• Recall from last lecture the example of finding an interval such that the mean cookie weight of the cookies in a randomly chosen bag of cookies has probability 0.95 of being in that interval.

⇒ We want to find $x_1$ and $x_2$ such that $P(x_1 < \bar{X} < x_2) = 0.95$.

Instead we can find $z$ such that $P(-z < Z < z) = 0.95$. 
From z-table we get $P(-1.96 < Z < 1.96) = 0.95$.

We use $\mu$ and $\sigma$ to convert this interval into one in terms of mean weight of the cookies in a bag of 32 cookies.

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

$\Rightarrow \bar{X} = Z\sigma + \mu$

$\Rightarrow P(-1.96\sigma + \mu < Z\sigma + \mu < 1.96\sigma + \mu)$

$= P(-1.96(0.088) + 11 < \bar{X} < 1.96(0.088) + 11)$

$= P(10.83 < \bar{X} < 11.17) = 0.95$

What if we don’t know the true population mean of a cookie, and want to learn it from the data (measurements of the sample)?

Inference: Our best guess (point estimate) of the population mean $\mu$ is the sample mean $\bar{x}$.

How good do we think this guess is? It depends on the data:
- How much data?
- How were they collected?
- How much variability in the data?

Example: Assume we know that the standard deviation of the weight of a cookie is 0.5g, but we don’t know the mean weight of a cookie. We get a bag (sample) of 32 cookies and find the average weight is 10.9g.

How confident are we in this estimate?

What would be an interval of plausible values?
→ Assuming cookie weights are approximately normally distributed (or using the CLT for the mean with $n \geq 30$), that $\sigma = 0.5$ is known and that we have a simple random sample, starting with a standard normal, **before** we observe $\bar{X}$ we know
\[
P(-1.96 \leq Z \leq 1.96) = 0.95
\]
\[
P(-1.96 \leq Z \leq 1.96) = P \left( -1.96 \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq 1.96 \right)
\]
\[
= P \left( -1.96 \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq 1.96 \right) = P \left( -1.96 \frac{\sigma}{\sqrt{n}} \leq \bar{X} - \mu \leq 1.96 \frac{\sigma}{\sqrt{n}} \right)
\]
\[
= P \left( 1.96 \frac{\sigma}{\sqrt{n}} \geq \mu - \bar{X} \geq -1.96 \frac{\sigma}{\sqrt{n}} \right) = P \left( -1.96 \frac{\sigma}{\sqrt{n}} \leq \mu - \bar{X} \leq 1.96 \frac{\sigma}{\sqrt{n}} \right)
\]
\[
= P \left( \bar{X} - 1.96 \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + 1.96 \frac{\sigma}{\sqrt{n}} \right) = 0.95
\]

** So, the **random** interval $\left( \bar{X} - 1.96 \frac{\sigma}{\sqrt{n}}, \bar{X} + 1.96 \frac{\sigma}{\sqrt{n}} \right)$ has a 95% probability of containing the true population mean, $\mu$!!

** 95% of the intervals constructed this way will contain the true population mean $\mu$!!

• An axiom (assumption) of Frequentist Statistics is that unknown parameters have a fixed, right answer. So once we observe $\bar{X}$, the interval is fixed, and the value of $\mu$ is fixed. So $\mu$ is either in the interval, or it isn’t, but we don’t know which.

★ The interval we get when we use the observed $\bar{X}$ is a called a **confidence interval**.

★ **Example:** For a sample bag of 32 cookies we have: $\bar{X} = 10.9$g and $\sigma = 0.5$g. Find the 95% CI for the mean weight of all cookies.

\[
\bar{X} - 1.96 \frac{\sigma}{\sqrt{n}} = 10.9 - 1.96 \frac{0.5}{\sqrt{32}} = 10.73
\]
\[
\bar{X} + 1.96 \frac{\sigma}{\sqrt{n}} = 10.9 + 1.96 \frac{0.5}{\sqrt{32}} = 11.07
\]

So $(10.73, 11.07)$ is a 95% CI for $\mu$.

★ **Technical Interpretation:** We are 95% confident that $\mu$ is in this interval.
• On average 95% of the intervals constructed this way will contain $\mu$, but we don’t know if this particular interval does contain it or not.

• Note that this is not a probability. The probability that $\mu$ is in this particular interval is 0 or 1, but we don’t know which.

• CI is an interval estimate for $\mu$. It provides a range of plausible values for $\mu$.

More general form:

$$(1 - \alpha)\text{CI} is \bar{x} \pm E, where E = z_{\alpha/2} \frac{\sigma}{\sqrt{n}} = \text{margin of error.}$$

$-z_{\alpha/2}$ is the $\frac{\alpha}{2}$ quantile of the standard normal distribution, i.e.

$$P(Z \leq -z_{\alpha/2}) = \frac{\alpha}{2}.$$
**Example:** Suppose a soda distributor is filling 20oz bottles and that from historical data, the sd of the contents of a bottle is known to be 0.03oz. Is the right amount of soda going into each bottle? Suppose a random sample of 34 bottles is found to have an average of 19.98oz. Find a 90% CI for the population mean contents.

\[ 1 - \alpha = 0.9 \Rightarrow \alpha = 0.1 \Rightarrow \frac{\alpha}{2} = 0.05 = P(Z < -1.645) \ (z\text{-table}) \]

So, \( z_{\alpha/2} = 1.645 \) and \( E = z_{\alpha/2} \frac{\sigma}{\sqrt{n}} = 1.645 \frac{0.03}{\sqrt{34}} = 0.0085 \)

\( \bar{x} \pm E = 19.98 \pm 0.0085 = (19.9715, 19.9885) \).

So (19.9715, 19.9885) is 90% CI for \( \mu \).

• Is it reasonable that 20oz are going to each bottle?

→ How do we determine the sample size needed for a desired margin of error?

... example continued: Suppose we want to be able to estimate soda contents with a margin of error of 0.001.

\[ 0.001 = E = z_{\alpha/2} \frac{\sigma}{\sqrt{n}} = 1.645 \frac{0.03}{\sqrt{n}} \]

\[ \Rightarrow \sqrt{n} = \frac{1.645 \cdot 0.03}{0.001} = 49.35 \Rightarrow n = (49.35)^2 = 2435.42 \]

* We need to round up, so that the margin of error is no larger than specified, so \( n = 2436 \).

**In general,**

\[ n = \left( \frac{z_{\alpha/2}\sigma}{E} \right)^2 \]

• Note: the sample size increases rapidly as the margin of error reduces.
Key Concepts!!!!!

- Confidence Interval