9. SAMPLING AND STATISTICAL INFERENCE

We often need to know something about a large population.

Eg: What is the average number of hours per day devoted to social media for all US residents?

It’s often infeasible to examine the entire population.

Instead, choose a small random sample and use the methods of statistical inference to draw conclusions about the population.

But how can any small sample be completely representative?
We can’t act as if statistics based on small samples are exactly representative of the entire population.

Why not just use the sample mean $\bar{x}$ in place of $\mu$?

For example, suppose that the average hours for 100 randomly-selected US residents was $\bar{x} = 9.34$. Can we conclude that the average hours for all US residents ($\mu$) is 9.34? Can we conclude that $\mu > 9$?

Fortunately, we can use probability theory to understand how the process of taking a random sample will blur the information in a population.

But first, we need to understand why and how the information is blurred.
Sampling Variability

Although the average social media hours per day for all US residents is a fixed number, the average of a sample of 100 residents depends on precisely which sample is taken. In other words, the sample mean is subject to “sampling variability”.

The problem is that by reporting $\bar{x}$ alone, we don’t take account of the variability caused by the sampling procedure. If we had polled different residents, we might have gotten a different average number of hours spent on social media per day.
In general, the characteristics of the observed distribution (mean, median, variance, range, IQR, etc.), change from sample to sample, and may never exactly match the population quantities.

To visualize properties of sampling distributions, we will use the following R demo (and some closely related ones).

[R Demo: CLT]
Statistical Inference: A body of techniques which use probability theory to help us to draw conclusions about a population on the basis of a random sample.

• Our conclusions will not always be correct. This problem is inevitable, unless we examine the entire population.

• We can, however, control the probability of making an error.

If we focus completely on what happened to us in our given sample, without putting it into the context of what might have happened, we can’t do statistical inference.

The success of statistical inference depends critically on our ability to understand sampling variability.
The Sampling Distribution of $\bar{X}$

Different samples lead to different values of $\bar{X}$.

But the sample was randomly selected!

Therefore, $\bar{X}$ is a random variable, taking different values depending on chance.

So $\bar{X}$ has its own distribution, called the sampling distribution.

[Sampling Lab Results]
The sampling lab results indicate that the sampling distribution of $\bar{X}$ is different from the distribution of the population. The sampling distribution has its own mean, variance, and shape, distinct from those of the population.

The sampling lab results show that the variance of $\bar{X}$ based on a sample of size 5 seems to be less than the variance of the population.

The average of the $\bar{x}$ values obtained seems quite close to the population mean.

We now give some precise definitions.
Random Sample

A random sample (of size $n$) from a finite population (of size $N$) is a sample chosen without replacement so that each of the $\binom{N}{n}$ possible samples is equally likely to be selected.

If the population is infinite, or, equivalently, if the sampling is done with replacement, a random sample consists of $n$ observations drawn independently, with replacement, from the population.

Hereafter, we assume that either the population is infinite, or else that $N$ is sufficiently large compared to $n$ that we can ignore the effects of having a finite population.
• Statistics (such as the sample mean $\bar{x}$) obtained from random samples can be thought of as random variables, and hence they have distributions, called **theoretical sampling distributions**.

• In order for our inferences to be valid, it is critical that we get a random sample, as defined above.

Suppose that a random sample, of size $n$, is taken from a population having mean $\mu$ and standard deviation $\sigma$.

Although $\mu$ and $\sigma$ are fixed numbers, their values are not known to us.
The Mean and Variance of $\bar{X}$

- Even though we will only take one sample in practice, we must remember that the sample was selected by a random mechanism.

- Therefore, $\bar{X}$ is a random variable! Its randomness is induced by the sampling procedure. If we had taken a different random sample, we might have gotten a different value for $\bar{x}$.

- Since $\bar{X}$ is a random variable, it must have a distribution. To draw valid inferences, we must take account of this sampling distribution, that is, we must think about all of the values that $\bar{x}$ might have taken (but didn’t).
Since all distributions have means and variances, the distribution of \( \bar{X} \) must also have a mean and a variance, denoted by \( \mu_{\bar{X}}, \sigma^2_{\bar{X}} \).

These quantities are given by the following simple expressions:

- \( \mu_{\bar{X}} = \mu \).
- \( \sigma^2_{\bar{X}} = \frac{\sigma^2}{n} \).

The formula \( \mu_{\bar{X}} = \mu \) shows that the sample mean is unbiased for the population mean. ("The mean of the mean is the mean").

Consider the histogram of our sample means from the sampling lab. The sample means seem to cluster around the population mean. The sample means average out to a value which is quite close to the mean of the population. If we had taken all possible samples, the corresponding sample means would average out to exactly \( \mu \).
The quantity $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ is called the standard error of the mean. It measures the extent to which sample means can be expected to fluctuate, or vary, due to chance.

- Note that $\sigma_{\bar{x}}$ increases with the population standard deviation, and decreases as the sample size increases.

For a given value of $\sigma$, the larger the sample size, the more tightly distributed $\bar{X}$ is around $\mu$ (and therefore the higher the probability that $\bar{X}$ will be close to $\mu$).
• For $n > 1$, the standard error of the mean is less than the standard deviation of the population. This demonstrates the benefits of averaging several observations instead of using just one.

Because of the square root in the denominator, we need to quadruple the sample size to double the reliability of $\bar{X}$.

In the sampling lab, the standard deviation of the sample means is (see Sampling Lab Results Handout).

This is noticeably smaller than the standard deviation of the population.
Eg 0: Suppose that the daily returns on Alphabet and Amazon are independent of each other and both have standard deviation .01. Then an equally weighted portfolio of the two stocks has a standard deviation of \( \frac{.01}{\sqrt{2}} = .007 \).

Diversification reduces risk.
The Central Limit Theorem

Although $\mu_{\bar{x}}$ and $\sigma_{\bar{x}}^2$ are easily calculated, the complete theoretical sampling distribution of $\bar{X}$ is often difficult to calculate. Fortunately, the **Central Limit Theorem** provides a useful approximation to this distribution.

**The Central Limit Theorem (CLT):** If $n$ is reasonably large, then the sampling distribution of $\bar{X}$ can be approximated closely with a normal distribution having mean $\mu_{\bar{x}}$ and variance $\sigma_{\bar{x}}^2$.

- The CLT says that if $n$ is not too small we can act as if $\bar{X}$ is normal, even if the population is not normal.
The CLT explains how $\bar{X}$ would behave under repeated sampling. Our understanding of this behavior allows us to draw conclusions about population means on the basis of sample means (statistical inference).

Without the CLT, inference would be much more difficult.

If the population is normal, then the sampling distribution of $\bar{X}$ is exactly $N(\mu, \sigma^2/\sqrt{n})$, for all $n$. This is the same distribution as given in the CLT, but here it holds exactly, not just approximately.

Note: Sampling lab shows that even when $n$ is as small as 5 and the population is highly skewed (non-normal), the sampling distribution of $\bar{X}$ is nearly normal, i.e., the CLT is a good approximation.
Eg 1: Ticketmaster is offering seats for next Monday’s concert at Madison Square Garden. Based on historical data, Ticketmaster knows that the number of tickets desired for such an event by a visitor to their website has a distribution with a mean of 2.4 and a standard deviation of 2.0.

Suppose there are currently 100 Ticketmaster visitors interested in purchasing tickets for next Monday’s concert. If only 250 tickets remain, what is the probability that all 100 people will be able to purchase the tickets they desire?
Eg 2: GMAT scores are normally distributed with a mean of 530 and a variance of 10,000.

A) What is the probability that the GMAT score of a randomly selected student falls between 500 and 640?

B) If a random sample of 10 students is taken, what is the probability that the sample mean GMAT score falls between 500 and 640?