Domača naloga 3

Alen Kahteran

7. 11. 2020

1. Želimo preveriti, ali je kovanec pošten. Naredili smo poizkus, kjer smo 100-krat vrgli kovanec in dobili, da je grb padel 61-krat. Denimo, da je vaša alternativna domneva $H_A : \pi > 0.5$. Odgovorite na spodnja vprašanja.

- Določite območje zavrnitve pri stopnji tveganja $\alpha = 0.05$.

```
qbinom(0.05, 100, 0.5, lower.tail=FALSE)
```

```
## [1] 58
```

Območje zavrnitve je torej $\{59, 60, \ldots, 99, 100\}$.

- Ali lahko na podlagi podatkov zavrnete ničelno domnevo v prid alternativni? Zakaj?

Da. Saj smo dobili 61 grbov, kar je v območju zavrnitve.

- Izračunajte tudi vrednost p.

```
pbinom(60, 100, 0.5, lower.tail = FALSE)
```

```
## [1] 0.0176001
```

- Kakšen statistični sklep sprejmete na podlagi izračunane vrednosti p? Zakaj?

Na podlagi izračunane vrednosti p lahko zavrnemo ničelno domnevo H_0, da je kovanec pošten, saj je $p < \alpha$.

- Zapišite vsebinski sklep

Verjetnost grba je večja od 0.5.

2. Preverite domnevo, da študenti **veterine** različno časa namenijo športu in gledanju televizije (datoteka `Ankete1011.txt`). Domnevo preverite pri stopnji tveganja $\alpha = 0.05$.

Najprej uredimo podatke.

```
data_full <- read.table("Ankete1011.txt", sep="\t", header=TRUE, dec=",")
data_full <- tibble(data_full)

data_full$Timestamp <- parse_date_time(data_full$Timestamp, c("dmY HM", "mdY HMS"))
data_full$Spol[data_full$Spol == "zenski"] <- "F"
data_full$Spol[data_full$Spol == "moski"] <- "M"
data_full$Visina <- as.numeric(gsub(",", ".", data_full$Visina))
data_full$Teza <- as.numeric(gsub(",", ".", data_full$Teza))
data_full$Cevelj <- as.numeric(gsub(",", ".", data_full$Cevelj))
data_full$Kajenje[data_full$Kajenje == "ne"] <- "N"
data_full$Kajenje[data_full$Kajenje == "da"] <- "Y"
data_full$Kajenje_koliko <- as.numeric(data_full$Kajenje_koliko)
data_full$Igrice[data_full$Igrice == "ne"] <- "N"
```

```
data_full$Igrice[data_full$Igrice == "da"] <- "Y"
data_full$Televizija <- as.numeric(gsub(",", ".", data_full$Televizija))
data_full$Knjige <- as.numeric(gsub(",", ".", data_full$Knjige))
data_full$Sport <- as.numeric(gsub(",", ".", data_full$Sport))

# checking rows with NA values
print(data_full[rowSums(is.na(data_full)) > 0, ], width=Inf)
```

```
## # A tibble: 2 x 17
##   Timestamp           Starost Spol  Visina  Teza Cevelj BarvaOci Kajenje
##   <dttm>                <int> <chr>  <dbl> <dbl>  <dbl> <chr>    <chr>
## 1 2010-10-30 18:10:34      19 M        169    68     42 rjava    N
## 2 2011-04-03 16:47:00      19 F        171    68     38 zelena   N
##   Kajenje_koliko Igrice Televizija Internet Knjige Sport Domace_zivali
##            <dbl> <chr>       <dbl>    <int>  <dbl> <dbl> <chr>
## 1              0 N               1       20     NA     3 Ne
## 2              0 N               6       15     NA     5 Riba
##   Studij           Fakulteta
##   <chr>            <chr>
## 1 Veterina         VF
## 2 Splosna medicina MF
```

```
# imputing missing values with median
data_full$Knjige[is.na(data_full$Knjige)] <- median(data_full$Knjige, na.rm=TRUE)
```

Omejimo se še na študente veterine

```
# get only students who study veterinary medicine.
data_vet <- data_full %>%
    filter(Studij=="Veterina")
```

- S katerim testom boste preverili domnevo?

T test za parne meritve (Odvisna vzorca).

- Z besedami zapišite ničelno domnevo.

študenti veterine v povprečju namenijo športu in gledanju televizije enako časa.

$\mu_{sp} = \mu_{tv}$ oz. $\mu_{sp} - \mu_{tv} = 0$

- Koliko je znašala vrednost p? Ali ničelno domnevo zavrnete?

```
t.test(data_vet$Sport, data_vet$Televizija, paired=TRUE)
```

```
##
##  Paired t-test
##
## data:  data_vet$Sport and data_vet$Televizija
## t = 1.5021, df = 43, p-value = 0.1404
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.4359706  2.9814251
## sample estimates:
## mean of the differences
##                1.272727
```

$p = 0.1404$

Ne, ničelne domneve ne moremo zavrniti saj je $p > \alpha$.

- Zapišite vsebinski sklep. Po potrebi lahko še kaj dodatno izračunate in komentirate.

H_0 ne moremo zavrniti. Torej ne moremo trditi da študentje veterine v povprečju različno časa namenijo športu kot televiziji. Razlika povprečij je $1.27h$, $p = 0.1404$ in $\alpha = 0.05$. Interval zaupanja je $[-0.436, 2.981]$. Torej če razlika povprečij ne bi bila v temu intervalu, bi lahko H_0 zavrnili.

- Komentirajte izpolnjenost predpostavk.

Predpostavka je, da je razlika vzorcev porazdeljena normalno. V našem primeru je velikost vzorca 44 (vidimo iz stopinj prostosti, ki so 43). Kar je po mojem mnenju dokaj malo za določanje normalnosti. Zato uporabimo Anderson-Darling test, ki za H_0 pravi da je vzorec normalno porazdeljen

```
library(nortest) # implementation of ad.test

ad.test(data_vet$Sport - data_vet$Televizija)
```

```
##
##  Anderson-Darling normality test
##
## data:  data_vet$Sport - data_vet$Televizija
## A = 0.57102, p-value = 0.1305
```

Torej na podlagi p vrednosti ne moremo ovreči H_0, tj. ne moremo reči da naš vzorec ni normalno porazdeljen.

Poglejmo še Shapiro-Wilk test, ki ravno tako preverja normalnost porazdelitve.

```
library(nortest) # implementation of ad.test

shapiro.test(data_vet$Sport - data_vet$Televizija)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data_vet$Sport - data_vet$Televizija
## W = 0.9532, p-value = 0.07242
```

Tu ravno tako ne moremo zavrniti ničelne hipoteze, ki ravno tako predpostavlja da je naša porazdelitev normalno porazdeljena.

3. Preverite domnevo, da študenti **veterine**, ki kadijo, več časa gledajo televizijo kot tisti, ki ne kadijo (datoteka `Ankete1011.txt`). Domnevo preverite pri stopnji tveganja $\alpha = 0.05$

- S katerim testom boste preverili domnevo?

t test za neodvisna vzorca.

- Z besedami zapišite ničelno domnevo.

$H_0 : \mu_k = \mu_{nk}$

Tj. Povprečna uporaba televizije pri kadilcih je enaka kot pri nekadilcih v populaciji študentov veterine.

- Koliko je znašala vrednost p? Ali ničelno domnevo zavrnete?

```
t.test(data_vet$Televizija ~ data_vet$Kajenje, paired=FALSE, var.equal=TRUE)
```

```
##
##  Two Sample t-test
##
## data:  data_vet$Televizija by data_vet$Kajenje
```

```
## t = -0.52883, df = 42, p-value = 0.5997
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -6.029962  3.525897
## sample estimates:
## mean in group N mean in group Y
##        4.414634        5.666667
```

```
t.test(data_vet$Televizija ~ data_vet$Kajenje, paired=FALSE, var.equal=FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  data_vet$Televizija by data_vet$Kajenje
## t = -0.26435, df = 2.0584, p-value = 0.8156
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -21.08644  18.58237
## sample estimates:
## mean in group N mean in group Y
##        4.414634        5.666667
```

Prvi test je ob predpostavki da je varianca enaka v obeh vzorcih, druga pa da ni. V obeh primerih je p vrednost več ko $\alpha = 0.05$ in na podlagi tega ničelne domneve ne moremo zavrniti.

- Zapišite vsebinski sklep testa. Po potrebi lahko še kaj dodatno izračunate in komentirate.

Na podlagi dobljenih p vrednosti ne moremo trditi da je povprečna uporaba televizije med kadilci in nekadilci različna.

- Komentirajte izpolnjenost predpostavk.

Predpostavka je, da sta obe porazdelitvi porazdeljeni normalno, kar težko trdimo da drži, saj v primeru kadilcev imamo samo 3 zapise. Vseeno opravimo Anderson-Darling in Shapiro-Wilk test za oba vzorca kjer pri obeh testih velja H_0 da je vzorec normalno porazdeljen.

```
# ad.test(data_vet$Televizija[data_vet$Kajenje == "Y"])
#
# returns an error, so this can't be run.
#
# Error in ad.test(data_vet$Televizija[data_vet$Kajenje == "Y"]) :
#   sample size must be greater than 7
ad.test(data_vet$Televizija[data_vet$Kajenje == "N"])
```

```
##
##  Anderson-Darling normality test
##
## data:  data_vet$Televizija[data_vet$Kajenje == "N"]
## A = 1.5512, p-value = 0.0004576
```

```
shapiro.test(data_vet$Televizija[data_vet$Kajenje == "Y"])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data_vet$Televizija[data_vet$Kajenje == "Y"]
## W = 0.84799, p-value = 0.2351
```

```r
shapiro.test(data_vet$Televizija[data_vet$Kajenje == "N"])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data_vet$Televizija[data_vet$Kajenje == "N"]
## W = 0.8883, p-value = 0.0007624
```

Kljub nedelovanju Anderson-Darling testa na kadilcih, vidimo da oba testa v primeru nekadilcev imata p vrednost manjšo od α, in tu lahko zavrnemo ničelno hipotezo da je vzorec nekadilcev normalno porazdeljen. V primeru kadilcev nam Shapiro-Wilk test vrne p vrednost večjo od α in zato ne moremo zavrniti H_0. Ampak, kot že omenjeno, je to najbrž zaradi treh zapisov. Posledično si lahko pogledamo še Mann-Whitneyev test, za katerega ni predpostavk (razen o neodvisnosti dogodkov). H_0 je v temu primeru da sta porazdelitvi uporabe televizije enaki pri kadilcih in nekadilcih v populaciji študentov veterine.

```r
wilcox.test(data_vet$Televizija ~ data_vet$Kajenje, paired=FALSE)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  data_vet$Televizija by data_vet$Kajenje
## W = 69, p-value = 0.7424
## alternative hypothesis: true location shift is not equal to 0
```

Na podlagi p vrednosti ne moremo zavrniti ničelne hipoteze.

4. V štirih državah smo preučevali povezanost med lastništvom živali (da/ne) in izbrano fakulteto (A/B), rezultati so povzeti v spodnji tabeli. Zanima nas, v kateri državi je povezanost med fakulteto in lastništvom živali najmočnejša.

		A1	B1	Vsota	A2	B2	Vsota	A3	B3	Vsota	A4	B4	Vsota
Živali	Da	60	40	100	6	4	10	6000	5200	11200	6	2	8
	Ne	40	60	100	4	6	10	4000	4800	8800	4	8	12
	Vsota	100	100	200	10	10	20	10000	10000	20000	10	10	20

- Izpolnite spodnjo tabelo: vrednost p je izračunana na podlagi primernega testa za primerjavo dveh neodvisnih deležev, stopnja tveganja $\alpha = 0.05$, primerjana deleža sta deleža lastnikov živali v posamezni fakulteti, stopnjo povezanosti ovrednotite sami.

Izračun stopnje povezanosti

– Cramérjev koeficient asociiranosti

```r
cramer_coef <- function(a, b, c, d){
    return(((a*d)-(b*c))/sqrt((a+b)*(c+d)*(a+c)*(b+d)))
}

# 1
cramer_coef(60, 40, 40, 60)
```

```
## [1] 0.2
```

```r
# 2
cramer_coef(6, 4, 4, 6)
```

```
## [1] 0.2
```

```r
# 3
cramer_coef(6000, 5200, 4000, 4800)
```

```
## [1] 0.0805823
```

```r
# 4
cramer_coef(6, 2, 4, 8)
```

```
## [1] 0.4082483
```

Država	1	2	3	4
Vrednost p	0.005	0.371	< 0.001	0.068
Sklep testa	H_0 zavrnemo	H_0 ne zavrnemo	H_0 zavrnemo	H_0 ne zavrnemo
Primerjana deleža	0.6 in 0.4	0.6 in 0.4	0.60 in 0.52	0.6 in 0.2
Stopnja povezanosti	0.2 - nizka	0.2 - nizka	0.08 - nizka	0.41 - srednja

- Ali je vrednost p primerna mera, s katero lahko povzamemo stopnjo povezanosti med spremenljivkami? Kratko utemeljite.

Ne. Statistična značilnost (p vrednost) ni enako kot stopnja povezanosti. Pri majhnih vzorcih se zna zgoditi, da je povezanost visoka, a statistično neznačilna. Pri velikih vzorcih pa se lahko zgodi celo, da je povezanost nizka, a je statistično zelo značilna.