Data Analytics I: Predictive Econometrics
Assignment IV

Lauritz Storch
Sebastian Maser
Ivan Medvedev
Jan Gutjahr

2022-12-20

Data Preparation and Package Installation

Download the data set drugs.RData[1] from Canvas. Load the data into R. Install and load the packages rpart and rpart.plot.

```
# # Clear R
# rm(list=ls())
# cat("\014")

## Set working directory to load data
setwd("~/Downloads/")

## Packages and Library
# Packages function
install_if_missing <- function(p) {
  if (p %in% rownames(installed.packages()) == F) {
    try(install.packages(p, dependencies = T))
  } else {
    cat(paste("Skipping already installed package:", p, "\n"))
  }
}

# Define packages
packages = c("rpart","rpart.plot","ggplot2")

# Install
invisible(sapply(packages, install_if_missing))
```

```
## Skipping already installed package: rpart
## Skipping already installed package: rpart.plot
## Skipping already installed package: ggplot2
```

```
# read library
for(pkg in packages){
  library(pkg, character.only = TRUE)
}
```

[1] Please make sure your device is connected to the internet and you are logged into Canvas to download the dataset.

```r
## Read in data
load("drugs.RData")
set.seed(21122022)
```

Task 1

Q: How large is the share of males who consume soft drugs (in percent)? (0.5 points)

```r
# Check for NAs in regarding columns
# sum(is.na(drugs$Gender))
# sum(is.na(drugs$Soft_Drug))

male_sdrugs <- sum(drugs$Gender == "male" & drugs$Soft_Drug == T)
sshare_male <- male_sdrugs/sum(drugs$Gender == "male")

cat("The share of males who consume soft drugs is: ",
    round(sshare_male*100,4), "%")
```

```
## The share of males who consume soft drugs is:  29.1771 %
```

Task 2

Q: How large is the difference between the share of male and female hard drug consumers (in percentage points)? (0.5 points)

```r
# Check for NAs in regarding columns
# sum(is.na(drugs$Hard_Drug))

male_hdrugs <- sum(drugs$Gender == "male" & drugs$Hard_Drug == T)
female_hdrugs <- sum(drugs$Gender == "female" & drugs$Hard_Drug == T)

hshare_male <- male_hdrugs/sum(drugs$Gender == "male")
hshare_female <- female_hdrugs/sum(drugs$Gender == "female")

diff_gender <- hshare_male - hshare_female

cat("The difference between the share of male and female hard
drug consumers is: ", round(diff_gender*100,4), " percentage points")
```

```
## The difference between the share of male and female hard
## drug consumers is:  2.7371  percentage points
```

Task 3

Q: Report the shares of young adults who consume soft drugs for each age group (16- 17 years, 18-19 years, and 20-24 years). Is soft drug consumption increasing or decreasing with age? (0.5 points)

```r
# Check for NAs in regarding columns
sum(is.na(drugs$Age))
```

```
## [1] 0
```

```r
age_span <- unique(drugs$Age)
for (i in seq(3)){
  assign(paste0("youth_share_sdrugs",i) ,
         sum(drugs$Age == age_span[i] & drugs$Soft_Drug == T)/
```

```
            sum(drugs$Age == age_span[i]))
}

cat("The share for 16-17 years is: ", round(youth_share_sdrugs1*100,4))
```

```
## The share for 16-17 years is:  48.4982
```

```
cat("The share for 18-19 years is: ", round(youth_share_sdrugs2*100,4))
```

```
## The share for 18-19 years is:  0
```

```
cat("The share for 20-24 years is: ", round(youth_share_sdrugs3*100,4))
```

```
## The share for 20-24 years is:  0
```

```
# Check whether the shares can be true for 18-19 and 20-24 years old
# sum(isTRUE(drugs$Soft_Drug[which(drugs$Age == "20-24 years")]))
# sum(isTRUE(drugs$Soft_Drug[which(drugs$Age == "18-19 years")]))
```

Answer: Soft drug consumption is decreasing with age. As the data has significantly less data for 18-19 years and 20-24 years old, it remains questionable whether the results are actually representative.

Task 4

Q: Tabulate the observations by earnings category and soft drug consumption. Perform a chi-squared test to evaluate whether soft drug consumption is independent of the earnings. Can you reject the independence hypothesis at a 5% significance level? (0.5 points)

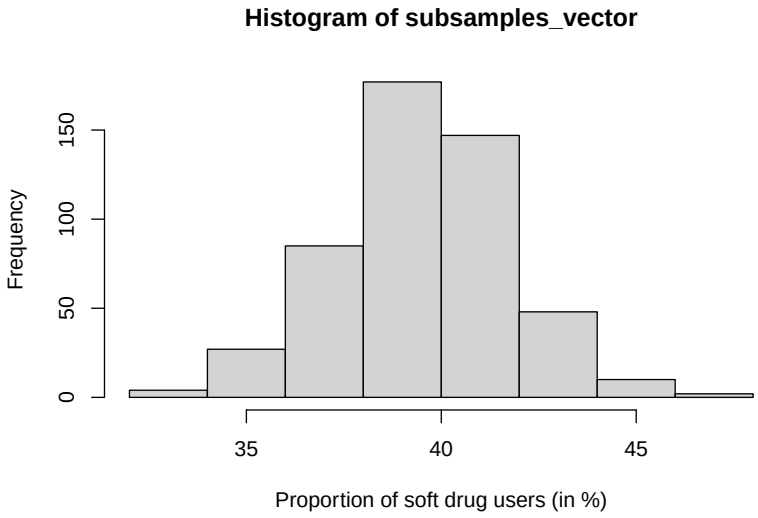
```
table_earn_sdrugs <- as.table(cbind(drugs$Earning,drugs$Soft_Drug))

chi_squared <- chisq.test(table_earn_sdrugs[,1],table_earn_sdrugs[,2])

cat("We reject the independence hypothesis at a 5% level. The p-value is ",
round(chi_squared$p.value,4)*100, "%")
```

```
## We reject the independence hypothesis at a 5% level. The p-value is  2.44 %
```

Task 5

Q: Draw 500 times a random subsample of 500 observations in your dataset and record the average soft drug consumption in each subsample. Draw a histogram of your results. Are the recorded subsample means close to the average drug consumption in the full sample? (1 point)

```
subsamples_vector <- rep(NA, 500)

for (i in 1:500){
  random <- sample(1:dim(drugs)[1], 500, replace = T)
  subsets <- drugs[random,]
  subsamples_vector[i] <- (sum(subsets$Soft_Drug == T)/500)*100
}
# Mean of all draws
subsample_mean <- mean(subsamples_vector)

# histogram plot
hist(subsamples_vector, xlab = "Proportion of soft drug users (in %)")
```




```
# Population mean
total_share_sdc <- sum(drugs$Soft_Drug == "TRUE")/nrow(drugs)

cat("The total share of soft drug consumers is",
round(total_share_sdc*100,4),"% in the entire population.
The mean of subsample average is ", round(subsample_mean,4), "% and, hence,
very close to the population mean. Although the average proportion of soft
drug users in many samples are concentrated around the true value of"
,round(total_share_sdc*100,4),"%
the subsample averages can also diverge strongly from the population value
(the tails in the histogram). According to the Central Limit theorem (CLT),
this is in line and we can expect an increasing concentration of subsample
means around the true value with increasing number of draws.")
```

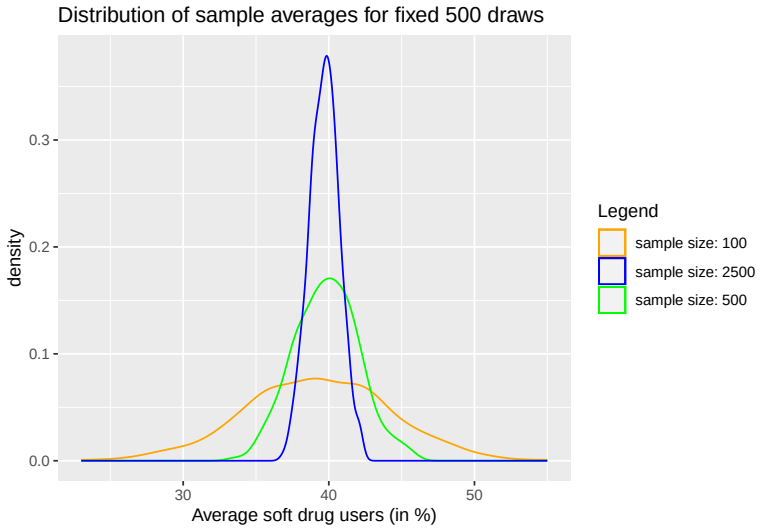
```
## The total share of soft drug consumers is 39.716 % in the entire population.
## The mean of subsample average is  39.6436 % and, hence,
## very close to the population mean. Although the average proportion of soft
## drug users in many samples are concentrated around the true value of 39.716 %
## the subsample averages can also diverge strongly from the population value
## (the tails in the histogram). According to the Central Limit theorem (CLT),
## this is in line and we can expect an increasing concentration of subsample
## means around the true value with increasing number of draws.
```

Task 6

Q: Write a function that allows to specify multiple subsample size and multiple number of draws. Your function should perform the same procedure as in the previous exercise, but for all combinations of specified subsample sizes and number of draws, and return the average drug consumption information of every single draw. Run your function for N runs=c(100, 500, 2500) and sample sizes=c(100, 500, 2500). Use the geom_density() function to draw two kernel density estimates of your results. In the first, fix the number or draws at 500 and visualize the density estimate for the three different subsample sizes. In the second, fix the subsample size at 500 and visualize the density estimate for the three different number of draws. What asymptotic behavior do you observe in each plot? (1 point)

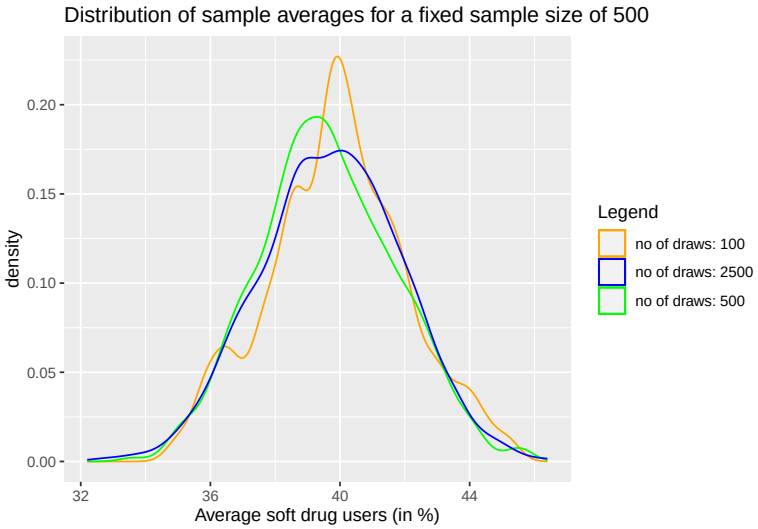
```
kernelbuilder <- function(sample_size, no_of_draws){
  random <- matrix(NA, ncol=sample_size, nrow=no_of_draws)
  subsets <- matrix(NA, ncol=sample_size, nrow=no_of_draws)
```

```
  for (i in 1:no_of_draws) {
    random[i,] <- sample(1:dim(drugs)[1], sample_size, replace = T)
    subsets[i,] <- as.logical(drugs[random[i,],8])
  }

  value <- round(100* rowSums(subsets)/sample_size, 2)

}

#fixed amount of draws plot
draws_fixed <- data.frame(as.data.frame(kernelbuilder(100, 500)),
                          as.data.frame(kernelbuilder(500, 500)),
                          as.data.frame(kernelbuilder(2500, 500)))
colnames(draws_fixed) = c("x","y","z")


colors <- c("sample size: 100" = "orange", "sample size: 500" = "green",
            "sample size: 2500" = "blue")

ggplot(data = draws_fixed) +
    geom_density(aes(x = x, color = "sample size: 100")) +
    geom_density(aes(x = y, color = "sample size: 500")) +
    geom_density(aes(x = z, color = "sample size: 2500")) +
    labs(x = "Average soft drug users (in %)",
         color = "Legend") +
    scale_color_manual(values = colors) +
ggtitle("Distribution of sample averages for fixed 500 draws")
```




```
#fixed sample size plot
samplesize_fixed <- data.frame(as.data.frame(kernelbuilder(500, 100)),
                          as.data.frame(kernelbuilder(500, 500)),
                          as.data.frame(kernelbuilder(500, 2500)))
colnames(samplesize_fixed) = c("x","y","z")


colors <- c("no of draws: 100" = "orange", "no of draws: 500" = "green",
            "no of draws: 2500" = "blue")
```

```
ggplot(data = samplesize_fixed) +
    geom_density(aes(x = x, color = "no of draws: 100")) +
    geom_density(aes(x = y, color = "no of draws: 500")) +
    geom_density(aes(x = z, color = "no of draws: 2500")) +
    labs(x = "Average soft drug users (in %)",
         color = "Legend") +
    scale_color_manual(values = colors) +
ggtitle("Distribution of sample averages for a fixed sample size of 500")
```



Answer: Increasing the sample size while keeping the number of draws at 500, the bigger the sample gets, the closer its' average will be to the true population parameter due to the Central Limit theorem (CLT). On the other hand, if we increase the amount of draws, but keep the sample size stable, the distribution of the subset averages does not converge to the true parameter. Thus, drawing more samples while keeping their size the same does not yield a lot of increase in accuracy.