Statistics 251: Lab 5 Pre-Reading
Sampling Distribution + CLT

Objectives:

- Understand
  - difference between a statistic & a parameter
  - a statistic can be a random variable
  - Central Limit Theorem (CLT)
- Look at an example
- Prepare for lab 5 by trying a pre-lab exercise

1 Introduction

1.1 Statistics & Parameters

What exactly is a 'statistic', versus a 'parameter'?

- A statistic is a number which describes the sample data. In other words, a statistic is a function of the sampled data, which typically generates a meaningful summary of the data. However, a statistic will vary from sample to sample.
  E.g. Sample mean, sample median, sample variance, sample maximum
- A parameter is a fixed number of some interest associated with a population.
  E.g. Population mean, population median, population variance, population maximum

Typically, we do not know what any of these population parameters are, and hence we have to estimate them with their associated statistics calculated from some sampled data. For example, an 'obvious' candidate for estimating the population mean is the sample mean.

The question we must ask is, "what ensures that the sample mean will be 'close' to the population mean? And can we quantify just how close it will be? "

This is where we make explicit a very fundamental concept in Statistics:

1.2 Statistics are Random Variables!

That is, statistics have distributions, expected values, variances and so forth, just like we commonly associate with other pre-defined random variables of interest. We typically call the distribution of a statistic a sampling distribution, as the distribution of a particular *statistic* will depend on how the data itself is distributed.

More intuitively, we can think of $\bar{X}$ as a random variable encompassing all possible sample means that could have been obtained from samples of size n, over the range of different samples that could have been collected from our population of interest. Finally, since this is a random variable, it will have associated density/distribution functions; i.e., the probabilities associated with getting a sample mean within a certain interval. For example, we are much more likely to observe sample means near the true population mean, rather than far away from their associated population mean.

1.3 Central Limit Theorem (CLT)

Now, we can learn why statisticians love the normal distribution so much. Typically, the exact sampling distribution for a given statistic may be difficult to calculate, especially if the distributions of the random variables being assessed are unknown. However, the central limit theorem (CLT) provides a very nice result for the distribution of sample means.

Sample Means are Approximately Normal! (for 'large' n)

More formally, lets suppose $X_1, X_2, \cdots, X_n$ are iid (identical & independently distributed) random variables (in other words, the data was obtained through random sampling), with unknown distributions, but some mean μ and variance σ^2. Then, by the central limit theorem,

$$\bar{X} \overset{approx.}{\sim} N(\mu, \frac{\sigma^2}{n}) \quad \text{for 'large' } n.$$

How large n needs to be depends on how close to 'Normal' the random variables X_i are. If the variables are symmetric with more probability mass concentrated near their mean, then we might require only $n \geq 5$. However, if the variables have very 'strange' distributions, i.e., long tails, multiple modes, strongly skewed distributions, we will require $n \geq 20$ or more. If we are given no information about the distribution of the random variables, or information about the distribution observed in the data collected, then it is usually best to be conservative and require $n \geq 20$.

Hence, by employing the CLT, we can say a lot about the sample means we observe from all kinds of random variables. As we'll see later, the sample mean is a rather nice variable to perform inference on; hence, having this result now will prove useful later. (Recall that, roughly speaking, inference is the process of using sample data to make conclusions about population parameters.)

1.4 Some Finer Points

A small note: we do not need to invoke the central limit theorem if we know the random variables $X_1, X_2, \cdots, X_n$ are already normal. The sample mean of a set of normal random variables will be normal as well, without application of the central limit theorem.

Another important thing to note: the results for the mean and variance of a sample mean hold regardless of what the distribution of the data is. The **central limit theorem** only tells us that the distribution of the sample mean will be normal: it is not required to derive the mean and variance of a sample mean.

Finally, please make sure you are clear on the distinction between the sample size, and the number of samples. **The central limit theorem depends on the sample size, not the number of samples!** For example, suppose we were to compare two histograms: one of 1000 samples of size 25, and one of 10000 samples of size 25. What would be the same? What would be different?

2 An Example

2.1 Problem

We are testing the load bearing capacity of lumber treated with a new chemical. Suppose that the mean capacity of lumber produced from this process is 1000 lbs, with a standard deviation of 75 lbs.

1) What is the probability that a randomly sampled unit of lumber will have a load-bearing capacity less than 960 lbs?

2) What is the probability that, based on a random sample of 30 units of lumber, a sample mean of less than 960 lbs is observed?

2.2 Solution

1) This question doesn't say anything about the distribution of load bearing capacity for this type of lumber... we dont have enough information to answer this question!

2) While we don't know the distribution of the load bearing capacity for a particular unit of lumber, we can use the CLT to get the distribution for the **sample mean of the load bearing capacity for n units of lumber** (provided n is large enough!)

Let X_i denote the load bearing strength of the ith piece of lumber. Then, we can write the sample mean as:

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_{30}}{30} = \frac{1}{30}\sum_{i=1}^{30} X_i \qquad \text{(Note the capital letters)}$$

And, since the distributions of X_i's is unknown, but we have a 'large' sample size ($n \geq 20$), by the **central limit theorem**,

$$\bar{X} \overset{approx.}{\sim} N\Big(\mu_{\bar{X}} = 1000, \sigma^2_{\bar{X}} = \frac{75^2}{30}\Big)$$

Next, we need to determine ($\bar{X} \leq 960$). This can be determined if we standardize $\bar{X}$:

subtract its mean, and divide out by its standard deviation.

$$\begin{aligned} P(\bar{X} \leq 950) &= P(\frac{\bar{X} - 1000}{75/\sqrt{30}} \leq \frac{960 \text{ - } 1000}{75/\sqrt{30}}) \\ &= P(Z \leq \text{-2.92 }) \\ &= \text{ 0.0018} \end{aligned}$$

3 Pre-Lab Exercise

3.1 A Simulation

Let's do a simulation to visualize the Central Limit Theorem.

Suppose the discussion board on the STAT 241/251 course webpage did not get utilized well throughout the term. However, on the day before the midterm, students made frequent posts. While scrolling through the posts, you notice that posts were made at a rate of one per minute, and the time between posts appear to follow an Exponential distribution.

To observe the shape of the probability density function, let's draw a random sample of size 500 and construct a histogram. (Hint: *prob=TRUE* in the *hist* command replaces the *frequency* on the vertical axis with the *density*).

```
x <- rexp(n = 500, rate = 1)
hist(x, prob=TRUE, breaks=20)       # Figure 1
```

(Note: Your histogram may not be identical to the one given here as this is a simulation).

We can add an exponential curve (with rate = 1) on the histogram and compare whether the histogram has an approximately exponential form or not.(Hint: use *curve* function, adjust *from* and *to* values from previous histogram horizontal axis).

```
curve( dexp(x, rate=1), from=0, to=7, add=TRUE)         # Figure 2
```

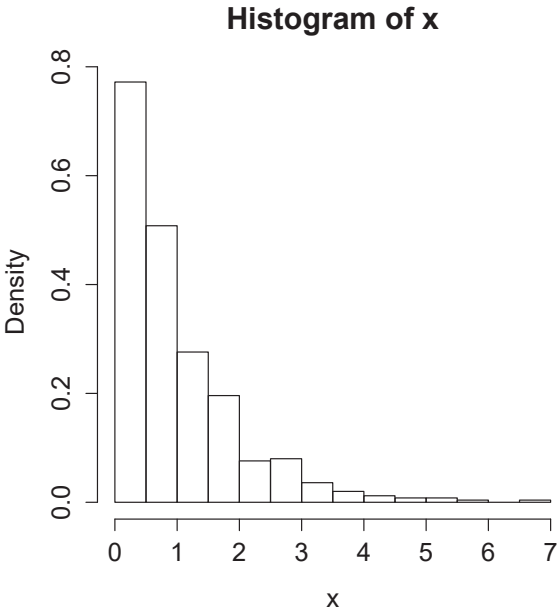

Figure 1: Histogram of one random sample.

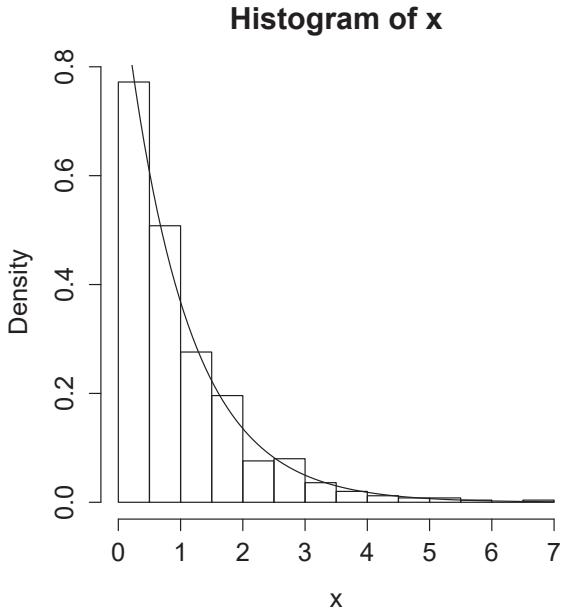

Figure 2: Exponential curve over the histogram.

Next, to observe the shape of the *distribution of sample means*, let's draw 1000 samples of size 50, find the *sample mean of each sample* and draw a histogram and density plot of those sample means.

```
n <- 50
N <- 1000

# Create a matrix which stores the 1000 samples of size 50
X <- matrix(data = rexp(N*n, rate = 1), nrow = N, ncol = n)

# Find the mean of each sample
xbar <- apply(X, MARGIN = 1, FUN = mean)
```

Recall that if you need to figure out how a particular function works, you can use ? command in R to take a look into the R documentation for that function. For example, *?apply* shows the following:

X	the array to be used.
MARGIN	a vector giving the subscripts which the function will be applied over. 1 indicates rows, 2 indicates columns, c(1,2) indicates rows and columns.
FUN	the function to be applied. (In the case of functions like +, %*%, etc., the function name must be backquoted or quoted).

We can draw a histogram and density plot of the 1000 sample means obtained above as follows:

```
hist(xbar, prob=TRUE, breaks=20)          # Figure 3
plot(density(xbar))                       # Figure 4
```

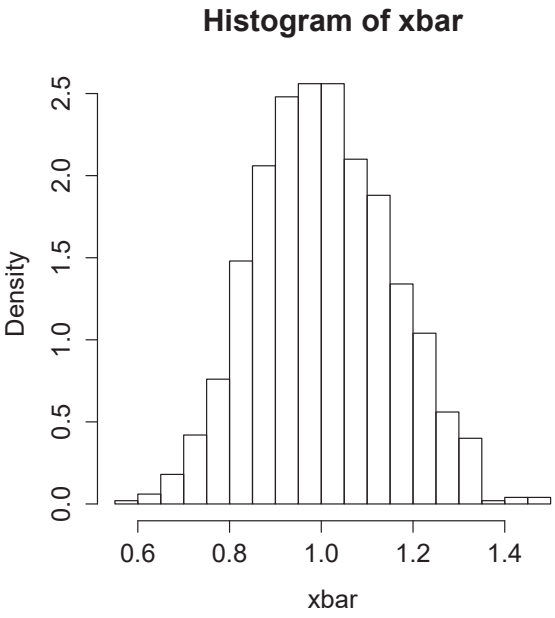

Figure 3: Histogram of the sample means.

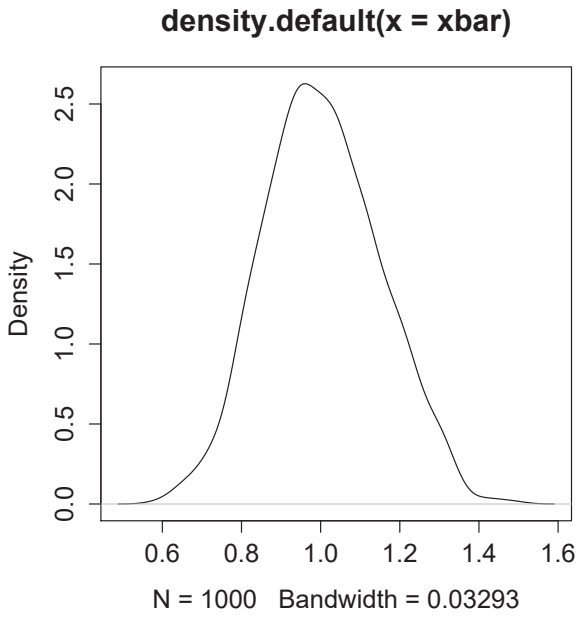

Figure 4: Density plot of the sample means.

What do you notice about the shape of the distribution of sample means, compared to the original distribution from which the samples were drawn? What can you conclude about the centre and spread of the distribution of sample means?

3.2 Food for Thought

Give some thought to the following questions. (Hint: You can use simulations similar to the above to find answers).

1. Would the resulting distribution of sample means look the same if the sample sizes were

   a) Smaller (try sample sizes of 5)?

   b) Larger (try sample sizes of 500)?

2. Would the resulting distribution of sample means look bell shaped if they were drawn from a distribution other than the Normal distribution? Try simulations using other distributions such as

   a) A Uniform distribution

   b) A Binomial distribution

3. In #2 above, what effect do you observe when you change the sample size?