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: Confidence Interval Answer Key

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 estimate 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

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

# Load the data
ddi <- read_ipums_ddi("ipumsi_00009.xml")
data_CI <- read_ipums_micro(ddi)
ipums_val_labels(data_CI$AGE2)
```

```
## # A tibble: 22 x 2
##      val lbl
##    <dbl> <chr>
##  1     1 0 to 4
##  2     2 5 to 9
##  3     3 10 to 14
##  4     4 15 to 19
##  5     5 15 to 17
##  6     6 18 to 19
##  7     7 18 to 24
##  8     8 20 to 24
##  9     9 25 to 29
## 10    10 30 to 34
## # ... with 12 more rows
```

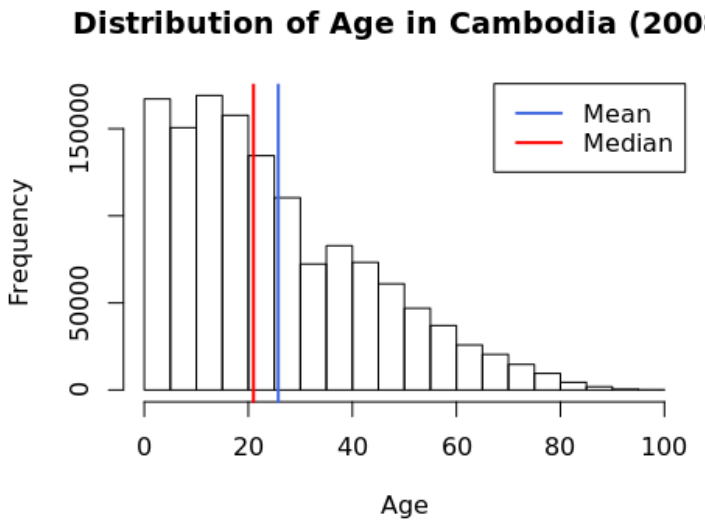
```
# Convert the following variables to factors
data_CI<-within(data_CI, COUNTRY<- as.factor(COUNTRY))
data_CI<-within(data_CI, YEAR<- as.factor(YEAR))
data_CI<-within(data_CI, SEX<- as.factor(SEX))
data_CI<-within(data_CI, EDATTAIN<- as.factor(EDATTAIN))
data_CI<-within(data_CI, AGE2<- as.factor(AGE2))


# Separate out data of Cambodia (COUNTRY = 116) and Portugal (COUNTRY = 620)
from the dataset
Cambodia<- subset(data_CI, COUNTRY == 116)
Portugal<- subset(data_CI, COUNTRY == 620)
```

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.

```
hist(Cambodia$AGE,main = "Distribution of Age in Cambodia (2008)",xlab="Age")
abline(v = mean(Cambodia$AGE),col = "royalblue",lwd = 2)
abline(v = median(Cambodia$AGE),col = "red",lwd = 2)
legend(x = "topright",c("Mean", "Median"), col = c("royalblue", "red"),lwd =
c(2,2))
```



The histogram is right skewed.

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

The population universe for BIRTHSLYR in Cambodia is females aged from 15 to 49 years.

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.

The population included in each missing value category is defined below:
1. Unknown (BIRTHSLYR = 8): Females age 15 to 49 who did not provide a valid response to the census question on births during the last year.
2. NIU (not in universe) (BIRTHSLYR = 9): All females under age 15 or over age 49 and all males.

BIRTHSLYR = 8 signifies unknown responses and if included then it will distort our results. BIRTHSLYR = 9 signifies persons who were not asked this census question. Hence, we should exclude observations with BIRTHSLYR value of 8 and 9. After exclusion of these observations, we observe that the sample size decreases by about 73%.

```
# Convert the BIRTHSLYR value of 8 (unknown responses) to missing values
Cambodia$BIRTHSLYR[Cambodia$BIRTHSLYR==8]<-NA
# Convert the BIRTHSLYR value of 9 (not in universe) to missing values
Cambodia$BIRTHSLYR[Cambodia$BIRTHSLYR==9]<-NA
# Remove the observations for which BIRTHSLYR value is either 8 or 9
newdata_Cambodia <- na.omit(Cambodia)

# Percentage (round off to 2 decimal places) change in total population after
exclusion of missing values
round(((dim(newdata_Cambodia)[1] - dim(Cambodia)[1])/dim(Cambodia)[1])*100,2)

## [1] -72.79
```

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

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 BIRTHSLYR in Cambodia is 4 or more.

❖ 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.

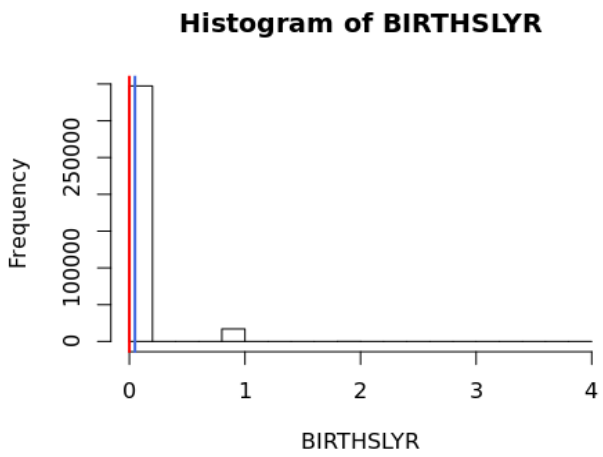
a) What assumptions are required to construct a confidence interval for the true mean? Check whether the assumptions are satisfied. If not, then how would you address the issue of violation of assumptions?

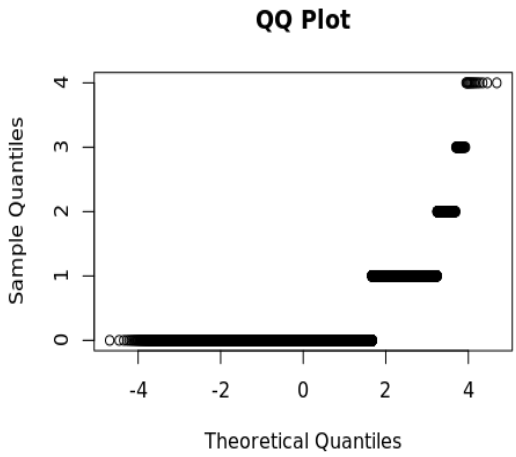
Assumptions:
1. Random sample
2. Population distribution is approximately normal

```
# QQ Plot
qqnorm(newdata_Cambodia$BIRTHSLYR, main = "QQ Plot")

# Histogram
hist(newdata_Cambodia$BIRTHSLYR, main = "Histogram of BIRTHSLYR", xlab = "BIR
THSLYR")
```

```
abline(v = mean(newdata_Cambodia$BIRTHSLYR),col = "royalblue",lwd = 2)
abline(v = median(newdata_Cambodia$BIRTHSLYR),col = "red",lwd = 2)
```





The assumption of random sample is satisfied. However, based on histogram (right skewed) and QQ plot we conclude that the assumption of normality is not satisfied. To construct a confidence interval for the true mean, t-distribution is used and we know that t-distribution is robust to non-normality when there are no outliers. Notice that in this dataset there is no outlier.

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.

$$\bar{x} \pm t_{\frac{\alpha}{2},n-1}\frac{s}{\sqrt{n}} \approx 0.04821 \pm 1.96 * \frac{0.2185}{\sqrt{364612}}$$

Here, α = 0.05. The required 95% confidence interval is (0.048, 0.049). This means that we are 95% confident that the true mean number of children born to Cambodian women in the year is between 0.048 and 0.049.

****** Notice that the confidence interval is very small due to large sample size.

```
# Method 1:

# xbar is computed using mean
xbar<- mean(newdata_Cambodia$BIRTHSLYR)
xbar

## [1] 0.04821015
```

```r
# Standard deviation
s<- sd(newdata_Cambodia$BIRTHSLYR)
s
```

```
## [1] 0.2185327
```

```r
# Number of rows gives n which is 364612
dim(newdata_Cambodia)
```

```
## [1] 364612      8
```

```r
t1<-qt(1-0.05/2,364611)
t1
```

```
## [1] 1.95997
```

```r
# Margin of error
moe1<- t1*(s/sqrt(364612))
moe1
```

```
## [1] 0.0007093336
```

```r
# Lower bound of the confidence interval
Lower_bound <- xbar - moe1
Lower_bound
```

```
## [1] 0.04750082
```

```r
round(Lower_bound,4)
```

```
## [1] 0.0475
```

```r
# Upper bound of the confidence interval
Upper_bound <- xbar + moe1
Upper_bound
```

```
## [1] 0.04891949
```

```r
round(Upper_bound,4)
```

```
## [1] 0.0489
```

```r
# Method 2: We can also get the confidence interval using two sided t.test
t.test(x=newdata_Cambodia$BIRTHSLYR, conf.level=.95, alternative="two.sided")
```

```
##
##  One Sample t-test
##
## data:  newdata_Cambodia$BIRTHSLYR
## t = 133.21, df = 364610, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  0.04750082 0.04891949
## sample estimates:
```

```
##  mean of x
## 0.04821015
```

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

The required 99s% confidence interval is (0.047, 0.049). This means that we are 99% confident that the true mean number of children born to Cambodian women in the year is between 0.047 and 0.049.

****** Notice that the confidence interval is very small due to large sample size. However, the width of confidence interval is larger that of 95% confidence interval.

```
# Confidence interval using two sided t.test
t.test(x=newdata_Cambodia$BIRTHSLYR, conf.level=.99, alternative="two.sided")

##
##  One Sample t-test
##
## data:  newdata_Cambodia$BIRTHSLYR
## t = 133.21, df = 364610, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 99 percent confidence interval:
##  0.04727793 0.04914238
## sample estimates:
##  mean of x
## 0.04821015
```

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

We observe that as the significance level decreases from 5% to 1% (that is confidence level increases from 95% to 99%), width of a confidence interval increases.

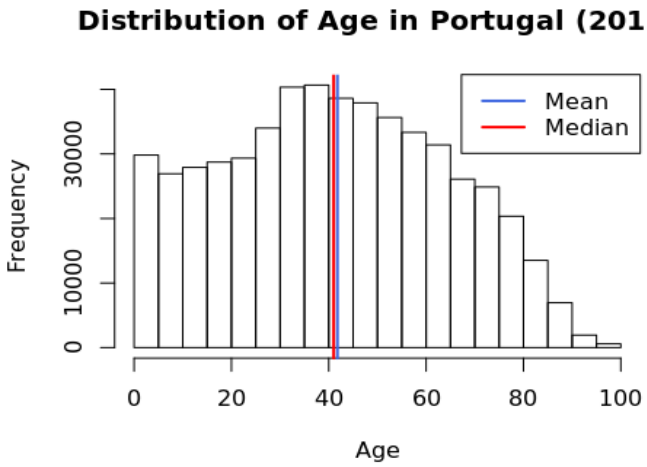
❖ Section IV

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.

```
hist(Portugal$AGE,main = "Distribution of Age in Portugal (2011)",xlab="Age")
abline(v = mean(Portugal$AGE),col = "royalblue",lwd = 2)
abline(v = median(Portugal$AGE),col = "red",lwd = 2)
```

```
legend(x = "topright",c("Mean", "Median"), col = c("royalblue", "red"),lwd =
c(2,2))
```



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?

All persons were 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?

$$\hat{p} = \frac{\text{Number of people who completed university}}{\text{Total number of people}} = \frac{62148}{528870} \approx 0.1175$$

Thus, the sample proportion of people who completed university is 0.1175.

```
# Sample size (n) is 528870
dim(Portugal)

## [1] 528870      8

n<- 528870

summary(Portugal$EDATTAIN)
```

```
##      0      1      2      3      4      9
##      0 235544 156225  74953  62148      0

university_completed <-62148

# Proportion
p<- university_completed/n
p

## [1] 0.1175109
```

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

The required assumption is that we have a random sample and the number of successes and failures are both at least 15. Number of successes = $n\hat{p} = 62148$ and number of failures = $n(1-\hat{p}) = 466722$. Hence, all the required assumptions are satisfied.

```
# Assumption Check
p*n

## [1] 62148

(1-p)*n

## [1] 466722
```

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

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \approx 0.1175 \pm 0.001$$

Here, α = 0.01. We are 99% confident that the true proportion of people from Portugal who completed university is between (0.116, 0.119).

****** Notice that the confidence interval is very small due to large sample size

```
# Method 1:

# Margin of error
z<- qnorm(0.99)
moe_p<- z*sqrt((p*(1-p))/n)
moe_p

## [1] 0.001030135
```

```r
# Lower bound of confidence interval
Lower_bound <- p - moe_p
Lower_bound

## [1] 0.1164808

round(Lower_bound,4)

## [1] 0.1165

# Uppper bound of confidence interval
Upper_bound <- p + moe_p
Upper_bound

## [1] 0.1185411

round(Upper_bound,4)

## [1] 0.1185

# Method 2: We can also get the confidence interval using prop.test
prop.test(x=university_completed, n=n, conf.level=0.99, alternative="two.side
d")

##
##  1-sample proportions test with continuity correction
##
## data:  university_completed out of n, null probability 0.5
## X-squared = 309490, df = 1, p-value < 2.2e-16
## alternative hypothesis: true p is not equal to 0.5
## 99 percent confidence interval:
##  0.1163742 0.1186573
## sample estimates:
##         p
## 0.1175109
```

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?

If only persons aged 15 years and older were asked about educational attainment in Portugal 2011 census instead of all persons, then the total sample size, which is the denominator in the equation of proportion, will decrease. However, the proportion of people who completed university will be the same as almost none would have completed university under the age of 15 years. Thus, the proportion of people completing university will be higher if only persons aged 15 years and older were asked about educational attainment.

❖ 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?

The required sample size is given by

$$n = \frac{\hat{p}(1-\hat{p})}{m^2} {z_{\alpha/2}}^2$$

Here, $z_{\alpha/2} = z_{0.05/2} = 1.96$ as we want a 95% confidence interval and $m = 0.01$. In this case, we do not have a prior guess of the proportion value. Hence, we choose $\hat{p} = 0.5$ as this will provide us with larger sample size as compared to any other value of $\hat{p}$.

$$n = \frac{0.5\,(1-0.5)}{(0.01)^2}(1.96)^2 = 9604$$

Thus, the required sample size is 9604.