylů $\left(H_0: \frac{\sigma_1^2}{\sigma_2^2} = 1, H_A: \frac{\sigma_1^2}{\sigma_2^2} \neq 1\right)$ + oboustranný 95% intervalový odhad poměru rozptylů, parametr `alternative` může nabývat hodnot {"two.sided", "less", "greater"}

Testy o shodě středních hodnot

- `t.test(my.data$vyska~my.data$skupina,mu=0,var.equal=TRUE,alternative="two.sided",conf.level=0.95)` # dvouvýběrový t-test ($H_0: \mu_1 - \mu_2 = 0, H_A: \mu_1 - \mu_2 \neq 0$) + oboustranný 95% intervalový odhad rozdílu středních hodnot pro homoskedasticitní data, parametr `alternative` může nabývat hodnot {"two.sided", "less", "greater"}
- `t.test(my.data$vyska~my.data$skupina,mu=0,var.equal=FALSE,alternative="two.sided",conf.level=0.95)` # Aspinové-Welchův test ($H_0: \mu_1 - \mu_2 = 0, H_A: \mu_1 - \mu_2 \neq 0$) + oboustranný 95% intervalový odhad rozdílu středních hodnot pro heteroskedasticitní data, parametr `alternative` může nabývat hodnot {"two.sided", "less", "greater"}

Testy o shodě mediánů

- `wilcox.test(my.data$vyska~my.data$skupina,mu=0,alternative="two.sided",conf.int=TRUE,conf.level=0.95)` # Mannův-Whitneyův test $\left(H_0: x_{0,5_1} - x_{0,5_2} = 0, H_A: x_{0,5_1} - x_{0,5_2} \neq 0\right)$ + oboustranný 95% intervalový odhad rozdílu mediánů, parametr `alternative` může nabývat hodnot {"two.sided", "less", "greater"}

Testy o shodě parametrů dvou binomických rozdělení

- `prop.test(c(10,20),c(1000,2100),conf.level=0.95,alternative="two.sided",conf.level=0.95)` # Pearsonův chí kvadrát test shody parametrů dvou binomických rozdělení s Yatesovou korekcí $\left(H_0: \pi_1 - \pi_2 = 0, H_A: \pi_1 - \pi_2 \neq 0, kde\ p_1 = \frac{10}{1000}, p_2 = \frac{20}{2100}\right)$ + oboustranný 95% intervalový odhad rozdílu parametrů dvou binomických rozdělení, parametr `alternative` může nabývat hodnot {"two.sided", "less", "greater"}

Statistická indukce pro tři a více nezávislé proměnné (standardní datový formát)

Kvantitativní proměnná dle tříd (pro každou statistickou jednotku jsou v souboru `my.data` uvedeny atributy `vyska(numeric)` a `skupina(factor)` – data ve standardním datovém formátu, atribut `skupina` nabývá tří nebo více variant)

Ověření normality

- `tapply(my.data$vyska,my.data$skupina,shapiro.test)` # test normality proměnné `vyska` pro každou variantu proměnné `skupina`

Vícevýběrové testy shody rozptylů

- `bartlett.test(my.data$vyska~my.data$skupina)` # Bartlettův test
- `leveneTest(my.data$vyska~my.data$skupina)` # Leveneův test (balíček car)

ANOVA

- `anova<-aov(my.data$vyska~my.data$skupina)` # do proměnné anova se uloží výsledky testu ANOVA
- `summary(anova)` # tabulka ANOVA (v proměnné `anova` musí být uložen výstup funkce `aov`)
- `TukeyHSD(anova)` # Tukeyho post hoc analýza (v proměnné `anova` musí být uložen výstup funkce `aov`)
- `plot(TukeyHSD(anova))` # grafická prezentace Tukeyho post hoc analýzy

- `oneway.test(my.data$vyska~my.data$skupina)` # ANOVA s Welchovou korekcí (pro heteroskedasticitní data s normálním rozdělením)

Kruskalův-Wallisův test

- `kruskal.test(my.data$vyska~my.data$skupina)` # Kruskalův-Wallisův test
- `dunn.test(my.data$vyska,my.data$skupina,method="bonferroni")` # Dunnové post hoc analýza s Bonferroniho korekcí (balíček dunn.test)

Analýza závislosti dvou kategoriálních proměnných (kontingenční tabulky)

Pro každou statistickou jednotku jsou v souboru `my.data` uvedeny atributy `vaha(factor)` a `nemoc(factor)` – data ve standardním datovém formátu, atribut `vaha` nabývá variant {nizka,normalni}, atribut `onemocneni` nabývá variant {ano,ne}

Tabulka sdružených četností

- `data<-table(my.data$vaha,my.data$nemoc,dnn=c("váha","onemocnění"))` # kontingenční tabulka (tabulka sdružených četností)

nebo

- `data<-matrix(c(12,23,45,54),nrow=2, byrow=FALSE)` # přímé zadání kontingenční tabulky (bez popisu řádků a sloupců)

vaha\nemoc	ano	ne
nizka	12	45
normalni	23	54

- `rownames(data)`<- c("nizka","normalni") # popisky řádků v kontingenční tabulce
- `colnames(data)`<- c("ano","ne") # popisky sloupců v kontingenční tabulce

Chí-kvadrát test nezávislosti (s Yatesovou korekcí)

- `reseni<-chisq.test(data)` # uložení výsledků chí-kvadrát testu nezávislosti s Yatesovou korekcí do proměnné `reseni`, v proměnné `data` musí být uložena tabulka sdružených četností
- `reseni` # výsledek testu (v proměnné `reseni` musí být uložen výstup funkce `chisq.test`)
- `reseni$observed` # pozorované četnosti O_i (v proměnné `reseni` musí být uložen výstup funkce `chisq.test`)
- `reseni$expected` # očekávané četnosti E_i (v proměnné `reseni` musí být uložen výstup funkce `chisq.test`)
- `reseni$residuals` # rozdíl mezi pozorovanými a očekávanými četnostmi $O_i - E_i$ (v proměnné `reseni` musí být uložen výstup funkce `chisq.test`)

Mozaikový graf

- `mosaicplot(data,color=TRUE,xlab="váha", ylab="onemocnění")` # mozaikový graf, v proměnné `data` musí být uložena tabulka sdružených četností

Míry kontingence

- `cramersV(data)` # Cramerovo V, v proměnné `data` musí být uložena tabulka sdružených četností (balíček `lsr`)

Speciální metody pro asociační tabulky

- `epi.2by2(data, conf.level=0.95)` # Chí-kvadrát test nezávislosti, bodový a 95% intervalový odhad poměru šancí a relativního rizika (balíček epiR), v proměnné `data` musí být uložena asociační tabulka v obvyklém formátu (1. řádek – exponovaná populace, 1. sloupec – výskyt jevu), **POZOR!** Tabulka musí být uložena jako datový typ matice (matrix)
- `data=data.matrix(data.z.data.frame)` # Do proměnné `data` (datový typ matrix) se uloží tabulka zapsaná v proměnné `data.z.data.frame` (datový typ data.frame)

Testy dobré shody

- `observed<-c(979,1002,1015,980,1040,984)` # uložení pozorovaných četností do proměnné `observed`
- `expected<-c(1/6,1/6,1/6,1/6,1/6,1/6)` # uložení očekávaných pravděpodobností do proměnné `expected`
- `chisq.test(observed,p=expected,rescale.p=TRUE)` # chí-kvadrát test dobré shody pozorovaných četností a teoretických pravděpodobností