Chapter 2 - Descriptive statistics

Dr. Alessandro Ruggieri

Contents

Intro

We study how labor market participation for women and men vary across countries.

We first load the data:

```
# import new data:hours worked rate across countries
data_lfp <- read.csv("EAP_2WAP_SEX_AGE_GEO_RT_A-filtered-2021-01-13.csv",  header = TRUE)
# rename columns
names(data_lfp) <- c("country", "label", "source",
                     "gender","age","area", "year",
                     "participation", "status")
```

We subset data for individuals of "Age=25+" and "Area=National":

```
# subset matrix according to age groups: occupations=Aggregate
data_lfp<-data_lfp[data_lfp$age == "Age (Youth, adults): 25+", ]
# subset matrix according to area groups: area=Aggregate
data_lfp<-data_lfp[data_lfp$area == "Area type: National", ]
# subset matrix according to time: year=201
data_lfp<-data_lfp[data_lfp$year == "2019", ]
```

We remove useless columns:

```
# remove label
data_lfp$label <- NULL
# remove source
data_lfp$source <- NULL
# remove info on status
data_lfp$status <- NULL
```

We select and extract data for men and women:

```
# subset matrix for men
data_lfp_men<-data_lfp[data_lfp$gender == "Sex: Male",]
# subset matrix for women
data_lfp_women<-data_lfp[data_lfp$gender == "Sex: Female",]
```

Finally, we merge back the cleaned data:

```r
# merge two data frames by country name and year
data_lfp_final <- merge(data_lfp_men,data_lfp_women,by=c("country","year"))
```

First, we can inspect the data using the functions head() and tails(), which will display the first and the last part of the data, respectively

```r
#
head(data_lfp_women, 4)
```

```
##                    country      gender                       age
## 268            Afghanistan Sex: Female Age (Youth, adults): 25+
## 538                 Angola Sex: Female Age (Youth, adults): 25+
## 808                Albania Sex: Female Age (Youth, adults): 25+
## 1078 United Arab Emirates Sex: Female Age (Youth, adults): 25+
##                       area year participation
## 268  Area type: National 2019          22.5
## 538  Area type: National 2019          86.2
## 808  Area type: National 2019          57.3
## 1078 Area type: National 2019          56.1
```

```r
#
tail(data_lfp_women, 4)
```

```
##             country      gender                       age                area
## 74788         Yemen Sex: Female Age (Youth, adults): 25+ Area type: National
## 75058 South Africa Sex: Female Age (Youth, adults): 25+ Area type: National
## 75328        Zambia Sex: Female Age (Youth, adults): 25+ Area type: National
## 75598      Zimbabwe Sex: Female Age (Youth, adults): 25+ Area type: National
##       year participation
## 74788 2019           6.8
## 75058 2019          57.4
## 75328 2019          80.7
## 75598 2019          82.9
```

Measures of centrality

To look at centrality, we can construct the following measures:

- the sample mean:

```r
# Compute the mean
avg_lfp_men<-mean(data_lfp_men$participation)
avg_lfp_women<-mean(data_lfp_women$participation)
# Print result
sprintf("Average participation of women: %f", avg_lfp_women)
```

```
## [1] "Average participation of women: 55.193571"
```

- the sample median:

```r
# Compute the median
med_lfp_men<-median(data_lfp_men$participation)
med_lfp_women<-median(data_lfp_women$participation)
# Print result
sprintf("Median participation of women: %f", med_lfp_women)
```

```
## [1] "Median participation of women: 56.300000"
```

- the sample mode:

```
# Compute the mode
#install.packages("modeest")
require(modeest)
```

```
## Loading required package: modeest
```

```
mode_lfp_men<-mfv(data_lfp_men$participation)
mode_lfp_women<-mfv(data_lfp_women$participation)
# Print result
sprintf("Women participation mode: %f", mode_lfp_women)
```

```
## [1] "Women participation mode: 58.000000"
```

Measures of dispersion

To look at dispersions, we can construct the following measures:

- the sample standard deviation:

```
# Compute the st.dev.
sd_lfp_men<-sd(data_lfp_men$participation)
sd_lfp_women<-sd(data_lfp_women$participation)
# Print result
sprintf("Dispersion in participation of women: %f", sd_lfp_women)
```

```
## [1] "Dispersion in participation of women: 15.922720"
```

- the coefficient of variation:

```
# Compute the coefficient of variation
cv_lfp_men<-sd_lfp_men/avg_lfp_men
cv_lfp_women<-sd_lfp_women/avg_lfp_women
# Print result
sprintf("Coefficient of variation for participation of women: %f", cv_lfp_women)
```

```
## [1] "Coefficient of variation for participation of women: 0.288489"
```

- the range:

```
# Compute the range
ramge_lfp_men<-range(data_lfp_men$participation)
range_lfp_women<-range(data_lfp_women$participation)
# Print result
sprintf("Lower bound participation of women: %f", range_lfp_women[1])
```

```
## [1] "Lower bound participation of women: 6.800000"
```

```
sprintf("Upper bound participation of women: %f", range_lfp_women[2])
```

```
## [1] "Upper bound participation of women: 93.300000"
```

- the interquartile range:

```
# Compute the iqr range
iqr_lfp_men<-IQR(data_lfp_men$participation)
iqr_lfp_women<-IQR(data_lfp_women$participation)
# Print result
sprintf("Interquartile range in participation of women: %f", iqr_lfp_women)
```

```
## [1] "Interquartile range in participation of women: 14.125000"
```

Notice that we can compute an overall summary of the variable using the function summary( )

```
# summarize participation for men
summary(data_lfp_men$participation)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   49.70   73.38   80.30   79.71   87.12   98.20
```

and for women

```
# summarize participation for women
summary(data_lfp_women$participation)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    6.80   48.48   56.30   55.19   62.60   93.30
```

It is also possible to use the function stat.desc( ) to compute descriptive statisticse:

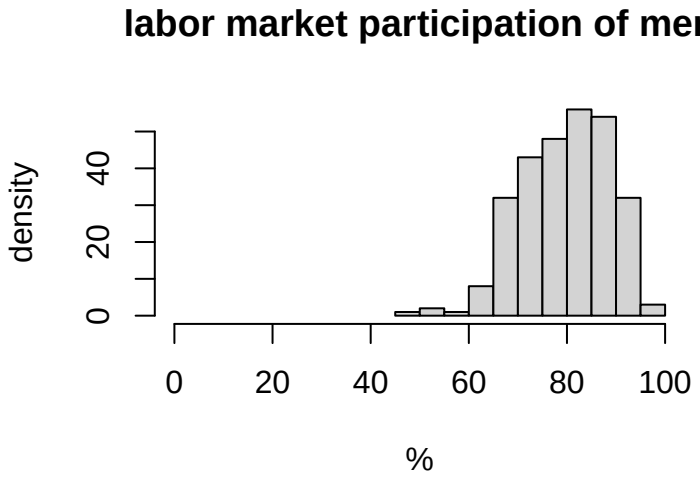
```
# summarize participation
#install.packages("pastecs")
library(pastecs)
res <- stat.desc(data_lfp_men$participation)
round(res, 2)
```

```
##      nbr.val     nbr.null       nbr.na          min          max        range
##       280.00         0.00         0.00        49.70        98.20        48.50
##          sum       median         mean      SE.mean CI.mean.0.95          var
##     22318.50        80.30        79.71         0.53         1.05        79.03
##      std.dev     coef.var
##         8.89         0.11
```

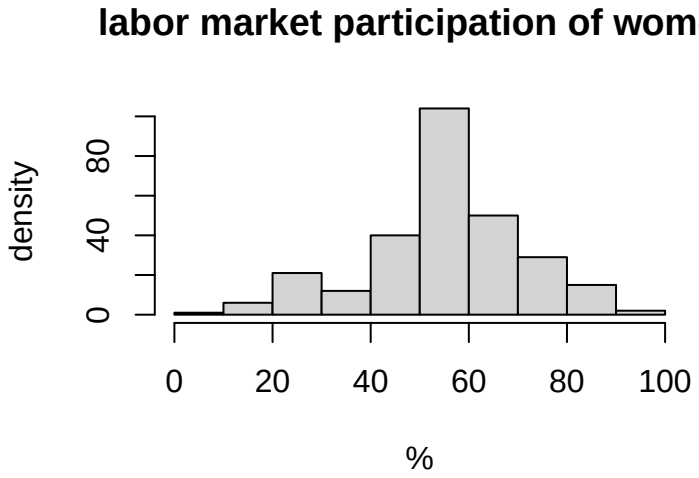
Frequency

To look at the entire distribution of participation rates for men and women, we can also plot their histograms:

```
# Create histogram of participation rate across countries
par(mfrow=c(1,1))
hist(data_lfp_men$participation, xlab="%", ylab="density",
     main="labor market participation of men",xlim=c(0,100))
```



```r
# Create histogram of participation rate across countries
hist(data_lfp_women$participation, xlab="%", ylab="density",
     main="labor market participation of women",xlim=c(0,100))
```



Measure of correlations

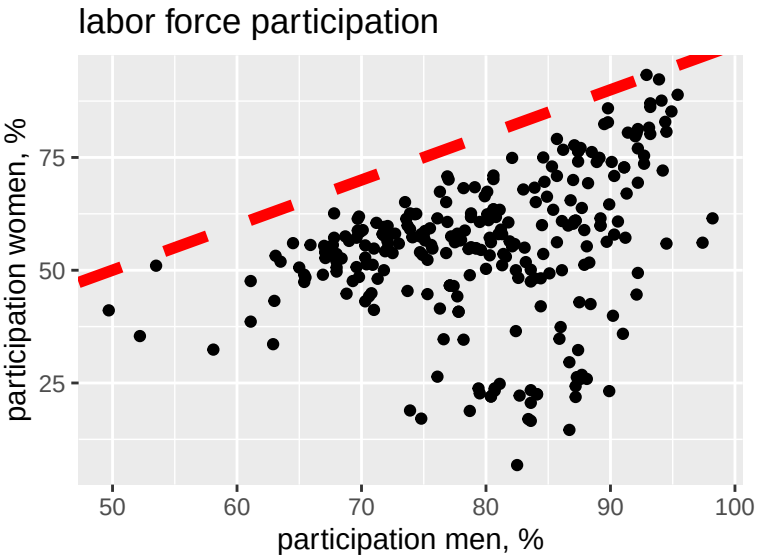
To look at how participation of men and women correlate across countries, we can construct their correlation coefficient. The correlation coefficient measures the association between two variables. Its value is bounded between -1 (perfect negative correlation: when x increases, y decreases) and +1 (perfect positive correlation: when x increases, y increases). A value closer to 0 suggests a weak relationship between the variables.

```r
# Correlation coefficient
cor(data_lfp_men$participation,data_lfp_women$participation)
```

```
## [1] 0.301275
```

Moreover, we can relate these two variables with a scatterplot:

```r
## load the package ggplot2
library(ggplot2)
# Create scatter of men and women participation
ggplot(data_lfp_final, aes(x = participation.x, y = participation.y)) +
  geom_point() +
  ggtitle("labor force participation") +
  geom_abline(intercept=0, slope = 1, linetype="dashed", color = "red", size=2) +
  xlab("participation men, %") + # for the x axis label
  ylab("participation women, %") # for the y axis label
```


labor force participation
participation women, %
75
50
25
50
60
70
80
90
100
participation men, %