Chapter 8: STATISTICAL INTERVALS FOR A SINGLE SAMPLE

Part 1: Introduction

Confidence interval for μ
with variance σ^2 known
Sample size for CI

Sections 8-1 to 8-2

- In the last chapter we looked at point estimators for parameters of interest.
  For example, $\bar{X}$ is a point estimate for μ.

- We ended the last chapter with a phrase:
  "moving beyond point estimates"

- Point estimates are a good start, but if we provide the client with an estimate, we should also give them some idea of the confidence in our estimate.

- For instance, more data gives more information. we will have more confidence in an estimate from an $n = 50$ sample, than an estimate from an $n = 3$ sample.

We know $\bar{x}$ most likely won't exactly equal μ, but maybe we can provide an interval around $\bar{x}$ such that we're 95% confident that the interval contains μ.

The confidence in an estimate is related to the size (or width) of such an interval.

Confidence Interval on the mean μ from a normal distribution with a known population variance σ^2

Section 8-2

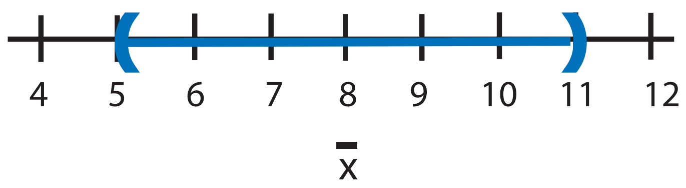
NOTE: The title of this section is very specific.

We're trying to estimate μ (so, μ is unknown) from a normally distributed pop'n with a known σ^2. Not terribly realistic, but this is where we start for learning the concept...

- To form an interval estimate, we consider our point estimate and give a 'cushion' around it for where the parameter is 'likely to fall'. For instance, if $\bar{X}$ is an estimator for μ, we would provide our client with an interval like:

$$[\bar{x} - cushion, \bar{x} + cushion]$$

Suppose $\bar{x} = 8$ and our cushion is 3



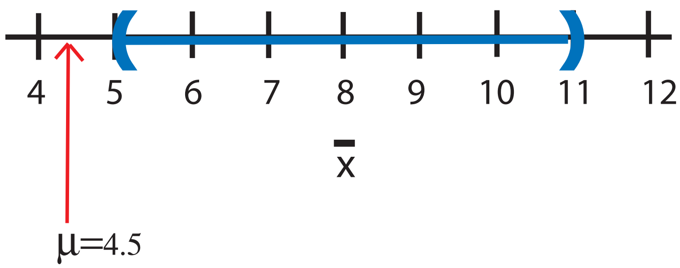
we would report the interval $[5, 11]$ as an interval estimate for μ and state that μ is likely to fall in this interval.

How do we calculate this "cushion"?
How 'likely' is it to fall in this interval?

- An interval estimate for a population parameter is called a <u>confidence interval</u>.

- We're not *certain* that the interval contains the unknown value of the parameter, but the methods we use to construct the interval will allow us to place a confidence level of parameter containment with our interval.

- We *could* miss the true parameter,



  but we try to protect against this by making the interval wide enough (but no wider than it needs to be for a certain level of containment confidence).

Computing the Confidence Interval
(i.e. getting a value for the $\pm$ cushion around $\bar{x}$)

- Consider a random sample $X_1, X_2, \ldots, X_n$ from a normal population with unknown mean μ and known variance σ^2.

- As before, we are still using $\bar{X}$ to estimate μ.

- We have already noted that

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

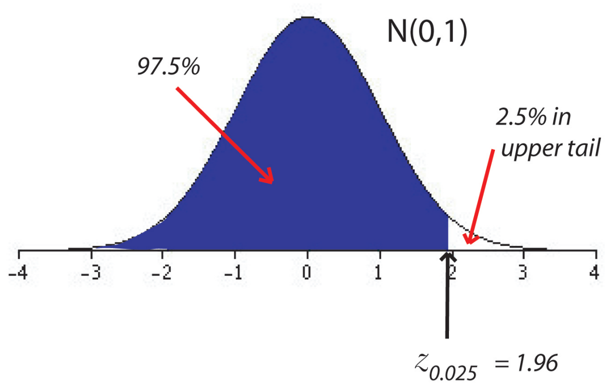
  has a standard normal, or $N(0, 1)$, distribution.

- Thus, we have

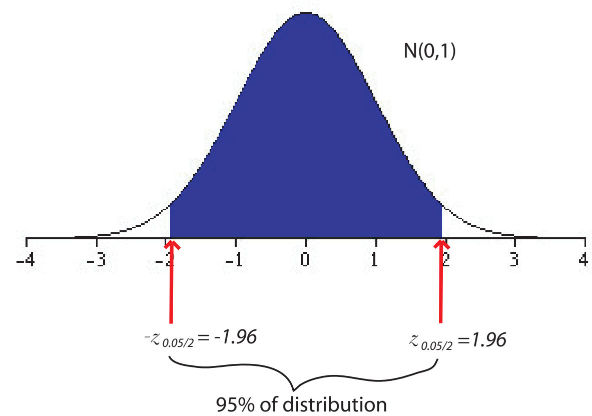
$$P(-z_{0.025} \leq Z \leq z_{0.025}) = 0.95$$
$$P(-z_{0.025} \leq \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \leq z_{0.025}) = 0.95$$

where $z_{0.025}$ is the 97.5^{th} percentile of the standard normal.

$z_{0.025}$ is the z-value such that 97.5% of the distribution is below and 2.5% is above it (an upper tail z-value).



- 95% of the standard normal falls between the values $(-z_{0.025}, z_{0.025})$:



- We can manipulate the following probability

$$P(-z_{0.025} \leq \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \leq z_{0.025}) = 0.95 \text{ to be}$$

$$P(\bar{X}-z_{0.025}\frac{\sigma}{\sqrt{n}} \leq \quad \mu \quad \leq \bar{X}+z_{0.025}\frac{\sigma}{\sqrt{n}}) = 0.95$$

NOW WE HAVE A 95% CONFIDENCE INTERVAL FOR μ !
(based on the known behavior of $\bar{X}$)

- We can state the lower and upper bounds of the 95% confidence interval as:

$$\text{Lower bound (L)} = \bar{X} - z_{0.025}\frac{\sigma}{\sqrt{n}}$$
$$\text{Upper bound (U)} = \bar{X} + z_{0.025}\frac{\sigma}{\sqrt{n}}$$

NOTE: $\bar{X}$ lies in the center of the confidence interval.

- **Example**: Fill weights of boxes.

The sample mean for the fill weights of 100 boxes is $\bar{x} = 12.05$. The population variance of the fill weights is known to be $(0.1)^2$. Find a 95% confidence interval for the population mean μ fill weight of the boxes.

ANS:

$$\begin{aligned} L &= \bar{x} - z_{0.025}\frac{\sigma}{\sqrt{n}} \\ &= 12.05 - 1.96\frac{0.1}{\sqrt{100}} = 12.0304. \end{aligned}$$

$$\begin{aligned} U &= \bar{x} + z_{0.025}\frac{\sigma}{\sqrt{n}} \\ &= 12.05 + 1.96\frac{0.1}{\sqrt{100}} = 12.0696. \end{aligned}$$

The 95% confidence interval for μ is

$$[12.0304,\ 12.0696]$$

We are 95% confident that the true parameter value lies in this interval.

Because σ^2 was very small and n was fairly large, we have a very narrow confidence interval for μ (which is good).

We can compute an interval with any confidence level

- The confidence level of choice is stated as $100(1-\alpha)\%$.
  - In the last example we computed a 95% or $100(1-0.05)\%$ confidence interval, and $\alpha = 0.05$.

- We can write the probability we saw earlier in general as:

$$P(-z_{\alpha/2} \leq \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}) = 1-\alpha$$

  this leads to the $100(1-\alpha)\%$ confidence interval for μ

$$P(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \quad \mu \quad \leq \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}) = 1-\alpha$$

- **100(1-α)% Confidence interval on the mean, variance known**

  If $\bar{x}$ is the sample mean of a random sample of size n from a normal population with known variance σ^2, a $100(1-\alpha)\%$ confidence interval for μ is given by

$$\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \quad \mu \quad \leq \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

  where $z_{\alpha/2}$ is the upper $100\alpha/2$ percentage point of the standard normal distribution (i.e. where there's $100\alpha/2$ in the upper tail).

- **Example**: Fill weights of boxes (same data as before).

The sample mean and for the fill weights of 100 boxes is $\bar{x}$=12.05. The population variance of the fill weights is known to be $(0.1)^2$. Find an **80**% confidence interval for the population mean μ fill weight of the boxes.

ANS: $\alpha = 0.20$

$$\begin{aligned} L &= \bar{x} - \boldsymbol{z_{0.10}} \frac{\sigma}{\sqrt{n}} \\ &= 12.05 - \mathbf{1.28} \frac{0.1}{\sqrt{100}} = \mathbf{12.0372}. \end{aligned}$$

$$\begin{aligned} U &= \bar{x} + \boldsymbol{z_{0.10}} \frac{\sigma}{\sqrt{n}} \\ &= 12.05 + \mathbf{1.28} \frac{0.1}{\sqrt{100}} = \mathbf{12.0628}. \end{aligned}$$

The 80% confidence interval for μ is
[12.0372, 12.0628]

We are 80% confident that the true parameter value lies in the interval [12.0372, 12.0628].

Compare these two confidence intervals:

The 80% confidence interval for μ is [12.0372, 12.0628]
(The width of this interval is 0.256)

The 95% confidence interval for μ is [12.0304, 12.0696]
(The width of this interval is 0.392)

All else being held constant, if you want to be more confident you capture μ, you'll have to make your *net* bigger. So the 95% CI is wider.

Looking at the form of the confidence interval:

$$\underbrace{\bar{x}}_{\substack{\text{Sample} \\ \text{mean}}} \pm \underbrace{z_{\alpha/2}}_{\substack{\text{multiplier} \\ \text{based on} \\ \text{confidence} \\ \text{level} \\ \uparrow \\ \text{changes} \\ \text{for different} \\ \text{\% CI}}} \cdot \underbrace{\frac{\sigma}{\sqrt{n}}}_{\substack{\text{value} \\ \text{based on} \\ \sigma \text{ and} \\ \text{sample size}}}$$

- More narrow CIs are desirable.

- How can this be achieved?

$$\underbrace{\bar{x}}_{\substack{\text{Sample} \\ \text{mean}}} \pm \underbrace{z_{\alpha/2}}_{\substack{\text{multiplier} \\ \text{based on} \\ \text{confidence} \\ \text{level}}} \cdot \underbrace{\frac{\sigma}{\sqrt{n}}}_{\substack{\text{value} \\ \text{based on} \\ \sigma \text{ and} \\ \text{sample size}}}$$

  - Increase your sample size (Good idea if possible)
  - Decrease σ? Not an option, it's fixed by original distribution
  - Decrease your confidence level? (Not a great idea. You reduce the CI width, but you're less likely to capture μ)

Interpreting the confidence interval

- Once the confidence interval is made (based on observed $\bar{x}$), it either <u>does</u> or <u>does not</u> contain the fixed unknown value μ

- For example, the 95% CI for box fill weights was:
$$[12.0304,\ 12.0696]$$
and the true pop'n mean either is or isn't in this interval

- The confidence interval level arises based on the <u>randomness</u> of the interval. BEFORE we collect the data, the CI is a <u>random interval</u> and it could take on many different values due to the *randomness* of $\bar{X}$.

- For a 95% CI, we are 95% confident that the true μ lies in the interval. This statement of confidence reflects the following...

  If we repeated this process 100 times (i.e. collect a sample, compute $\bar{x}$, compute the CI), 95 out of 100 times we will capture the true μ on average.

  The confidence relates to the method used to calculate the CI. We don't know if our CI captured μ or not (μ is unknown), but using the same method, 95 out of 100 times I'll get it (on average).

- See confidence interval applet website:
http://statweb.calpoly.edu/chance/applets/ConfSim/ConfSim.html

Choice of Sample Size

- The length of the CI is a measure of precision of estimation.
- We've seen that the precision is related to sample size n.
- You can get better precision with a large sample (all else being held constant).
- What sample size should you choose? (when you CAN choose)

Let E be the error in estimating μ

E=$|\bar{x} - \mu|$ (distance of observed $\bar{x}$ from target)

This distance, $|\bar{x} - \mu|$, is less than or equal to $(z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}})$ with confidence $100(1-\alpha)$.

$\{$ Based on the behavior of $\bar{X}$ $\}$

So choose n to provide a certain bound on the error with confidence $100(1-\alpha)$.

$$\bar{x} \pm \underbrace{z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}}_{\uparrow}$$

choose this CI half-width (or cushion) to be what you want, and this will be associated with a particular n

– Letting E be the chosen amount of error,

$$\text{E} = z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

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

– Choosing n equal to this value will give us a $100(1-\alpha)\%$ confidence that the error $|\bar{x} - \mu|$ will be less than or equal to E.

• **Example**: The fill weight example

In the fill weight example discussed earlier, how many boxes must be sampled to obtain a 99% confidence interval of full width ± 0.012 oz.?

ANS: $\sigma = 0.1$ from before

$$n = \left(\frac{z_{\alpha/2}\,\sigma}{E}\right)^2 = \left(\frac{z_{0.01/2}\;0.1}{0.012}\right)^2$$

$$= \left(\frac{2.58 \times 0.1}{0.012}\right)^2 = 462.25$$

We can't sample a fraction of a box, so we round-up to keep our confidence level, thus the required sample size is n=463.

NOTE: Be careful whether you're referring to the half-width of a CI (the cushion up or down), or the full width of the CI which is 2×cushion.

One-sided Confidence Bounds

- Occasionally, you may be interested in finding a bound for μ on only one side.

- If you want an upper-confidence bound, the lower bound is $-\infty$.

- If you want an lower-confidence bound, the upper bound is $+\infty$.

- In this case, when computing the CI, we don't split the α in two, because the α probability (or area) will remain in only one tail.

- **One-sided Confidence Bounds on μ, variance known**

A $100(1-\alpha)\%$ upper-confidence bound for μ is

$$\mu \leq \bar{x}+z_\alpha\sigma/\sqrt{n} \quad \Rightarrow \quad (-\infty,\ \bar{x}+z_\alpha\sigma/\sqrt{n}).$$

(This gives an upper bound on μ)

A $100(1-\alpha)\%$ lower-confidence bound for μ is

$$\bar{x}-z_\alpha\sigma/\sqrt{n} \leq \mu \quad \Rightarrow \quad (\bar{x}-z_\alpha\sigma/\sqrt{n},\ +\infty).$$

(This gives a lower bound on μ)

Confidence Interval on the mean μ from any population (not necessarily normal) with an unknown population variance σ^2 when the sample size is LARGE
Section 8-2.5

- As stated earlier, the previous confidence intervals were based a sample from an original population that was normally distributed with an unknown mean μ and a known σ^2.
- Since the original pop'n was normal, we didn't need a large sample size for the central limit theorem to kick-in to have $\bar{X}$ be normal.
- If n is large enough and σ^2 known, we can use the central limit theorem (CLT) to get our nice behavior for $\bar{X}$ with
$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$
and Z being approximately standard normal, or $N(0, 1)$, even if the original pop'n was not normal.

- But we also don't know σ^2, so we have to use our plug-in estimator of S^2.
- Now, if n is large enough, say $n \geq 60$ (my choice), or $n \geq 40$ (book choice), then we can overcome both...

  an unknown σ^2, and a
  non-normal original population

  and $\frac{\bar{X}-\mu}{S/\sqrt{n}}$ will still be approximately $N(0, 1)$.
- Getting $n \geq 30$ gets you the CLT result of $\bar{X}$ being normal, but you have to have an even higher n to use the previously stated CI methods because you had to estimate σ^2 with S^2 (so there's added uncertainty or variability in your value).