**Statistical Inference**

In this lab, we will explore inferential statistics. We will start with sampling distribution, and continue with central limit theorem, confidence interval and hypothesis testing.

**1 Sampling Distribution**

In this section, we will use a dataset called ames. It is a real estate data from the city of Ames, Iowa, USA. The details of every real estate transaction in Ames is recorded by the City Assessor’s office. Our particular focus for this lab will be **all residential home sales in Ames between 2006 and 2010**. This collection represents our **population** of interest.

Let’s load the data first and have a quick look at the variables.

ames <- read.csv("http://bit.ly/315N5R5")

glimpse(ames) *# you need dplyr to use this function*

There are quite a few variables in the data set, enough to do a very in-depth analysis. For this lab, we’ll restrict our attention to just two of the variables:

* the above ground living area of the house in square feet (Gr.Liv.Area)
* sale price (SalePrice) in U.S. Dollars.

To save some effort throughout the lab, create two objects with short names that represent these two variables.

area <- ames$Gr.Liv.Area

price <- ames$SalePrice

head(area, n=10) *#show first 10 observations*

head(price, n=10) *#show first 10 observations*

Let’s look at the distribution of area in our population of home sales by calculating a few summary statistics and making a histogram.

length(area) *#how many observations in the vector?*

any(is.na(area)) *#is there any NA in the vector area?*

area.pop.sd<-sqrt(sum((area - mean(area))^2)/(2930)) *# Population standard deviation*

area.pop.sd

summary(area)

The above ground living area of the house variable contains 2930 distinct observations and there are not missing values. We see that the distribution of above ground living area of the house population variable has a mean of 1499.69 sq.ft.(square feet), a median of 1442 sq.ft. and a population standard deviation of 505.42 sq.ft. Observations in the dataset range from a minimum of 334sq.ft. for the smallest house to a maximum of 5642 sq.ft. for the biggest house. The histogram of the variable shows a positive (right) skew of the population data.

hist(area,

main = "Histogram of above ground living area",

xlab = "Above ground living area (sq.ft.)",

)

In this exercise, we have access to the **entire population** of sales transactions, but this is rarely the case in real life. Gathering information on an entire population is often extremely costly or impossible. Because of this, we often take a sample of the population and use that to understand the properties of the population.

For our learning purpose, having the entire population of interest can help us understand the relation between population, samples, and sampling distribution. In this lab we will take smaller samples from the full population.

area <- ames$Gr.Liv.Area *# create new dataset containing only variable 'Gr.Liv.Area' from dataset 'ames'*

samp1 <- sample(area, 50) *#take a random sample of 50 observations from the dataset 'area'*

mean(samp1) *# mean of the sample distribution for area. Note difference from population mean.*

This command collects a simple random sample of size 50 from the ames dataset, which is assigned to samp1. This is like going into the City Assessor’s database and pulling up the files on 50 random home sales. Working with these 50 files would be considerably simpler than working with all 2930 home sales.

**Exercise**

1. Using your RStudio, take a sample, also of size 50, and call it samp1. Take a second sample of size 1000, and call it samp2. How does the mean of samp1 compare with the mean of samp2?
2. Suppose we took two more samples, one of size 1500 and another of size 2000. Which would you think would provide a more accurate estimate of the population mean?

Not surprisingly, every time we take another **random** sample, we get a different sample mean. It’s useful to get a sense of just how much variability we should expect when estimating the population mean this way. The distribution of sample means, called the sampling distribution, can help us understand this variability. In this lab, because we have access to the population, we can build up the sampling distribution of the sample mean by repeating the above steps many times (also called “drawing samples from the population”). Here we will generate 5000 samples of size 50 from the population, calculate the mean of each sample, and store each result in a vector called sample\_means50. We will then plot the histogram of this sampling distribution.

area <- ames$Gr.Liv.Area

sample\_means50 <- rep(NA, 5000) *#initialise a vector*

**for**(i **in** 1:5000){ *# use of a loop function to draw a random sample 5000 times*

samp <- sample(area, 50)

sample\_means50[i] <- mean(samp)

}

hist(sample\_means50, breaks = 25,

main = "Sampling distribution of sample mean for Above ground living area",

xlab = "Means (sq.ft.)") *#Histogram of the 5000 samples (sampling distribution of the samples mean)*

Try running the code a number of times. What do you observe? Each time you run the code, you see slightly different sampling distribution of mean.

What happens to the distribution if we instead collected 50,000 sample means? Try it! Change the number of draws on the code above and see how the shape of distribution changes.

(**Hint:** Check the normality of the distribution.)

## 2 Sample Size and Sampling Distribution

Mechanics aside, let’s return to the reason we used a for loop in [Sampling Distribution](https://bookdown.org/mrenna/statbook/statistical-inference.html#sampling-distribution) section: to compute a sampling distribution, specifically, this one.

area <- ames$Gr.Liv.Area

sample\_means50 <- rep(NA, 5000)

**for**(i **in** 1:5000){

samp <- sample(area, 50)

sample\_means50[i] <- mean(samp)

}

hist(sample\_means50)

The sampling distribution that we computed tells us much about estimating the average living area in homes in Ames. Because the sample mean is an unbiased estimator, the sampling distribution is centred at the true average living area of the the population, and the spread of the distribution indicates how much variability is induced by sampling only 50 home sales.

To get a sense of the effect that sample size has on our distribution, let’s build up two more sampling distributions: one based on a sample size of 10 and another based on a sample size of 100 from a population size of 5000.

area <- ames$Gr.Liv.Area

sample\_means10 <- rep(NA, 5000)

sample\_means100 <- rep(NA, 5000)

**for**(i **in** 1:5000){

samp <- sample(area, 10)

sample\_means10[i] <- mean(samp)

samp <- sample(area, 100)

sample\_means100[i] <- mean(samp)

}

Here we’re able to use a single for loop to build two distributions by adding additional lines inside the curly braces. Don’t worry about the fact that samp is used for the name of two different objects. In the second command of the for loop, the mean of samp is saved to the relevant place in the vector sample\_means10. With the mean saved, we’re now free to overwrite the object samp with a new sample, this time of size 100. In general, anytime you create an object using a name that is already in use, the old object will get replaced with the new one.

To see the effect that different sample sizes have on the sampling distribution, let’s plot the three distributions on top of one another.

area <- ames$Gr.Liv.Area

sample\_means10 <- rep(NA, 5000)

sample\_means50 <- rep(NA, 5000)

sample\_means100 <- rep(NA, 5000)

**for**(i **in** 1:5000){

samp <- sample(area, 10)

sample\_means10[i] <- mean(samp)

samp <- sample(area, 50)

sample\_means50[i] <- mean(samp)

samp <- sample(area, 100)

sample\_means100[i] <- mean(samp)

}

par(mfrow = c(3, 1)) *# this creates 3 rows and 1 column for graphs*

xlimits <- range(sample\_means10)

hist(sample\_means10, breaks = 25, xlim = xlimits)

hist(sample\_means50, breaks = 25, xlim = xlimits)

hist(sample\_means100, breaks = 25, xlim = xlimits)

The first command specifies that you’d like to divide the plotting area into 3 rows and 1 column of plots (to return to the default setting of plotting one at a time, use par(mfrow = c(1, 1)). The breaks argument specifies the number of bins used in constructing the histogram. The xlim argument specifies the range of the x-axis of the histogram, and by setting it equal to xlimits for each histogram, we ensure that all three histograms will be plotted with the same limits on the x-axis. Here, I set it to be the same as the histogram which has more spread on the x-axis.

Your turn! Try different sample sizes and see what happens to the distribution.

* When the sample size is larger, what happens to the center? What about the spread?

1. The center moves to the right; standard deviation gets smaller
2. The center approaches to the true mean; the standard deviation does not change
3. The center approaches to the true mean; the standard deviation gets smaller

**Exercise** So far, we have only focused on estimating the mean living area in homes in Ames. Now you’ll try to estimate the mean home price and do the same analyses we did above for living area using RStudio. Specifically,

* Take a random sample of size 50 from price. Using this sample, what is your best point estimate of the population mean?
* Since you have access to the population, simulate the sampling distribution for mean of x\_price by taking 5000 samples from the population of size 50 and computing 5000 sample means. Store these means in a vector called sample\_means50. Plot the data, then describe the shape of this sampling distribution. Based on this sampling distribution, what would you guess the mean home price of the population to be? Finally, calculate and report the population mean.
* Change your sample size from 50 to 1500, then compute the sampling distribution using the same method as above, and store these means in a new vector called sample\_means1500. Describe the shape of this sampling distribution, and compare it to the sampling distribution for a sample size of 50. Based on this sampling distribution, what would you guess to be the mean sale price of homes in Ames?
* Of the sampling distributions from 2 and 3, which has a smaller spread? If we’re concerned with making estimates that are more often close to the true value, would we prefer a distribution with a large or small spread?