LAB 2

Objectives

- Insert a population data and summarize its statistics.
- Withdraw samples from the population.
- Investigate the sampling distribution of the sample mean for different sample sizes.
- Investigate the relation between the sample size and the standard error (sampling variability of the mean).

STEP 1: the step's objective is to install the required packages

```
install.packages("dplyr", dependencies = TRUE)
install.packages("ggplot2", dependencies = TRUE)
install.packages ("statsr", dependencies = TRUE)
```

STEP 2: the step's objective is to load the required packages

```
library(statsr)
library(dplyr)
library(ggplot2)
```

STEP 3: the step's objective is to load Population data

```
data(ames)
```

STEP 4: the step's objective is to summarize the statistics of population

```
ames %>%
summarise(mu = mean(area), pop_med = median(area),sigma
= sd(area), pop_iqr = IQR(area),
pop_min = min(area), pop_max = max(area),
pop_q1 = quantile(area, 0.25),  # first quartile, 25th percentile
pop_q3 = quantile(area, 0.75))  # third quartile, 75th percentile
```

```
# A tibble: 1 x 8
     mu pop_med sigma pop_iqr pop_min pop_max pop_q1 pop_q3
  <dbl>   <dbl> <dbl>   <dbl>   <int>   <int>  <dbl>  <dbl>
1 1500.    1442  506.    617.     334    5642   1126  1743.
```

STEP 5: the step's objective is to take a random sample form the population

- In this lab we have access to the entire population, 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.
- If we were interested in estimating the mean living area in Ames based on a sample, we can use the following command to survey the population.

```
sampl <- ames %>%
sample_n(size = 50)
```

- This command collects a simple random sample of `size` 50 from the `ames` dataset, which is assigned to `sampl`. 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.

STEP 6: the step's objective is to summarize the statistics of random sample

```
sampl %>%
  summarise(mu_sample = mean(area), sample_med = median(area),
            sigma_sample = sd(area), sample_iqr = IQR(area),
            sample_min = min(area), sample_max = max(area),
            sample_q1 = quantile(area, 0.25),  # first quartile, 25th percentile
            sample_q3 = quantile(area, 0.75))  # third quartile, 75th percentile
```

```
# A tibble: 1 x 8
  mu_sample sample_med sigma_sample sample_iqr sample_min sample_max sample_q1 sample_q3
      <dbl>      <dbl>        <dbl>      <dbl>      <int>      <int>     <dbl>     <dbl>
1     1459.      1444.         477.       614.        848       3112     1093.     1707.
```

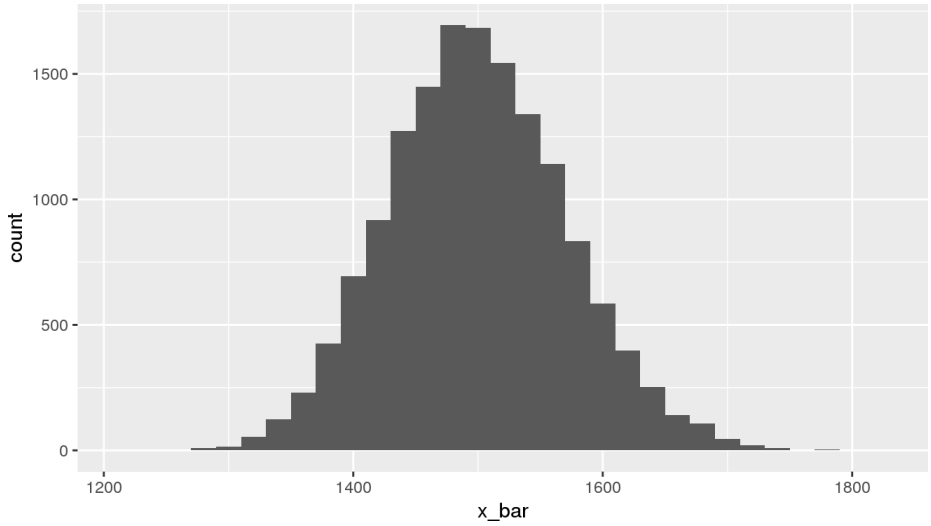
Discussion

1. In the given population, by sampling less than 3% of the population, does the sample mean give a pretty good estimate of the average living area. **TRUE**
2. Suppose we took two more samples, one of size 100 and one of size 1000. Which would you think would provide a more accurate estimate of the population mean? **1000 provides more accurate estimation of the population mean.**
3. Does the distribution of the samples mean looks like the normal distribution? **TRUE**

STEP 8: the step's objective is to build up the sampling distribution for the sample mean

- by repeating the above steps many times. Here we will generate 15,000 samples and compute the sample mean of each. Note that we are sampling with replacement, `replace = TRUE` since sampling distributions are constructed with sampling with replacement.

```
sample_means50 <- ames %>%
rep_sample_n(size = 50, reps = 15000, replace = TRUE) %>%
summarise(x_bar = mean(area))
ggplot(data = sample_means50, aes(x = x_bar)) + geom_histogram(binwidth = 20)
```



STEP 9: the step's objective is to notice the relation between the average mean and variability for the distribution of the samples means and compare them with the population mean and its standard deviation

```
# Avarage mean.
mean(sample_means50$x_bar)
```

```
# Stardard error.
sd(sample_means50$x_bar)
```

```
> mean(sample_means50$x_bar)
[1] 1499.177
> # Stardard error.
> sd(sample_means50$x_bar)
[1] 70.9654
```

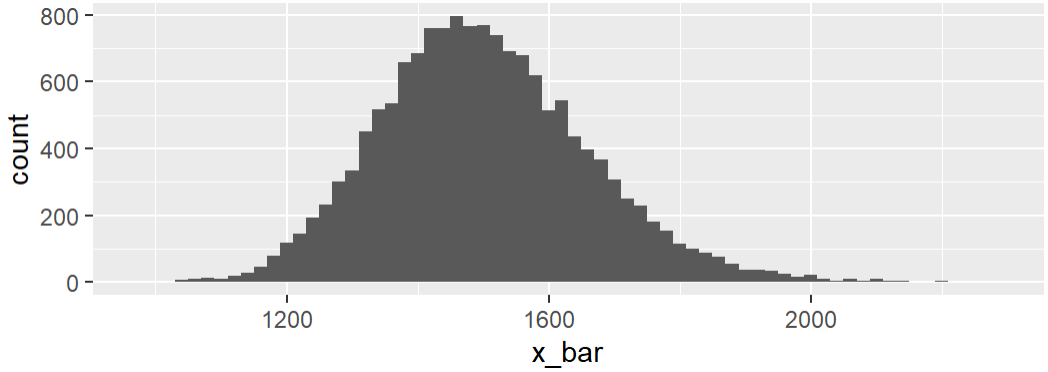
NOTE: The average mean from `sample_means50` is very close to the population mean 1499.69, as expected. Finally, the standard error is approximately the population standard deviation (505.51) divided by $\sqrt{50}$, as expected. According to the centra limit theorem

- 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 centered at the true average living area of the population, and the spread of the distribution indicates how much variability is induced by sampling only 50 home sales.
- In the remainder of this LAP we will work on getting a sense of the effect that sample size has on our sampling distribution.

STEP 10: the step's objective is to sample the population by different sample sizes

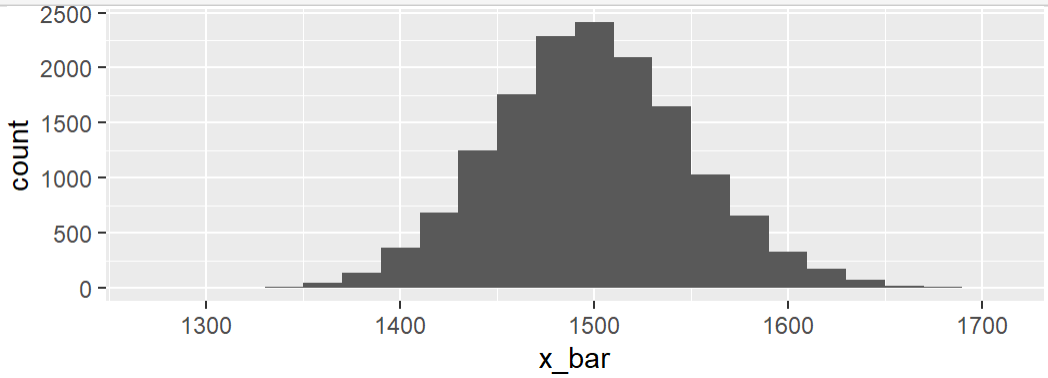
First the n=10

```
sample_means10 <- ames %>%
rep_sample_n(size = 10, reps = 15000, replace = TRUE) %>%
summarise(x_bar = mean(area))
ggplot(data = sample_means10, aes(x = x_bar)) + geom_histogram(binwidth = 20)
```



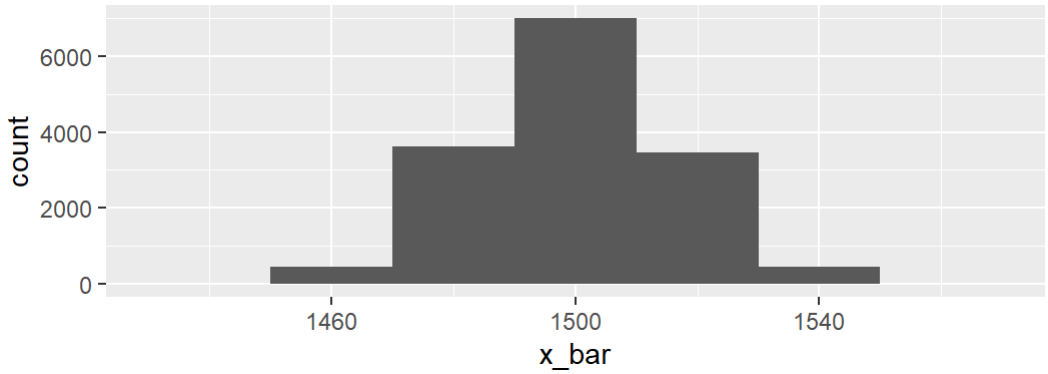
second the n=100

```
sample_means100 <- ames %>%
rep_sample_n(size = 100, reps = 15000, replace = TRUE) %>%
summarise(x_bar = mean(area))
ggplot(data = sample_means100, aes(x = x_bar)) + geom_histogram(binwidth = 20)
```



Finally, the n=1000

```
sample_means1000 <- ames %>%
rep_sample_n(size = 1000, reps = 15000, replace = TRUE) %>%
summarise(x_bar = mean(area))
ggplot(data = sample_means1000, aes(x = x_bar)) + geom_histogram(binwidth = 20)
```



Discussion

1. As the sample size increases, the variability of the sampling distribution decreases **TRUE**

- It makes intuitive sense that as the sample size increases, the center of the sampling distribution becomes a more reliable estimate for the true population mean.

Function	Description
dnorm	Normal density (Probability Density Function)
pnorm	Normal distribution (Cumulative Distribution Function)
qnorm	Quantile function of the Normal distribution
rnorm	Normal random number generation

Exercises:

E1: Take a random sample of size 50 from `price`. Using this sample, what is your best point estimate of the population mean?

E2: Since you have access to the population, simulate the sampling distribution for $\bar{x}_{price}$ 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?

E3: Change your sample size from 50 to 150, then compute the sampling distribution using the same method as above, and store these means in a new vector called `sample_means150`. 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?

E4: Take a sample of size 15 from the population and calculate the mean `price` of the homes in this sample. Using this sample, what is your best point estimate of the population mean of prices of homes?

E5: Since you have access to the population, simulate the sampling distribution for $\bar{x}_{price}$ by taking 2000 samples from the population of size 15 and computing 2000 sample means. Store these means in a vector called `sample_means15`. 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

E6: Change your sample size from 15 to 150, then compute the sampling distribution using the same method as above, and store these means in a new vector called `sample_means150`. Describe the shape of this sampling distribution and compare it to the sampling distribution for a sample size of 15. Based on this sampling distribution, what would you guess to be the mean sale price of homes in Ames?