Teaching Statistics with Real World Data from IPUMS

Answer key to basic statistical concept exercises in R, using real-world census microdata from the IPUMS International database.

Exercise: Exploring Data Answer Key

Topics Covered:

- Frequency tables
- Visual Summaries (pie chart and bar graph)
- Numerical Summaries (mean, median, IQR, standard deviation, boxplot)

Required dataset: IPUMS-International

Required variables:

1. COUNTRY
2. YEAR
3. AGE (age)
4. SEX
5. MARST (marital status)
6. EDATTAIN (educational attainment)
7. CHBORN (children ever born)

[The only preselected variables that are needed in this exercise are COUNTRY and YEAR. Make sure to remove all of the other preselected variables by unchecking the blue boxes next to them. This will reduce the size of your data file and also make it easier to view the data in R.]

Recommended samples:

1. Kenya [2009]
2. Philippines [2010]
3. Romania [2011]
4. Tanzania [2012]

Sample selection instructions:

- Limit sample to 10,000 households per country.
- Select "Customize Sample Sizes" in the "Extract Request" page and type 10 in the box under households for each of each of the counties. *[Note that the sample size is in 1000s]*.

• Section I

```r
library(ipumsr)
library(dplyr)
library(ggplot2)

# Load the data
ddi_exp <- read_ipums_ddi("ipumsi_00021.xml")
data_exp <- read_ipums_micro(ddi_exp)

# Convert the following variables to factors
data_exp<-within(data_exp, COUNTRY<- as.factor(COUNTRY))
data_exp<-within(data_exp, YEAR<- as.factor(YEAR))
data_exp<-within(data_exp, MARST<- as.factor(MARST))
data_exp<-within(data_exp, SEX<- as.factor(SEX))
data_exp<-within(data_exp, EDATTAIN<- as.factor(EDATTAIN))
```

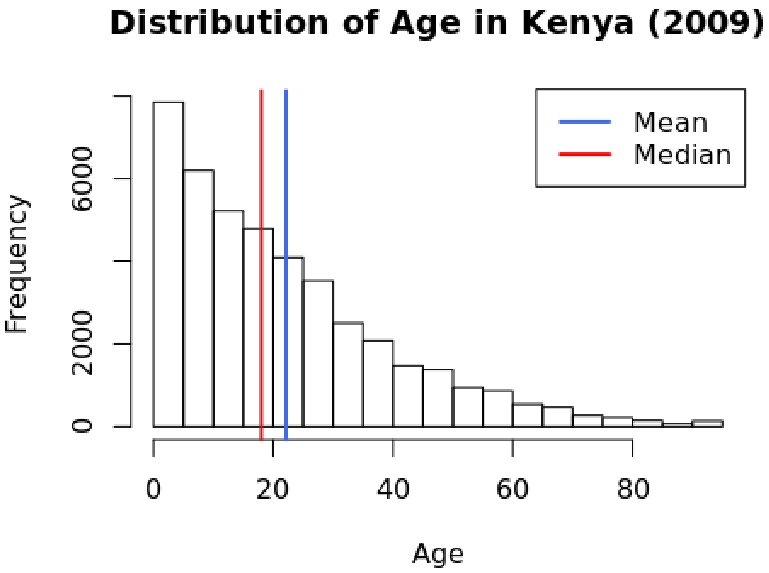
1. Complete the following frequency table:

	Sex	
Country	Male (1)	Female (2)
Kenya (404)	21,199	21,614
Philippines (608)	23,044	22,581
Romania (642)	12,960	13,736
Tanzania (834)	22,700	24,129

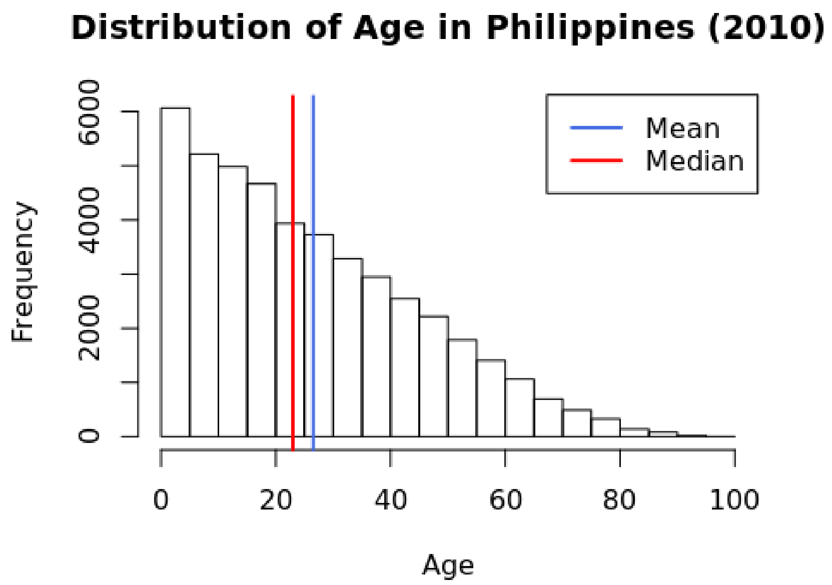
```r
# Cross tabulate
tab1<- xtabs(~COUNTRY+SEX, data = data)
# Create contingency table
ftable(tab1)

##         SEX       1       2
## COUNTRY
## 404         1904617 1937318
## 608         4744268 4666988
## 642          966637 1025287
## 834         2171639 2326383
```

2. Show the age distribution in each of the countries by creating histograms. Also, describe the type of distribution (left skewed/right skewed/ uniform/ unimodal/ bimodal, etc.) and draw lines on histogram to show the mean and median age.

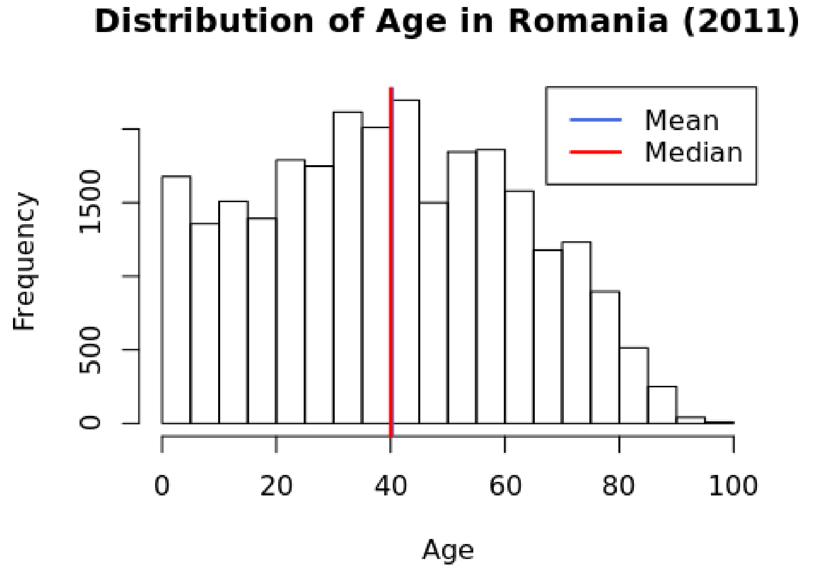
```
# Age distribution by country

# Choose subset of the entire data such that we have data only for Kenya
Kenya <- subset(data_exp, COUNTRY == 404)
# Draw histogram
hist(Kenya$AGE,main = "Distribution of Age in Kenya (2009)",xlab="Age")
# Add line to show the mean age
abline(v = mean(Kenya$AGE),col = "royalblue",lwd = 2)
abline(v = median(Kenya$AGE),col = "red",lwd = 2)
legend(x = "topright",c("Mean", "Median"), col = c("royalblue", "red"),lwd =
c(2,2))
```



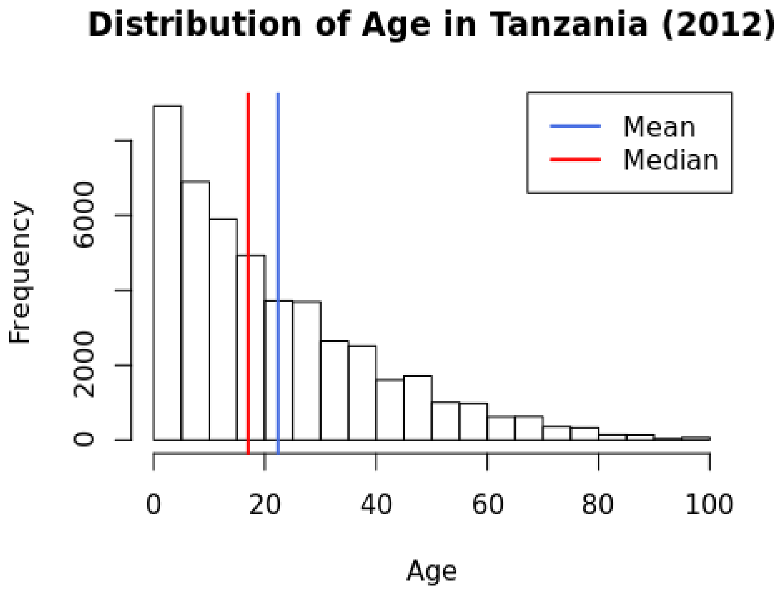
```
Philippines<- subset(data_exp, COUNTRY == 608)
hist(Philippines$AGE,main = "Distribution of Age in Philippines (2010)",xlab=
"Age")
abline(v = mean(Philippines$AGE),col = "royalblue",lwd = 2)
abline(v = median(Philippines$AGE),col = "red",lwd = 2)
legend(x = "topright",c("Mean", "Median"), col = c("royalblue", "red"),lwd =
c(2,2))
```



```
Romania<- subset(data_exp, COUNTRY == 642)
hist(Romania$AGE,main = "Distribution of Age in Romania (2011)",xlab="Age")
abline(v = mean(Romania$AGE),col = "royalblue",lwd = 2)
abline(v = median(Romania$AGE),col = "red",lwd = 2)
legend(x = "topright",c("Mean", "Median"), col = c("royalblue", "red"),lwd =
c(2,2))
```



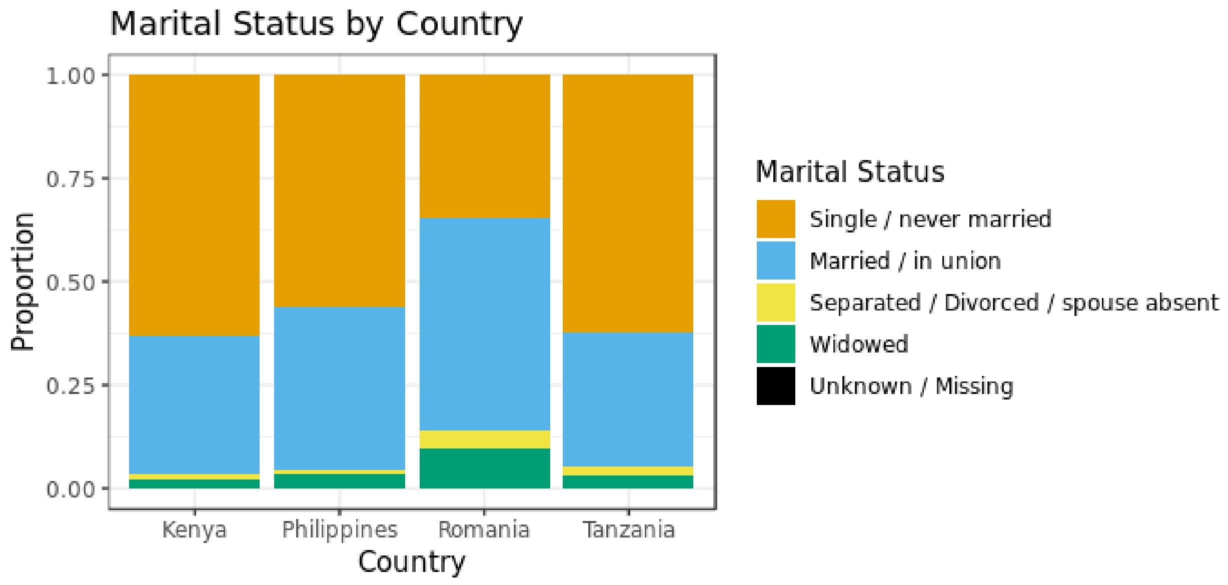
```
Tanzania<- subset(data_exp, COUNTRY == 834)
hist(Tanzania$AGE,main = "Distribution of Age in Tanzania (2012)",xlab="Age")
abline(v = mean(Tanzania$AGE),col = "royalblue",lwd = 2)
abline(v = median(Tanzania$AGE),col = "red",lwd = 2)
legend(x = "topright",c("Mean", "Median"), col = c("royalblue", "red"),lwd =
c(2,2))
```



We observe that the distribution of age for Kenya, Philippines and Tanzania is right-skewed. However, for Romania the distribution of age is approximately symmetric.

3. Create a stacked bar chart of marital status by country.

```
# Marital Status by Country

# Customized color scheme
new_color1<- c("#E69F00", "#56B4E9", "#F0E442",  "#009E73","#000000")
# Create a country label to write the country names instead of country code
country_label <- c("Kenya", "Philippines", "Romania", "Tanzania")

# Draw stacked bar chart of marital status by country using ggplot
ggplot(data_exp, aes(x = COUNTRY, fill = MARST)) + geom_bar(position = "fill"
) +
  # White plot background with grid lines
  theme_bw() +
  # Change the colors of the plot and write the complete marital status in le
gend instead of the codes used torepresent different marital status
  scale_fill_manual(values=new_color1, labels = c("Single / never married", "
Married / in union", "Separated / Divorced / spouse absent", "Widowed", "Unkn
own / Missing"), name = "Marital Status") +
  # Add title of the plot
  ggtitle("Marital Status by Country") +
  # Label x axis as country
  xlab("Country") +
  # Label y axis as proportion
  ylab("Proportion") +
  # In x axis write the country names instead of country code
  scale_x_discrete(labels = country_label)
```



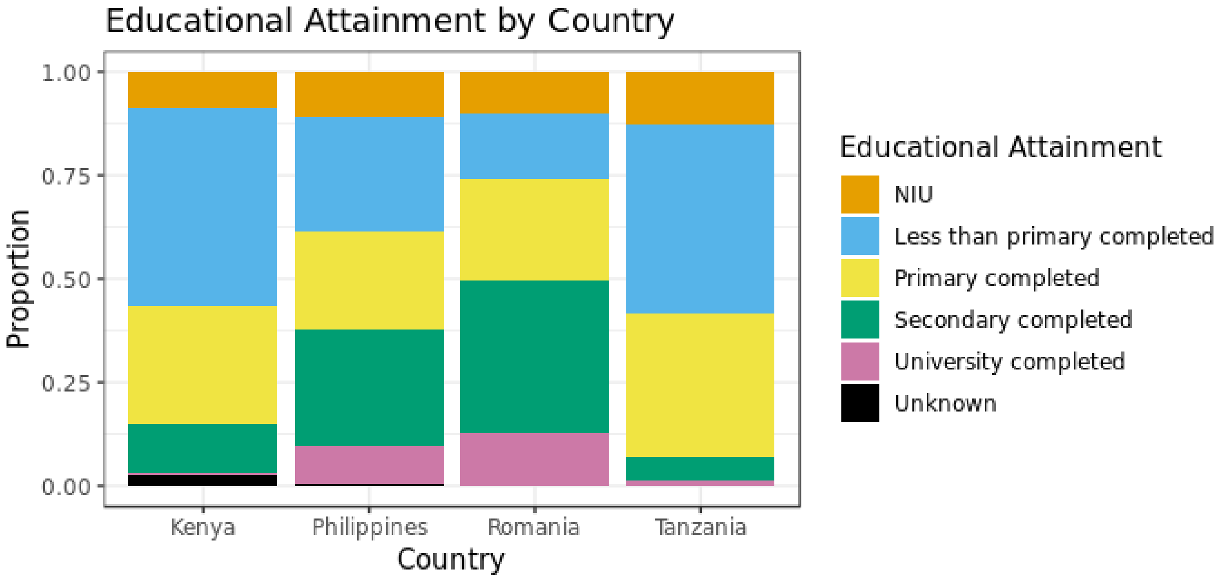
4. Complete the following frequency table:

	Marital Status				
Country	Single/never married (1)	Married/in union (2)	Separated/ Divorced/ spouse absent (3)	Widowed (4)	Unknown/ Missing (9)
Kenya (404)	26,994	14,215	606	998	0
Philippines (608)	25,643	17,941	463	1,550	28
Romania (642)	9,237	13,708	1,171	2,563	17
Tanzania (834)	29,152	15,301	907	1,469	0

```
# Cross tabulate
tab2<- xtabs(~COUNTRY+MARST, data = data_exp)
# Create contingency table
ftable(tab2)

##          MARST     1     2     3     4     9
## COUNTRY
## 404            26994 14215   606   998     0
## 608            25643 17941   463  1550    28
## 642             9237 13708  1171  2563    17
## 834            29152 15301   907  1469     0
```

5. Create a stacked bar chart of educational attainment by country.

```
# Educational Attainment

# Customized color scheme
new_color2<- c("#E69F00", "#56B4E9", "#F0E442","#009E73", "#CC79A7","#000000"
)

# Draw stacked bar chart of marital status by country using ggplot
ggplot(data_exp, aes(x = COUNTRY, fill = EDATTAIN)) + geom_bar(position = "fi
ll") +
  # White plot background with grid lines
  theme_bw() +
  # Change the colors of the plot and write the various categories of educati
onal attainment    in the legend instead of the codes used to them
  scale_fill_manual(values=new_color2, labels = c("NIU", "Less than primary c
ompleted", "Primary completed","Secondary completed", "University completed",
"Unknown"), name = "Educational Attainment") +
  # Add title of the plot
  ggtitle("Educational Attainment by Country") +
  # Label x axis as country
  xlab("Country") +
  # Label y axis as proportion
  ylab("Proportion") +
  # In x axis write the country names instead of country code (country_label
is defined in the   answer to question 3)
  scale_x_discrete(labels= country_label)
```



6. Complete the following frequency table (write the percentage of individuals who have attained an education level in a country in bracket):

		Educational attainment					
Country	NIU (0)	Less than primary completed (1)	Primary completed (2)	Secondary completed (3)	University completed (4)	Unknown (9)	Total
Kenya (404)	3,833 (8.95 %)	20,411 (47.67 %)	12,122 (28.31 %)	5,030 (11.75 %)	369 (0.86 %)	1,048 (2.45 %)	42813 (100 %)
Philippines (608)	5,049 (11.07 %)	12,558 (27.52 %)	10,844 (23.77 %)	12,759 (27.96 %)	4,242 (9.3 %)	173 (0.38 %)	45625 (100 %)
Romania (642)	2,733 (10.24 %)	4,172 (15.63 %)	6,546 (24.52 %)	9,891 (37.05 %)	3,354 (12.56 %)	0	26696 (100 %)
Tanzania (834)	5,886 (12.57 %)	21,438 (45.78 %)	16,231 (34.66 %)	2,645 (5.65 %)	629 (1.34 %)	0	46829 (100 %)

```
# Cross tabulate
tab3<- xtabs(~COUNTRY+EDATTAIN, data = data_exp)
# Create contingency tables
ftable(tab3)

##          EDATTAIN     0     1     2     3     4     9
## COUNTRY
## 404                3833 20411 12122  5030   369  1048
## 608                5049 12558 10844 12759  4242   173
## 642                2733  4172  6546  9891  3354     0
## 834                5886 21438 16231  2645   629     0

# Percentage of individuals who have attained an education level in a country
data_exp %>% count(COUNTRY,EDATTAIN) %>% group_by(COUNTRY)%>% mutate(Percent
= round(prop.table(n),4)*100)

## # A tibble: 22 x 4
## # Groups:   COUNTRY [4]
##    COUNTRY EDATTAIN     n Percent
##    <fct>   <fct>    <int>   <dbl>
##  1 404     0         3833    8.95
##  2 404     1        20411   47.7
##  3 404     2        12122   28.3
##  4 404     3         5030   11.8
##  5 404     4          369    0.86
##  6 404     9         1048    2.45
##  7 608     0         5049   11.1
##  8 608     1        12558   27.5
##  9 608     2        10844   23.8
## 10 608     3        12759   28.0
## # ... with 12 more rows

# Total number of individuals in each of the country (last column in the tabl
e)
data_exp %>% count(COUNTRY,EDATTAIN) %>% group_by(COUNTRY)%>%summarise(sum(n)
)

## # A tibble: 4 x 2
##   COUNTRY `sum(n)`
##   <fct>      <int>
## 1 404        42813
## 2 608        45625
## 3 642        26696
## 4 834        46829
```

• Section II

1. What are the source variables for CHBORN?

The source variables for CHBORN of the four countries are as follows:

1. Kenya: KE2009A_CHBORNM which refers to the number of boys who were born alive and KE2009B_CHBORNF which refers to the number of girls who were born alive.
2. Philippines: PH2010A_CHBORN which refers to the number of children born alive
3. Romania: RO2011A_CHBORN which refers to the number of children ever born
4. Tanzania: TZ2012A_CHBORN which refers to the number of children ever born

2. What is meant by top code? What are the values of top codes of CHBORN for the different countries? How might it affect comparability between the countries?

Top code is the upper limit of the variable, i.e., all observations having value greater than this upper limit are grouped together. This may be done when we have sparse cases for high values of a variable. The top code for CHBORN in Kenya and Tanzania is 15 or more, in Philippines it is 17 or more and in Romania it is 16 or more. In general, in the entire sample the proportion of observations with the top code is quite less and hence it would not affect the comparability between the countries. However, if the total sample size is quite small then it may affect the comparability between the countries.

3. What are the major differences between how marital status was collected in the four countries? (Hint: Look at the questionnaires and questionnaire instruction)

The major difference between the ways in which the marital status was collected in the four countries is reflected by the various categories of the marital status considered in each of the country. In addition, we observe that there is inconsistency in the age limit of the persons who are coded as “Never married/single”. For example, in Philippines persons younger than 12 years are coded in this category whereas in Romania persons younger than 14 years are coded as such.

Section III

1. What does the value CHBORN = 99 signify? Should the observations with CHBORN value of 99 be included or excluded from calculation?

Persons who were not asked this census question because they are outside the population for which this question was intended (in this case all males and female children) receive a code of 99. They are not part of the population for which we are calculating summary statistics hence they should be excluded.

2. Based on your answer to (Section II: part 1 – include/exclude) complete the following frequency table:

Country	Observations
Kenya (404)	14,179
Philippines (608)	11,820
Romania (642)	12,065
Tanzania (834)	15,637

```
#Remove value which are NIU (not in universe) - CHBORN
data_exp$CHBORN[data_exp$CHBORN==99]<-NA
newdata_exp <- na.omit(data_exp)

# Number of observations by country
newdata_exp %>% group_by(COUNTRY)%>% summarise(Observations = n())
## # A tibble: 4 x 2
##   COUNTRY Observations
##   <fct>          <int>
## 1 404            14179
## 2 608            11820
## 3 642            12065
## 4 834            15637
```

3. Note that the observations in the previous question comprises of only females. Compute the percentage change in the number of females after your decision in part 1 (number of observations) for each country using your answer in Section II: part 2 and Section I: part 1.

We observe that after excluding the observations for which the value of CHBORN is 99, there is a decrease in the number of females in our current sample as compared to our original sample. In Kenya, the number of females decrease by about 34 percentage, in Philippines by about 48 percentage, in Romania by about 12 percentage and in Tanzania by about 35 percentage.

```
# Round to the nearest two decimal places
# Kenya - 404
round((14179 - 21614)/21614,2)

## [1] -0.34

# Philippines - 608
round((11820 - 22581)/22581,2)

## [1] -0.48

# Romania - 642
round((12065 - 13736)/13736,2)

## [1] -0.12

# Tanzania - 834
round((15637 - 24129)/24129,2)

## [1] -0.35
```

4. What is the population universe for CHBORN in each country? That is, in each census, who was asked this question?

The population universe for CHBORN in each of the country is as follows:

1. Kenya : Females aged 12 or more
2. Philippines: Females aged from 15 to 49
3. Romania: Females aged 15 or more
4. Tanzania: Females aged 12 or more

5. To make an accurate comparison of children ever born across these four countries, which cases should you drop from your dataset?

To make an accurate comparison of children ever born across these four countries we need to consider only females aged from 15 to 49 from all the countries and exclude all other observations.

6. Based on your answer to (Section III: part 5) complete the following frequency table:

Country	Observations
Kenya (404)	10,479
Philippines (608)	11,820
Romania (642)	6,219
Tanzania (834)	11,421

```
# For comparable universes - select individuals aged 15 - 49 (Philippines - 4
9 upper limit)
newdata_exp1 <- subset(newdata_exp, AGE >= 15 & AGE<= 49)

# Number of observations by country
newdata_exp1 %>% group_by(COUNTRY)%>% summarise(Observations = n())

## # A tibble: 4 x 2
##   COUNTRY Observations
##   <fct>          <int>
## 1 404            10479
## 2 608            11820
## 3 642             6219
## 4 834            11421
```

7. Now compute the percentage change in the number of females for each country using your answer in Section III: part 6 and Section I: part 1.

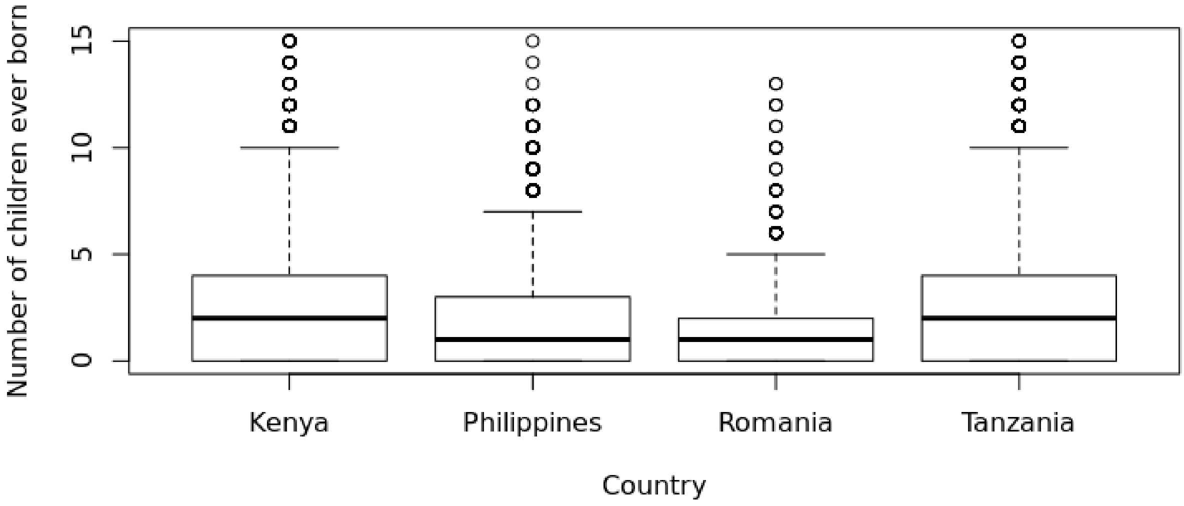
We observe that after excluding the observations for which the value of CHBORN is 99 and considering females aged from 15 to 49 years, there is a decrease in the number of females in our current sample as compared to our original sample. In Kenya, the number of females decreases by about 52 percentage, in Philippines by about 48 percentage, in Romania by about 55 percentage and in Tanzania by about 53 percentage. (Notice that for Philippines the answer here is same as that of in part 3 because the universe here is females aged from 15 to 49).

```
# Round to the nearest two decimal places
# Kenya - 404
round((10479 - 21614)/21614,2)

## [1] -0.52
```

```
# Philippines - 608
round((11820 - 22581)/22581,2)

## [1] -0.48

# Romania - 642
round((6219 - 13736)/13736,2)

## [1] -0.55

# Tanzania - 834
round((11421 - 24129)/24129,2)

## [1] -0.53
```

8. Create a boxplot representing the number of children born (CHBORN) to women by country.

```
#Create a country label to write the country names instead of country code
country_label <- c("Kenya", "Philippines", "Romania", "Tanzania")
#Remove the labels of CHBORN (as otherwise boxplot function in R cannot be us
ed)
newdata_exp1$CHBORN<- zap_labels(newdata_exp1$CHBORN)
#Create boxplot
boxplot(CHBORN ~ COUNTRY, data = newdata_exp1, ylab = "Number of children eve
r born", xlab = "Country", names=country_label)
```



9. For each of the countries in your dataset, find mean, median, Q1, Q3, interquartile range [IQR] and standard deviation (two decimal points) of the total number of children ever born [CHBORN].

Country	Mean*	Median*	Q1	Q3	IQR	SD
Kenya (404)	3	2	0	4	4	2.72
Philippines (608)	2	1	0	3	3	2.05
Romania (642)	1	1	0	2	2	1.25
Tanzania (834)	3	2	0	4	5	2.84

** Note that [CHBORN] cannot be in fraction, hence round it to the next integer.*

```
# Mean, median, Q1 and Q3
tapply(newdata_exp1$CHBORN,newdata_exp1$COUNTRY,summary)

## $`404`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    0.00    2.00    2.53    4.00   15.00
##
## $`608`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.000   1.000   1.674   3.000  15.000
##
## $`642`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    0.00    1.00    1.13    2.00   13.00
##
## $`834`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.000   2.000   2.863   4.000  15.000

# Standard deviation
tapply(newdata_exp1$CHBORN,newdata_exp1$COUNTRY,sd)

##      404      608      642      834
## 2.718765 2.046081 1.254504 2.840714
```