Teaching Statistics with Real World Data from IPUMS

These exercises cover basic statistical concepts, guiding students through real world examples, using R and 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: Confidence Interval

Topics covered:

- Unbiased estimators ($\bar{x}$ and $\hat{p}$)
- Interpretation of intervals
- Margin of error and standard error
- Confidence interval for population proportion
- Confidence interval for population mean
- t-distribution

Required dataset: IPUMS-International

Required variables:

1. COUNTRY
2. YEAR
3. BIRTHSLYR (number of births last year)
4. EDATTAIN (educational attainment)

[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. Cambodia [2008]
2. Portugal [2011]

❖ Section I

1. We want a biased point estimator because it does not tend to underestimate or overestimate the true parameter.

a) True
b) False

2. If a statistic is unbiased, then the difference between the sampling distribution and the value of the true parameter is 0.

a) True
b) False

3. Which type of statistic do we prefer to work with when conducting confidence intervals and later on with hypothesis tests?

a) Biased with a small standard error
b) Biased with a large standard error
c) Unbiased with a small standard error
d) Unbiased with a large standard error

4. Which of the following is the definition for the margin of error (MOE)?

a) The margin of error measures how accurate a point estimate is likely to be in estimating a parameter.
b) The margin of error measures how accurate a point estimate is likely to be in estimating a statistic.
c) The margin of error measures how accurate a confidence interval is likely to be in estimating a parameter.
d) The margin of error measures how accurate a confidence interval is likely to be in estimating a statistic.

5. Suppose you have a confidence interval with a point estimate of 2.5 and a MOE of 0.06. Now suppose the MOE increases to 0.15. What happens to the interval and the accuracy of our estimate?

a) The interval decreases and the accuracy of our estimate decreases.
b) The interval decreases and the accuracy of our estimate increases.
c) The interval increases and the accuracy of our estimate decreases.
d) The interval increases and the accuracy of our estimate increases.

6. Select all of the following that are true about the t-distribution.

a) The t-distribution has wider/fatter tails than the normal distribution. (It has a larger spread.)
b) The shape and spread of the t-distribution does not depend on the degrees of freedom.
c) The t-distribution is bell-shaped.
d) The t-distribution is symmetric about 1.
e) The t-distribution is symmetric about 0.

7. Why does the t-distribution get closer to the normal distribution as the degrees of freedom increases?

a) The mean estimate gets better as n decreases.
b) The mean estimate gets better as n increases.
c) The standard deviation estimate gets better as n decreases.
d) The standard deviation estimate gets better as n increases.

❖ Section II

Consider only dataset of Cambodia for Section II and Section III

1. Show the age distribution by creating histogram. Draw lines on histogram to show the mean and median age.

2. What is the population universe for the variable - BIRTHSLYR? That is, in the census, who was asked this question?

3. Examine the missing values for BIRTHSLYR. Define the population included in each missing value category. Should these values be included or excluded for the analysis and why? If you decide to exclude the observations then compute the percentage change of the sample size.

4. What does Top Codes represent? Is the variable BIRTHSLYR top coded?

❖ Section III

Laila is interested in the number of births that occur in Cambodia every year. Using the 2008 census, she calculates the mean number of children born in the year to a woman before the census year.
a) What assumptions are required to construct a confidence interval for the true mean? Check whether the assumptions are satisfied. If not, then how would you address the issue of violation of assumptions?

b) Create and interpret a 95% confidence interval for the true mean number of children born to Cambodian women (aged 15 to 49) in the year before the 2008 census.

c) Now create and interpret a 99% confidence interval for the true mean number of children Cambodian women had in the last year.

d) What is the relation between significance level and width of a confidence interval?

❖ Section IV

Consider only dataset of Portugal for Section IV.

1. Show the age distribution by creating histogram. Draw lines on histogram to show the mean and median age.

2. Ma used the Portugal 2011 census to calculate the proportion of individuals that had completed University [EDATTAIN].
a) Who was asked about educational attainment in the Portugal 2011 census?

b) 99% confidence interval for the true proportion of people from Portugal who completed university.

(i) What is the sample proportion of people who completed university?

(ii) Can we use the data to calculate a valid confidence interval? That is, check the required assumption.

(iii) Create and interpret the 99% confidence interval for the true proportion of people from Portugal who completed university.

c) Would you expect the proportion of people completing university to be lower or higher if only persons aged 15 years and older were asked about educational attainment in the Portugal 2011 census?

❖ Section V

Manny wishes to draw a sample of Ghana 2010 census data in order to estimate the proportion of people in the population who have a disability. How many people should Manny include in his sample in order to be 95% confident that the margin of error is within 0.01 of the true proportion?