# STAT 100B Lab 3 Confidence intervals

First Name __________ Last Name ___________SID (last 4 digits only)_____

## Sampling from Ames, Iowa

If you have access to data on an entire population, say the size of every house in Ames, Iowa, it's straight forward to answer questions like, "How big is the typical house in Ames?" and "How much variation is there in sizes of houses?". If you have access to only a sample of the population, as is often the case, the task becomes more complicated. What is your best guess for the typical size if you only know the sizes of several dozen houses? This sort of situation requires that you use your sample to make inference on what your population looks like.

## The data

In the previous lab, `Sampling Distributions` , we looked at the population data of houses from Ames, Iowa. Let's start by loading that data set.

```
download.file("http://www.openintro.org/stat/data/ames.RData", destfile = "ames.RData")
load("ames.RData")
```

In this lab we'll start with a simple random sample of size 60 from the population. Specifically, this is a simple random sample of size 60. Note that the data set has information on many housing variables, but for the first portion of the lab we'll focus on the size of the house, represented by the variable `Gr.Liv.Area` .

```
population <- ames$Gr.Liv.Area
samp <- sample(population, 60)
```

**Exercise 1** Check the distribution of your sample. (1) To build a histogram of your sample, which R code could you use? (2pts) (2) Describe the distribution of your sample. (1pt) (3) What would you say is the typical size within your sample? (1pt) (4) Also state precisely what you interpreted "typical" to mean.(1pt)

2.Variability from sample to sample. (1)Would you expect another student to obtain a distribution identical to yours?________(Yes/ No) (2pts). (2) Would you expect it to be similar?________(Yes/ No) (2pts) (3) Why or why not? (2pts)

## Confidence intervals

One of the most common ways to describe the typical or central value of a distribution is to use the mean. In this case we can calculate the mean of the sample using,

```
sample_mean <- mean(samp)
```

Return for a moment to the question that first motivated this lab: based on this sample, what can we infer about the population? Based only on this single sample, the best estimate of the average living area of houses sold in Ames would be the sample mean, usually denoted as $\bar{x}$ (here we're calling it `sample_mean`). That serves as a good *point estimate* but it would be useful to also communicate how uncertain we are of that estimate. This can be captured by using a *confidence interval*.

We can calculate a 95% confidence interval for a sample mean by adding and subtracting the critical value (`cv`) times the standard error (`se`) to the point estimate. Note that the `R` function `qnorm` finds the lower tail percentile from the standard normal distribution.

```r
se <- sd(samp) / sqrt(60)
cv <- -qnorm(0.05/2)
lower <- sample_mean - cv * se
upper <- sample_mean + cv * se
c(lower, upper)
```

```
## [1] 1324.447 1569.920
```

This is an important inference that we've just made: even though we don't know what the full population looks like, we're 95% confident that the true average size of houses in Ames lies between the values *lower* and *upper*. There are a few conditions that must be met for this interval to be valid.

**Exercise 2**

Check the confidence interval you calcuated. (1) Your confidence interval is _____________ (1pt). (2) Does it capture the true average size of houses in Ames? (1pt) _______ (Yes/No)

**Exercise 3**

For the confidence interval to be valid, the sample mean must be normally distributed and have standard error $s/\sqrt{n}$. What conditions must be met for the normality assumption to be true? (4pts)

# What does "95% confidence" mean

In this case we have the luxury of knowing the true population mean since we have data on the entire population. This value can be calculated using the following command:

```r
pop_mean <- mean(population)
print(pop_mean)
```

```
## [1] 1499.69
```

Using R, we're going to recreate many samples to learn more about how sample means and confidence intervals vary from one sample to another. *Loops* come in handy here (If you are unfamiliar with loops, review the Sampling Distribution Lab (http://htmlpreview.github.io/?https://github.com/andrewpbray/oiLabs/blob/master/sampling_distributions/sampling_distributions.html)).

Here is the rough outline:

1. Obtain a random sample.
2. Calculate and store the sample's mean and standard deviation.
3. Repeat steps (1) and (2) 50 times.
4. Use these stored statistics to calculate many confidence intervals.

But before we do all of this, we need to first create empty vectors where we can save the means and standard deviations that will be calculated from each sample. And while we're at it, let's also store the desired sample size as `n`.

```
samp_mean <- rep(NA, 50)
samp_sd <- rep(NA, 50)
n <- 60
```

Now we're ready for the loop where we calculate the means and standard deviations of 50 random samples.

```
num_ci <- 50
for(i in 1:num_ci){
  samp <- sample(population, n) # obtain a sample of size n = 60 from the population
  samp_mean[i] <- mean(samp)    # save sample mean in ith element of samp_mean
  samp_sd[i] <- sd(samp)        # save sample sd in ith element of samp_sd
}
```

Lastly, we construct the confidence intervals.

```
cv <- -qnorm(0.05/2)
lower_vector <- samp_mean - cv * samp_sd / sqrt(n)
upper_vector <- samp_mean + cv * samp_sd / sqrt(n)
```

Lower bounds of these 50 confidence intervals are stored in `lower_vector`, and the upper bounds are in `upper_vector`. Let's view the first interval.

```
c(lower_vector[1], upper_vector[1])
```

```
## [1] 1362.945 1591.189
```

**Exercise 4** Is the population mean in your first confidence interval? (1) The the first interval you obtained is _________(1pt). (2) Does your confidence interval capture the population mean? __________(Yes/No) (1pt)

Using the following function (which was downloaded with the data set), plot all intervals.

```
plot_ci(lower_vector, upper_vector, mean(population))
```

You can either look on the graph and count or use the following R code to check what proportion of the confidence intervals capture the true mean.

```
sum( (lower_vector <= pop_mean) & (upper_vector >= pop_mean)) / num_ci
```

```
## [1] 0.94
```

**Exercise 5** What does 95% confidence mean? (1) What proportion of your confidence intervals include the true population mean? (1pt) (2) Is this proportion exactly equal to the confidence level? If not, explain why. (2pts)

# On your own

1 Suppose we'd like to have a 98% confidence interval. What is the appropriate critical value (z value)? (1) Show your R code to find critical value (1pt). (2) The appropriate critical value is _______.(1pt)

2 Construct 50 confidence intervals at the confidence level of 98%. You do not need to obtain new samples, simply calculate new intervals based on the sample means and standard deviations you have already collected. (1) Show your R code to calculate the new intervals (4pts). (2) Using the `plot_ci` function, plot all intervals. Your R code to plot the intervals (2pts). (3) Calculate the proportion of intervals that inclue the true population mean. Show your R code below (5pts)