Definition: **Parameter** = a number that characterizes a population (example: population mean \( \mu \)) – it’s typically **unknown**.

**Statistic** = a number that characterizes a sample

(example: sample mean \( \bar{X} \)) – we can calculate it from our sample data.

We use the sample mean \( \bar{X} \) to estimate the population mean \( \mu \).

Suppose we take a sample and calculate \( \bar{X} \).

\[ \text{\checkmark Probably not.} \]

Will \( \bar{X} \) equal \( \mu \)? Will \( \bar{X} \) be close to \( \mu \)? **We hope so.**

Suppose we take another sample and get another \( \bar{X} \).

Will it be same as first \( \bar{X} \)? Will it be close to first \( \bar{X} \)?

\[ \text{\checkmark Probably not.} \]

\[ \text{(Probably, depends on sample size)} \]

- What if we took **many repeated samples** (of the same size) from the same population, and each time, calculated the sample mean?

What would that set of \( \bar{X} \) values look like?

The **sampling distribution** of a statistic is the distribution of values of the statistic in all possible samples (of the same size) from the same population.
Consider the sampling distribution of the sample mean $\bar{X}$ when we take samples of size $n$ from a population with mean $\mu$ and variance $\sigma^2$.

Picture:

The sampling distribution of $\bar{X}$ has mean $\mu$ and standard deviation $\sigma/\sqrt{n}$.

Notation:

$\mu_{\bar{X}} = \mu \iff E(\bar{X}) = \mu$

$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$

**Point Estimator:** A statistic which is a single number meant to estimate a parameter.

It would be nice if the average value of the estimator (over repeated sampling) equaled the target parameter.

An estimator is called **unbiased** if the mean of its sampling distribution is equal to the parameter being estimated.
Examples: $E(\bar{X}) = \mu$, so $\bar{X}$ is an unbiased estimator of $\mu$.

$E(s^2) = \sigma^2$, so $s^2$ is an unbiased estimator of $\sigma^2$.

$E(s) \neq \sigma$, so $s$ is a biased estimator of $\sigma$.

Another nice property of an estimator: we want the spread of its sampling distribution to be as small as possible.

The standard deviation of a statistic’s sampling distribution is called the standard error of the statistic.

The standard error of the sample mean $\bar{X}$ is $\sigma/\sqrt{n}$.

Note: As the sample size gets larger, the spread of the sampling distribution gets smaller.

When the sample size is large, the sample mean varies less across samples. $\leftarrow$ good

Evaluating an estimator:
(1) Is it unbiased?
(2) Does it have a small standard error?
Central Limit Theorem

We have determined the center and the spread of the sampling distribution of $\bar{X}$. What is the shape of its sampling distribution?

Case I: If the distribution of the original data is normal, the sampling distribution of $\bar{X}$ is normal. (This is true no matter what the sample size is.)

Case II: Central Limit Theorem: If we take a random sample (of size $n$) from any population with mean $\mu$ and standard deviation $\sigma$, the sampling distribution of $\bar{X}$ is approximately normal, if the sample size is large.

$$\bar{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}}) \text{ if } n \text{ is large}$$

How large does $n$ have to be?
Our rule of thumb: If $n \geq 30$, we can apply the CLT result.

Pictures: See R code when sampling from an exponential distribution.

As $n$ gets larger, the closer the sampling distribution looks to a normal distribution.
Why is the CLT important? Because when $\bar{X}$ is (approximately) normally distributed, we can answer probability questions about the sample mean.

Standardizing values of $\bar{X}$:

If $\bar{X}$ is normal with mean $\mu$ and standard deviation $\sigma/\sqrt{n}$, then

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

has a standard normal distribution.

Example: Suppose we’re studying the failure time (at high stress) of a certain engine part. The failure times have a mean of 1.4 hours and a standard deviation of 0.9 hours.

If our sample size is 40 engine parts, then what is the sampling distribution of the sample mean?

\[ \text{CLT applies (since } n=40) \]

\[ \bar{X} \sim N(1.4, \frac{0.9}{\sqrt{40}}) \]

\[ \Rightarrow \bar{X} \sim N(1.4, 0.1423) \]
What is the probability that the sample mean will be greater than 1.5?

\[
P(\bar{X} > 1.5) \approx P(Z > 0.70)
\]

\[
= 0.242
\]

\[
\bar{X} = 1.5
\]

\[
\Rightarrow Z = \frac{1.5 - 1.4}{0.9/\sqrt{40}}
\]

\[
= \frac{1.5 - 1.4}{0.1423} = 0.70
\]

Example: Suppose lawyers' salaries have a mean of $90,000 and a standard deviation of $30,000 (highly skewed). Given a sample of lawyers, can we find the probability the sample mean is less than $100,000 if \( n = 5 \)? No If \( n = 30 \) Yes, we can use CLT.

\[
P(\bar{X} < 100,000) \approx P(Z < 1.83) = 0.9664
\]

\[
\bar{X} = 100,000 \Rightarrow Z = \frac{100,000 - 90,000}{30000/\sqrt{30}} = 1.83
\]
Other Sampling Distributions

In practice, the population standard deviation $\sigma$ is typically unknown.

We estimate $\sigma$ with $s$.

$$\frac{\bar{X} - \mu}{s / \sqrt{n}}$$

But the quantity $\frac{\bar{X} - \mu}{s / \sqrt{n}}$ no longer has a standard normal distribution.

Its sampling distribution is as follows:
• If the data come from a normal population, then the statistic $T = \frac{\bar{X} - \mu}{s / \sqrt{n}}$ has a t-distribution ("Student’s t") with $n - 1$ degrees of freedom (the parameter of the t-distribution).

• The t-distribution resembles the standard normal (symmetric, mound-shaped, centered at zero) but it is more spread out.
• The fewer the degrees of freedom, the more spread out the t-distribution is.
• As the d.f. increase, the t-distribution gets closer to the standard normal.

Picture:
Table \( \text{III} \) gives values of the t-distribution with specific areas to the right of these values:

Verify:
In \( t \)-distribution with 3 d.f., area to the right of \( 3.182 \) is \( .025 \). (Notation: For 3 d.f., \( t_{.025} = 3.182 \) )

In \( t \) with 14 d.f., area to the right of \( 1.761 \) is \( .05 \).

For 14 d.f., \( t_{.05} = 1.761 \)

In \( t \) with 25 d.f., area to the right of \( \_\_\_\_\_\_ \) is \( .999 \).
The $\chi^2$ (Chi-square) Distribution

Suppose our sample (of size $n$) comes from a normal population with mean $\mu$ and standard deviation $\sigma$.

Then $\frac{(n-1)s^2}{\sigma^2}$ has a $\chi^2$ distribution with $n - 1$ degrees of freedom.

- The $\chi^2$ distribution takes on positive values.
- It is skewed to the right.
- It is less skewed for higher degrees of freedom.
- The mean of a $\chi^2$ distribution with $n - 1$ degrees of freedom is $n - 1$ and the variance is $2(n - 1)$.

**Fact:** If we add the squares of $n$ independent standard normal r.v.'s, the resulting sum has a $\chi^2_n$ distribution.

Note that $\frac{(n-1)s^2}{\sigma^2} = \frac{n-1}{\sigma^2} \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n-1}$

$= \frac{1}{\sigma^2} \sum_{i=1}^{n} (X_i - \bar{X})^2 = \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{\sigma^2} = \sum_{i=1}^{n} \left( \frac{X_i - \bar{X}}{\sigma} \right)^2$

- If we had used $\mu$ rather than $\bar{X}$, this would be $n$ independent standard normal r.v.'s, squared and added up would have a $\chi^2_n$ distribution.

We sacrifice one d.f. by estimating $\mu$ with $\bar{X}$, so it is $\chi^2_{n-1}$. 
Table IV gives values of a $\chi^2$ r.v. with specific areas to the right of those values.

Examples:

For $\chi^2$ with 6 d.f., area to the right of $2.20^4$ is .90.

For $\chi^2$ with 6 d.f., area to the right of $12.59$ is .05.

For $\chi^2$ with 80 d.f., area to the right of $96.58$ is .10.
The F Distribution

\[ \frac{\chi^2_{n_1-1}}{n_1-1} / \left( \frac{\chi^2_{n_2-1}}{n_2-1} \right) \]

The quantity \( \frac{\chi^2_{n_1-1}}{n_1-1} / \left( \frac{\chi^2_{n_2-1}}{n_2-1} \right) \) where the two \( \chi^2 \) r.v.'s are independent, has an F-distribution with \( n_1 - 1 \) "numerator degrees of freedom" and \( n_2 - 1 \) denominator degrees of freedom.

So, if we have samples (of sizes \( n_1 \) and \( n_2 \)) from two normal populations, note:

\[ \frac{(n_1-1) S_1^2}{\sigma_1^2 (n_1-1)} \]

\[ = \]

\[ \frac{n_2-1) S_2^2}{\sigma_2^2 (n_2-1)} \]

\( \frac{S_1^2 / \sigma_1^2}{S_2^2 / \sigma_2^2} \)

has an F-distribution with \( (n_1 - 1, n_2 - 1) \) d.f.
Table V gives values of F r.v. with area .10 to the right.
Table VII gives values of F r.v. with area .05 to the right.
Table VIII gives values of F r.v. with area .025 to the right.
Table IX gives values of F r.v. with area .01 to the right.

Verify:

For F with (3, 9) d.f., 2.81 has area 0.10 to right.

For F with (15, 13) d.f., 3.82 has area 0.01 to right.

• These sampling distributions will be important in many inferential procedures we will learn.