STAT 350 Lecture 6: Confidence Intervals

Interval Estimation of Unknown Parameter

(Chapter 9 of WMMY)

Outline

Recall

Theorem (CLT-Sampling Distribution)

If $X_1, \ldots, X_n$ are I.I.D. RVs with mean μ and variance $\sigma^2 < \infty$, then for large n the sample mean $\overline{X}_n$ is approximately Normal. In particular, for large n,

$$\overline{X}_n \sim \mathrm{N}(\mu, \sigma^2/n).$$

1. What is $P(|\overline{X}_n - \mu| \leq 1.96\sigma/\sqrt{n}) =?$

Three Important Interpretations

1. Probabilistic Interpretation: given n, μ and σ, for the random variable $\overline{X}_n$, we are 95% sure that it will be between $[\mu - 1.96\sigma/\sqrt{n}, \mu + 1.96\sigma/\sqrt{n}]$.
2. Frequentist Interpretation: Repeat collecting lots of $\overline{X}_n$s, there are 95% of them will be $[\mu - 1.96\sigma/\sqrt{n}, \mu + 1.96\sigma/\sqrt{n}]$.
3. Statistical Interpretation: if μ is unknown,

$$
\begin{aligned}
|\overline{X}_n - \mu| \le 1.96\frac{\sigma}{\sqrt{n}} \Leftrightarrow \mu - 1.96\frac{\sigma}{\sqrt{n}} \le \overline{X} \le \mu + 1.96\frac{\sigma}{\sqrt{n}} \\
\Leftrightarrow \overline{X}_n - 1.96\frac{\sigma}{\sqrt{n}} \le \mu \le \overline{X}_n + 1.96\frac{\sigma}{\sqrt{n}}.
\end{aligned}
$$

We're 95% sure that μ will be $[\overline{X}_n - 1.96\frac{\sigma}{\sqrt{n}}, \overline{X}_n + 1.96\frac{\sigma}{\sqrt{n}}]$. Formally, by repeating collection of $\overline{X}_n$s, among all of such intervals, 95% of them will contain the population parameter μ.

Outline

Confidence Interval (CI)

Goal: we want to have an interval to say the we are 95% sure the true population parameter is in there. This interval is then called 95% Confidence Interval.

Statistical Interpretation of CLT

We're 95% sure that μ will be $[\overline{X}_n - 1.96\frac{\sigma}{\sqrt{n}}, \overline{X}_n + 1.96\frac{\sigma}{\sqrt{n}}]$.

- We think the observed sample mean $\overline{X}_n$ should be a good point estimation for the population mean μ. $\Rightarrow$ **the center of the interval**
- The amount around the center, which is called **margin of error** ($ME = 1.96\frac{\sigma}{\sqrt{n}}$), should depend on the following two things:
  1. the variability of $\overline{X}_n$, which is called **standard error** ($SE = \frac{\sigma}{\sqrt{n}}$);
  2. the **confidence level** $1 - \alpha$=95%, which correspond to the multiplier $z^*_{\alpha/2} = 1.96$. α is called **significance level**

Specifics

- Suppose data $X_1, \ldots, X_n$ are available from a population with unknown mean μ but known std dev σ.
- Choose significance level $\alpha \in (0, 1)$; the quantity $1 - \alpha$ represents our *confidence level*.
- Roughly, $1 - \alpha$ measures how sure we want to be that our interval contains μ.
- The $100(1 - \alpha)\%$ CI for μ is given by

$$\overline{X}_n \pm \underbrace{z^*_{\alpha/2} \frac{\sigma}{\sqrt{n}}}_{\text{ME}},$$

where $z^*_{\alpha/2}$ is the $100(1 - \alpha/2)$ percentile of $\mathrm{N}(0, 1)$.

Relationship between α and $z^*_{\alpha/2}$

- For given α, you can find $z^*_{\alpha/2}$ by using command "invNorm" in the calculator.
- Here are the most common pairs $(1-\alpha, z^*_{\alpha/2})$:

$1-\alpha$	0.90	0.95	0.99
$z^*_{\alpha/2}$	1.645	1.960	2.576

- In the figure, $C = 1 - \alpha$.

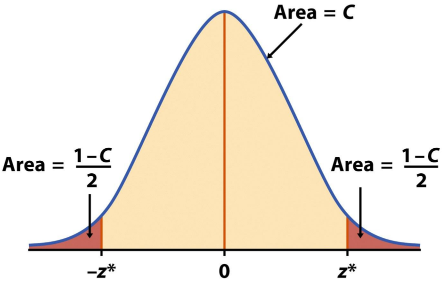


Figure 6-4
Introduction to the Practice of Statistics, Fifth Edition

Drinking example (a)

A questionnaire of drinking habits was given to a random sample of fraternity members, and each was asked to report the number of beers consumed in the past month. The sample of 30 students resulted in an average of 22 beers consumed. (Assume the population standard deviation is $\sigma = 9$ beers.) Give a 90% CI for the mean number of beers consumed by fraternity members in the past month.

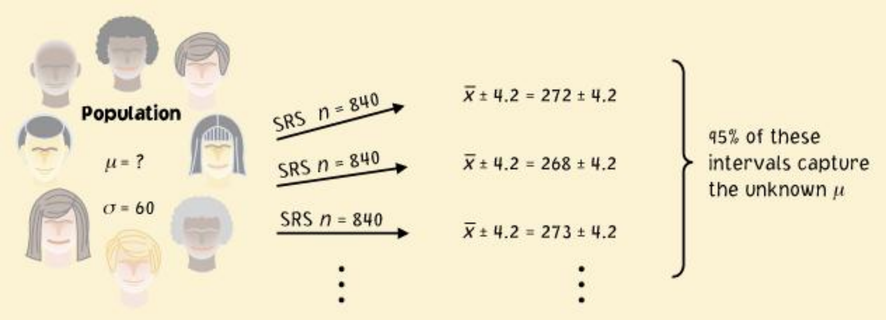
Revisit the Interpretation

95% CI: $\mu \in [2.5, 4.3]$

- **Very roughly speaking:** We're 95% sure that the true population parameter μ will be between 2.5 and 4.3.
- But what does the "95%" mean?
- **Formally:**

  If we take lots of samples and calculate CIs for each, then 95% of those CIs will contain μ.
- In other words, the confidence statement is about the *procedure/formula* and not about a particular interval.

Illustration



Drinking example (b)

Is it true that 90% of the fraternity members each month drink the number of beers that lie in the interval found in part (a) above?

> *No, this is the CI for the population mean, not for individual population members. If we take many 30-frat member samples and make a CI for each of these samples, then 90% of these CIs will contain the true mean number of beers consumed per month by a fraternity member.*

Accuracy of CI

- The accuracy of the CI depends on the length of the CI, the narrower the CI, the more accurate of the CI.
- The accuracy is controlled by $ME = z^*_{\alpha/2}\frac{\sigma}{\sqrt{n}}$.
- We see that three quantities play a role:
  - $z^*_{\alpha/2}$ which only depends on the confidence level $1-\alpha$
  - σ, the population standard deviation
  - n, the sample size.
- The standard deviation σ is fixed, so we focus on the other two.

Changing n and α

- Increasing the sample size n will decrease the ME and shrink the CI.

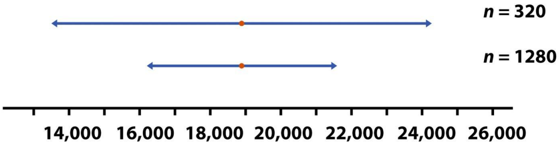


Figure 6-5
Introduction to the Practice of Statistics, Fifth Edition

- Decreasing the confidence level $1 - \alpha$ (or, equivalently, increasing α) will decrease the ME and shrink the CI.

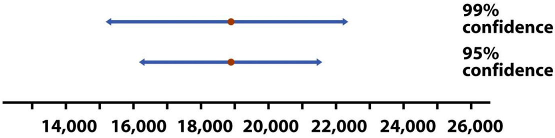


Figure 6-6
Introduction to the Practice of Statistics, Fifth Edition

Drinking example (c)

Find the MEs for 95% and 99% CIs and compare them to that of the 90% CI from part (a).

Sample size problems

- Often, for fixed α, there is a certain ME that's desired – the goal is to choose n large enough.
- For example, in election polls you see things like "Candidate A has 30% $\pm$ 2% of the popular vote." In these studies, the sample size n that was needed to achieve $\pm 2\%$ was determined ahead of time.
- Write out the formula $\text{ME} = z^*_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$.
- If we want to achieve the given ME, we can just solve for n:

$$n = \left(\frac{z^*_{\alpha/2}\sigma}{ME} \right)^2 .$$

- Typically, this calculation will produce a decimal number – round UP to get the necessary sample size.

Drinking example (d)

How many fraternity members would you need to sample if you want the ME of your 90% CI to be 1.0?

Outline

5 Different Cases

1. without normality, $n \geq 30$ and σ is given
2. without normality, $n \geq 30$ and s is given
3. X_i is Normal, σ is given
4. X_i is Normal, s is given
5. X_i is Bernoulli, $n \geq 30$

sampling distribution for the 5 cases:

1. $\frac{(\overline{X}_n - \mu)}{\sigma/\sqrt{n}} \sim \mathrm{N}(0,1) \Rightarrow \overline{X}_n \pm z^*_{\alpha/2}\sigma/\sqrt{n}$
2. $\frac{(\overline{X}_n - \mu)}{s/\sqrt{n}} \sim \mathrm{N}(0,1) \Rightarrow \overline{X}_n \pm z^*_{\alpha/2}s/\sqrt{n}$
3. $\frac{(\overline{X}_n - \mu)}{\sigma/\sqrt{n}} \sim \mathrm{N}(0,1) \Rightarrow \overline{X}_n \pm z^*_{\alpha/2}\sigma/\sqrt{n}$
4. $\frac{(\overline{X}_n - \mu)}{s/\sqrt{n}} \sim \mathrm{t}(n-1) \Rightarrow \overline{X}_n \pm t^*_{\alpha/2}s/\sqrt{n}$
5. $\frac{\hat{p} - p}{\sqrt{\hat{p}(1-\hat{p})/n}} \sim \mathrm{N}(0,1) \Rightarrow \hat{p} \pm z^*_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}$

Note $t^*_{\alpha/2}$, the $100(1-\alpha/2)$ percentile of a Student-t distribution with $n-1$ degrees of freedom. Sometimes it is written as $t^*_{\alpha/2}(n-1)$.

CIs for means – general

$$\boxed{\text{Normal}}\begin{cases}\boxed{\text{Yes}}\begin{cases}\boxed{\sigma\text{ given}} : \boxed{\text{ZInterval}}\\ \boxed{s\text{ given}} : \boxed{\text{TInterval}}\end{cases}\\ \boxed{\text{No}}\begin{cases}\boxed{\text{Bernoulli}}\begin{cases}\boxed{n\ge 30} : \boxed{\text{1PropZInterval}}\\ \boxed{n<30} : \boxed{\text{No method}}\end{cases}\\ \boxed{\text{Not Bernoulli}}\begin{cases}\boxed{n\ge 30} : \boxed{\text{ZInterval}}\ (\boxed{\text{no matter } s \text{ or } \sigma})\\ \boxed{n<30} : \boxed{\text{No method}}\end{cases}\end{cases}\end{cases}$$

Example 1

An agricultural expert performs a study to measure the yield of a tomato field which is assumed to be normally distributed with a standard deviation of 12.75 bushels. Studying 100 plots of land, she finds that the average yield is 34 bushels. Find a 95% CI for the unknown population mean tomato yield.

Example 2

How accurate are radon detectors sold to homeowners? To answer this question, researchers placed 12 detectors in a chamber that exposed them to 105 picocuries per liter. The detector readings are

91.9	97.8	111.4	122.3	105.4	95.0
103.8	99.6	119.3	104.8	101.7	96.6

(From software: $\overline{x} = 104.13$ and $s = 9.40$.) Assume the radon reading for this brand of detector is normally distributed. Construct a 90% CI for the true mean radon reading for this brand of detector.

Example 3

Information on a large packet of seeds indicates that the germination rate is $p_0=92\%$. Some researchers claim that the proportion of germination should be higher than that.

1. Based on the indication of the information, what's the probability that more than 95% of the 300 seeds in the packet will germinate?
2. Now a random sample containing 300 seeds with 290 germinated, is chosen to test this claim. Based on the sample, construct a 90% confidence interval for the population proportion of the seeds germinated in one packet. Please interpret the interval. If some researchers want to conduct a large-scale study, hoping to estimate the proportion to within 1% with 98% confidence, how many seeds must contain in the sample?

Example 3

solution:

1. Due to CLT,
   $\hat{p} \sim \mathrm{N}(p_0, p_0(1-p_0)/n) = \mathrm{N}(0.92, 0.92 \times 0.08/300)$. Hence

   $\mathrm{P}(\hat{p} > 0.95) = \mathrm{normalcdf}(.95, 9^9, .92, \sqrt{0.92 \times 0.08/300}) = 0.028.$

2. Now $\hat{p} = 290/300 = 0.967$.
   - 90% CI: $\hat{p} \pm z^*_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n} =$
     $0.967 \pm 1.64 \times \sqrt{\frac{0.967*0.033}{300}} = (0.950, 0.984)$.
   - We're 90% sure the germination rate of the seed packet is between 95% and 96.7%.
   - By $\mathrm{ME} \leq \mathrm{ME}_0 = 1\%$ with CL=98% (which corresponds to $z^*_{\alpha/2} = 2.33$), we have $\mathrm{ME} = z^*_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n} \leq \mathrm{ME}_0$ which leads to the formula for the sample size in the proportion case:

     $$n \geq \frac{(z^*_{\alpha/2})^2\hat{p}(1-\hat{p})}{\mathrm{ME}_0^2}.$$(If without initial guess for p, we use $\hat{p} = 0.5$.)

     Hence in this case: $n \geq \frac{2.33^2 \times 0.967 \times 0.033}{0.01^2} = 1732.4$, which leads to $n \geq 1733$.

Outline

Matched pairs example

A random sample of $n = 8$ golfers each hit both of the Brand A and Brand B golf balls, and the distances (in yards) was measured (see the table below). Assume that the Brand A distances and the Brand B distances are approximately Normal. Can the Brand A golf ball be hit further, on average, than the Brand B ball?

Golfer	Ball A dist	Ball B dist
1	265	252
2	272	276
3	246	243
4	260	246
5	274	275
6	263	246
7	255	244
8	258	245

Matched pairs example – cont.

- The two random variables X_1 (pre) and X_2 (post) are dependent and matched
- In this case, it makes sense to work on the *difference* $D = X_2 - X_1 = X^{(\text{post})} - X^{(\text{pre})}$ directly.
- We can then view the observed $d_1, \ldots, d_{10}$ as a random sample from a Normal population with unknown mean and variance.
- Now the question becomes a one-sample T problem, and we proceed just as before. Then the CI for mean of difference, μ_D, is:

$$\bar{d} \pm t^*_{\alpha/2}(n-1)\frac{s_d}{\sqrt{n}},$$

  where $\bar{d}$ is sample mean of difference, s_d is sample std. dev of difference, n is number of pairs.
- Solve the Matched Pairs Example

Another Golfer Example

Researchers want to know whether Brand A golf ball can be hit further, on average, than the Brand B ball. 16 golfers were recruited and each person was randomly assigned to Brand A or Brand B. Each golfer hit one ball taken from the assigned brand, and the distance (in yards) was measured (see table below). Assume that the Brand A distances and the Brand B distances are approximately Normal. Can the Brand A golf ball be hit further, on average, than the Brand B ball?

Golfer	Ball A dist	Golfer	Ball B dist
1	265	9	252
11	272	10	276
3	246	2	243
4	260	12	246
16	274	7	275
6	263	14	246
13	255	15	244
8	258	5	245

Independent samples

- Two independent populations in question: $\mathrm{N}(\mu_1, \sigma_1^2)$ and $\mathrm{N}(\mu_2, \sigma_2^2)$.
- Take samples $X_{11}, \ldots, X_{1n_1}$ and $X_{21}, \ldots, X_{2n_2}$ independently from the two normal populations.
- Note the possibility of samples of different sizes (n_1 and n_2).
- **Goal:** Build a CI for $\mu_1 - \mu_2$.
- For example:
  - Two populations: male and female salaries on Wall Street.
  - Do males get paid more than females on average?
- Distribution of $\overline{X}_1 - \overline{X}_2$?
  - σ_1 and σ_2 known
  - σ_1 and σ_2 unknown but equal

Details

There are two general cases that can arise, depending on what is known about the population standard deviations σ_1 and σ_2.

1. Case: σ_1 and σ_2 known. In this case, CIs can be made just like before using the fact that

$$\overline{X}_1 - \overline{X}_2 \sim \mathrm{N}(\mu_1 - \mu_2, \sigma_1^2/n_1 + \sigma_2^2/n_2).$$

   Then $(1-\alpha)100\%$ CI is:

$$\overline{X}_1 - \overline{X}_2 \pm z^*_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}.$$

2. Case: σ_1 and σ_2 unknown. We will work only under the assumption that $\sigma_1 = \sigma_2$, denoted by σ. And by replace the common σ with its estimation, say s_{pool} (pooled standard deviation), we have

$$\frac{[\overline{X}_1 - \overline{X}_2] - [\mu_1 - \mu_2]}{s_{pool}\sqrt{1/n_1 + 1/n_2}} \sim t(n_1 + n_2 - 2).$$

   Then $(1-\alpha)100\%$ CI is:

$$(\overline{X}_1 - \overline{X}_2) \pm t^*_{\alpha/2}(n_1 + n_2 - 2)\, s_{\text{pool}}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}.$$

Pooled standard deviation

- Since we're assuming that $\sigma_1 = \sigma_2$, it makes sense to somehow combine the two sample std devs s_1 and s_2.
- The intuition is that s_1 and s_2 are both estimating the same thing, so putting them together effectively increases the sample size, making the combined estimate better.
- The pooled estimate of the common variance σ^2 is

$$s_{\text{pool}}^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}.$$

- This is just a weighted average of s_1^2 and s_2^2 – if $n_1 = n_2$, then s_{pool}^2 is just the usual average of s_1^2 and s_2^2.
- The pooled standard deviation is $s_{\text{pool}} = \sqrt{s_{\text{pool}}^2}$.

Two-sample CI

Summary for 2 independent normal sample case:

- If σ_1 and σ_2 are known, we would use

$$(\overline{X}_1 - \overline{X}_2) \pm z^*_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}.$$

- If σ_1 and σ_2 are unknown but equal, with the pooled standard deviation we get

$$(\overline{X}_1 - \overline{X}_2) \pm t^*_{\alpha/2}(n_1 + n_2 - 2)\, s_{\text{pool}}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}.$$

- Note the degrees of freedom in the t distribution.

CIs for 2 sample means – general

$$\boxed{\text{Normal}} \begin{cases} \boxed{\text{Yes}} \begin{cases} \boxed{\sigma_1 \text{ and } \sigma_2 \text{ given}} : \boxed{\text{2-SampZInt}} \\ \boxed{\sigma_1 = \sigma_2 \text{ unknown}} : \boxed{\text{2-SampTInt}} \ (\boxed{\text{Pooled}}) \end{cases} \\ \boxed{\text{No}} \begin{cases} \boxed{n_1, n_2 \geq 30} : \boxed{\text{2-SampZInt}} \ (\boxed{s_1, s_2 \text{ or } \sigma_1, \sigma_2 \text{ given}}) \\ \boxed{n < 30} : \boxed{\text{No method}} \end{cases} \end{cases}$$

Example

There are two different techniques a manufacturer can use to produce batteries. Random samples of 12 and 14 batteries produced using Techniques 1 and 2, respectively, are taken and the capacities (in ampere hours) (normally distributed with common variance) are measured on each. A summary of the data is as follows:

$$\overline{x}_1 = 143.0, \quad s_1 = 7.11 \quad \text{and} \quad \overline{x}_2 = 135.8, \quad s_2 = 6.92.$$

Find a 90% CI for $\mu_1 - \mu_2$, the difference in mean capacities for the two manufacturing techniques.

Example

A certain change in a process for manufacturing component parts is being considered. Samples are taken under both the existing and the new process so as to determine if the new process results in an improvement. If 75 of 1500 items from the existing process are found to be defective and 80 of 2000 items from the new process are found to be defective, find a 90% confidence interval for the true difference in the proportion of defectives between the existing and the new process.