1. Domača naloga

Uvod v statistiko

Alen Kahteran

9. 10. 2020

Contents

Naloga #1

- Naredili ste pet meritev in dobili povprečje x, mediano 20 in modus 10, kjer je x vsota števk vaše vpisne številke. Izmislite si podatke, ki ustrezajo tem opisnim statistikam.

```
vpis_st <- "64200421"

vsota_stevk <- sapply(strsplit(vpis_st, ""), function(x) sum(as.numeric(x)))
vsota_stevk
```

```
## [1] 19
```

Torej potrebujemo 5 številk, ki imajo povprečje 19, mediano 20 in modus 10.

```
# izmisljene meritve
mer5 <- c(10,10,20,25,30)

# povprecje
mean(mer5)
```

```
## [1] 19
```

```
# mediana
median(mer5)
```

```
## [1] 20
```

```
# funkcija za izracun modusa
get_modus <- function(vec) {
   uniqvec <- unique(vec)
   uniqvec[which.max(tabulate(match(vec, uniqvec)))]
}

# modus
get_modus(mer5)
```

```
## [1] 10
```

Naloga #2

```
# preberemo podatke
data_full <- read.table("Ankete1011.txt", sep="\t", header=TRUE)
```

```
# spremenimo v tibble (zato je potreben Tidyverse paket)
data_full <- tibble(data_full)
data_full
```

```
## # A tibble: 140 x 17
##    Timestamp Starost Spol  Visina Teza  Cevelj BarvaOci Kajenje Kajenje_koliko
##    <chr>       <int> <chr> <chr>  <chr>  <int> <chr>    <chr>            <int>
##  1 10.07.20~      20 zens~ 170    62        41 rjava    ne                   0
##  2 10.07.20~      19 moski 188    75        46 plava    ne                   0
##  3 10.07.20~      20 zens~ 156    54        38 zelena   ne                   0
##  4 10.07.20~      20 zens~ 165    77        39 rjava    ne                   0
##  5 10.07.20~      18 zens~ 170    52        38 plava    ne                   0
##  6 10.08.20~      18 zens~ 167    53        38 crna     ne                   0
##  7 10.08.20~      19 moski 180    69        43 plava    ne                   0
##  8 10.09.20~      19 zens~ 173    63        40 zelena   da                   1
##  9 10.10.20~      19 moski 182    80        45 rjava    ne                   0
## 10 10.10.20~      19 zens~ 165    92        44 crna     ne                   0
## # ... with 130 more rows, and 8 more variables: Igrice <chr>, Televizija <chr>,
## #   Internet <int>, Knjige <chr>, Sport <chr>, Domace_zivali <chr>,
## #   Studij <chr>, Fakulteta <chr>
```

Kot vidimo so nekatere spremenljivke napačnega tipa. Potrebno jih je korektno spremeniti.

```
# pravilno spremenimo zapis datumov v datetime objekt
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))

# preverjanje vseh unikatnih zapisov v spremenljivki Domace_zivali
unique(unlist(strsplit(data_full$Domace_zivali, ", ")))
```

```
## [1] "Macka"   "Ptic"    "Da"      "drugo"   "Pes"     "Godalec" "Riba"
## [8] "Ne"
```

```
# dodajanje novih stolpcev kjer bodo logicne vrednosti "Y" ali "N" glede na to
# ali ima nekdo doloceno zival ali ne, poleg tega se spremenljivka Domace_zivali
# spremeni v logicno vrednost "Y" ali "N", kjer ce ima usaj eno zival ima "Y" drugace "N"
data_full <- data_full %>%
    mutate(Macka=NA,
           Ptic=NA,
           Pes=NA,
           Glodalec=NA,
           Riba=NA,
           Drugo=NA) %>%
```

```
    mutate(Domace_zivali_temp = Domace_zivali) %>%
    mutate(Domace_zivali = if_else(Domace_zivali == "Ne", "N", "Y"))

data_full <- data_full %>%
    mutate(., Macka    = if_else(str_detect(.$Domace_zivali_temp, "Macka"),    "Y", "N")) %>%
    mutate(., Ptic     = if_else(str_detect(.$Domace_zivali_temp, "Ptic"),     "Y", "N")) %>%
    mutate(., Pes      = if_else(str_detect(.$Domace_zivali_temp, "Pes"),      "Y", "N")) %>%
    mutate(., Glodalec = if_else(str_detect(.$Domace_zivali_temp, "Godalec"), "Y", "N")) %>%
    mutate(., Riba     = if_else(str_detect(.$Domace_zivali_temp, "Riba"),     "Y", "N")) %>%
    mutate(., Drugo    = if_else(str_detect(.$Domace_zivali_temp, "Drugo"),    "Y", "N")) %>%
    select(-Domace_zivali_temp)

# Preverjanje vrstic ce kateri stolpec nima zapisa
print(data_full[rowSums(is.na(data_full)) > 0, ], width=Inf)
```

```
## # A tibble: 2 x 23
##   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 N
## 2              0 N               6       15     NA     5 Y
##   Studij           Fakulteta Macka Ptic  Pes   Glodalec Riba  Drugo
##   <chr>            <chr>     <chr> <chr> <chr> <chr>    <chr> <chr>
## 1 Veterina         VF        N     N     N     N        N     N
## 2 Splosna medicina MF        N     N     N     N        Y     N
```

```
# nadomescanje manjkajocih vrednosti z mediano
data_full$Knjige[is.na(data_full$Knjige)] <- median(data_full$Knjige, na.rm=TRUE)

print(data_full, width=Inf)
```

```
## # A tibble: 140 x 23
##    Timestamp           Starost Spol  Visina  Teza Cevelj BarvaOci Kajenje
##    <dttm>                <int> <chr>  <dbl> <dbl>  <dbl> <chr>    <chr>
##  1 2010-07-10 10:01:00      20 F        170    62     41 rjava    N
##  2 2010-07-10 11:02:00      19 M        188    75     46 plava    N
##  3 2010-07-10 13:18:00      20 F        156    54     38 zelena   N
##  4 2010-07-10 14:53:00      20 F        165    77     39 rjava    N
##  5 2010-07-10 15:27:00      18 F        170    52     38 plava    N
##  6 2010-08-10 17:31:00      18 F        167    53     38 crna     N
##  7 2010-08-10 18:18:00      19 M        180    69     43 plava    N
##  8 2010-09-10 13:00:00      19 F        173    63     40 zelena   Y
##  9 2010-10-10 10:38:00      19 M        182    80     45 rjava    N
## 10 2010-10-10 14:13:00      19 F        165    92     44 crna     N
##    Kajenje_koliko Igrice Televizija Internet Knjige Sport Domace_zivali Studij
##             <dbl> <chr>       <dbl>    <int>  <dbl> <dbl> <chr>         <chr>
##  1              0 N             5          9   15      10 Y             Veterina
##  2              0 Y            10         25    1       7 Y             Veterina
##  3              0 N             0         10   10       1 Y             Veterina
##  4              0 N            15         10   20      10 Y             Veterina
##  5              0 N             7         16   10       2 Y             Veterina
```

```
##  6              0 Y             4          10   15       9 Y             Veterina
##  7              0 N             2.5        11    1.5    10 Y             Veterina
##  8              1 N             2           5   10       2 Y             Veterina
##  9              0 N             5           7    4       3 Y             Veterina
## 10              0 N             0           1   24.5     0 Y             Veterina
##    Fakulteta Macka Ptic  Pes   Glodalec Riba  Drugo
##    <chr>     <chr> <chr> <chr> <chr>    <chr> <chr>
##  1 VF        Y     Y     N     N        N     N
##  2 VF        N     N     Y     N        N     N
##  3 VF        Y     N     Y     N        N     N
##  4 VF        Y     Y     N     Y        N     N
##  5 VF        Y     N     Y     N        N     N
##  6 VF        N     N     Y     N        N     N
##  7 VF        Y     N     Y     N        N     N
##  8 VF        Y     Y     Y     N        Y     N
##  9 VF        N     N     N     Y        N     N
## 10 VF        N     N     N     N        N     N
## # ... with 130 more rows
```

Po obdelavi podatkov, izberemo 40 zapisov kot omenjeno v navodilih.

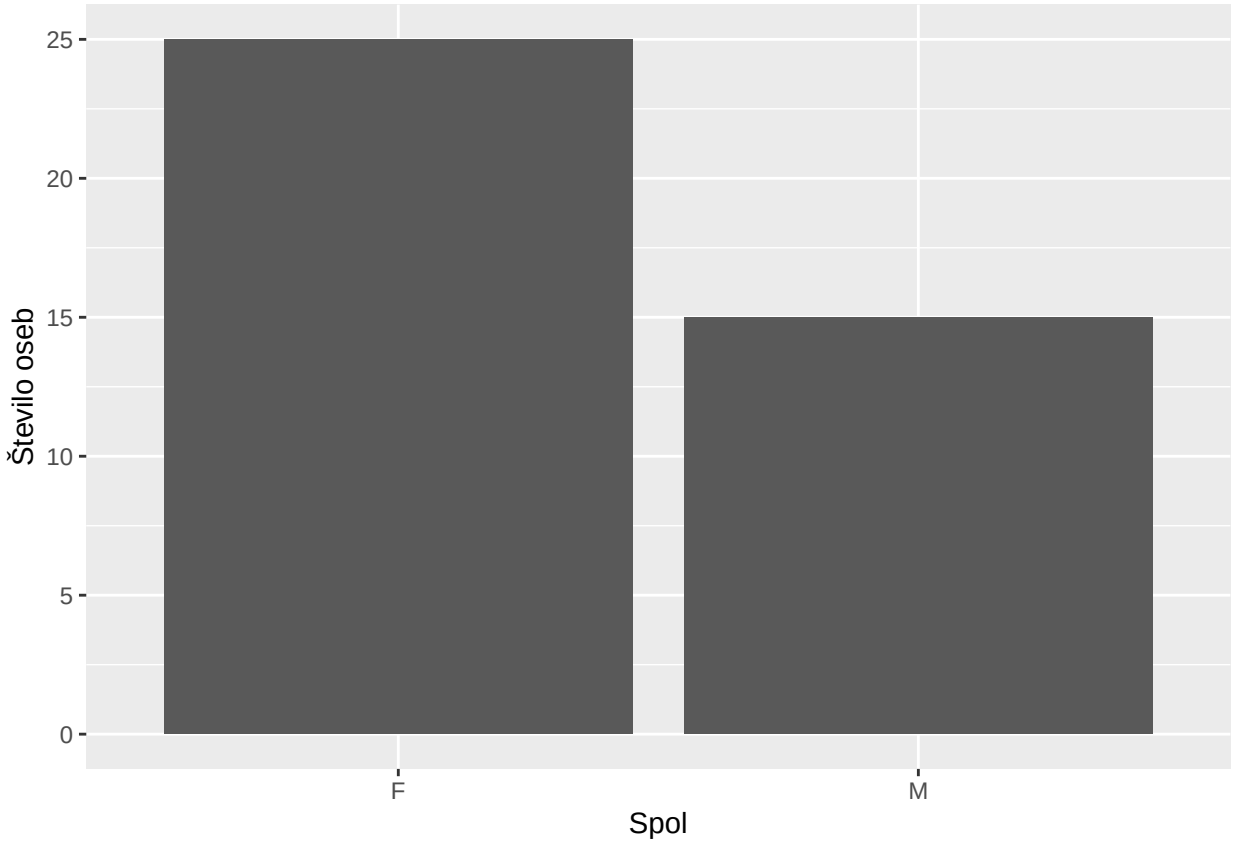
```
# nastavimo seme
set.seed(19)

data_40 <- sample_n(data_full, 40)
```

- Kakšen je delež moških na vašem vzorcu? Porazdelitev spremenljivke spol za vaš vzorec prikažite v ustreznem grafu, kjer naj bo razvidna absolutna frekvenca.

```
male_num <- nrow(data_40[data_40$Spol == "M", ])
total_num <- nrow(data_40)
print(male_num / total_num)
```

```
## [1] 0.375
```

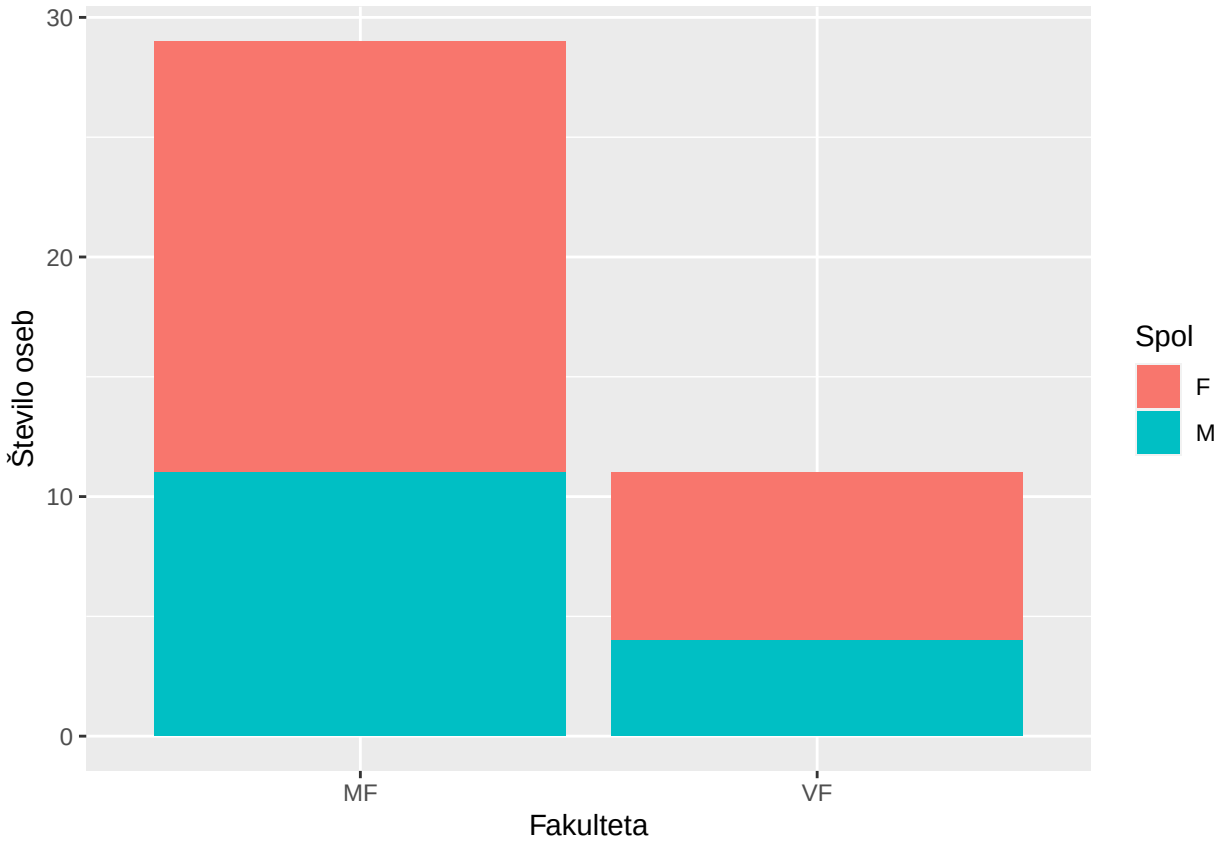
```
g <- ggplot(data_40, aes(Spol))

g + geom_bar() + ylab("Število oseb")
```



- Porazdelitev spremenljivke fakulteta prikažite po spolu. V grafu naj bodo razvidne relativne frekvence znotraj vrednosti spremenljivke fakulteta. Komentirajte morebitne razlike, ki jih grafično opazite na vašem vzorcu.

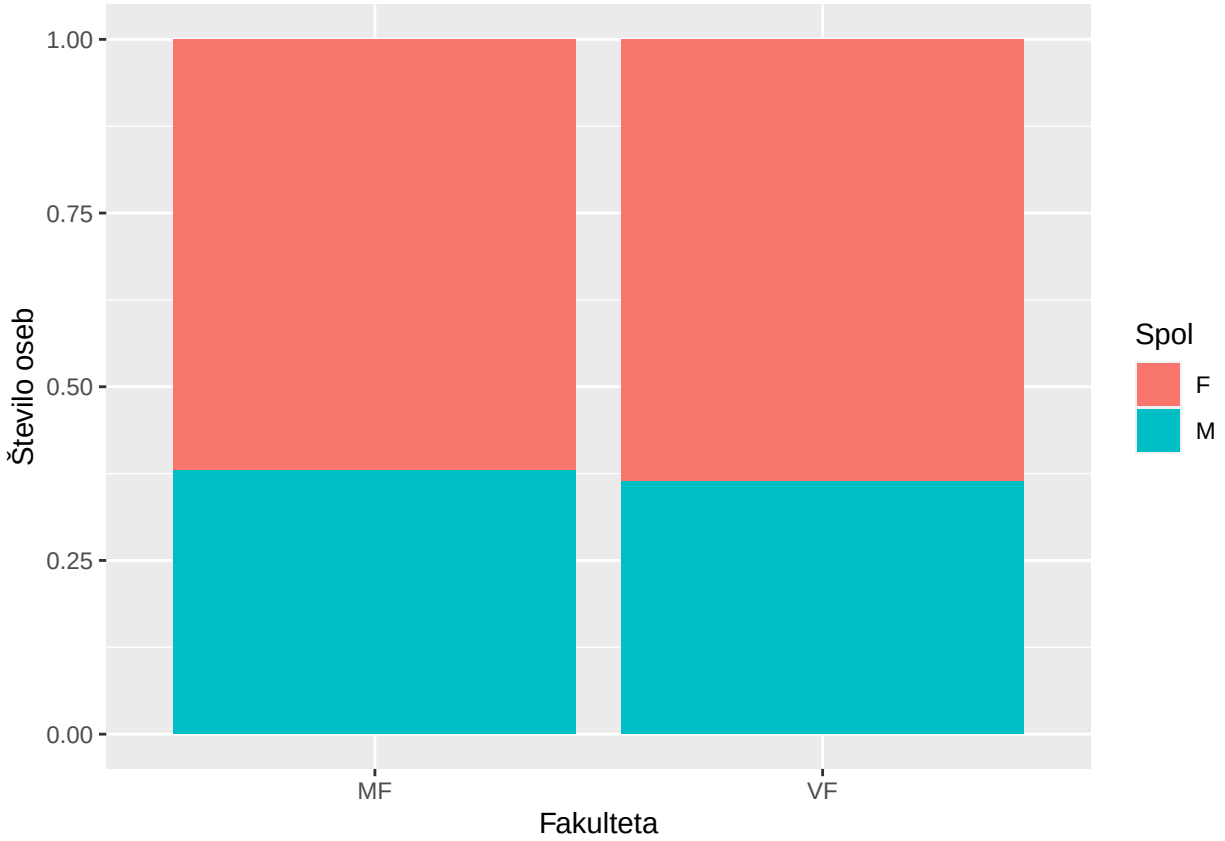
Naslednji graf prikazuje absolutno frekvenco oseb na različnih fakultetah razdeljeno na spol.

```
g <- ggplot(data_40, aes(Fakulteta))
g + geom_bar(aes(fill = Spol)) + ylab("Število oseb")
```



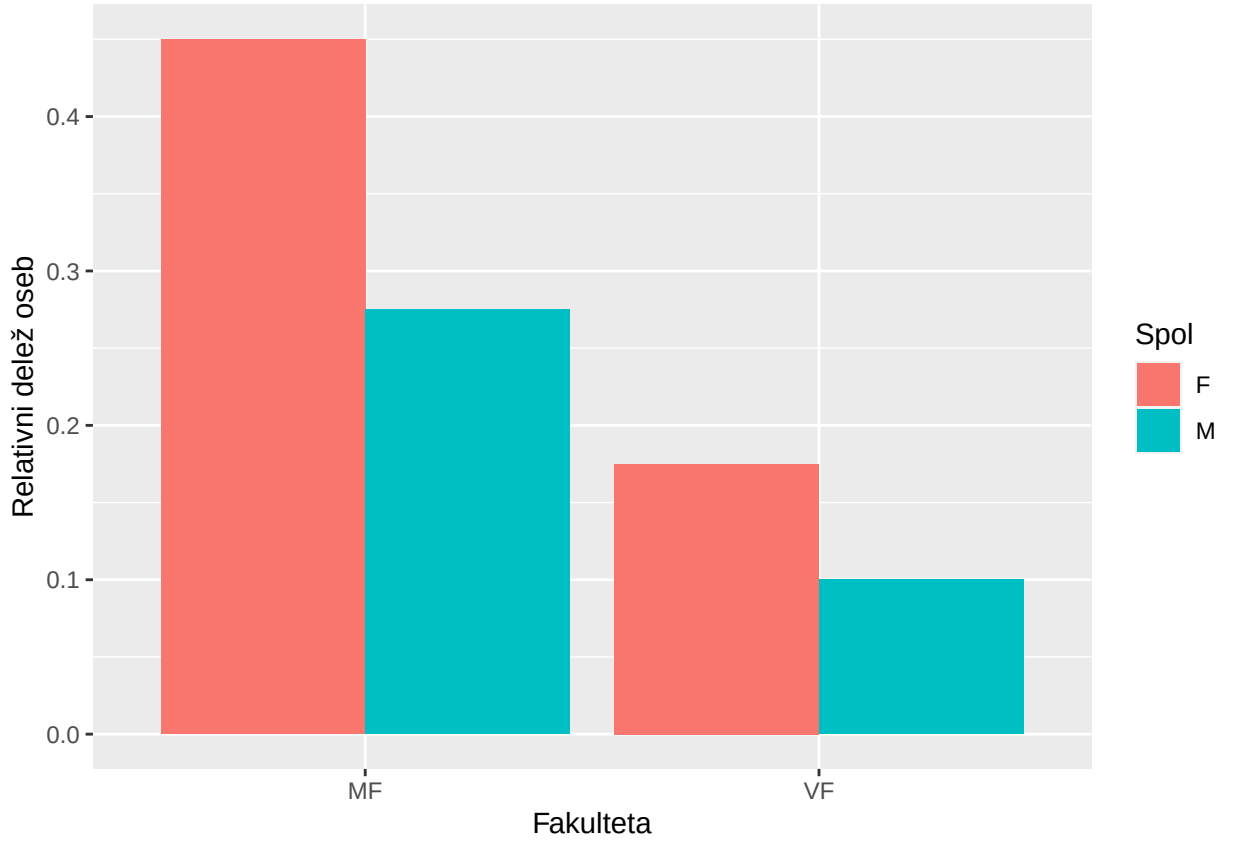
Naslednji graf prikazuje delež spola v našem vzorcu na različnih fakultetah.

```
# dodatek 16.10.2020
# če želimo primerjati spol po fakulteti
g <- ggplot(data_40, aes(Fakulteta))
g + geom_bar(aes(fill = Spol), position="fill") + ylab("Število oseb")
```



Naslednji graf pa prikazuje relativni delež vseh študentov po fakultetah in spolu.

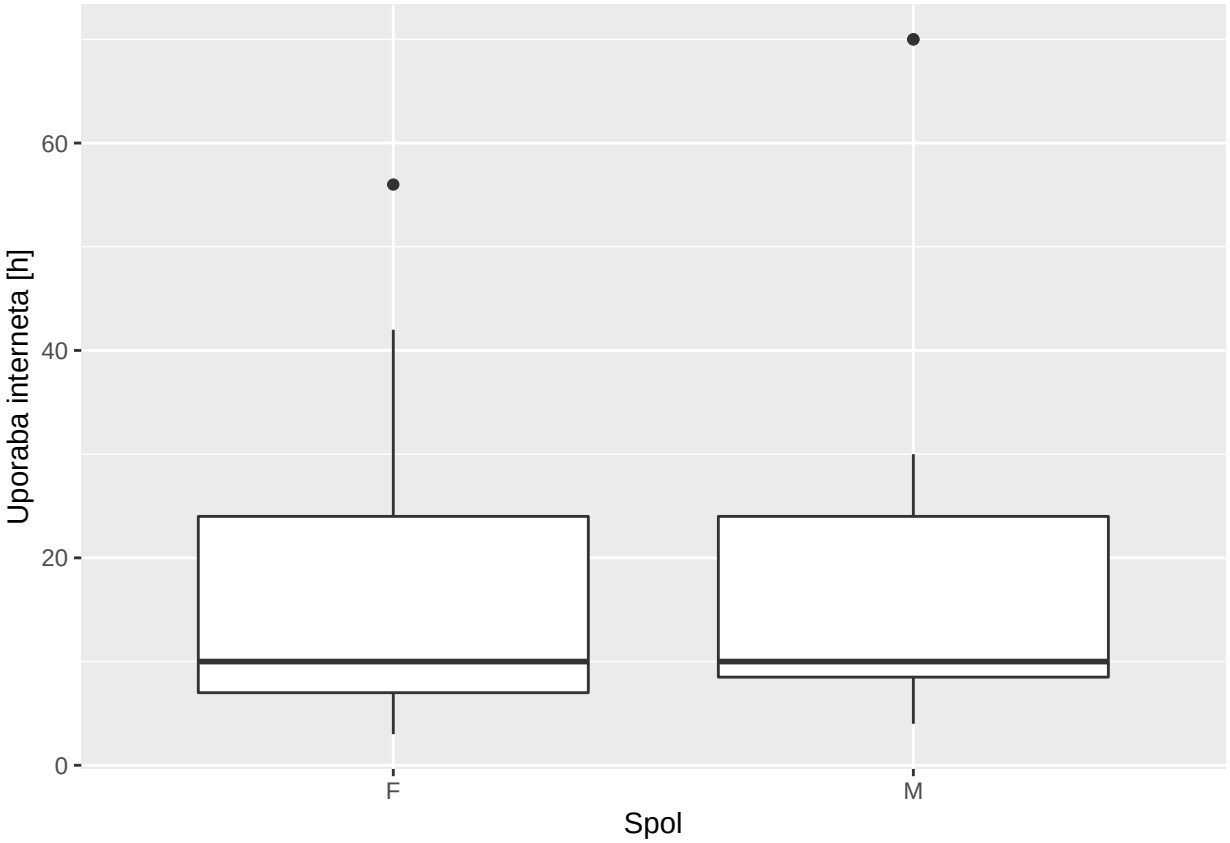
```
# dodatek 16.10.2020
df_temp <- data_40 %>%
  group_by(Fakulteta, Spol) %>%
  summarise(Total=n()) %>%
  ungroup() %>%
  mutate(freq=Total / sum(Total))
g <- ggplot(df_temp, aes(x=Fakulteta, y=freq))
g + geom_bar(aes(fill=Spol),
             position="dodge",
             stat = "identity") +
  ylab("Relativni delež oseb")
```



Večjih razlik v deležu moških (ali žensk) na fakultetah ne opazim. Na obeh fakultetah je približno 40% moških oz. 60% žensk. Opazno je le to, da je v mojem vzorcu skoraj trikrat toliko oseb vpisanih na Medicinsko fakulteto kot na Veterinarsko fakulteto.

- Z okvirjem z ročaji prikažite razliko v uporabi interneta med moškimi in ženskami. Komentirajte morebitne razlike med spoloma. Ločeno izračunajte vse vrednosti, ki so navedene v grafu in jih interpretirajte.

```
b <- ggplot(data_40, aes(x=Spol, y=Internet))
b + geom_boxplot() + ylab("Uporaba interneta [h]")
```




```
# 0 predstavlja minimum, 1 maximum. 0.25 spodnji kvartil, 0.5 je mediana, 0.75 zgornji kvartil
# type=4 pri izracunu uporabi linearno interpolacijo, ce to potrebuje.
quantile(data_40$Internet[data_40$Spol == "M"], probs=c(0, 0.25, 0.5, 0.75, 1), type=4)
```

```
##   0%  25%  50%  75% 100%
##  4.0  6.5 10.0 22.0 70.0
```

```
quantile(data_40$Internet[data_40$Spol == "F"], probs=c(0, 0.25, 0.5, 0.75, 1), type=4)
```

```
##   0%  25%  50%  75% 100%
##  3.0  5.5 10.0 22.0 56.0
```

```
# lahko preverimo
min(data_40$Internet[data_40$Spol == "M"])
```

```
## [1] 4
```

```
max(data_40$Internet[data_40$Spol == "M"])
```

```
## [1] 70
```

```
median(data_40$Internet[data_40$Spol == "M"])
```

```
## [1] 10
```

```
min(data_40$Internet[data_40$Spol == "F"])
```

```
## [1] 3
```

```
max(data_40$Internet[data_40$Spol == "F"])
```

```
## [1] 56
```

```
median(data_40$Internet[data_40$Spol == "F"])
```

```
## [1] 10
```

Iz samega grafa se težko kaj razbere, saj večjih razlik ni. Pri ženskah se opazi nekoliko daljši zgornji rep porazdelitve. Ter to, da je osamelec pri moški porazdelitvi veliko dlje od ostale populacije moških.

Spodnja točka ročaja v obeh primerih predstavlja tudi minimum, saj ni nobene točke izven. Na zgornji strani pa točka predstavlja maximum, konec ročaja pa najbližjo manjšo vrednost od vrednosti tretjega kvartila + 1.5 interkvartilnega razmika (interkvartilni razmik: $IQR = zgornji - spodnji$, kjer $zgornji$ predstavlja vrednost zgornjega kvartila in $spodnji$ predstavlja vrednost spodnjega kvartila)

da samostojno izračunamo mejo zgornjega ročaja je potrebno najprej izračunati IQR

```
IQR(data_40$Internet[data_40$Spol == "M"])
```

```
## [1] 15.5
```

```
IQR(data_40$Internet[data_40$Spol == "F"])
```

```
## [1] 17
```

Nato uporabimo zgornje podatke za zgornji kvartil, da izračunamo mejo za katero bomo gledali naslednjo manjšo vrednost.

```
f_meja <- 22 + 15.5*1.5
m_meja <- 22 + 17*1.5
f_meja
```

```
## [1] 45.25
```

```
m_meja
```

```
## [1] 47.5
```

Torej v moškem primeru nas zanima prvo število, ki je manjše od 47,5. V ženskem primeru pa nas zanima prvo število, ki je manjše od 45,5.

```
max(data_40[(data_40$Internet <= m_meja) & (data_40$Spol == "M"), ]$Internet)
```

```
## [1] 30
```

```
max(data_40[(data_40$Internet <= m_meja) & (data_40$Spol == "F"), ]$Internet)
```

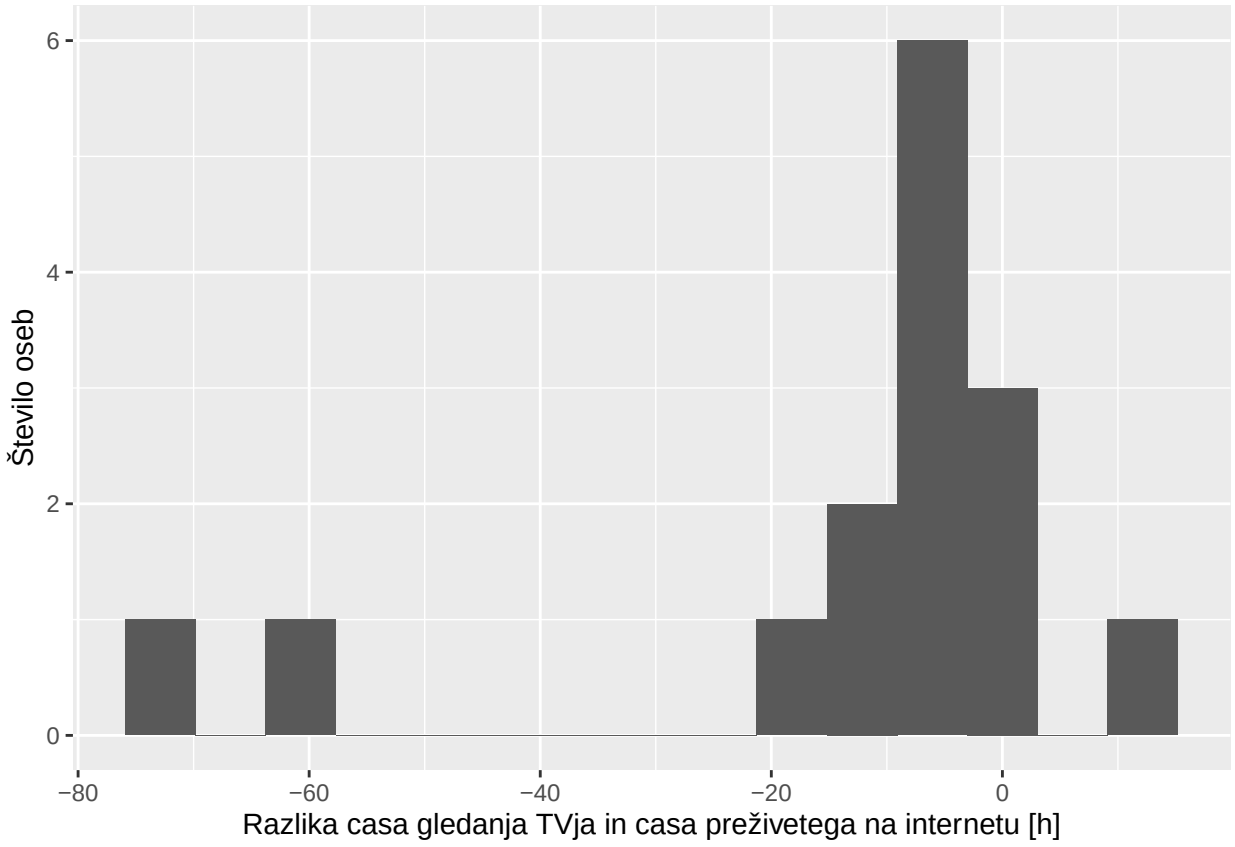
```
## [1] 42
```

Iz slike je razvidno da se številki ujemata.

Izberite podatke zgolj za moške. V ustreznem grafu prikažite porazdelitev razlike med številom ur gledanja televizije in uporabo interneta. Izračunajte ustrezno mero srednje vrednosti in razpršenosti ter ju interpretirajte.

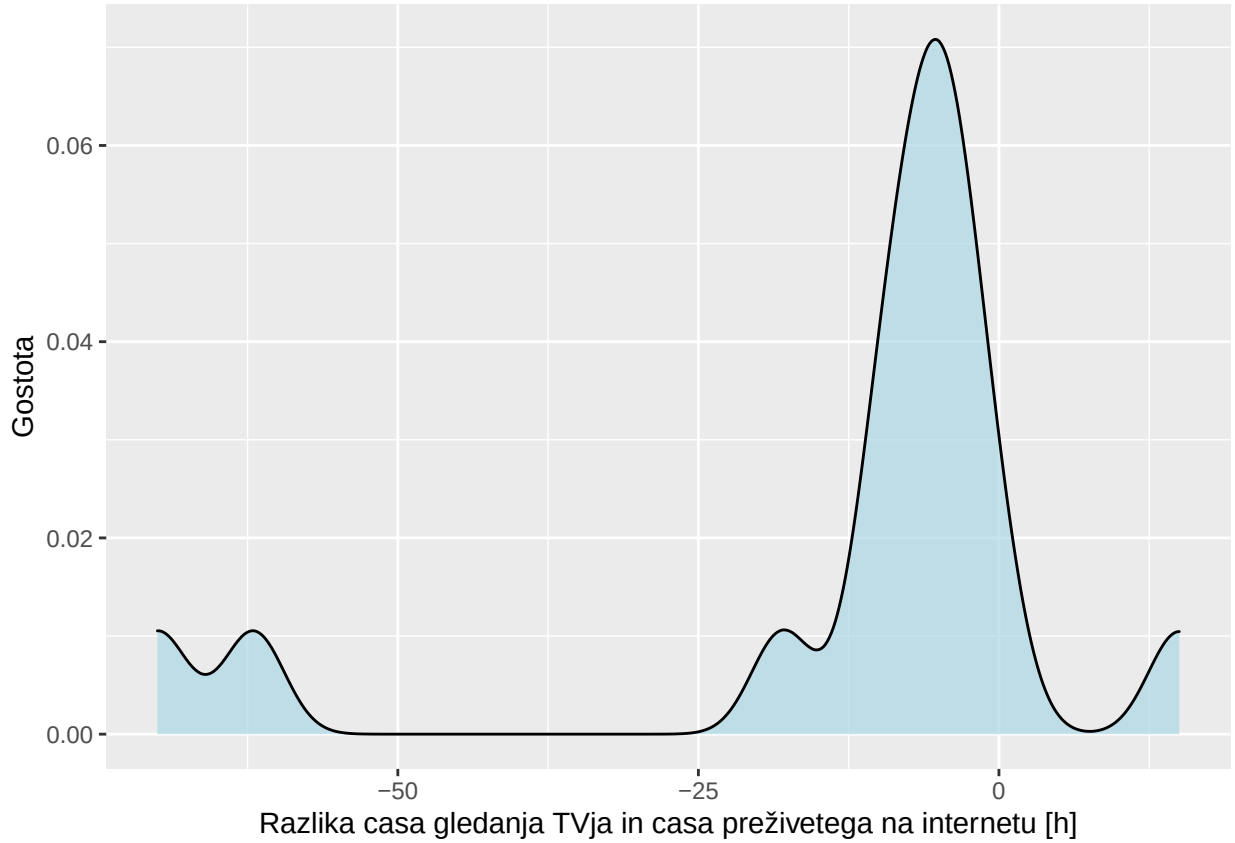
```
# izberimo podatke le za moške
data_40 <- data_40 %>% mutate(diff_TV_Int = Televizija - Internet)
data_40_m <- data_40[data_40$Spol == "M", ]

# histogram
h <- ggplot(data_40_m, aes(diff_TV_Int))
h + geom_histogram(bins=15) +
```

```
    ylab("Število oseb") +
    xlab("Razlika časa gledanja TVja in časa preživetega na internetu [h]")
```



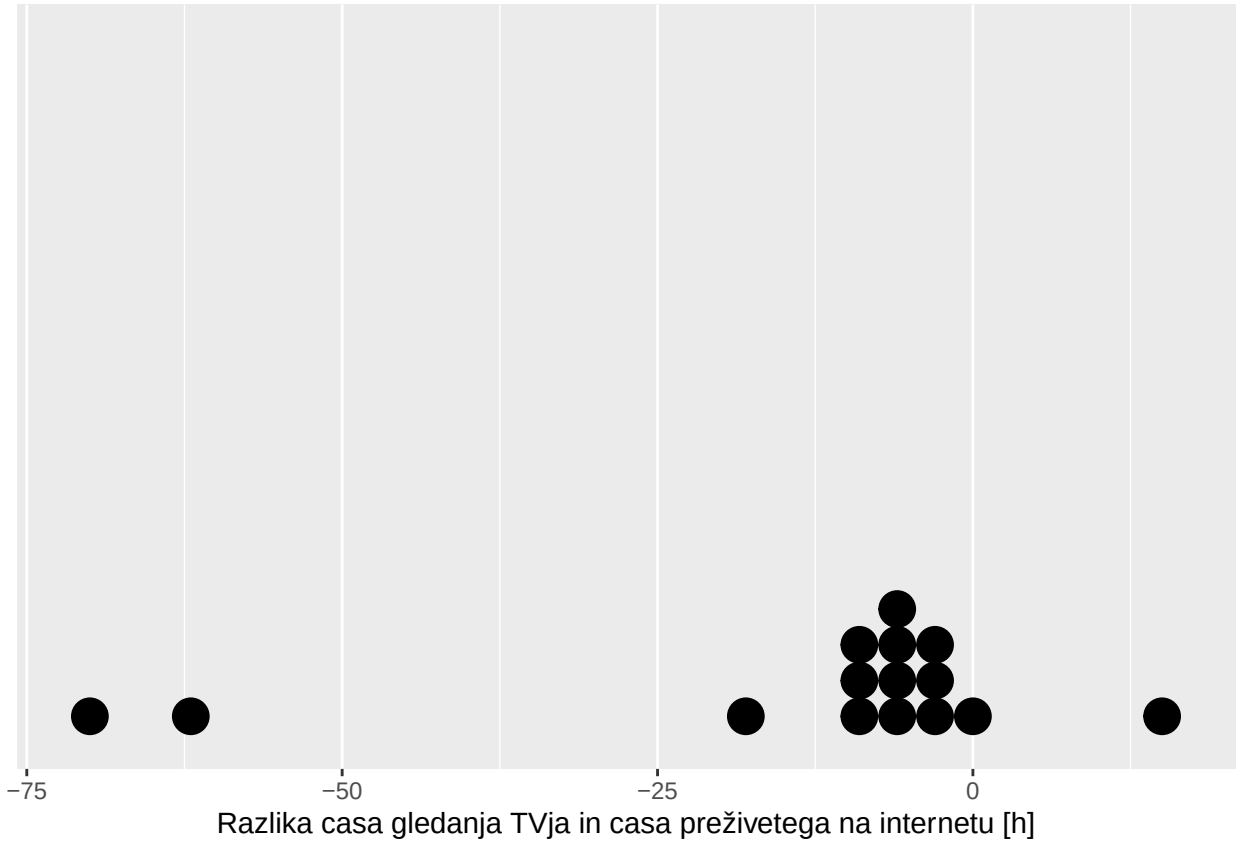
Še prikaz gostote na drugačna načina. `geom_dotplot()` je v temu primeru mogoče še najbolj uporaben, saj je meritev zelo malo, in je lepo razvidno, kako je razporejenih 15 ljudi.

```
# dodatek 16. 10. 2020
h + geom_density(col="black", fill="lightblue", alpha=0.7) +
  ylab("Gostota") +
  xlab("Razlika časa gledanja TVja in časa preživetega na internetu [h]")
```




```
h + geom_dotplot() +
  ylab("Gostota") +
  xlab("Razlika časa gledanja TVja in časa preživetega na internetu [h]") +
  scale_y_continuous(NULL, breaks = NULL)
```

```
## `stat_bindot()` using `bins = 30`. Pick better value with `binwidth`.
```



Izračun mer srednjih vrednosti ter razpršenosti.

Vsaka nosi nekaj informacij, zato sem jih izračunal več. V primeru da bi bila porazdelitev idealno normalna, bi bilo nekaj vrednosti enakih. Ker je videti da je večina primerov negativnih, se da iz tega razbrati da večina moških študentov iz vzorca preživi več časa za internetom kot za televizijo.

```
# srednje vrednosti
mean(data_40_m$diff_TV_Int)
```

```
## [1] -13
```

```
median(data_40_m$diff_TV_Int)
```

```
## [1] -6
```

```
get_modus(data_40_m$diff_TV_Int)
```

```
## [1] -5
```

```
# mere razprsenosti
IQR(data_40_m$diff_TV_Int) # interkvartilni razmik
```

```
## [1] 6.5
```

```
sd(data_40_m$diff_TV_Int) # standardni odklon
```

```
## [1] 22.65581
```

```
var(data_40_m$diff_TV_Int) # varianca
```

```
## [1] 513.2857
```

```r
mad(data_40_m$diff_TV_Int) # povprecni absolutni odklon
```

```
## [1] 5.9304
```

- Za skupino študentov Veterinarske fakultete in študentov dentalne medicine, ki imajo najbolj pogosto barvo oči na vašem začetnem vzorcu, izračunajte ustrezno mero srednje vrednosti in razpršenosti za število ur, ki ga študenti namenijo uporabi interneta.

```r
# da preverimo katera je najbolj pogosta barva oci
table(data_40$BarvaOci) # rjava
```

```
##
##   crna  drugo  plava  rjava zelena
##      1      3     12     15      9
```

```r
# izbor studentov dentalne medicine in veterine ki imajo rjave oci
data_vet_dent <- data_40 %>%
  filter(BarvaOci == "rjava" &
           Studij %in% c("Dentalna medicina", "Veterina"))

# srednje vrednosti
mean(data_vet_dent$Internet)
```

```
## [1] 9.25
```

```r
median(data_vet_dent$Internet)
```

```
## [1] 9.5
```

```r
get_modus(data_vet_dent$Internet)
```

```
## [1] 10
```

```r
# mere razprsenosti
IQR(data_vet_dent$Internet) # interkvartilni razmik
```

```
## [1] 4.5
```

```r
sd(data_vet_dent$Internet) # standardni odklon
```

```
## [1] 3.918819
```

```r
var(data_vet_dent$Internet) # varianca
```

```
## [1] 15.35714
```

```r
mad(data_vet_dent$Internet) # povprecni absolutni odklon
```

```
## [1] 5.1891
```

- Kakšen delež študentov ima na vašem vzorcu manj kot 71 kg? Kakšen delež ima težo med 80 in 90 kg?

```r
# delez tistih z manj kot 71 kg
nrow(data_40 %>% filter(Teza < 71))/nrow(data_40)
```

```
## [1] 0.725
```

```r
# delez tistih med 80 in 90 kg
nrow(data_40 %>% filter(Teza > 80, Teza < 90))/nrow(data_40)
```

```
## [1] 0.075
```

Naloga #3

- S svojimi besedami razložite neodvisnost dogodkov. Na primeru dveh konkretnih spremenljivk razložite neodvisnost spremenljivk na laičen, zanimiv način.

neodvisnost dogodkov pomeni, da v primeru dveh dogodkov, ali se en dogodek zgodi oz. ne zgodi, ne vpliva na to ali se bo oz. ne bo zgodil tudi drugi dogodek.

Kot primer lahko vzamemo tudi podatke študentov. Kakšno barvo oči ima nekdo ne vpliva na to ali bo oz. ne bo igral igrice, ali bo ali ne bo imel domačih živali, itd.

Po drugi strani pa lahko vidimo tudi odvisnost spremenljivk na več načinov. To ali nekdo kadi definitivno vpliva na to koliko pokadi, saj če ne kadi (spremenljivka kajenje enaka "N") bo Kajenje_koliko spremenljivka zagotovo 0. Drug pogled je lahko tudi takšen, če nekdo več časa preživi na internetu bo preprosto imel manj časa možnost preživeti ob knjigah ali športu. Malce idealiziran primer, vendarle, če ima oseba maksimalno 100h prostega časa na teden, in zapravi 50h prostega časa da preživi na internetu, 20h ob športu, 10h za televizijo, mu ostane le še največ 20h časa ob knjigah.

Če se še umaknemo iz danih podatkov se neodvisnost dogodkov vidi v še dokaj enostavnem primeru za razumeti. Met kocke in dobljeno št. pik na kocki ne vpliva na to kakšno karto bomo za tem izvlekli iz kupa 52-ih kart. V primeru da bi met kocke vplival na izvlečeno karto, bi lahko vedeli kakšno karto bomo izvlekli še preden to storimo. To bi povzročilo precej velik nemir v kazino industriji.

Naloga #4

- S svojimi besedami razložite, kaj pomeni reprezentativen vzorec iz populacije. Navedite konkretna primera vzorcev, kjer je eden reprezentativen, drugi pa ne. Primera naj bosta realna, zanimiva.

Reprezentativen vzorec pomeni, da ko vzamemo nek vzorec (del celotne populacije) iz celotne populacije, mora porazdelitev tega vzorca biti čim bolj podobna porazdelitvi celotne populacije (po vseh spremenljivkah). Pred kratkim je bila zelo aktualna COVID-19 raziskava, kjer so izbrali 3000 kandidatov (Okrog 1300 jih je pristalo na raziskavo), ki bi lahko sodelovali v raziskavi glede prekuženosti celotne populacije. Teh 3000 kandidatov so morali izbrati po več različnih lastnostih, tako da so sledili lastnostim iz celotne populacije. Kot primer lahko gledamo kakšno starostno porazdelitev mora imeti vzorec. Pač delež izbranih kandidatov v vzorcu v določeni starostni skupini (ne vem kakšne so bile skupine, ampak kot primer 0-10, 11-20, 21-30, ...) mora biti čimbolj enak deležu starostne skupine v celotni populaciji. Podobno velja glede gostote poselitve v Sloveniji. Več kandidatov bo iz Ljubljane, saj je tudi v Sloveniji tam največ prebivalcev. Itd.

Nereprezentativen vzorec bi bil, če ta ne predstavlja celotne populacije. Primer je izmišljen in ne vem če drži, vendar se pretvarjajmo da drži. Celotna populacija v temu primeru so vsi avtomobili registrirani v Sloveniji. Če je naš vzorec avti lastnikov, ki živijo ob obali, ta ne predstavlja celotne populacije, saj lastniki avtomobilov, ki živijo ob obali, imajo ponavadi toplejše vreme kot drugod po Sloveniji. Posledično kupijo več avtov brez strehe. Torej je delež avtov brez strehe ob obali večji, kot delež v celotni populaciji. Zagotovo tudi ta vzorec ne predstavlja avtov lastnikov, ki živijo v bolj gorskem svetu, saj bodo ti lastniki imeli večji delež avtomobilov s štirikolesnim pogonom (saj to potrebujejo, zaradi več makadamskih in strmih cest ter večje verjetnosti da bo snežilo) in seveda manjši delež avtov brez strehe. Na podoben način kot zgoraj (COVID raziskava) bi bilo potrebno izbrati ustrezno količino avtov glede na regijo, kjer lastniki živijo, da bi se dobil reprezentativen vzorec.