# Teaching Statistics with Real World Data from IPUMS

These exercises cover basic statistical concepts, guiding students through real world examples, using R and real-world census microdata from the IPUMS International database. See other IPUMS resources for instructions and a tutorial about accessing IPUMS data, and for a guide to the IPUMS R package for reading IPUMS data extracts into R.

## Exercise: Exploring Data

**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. MARST (marital status)
5. EDATTAIN (educational attainment)
6. 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 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

1. Complete the following frequency table:

| | Sex | |
|---|---|---|
| **Country** | Male (1) | Female (2) |
| Kenya (404) | | |
| Philippines (608) | | |
| Romania (642) | | |
| Tanzania (834) | | |

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.

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

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) | | | | | |
| Philippines (608) | | | | | |
| Romania (642) | | | | | |
| Tanzania (834) | | | | | |

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

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) | 348786 (9.08%) | | | | | | |
| Philippines (608) | | | | | | | |
| Romania (642) | | | | | | | |
| Tanzania (834) | | | | | | | |

---

## • Section II

1. What are the source variables for CHBORN?

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?

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

---

## • Section III

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

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

| Country | Observations |
|---|---|
| Kenya (404) | |
| Philippines (608) | |
| Romania (642) | |
| Tanzania (834) | |

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.

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

5. To make an accurate comparison of children ever born across these four countries are there any cases which should be dropped from the dataset? If yes, then mention the criteria.

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

| Country | Observations |
|---|---|
| Kenya (404) | |
| Philippines (608) | |
| Romania (642) | |
| Tanzania (834) | |

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.

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

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

| Country | Mean | Median | Q1 | Q3 | IQR | SD |
|---|---|---|---|---|---|---|
| Kenya (404) | | | | | | |
| Philippines (608) | | | | | | |
| Romania (642) | | | | | | |
| Tanzania (834) | | | | | | |