MATH 321: Mathematical Statistics

Lecture 6: Confidence Intervals

Chapter 7: Interval Estimation (7.1 - 7.4)

Introduction

Estimating parameters

- Point estimates
  - Using a point estimator $\hat{\theta}$ to estimate a parameter θ.
  - It is our single best guess.
  - Usually the point estimates do not equal the parameter because of sampling variability.
- Interval estimates
  - Give a range for what we think the population parameter is.
  - Takes into account sampling variability.

Constructing confidence intervals

Interval estimators / confidence intervals

- Definition: An **interval estimator** or **confidence interval** is a rule specifying the method for using the sample data to calculate two numbers that form the endpoints of the interval.

$$[L(\mathbf{X}), U(\mathbf{X})]$$

  Once $\mathbf{X} = \mathbf{x}$ is observed, the **interval estimate** is then $L(\mathbf{x})$ and $U(\mathbf{x})$ and the inference $L(\mathbf{x}) \leq \theta \leq U(\mathbf{x})$ is made.

- Ideally, the resulting interval will have two properties:
  1. It will contain the target parameter θ.
  2. It will be relatively narrow.

- Notes about properties:
  - The endpoints $L(\mathbf{X})$ and $U(\mathbf{X})$ (called the **lower and upper confidence limits**) of the interval are functions of the sample, which means they will vary randomly from sample to sample.

    Thus, the length and location of the interval are random quantities.

    ↳ $U(x) - L(x)$ → $\hat{\theta} = \frac{L(x) + U(x)}{2}$ (midpoint)
  - Because of this, we cannot be certain that the (fixed) target parameter θ will fall between the endpoints of any single interval calculated from a single sample.

    This being the case, our objective is to find an interval estimator capable of generating narrow intervals that have a high probability of enclosing θ.

- The probability that a (random) confidence interval will enclose θ (a fixed quantity) is called the **confidence coefficient**:

$$P\big(L(\mathbf{X}) \le \theta \le U(\mathbf{X})\big) = 1 - \alpha$$

  where α is called the **significance level**.

  Thus $[L(\mathbf{X}), U(\mathbf{X})]$ is called a **$100(1-\alpha)\%$ confidence interval for θ**.

Constructing confidence intervals

- All of the confidence intervals we will build start from this general setup and use properties of normal distributions or the central limit theorem to get the final interval of interest.
- Setup: Let $\hat{\theta}$ be an unbiased point estimator for parameter θ and $\sigma_{\hat{\theta}}$ be the standard deviation of the sampling distribution of $\hat{\theta}$ (this is often called the **standard error** of $\hat{\theta}$).

  Based on the scenario, if $\hat{\theta}$ is normally distributed, the quantity

$\hat{\theta} \overset{\text{approx}}{\sim} N(\theta, \sigma_{\hat{\theta}})$ $\Longrightarrow$ $Z = \dfrac{\hat{\theta} - \theta}{\sigma_{\hat{\theta}}} \sim \text{Normal}\,(0, 1)$

- Then to find a confidence interval for θ that possesses a confidence coefficient equal to $1 - \alpha$, we just need to select two values in the tails of this distribution, $-z_{\alpha/2}$ and $z_{\alpha/2}$ (these are called **critical values**), such that $P(-z_{\alpha/2} \le Z \le z_{\alpha/2}) = 1 - \alpha$. → Goal

  ☆ Then because we seek an interval estimator for θ, we just have to substitute in $\frac{\hat{\theta} - \theta}{\sigma_{\hat{\theta}}}$ for Z and rearrange to isolate θ in the middle.

  $P(a \le \theta \le b) = 1 - \alpha$

$$1-\alpha = P(-z_{\alpha/2} \le Z \le z_{\alpha/2})$$

$$= P\left(-z_{\alpha/2} \le \frac{\hat{\theta}-\theta}{\sigma_{\hat{\theta}}} \le z_{\alpha/2}\right)$$

$$= P(-z_{\alpha/2}\,\sigma_{\hat{\theta}} \le \hat{\theta}-\theta \le z_{\alpha/2}\,\sigma_{\hat{\theta}})$$

$$= P(-\hat{\theta} - z_{\alpha/2}\,\sigma_{\hat{\theta}} \le -\theta \le -\hat{\theta} + z_{\alpha/2}\,\sigma_{\hat{\theta}})$$

$$= P(\hat{\theta} - z_{\alpha/2}\,\sigma_{\hat{\theta}} \le \theta \le \hat{\theta} + z_{\alpha/2}\,\sigma_{\hat{\theta}})$$

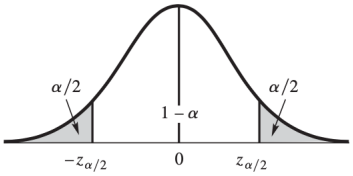
probability statement

$$\Rightarrow 100(1-\alpha)\%\ CI = \left[\hat{\theta} - z_{\alpha/2}\,\sigma_{\hat{\theta}},\ \hat{\theta} + z_{\alpha/2}\,\sigma_{\hat{\theta}}\right]$$

Corresponding interval

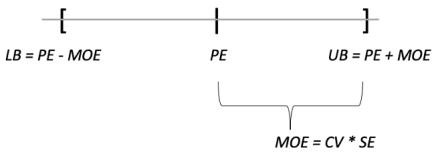
$$= \hat{\theta} \pm z_{\alpha/2}\,\sigma_{\hat{\theta}}$$

$$= PE \pm MOE$$



- Thus, we can summarize any (two-sided) confidence interval with

  CI = Point Estimate ± Margin of Error



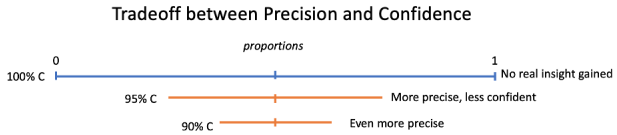
  - Point Estimate (PE) is the best guess; at the center of the interval.
  - Margin of Error (MOE) = Critical Value (CV) × Standard Error (SE).

    SE (standard deviation of the statistic) measures sampling error.

    % Confident is determined by confidence level set and incorporated via the critical value (CV).

- Recall the two goals of confidence intervals: (1) capture the parameter of interest and (2) be precise (smaller MOE = narrower interval).
  - The location (center) of the interval is determined by the data
  - The precision (MOE) is determined by the data (via the standard error) AND by the researcher (via the confidence level).

- All else equal, here is how the researcher can affect the precision of intervals:
  - Larger sample size $n \rightarrow$ smaller interval (smaller standard error)
  - More confident $\rightarrow$ larger interval (Larger critical value)



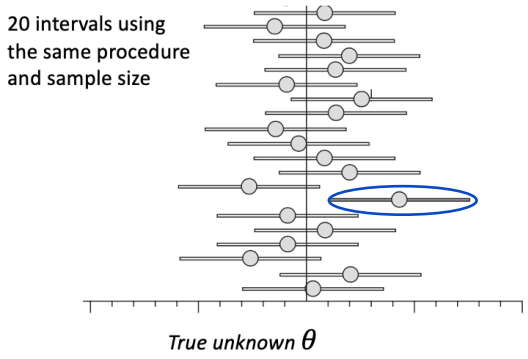
Interpreting confidence intervals

- Interpretation
  - General Structure

  I am % confident that the true/population parameter + context is between lower bound and upper bound.
  - Example: Suppose 95% CI = [24, 30]

    I am 95% confident that the true (population) mean of all Indiana ACT test scores is between 24 and 30.
  - When interpreting CIs: Make sure to mention what θ represents in context and keep in mind that the interval is giving us the range of plausible values for θ.

- Confidence coefficient
  - "95% confident": This tells us that in repeated sampling, approximately 95% of all intervals of the form $\hat{\theta} \pm 1.96\,\sigma_{\hat{\theta}}$ include θ.



$\frac{19}{20} = 0.95$ captured θ

How to use probability

  - Be careful with using "confidence" and "probability" interchangeably.

    AFTER collecting data: Misuse ✗

    ex) $P(24 \le \mu \le 30) = 1$ or 0

    ALL fixed numbers ⇒ T/F

    BEFORE collecting data: Correct use ✓

    $P(L(X) \le \mu \le U(X)) = 0.95$

    functions of sample ⇒ Probability ✓
  - For a particular sample, this interval either does or does not contain the parameter θ, but we never know.

How to use confidence

  - However, we are "95% confident" that the interval contains the parameter because the procedure that generated it yields intervals that do capture the true parameter in approximately 95% of the time that the procedure is used.

95% confident $\mu \in [24, 30]$ ✓

Confidence intervals for proportions

Introduction

- Often we want to estimate population proportions or the difference in proportions. For example
  - Proportion of voters in favor of an issue, proportion of students that graduate college, proportion of the population in a certain interval of values (success / fail perspective on a numeric variable), etc. — one sample
  - Difference in polling position for two candidates, difference in graduation rates for students involved in clubs vs not, etc. — two samples
- We can compute confidence intervals for one proportion or the difference in two proportions.

Confidence intervals for one proportion

$Y_i \overset{iid}{\sim} \text{Bernoulli}(p)$

- Setup: If we observe n independent Bernoulli trials, each with success probability p, then

  $X = \sum Y_i \implies X \sim \text{Binomial}(n, p)$

  Thus X represents the number of successes in the n trials.
- Now we are interested in the parameter $\theta = p$

  The unbiased estimator is the sample proportion $\hat{p} = \frac{X}{n} = \bar{Y}$
- Main result to form the interval (just need to meet conditions):

  Conditions: $np \geq 5$ + $n(1-p) \geq 5 \implies$ Expect at least 5 successes and 5 failures

Bernoulli

$E(X) = p$

$V(X) = p(1-p)$

1) Normal approximation to the binomial

$X = \sum Y_i \implies X \overset{\text{approx}}{\sim} N(np, np(1-p))$ by CLT

$\implies \frac{1}{n} X \overset{\text{approx}}{\sim} N\left(p, \frac{p(1-p)}{n}\right)$

$\hookrightarrow V(\frac{1}{n}X) = \frac{1}{n^2} V(X)$

2) CLT for Bernoulli mean

$\bar{Y} \overset{\text{approx}}{\sim} N\left(p, \frac{p(1-p)}{n}\right)$ by CLT

- Thus we can construct the approximate $100(1-\alpha)\%$ confidence interval for p with

$\frac{\hat{\theta} - \theta}{\sigma_{\hat{\theta}}} = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \approx z \overset{\text{approx}}{\sim} N(0,1)$

$\implies 1 - \alpha = P\left(\hat{p} - z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} \leq p \leq \hat{p} + z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}\right)$

$\implies \hat{p} \pm z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$

$\star \rightarrow$ substitute $\hat{p}$ for p in $\sigma_{\hat{p}}$ because unknown

Note that the unknown parameter p appears in both endpoints of the interval, so we do the obvious thing and substitute $\hat{p}$ into the standard error $\sigma_{\hat{p}}$.

- Example: Let p equal the proportion of triathletes who suffered an overuse injury during the past year. Out of 330 triathletes who responded to a survey, 167 indicated that they suffered such an injury during the past year.

  (a) Use these data to give a point estimate of p and to find an approximate 90% confidence interval for p.

Check conditions

$n\hat{p} = 330\left(\frac{167}{330}\right) = 167 \geq 5$ ✓ (↳ x)

$n(1-\hat{p}) = n - x = 163 \geq 5$ ✓

⇒ just check number of successes & failures in sample

→ 90% CI $\approx \hat{p} \pm z_{0.05}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \frac{167}{330} \pm 1.645\sqrt{\frac{0.506(0.494)}{330}}$

(↳ $\alpha = 0.10$; ↳ ≈ 0.506; ↳ $\text{Inv.Norm}(\text{area} = 0.05, \mu = 0, \sigma = 1)$)

$= [0.461,\ 0.551]$

→ We are 90% confident that the true proportion of triathletes who suffered an overuse injury is between 46.1% and 55.1%.

  (b) Do you think that the 330 triathletes who responded to the survey may be considered a random sample from the population of triathletes?

  This is an example of self selection bias (aka voluntary response bias). There is a whole branch of statistics related to surveys and how to collect data while minimizing bias in the sample.

Confidence intervals for difference of two proportions

- Setup: Same setup as for one proportion, now just two samples:

$Y_1 \sim \text{Bernoulli}(p_1) \longrightarrow X_1 = \sum Y_{1,i} \sim \text{Bin}(n_1, p_1)$

$Y_2 \sim \text{Bernoulli}(p_2) \longrightarrow X_2 = \sum Y_{2,i} \sim \text{Bin}(n_2, p_2)$

- Now we are interested in the parameter $\theta = p_1 - p_2$

  The unbiased estimator is difference in sample proportions $\hat{p}_1 - \hat{p}_2 = \frac{X_1}{n_1} - \frac{X_2}{n_2} = \bar{Y}_1 - \bar{Y}_2$

- Just need to check the conditions first before constructing the desired interval.

- Forming the interval → Conditions: $n_1 p_1 \geq 5$, $n_1(1-p_1) \geq 5$ + $n_2 p_2 \geq 5$, $n_2(1-p_2) \geq 5$

  Result:

$$\frac{X_1}{n_1} - \frac{X_2}{n_2} = \bar{Y}_1 - \bar{Y}_2 \overset{\text{approx}}{\sim} \text{Normal}\left(\mu = p_1 - p_2,\ \sigma^2 = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}\right)$$

- Thus we can construct the $100(1-\alpha)\%$ confidence interval for $p_1 - p_2$ with

$$\frac{\hat{\theta} - \theta}{\sigma_{\hat{\theta}}} = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \approx z \overset{\text{approx}}{\sim} \text{Normal}(0,1)$$

(same thing, just add MOE)

$$\Rightarrow 1-\alpha \approx P\left(\hat{p}_1 - \hat{p}_2 - z_{\alpha/2}\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} \leq p_1 - p_2 \leq \hat{p}_1 - \hat{p}_2 + z_{\alpha/2}\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}\right)$$

$$\Rightarrow \hat{p}_1 - \hat{p}_2 \pm z_{\alpha/2}\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

★ Substitute $\hat{p}_1$ & $\hat{p}_2$ for p_1 & p_2 in $\sigma_{\hat{p}_1 - \hat{p}_2}$

Again, we will estimate the standard error using the respective sample proportions.

- Example: Two detergents were independently tested for their ability to remove stains of a certain type. An inspector judged the first detergent to be successful on 83 out of 100 independent trials and the second one to be successful on 42 out of 79 independent trials.

  Find a 98% confidence interval for the difference in the probability in removing stains of the two detergents and state the conclusion.

check conditions

$x_1 = 83 \geq 5$ ✓
$n_1 - x_1 = 17 \geq 5$ ✓
$x_2 = 42 \geq 5$ ✓
$n_2 - x_2 = 37 \geq 5$ ✓

$\Rightarrow$ 98% CI $\approx \hat{p}_1 - \hat{p}_2 \pm z_{0.01} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$

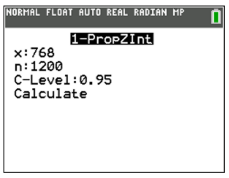
$= (0.83 - 0.53) \pm 2.326 \sqrt{\frac{0.83(0.17)}{100} + \frac{0.53(0.47)}{79}}$

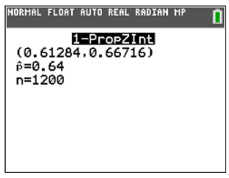
$= [0.141,\ 0.459]$

Calculator session (different example)

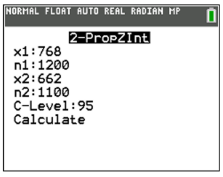
STAT > TESTS >

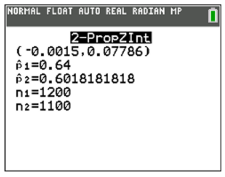
One proportion CI: p





Two proportion CI: $p_1 - p_2$





Confidence intervals for means

Confidence intervals for means

- Now we are interested in the parameter $\theta =$

  All we have to do is use the unbiased estimator for and the correct standard error $\sigma_{\hat{\theta}}$, then apply those to the final confidence interval shown at the beginning.

- Variables that affect the formation of our confidence intervals:
  - Sample size (large or small)
  - Population distribution X (normal or not normal)
  - Population variance σ^2 (known or unknown)
- We will simplify the scenarios and just think about large or small samples, and add notes about when the intervals are approximate or exact.

Large sample confidence intervals

- Suppose $X_1, \ldots, X_n$ are a random sample with **"large" n** from some distribution X with unknown variance σ^2.

$$\rightarrow \frac{\hat{\theta} - \theta}{\sigma_{\hat{\theta}}} = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \approx z$$

$\rightarrow \bar{x} \overset{\text{approx}}{\sim} N(\mu, \frac{\sigma^2}{n})$ $\overset{\text{approx}}{\sim}$ normal $(0,1)$

Substitute sample standard deviation $s = \sqrt{\frac{1}{n-1}\sum(x_i - \bar{x})^2}$ for σ when unknown

$$\rightarrow 1 - \alpha \approx P\left(\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

$$\Rightarrow \bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

- How large must n be / goodness of the approximation:

  Because of the unknowns (distribution and variance), we have approximately $100(1-\alpha)\%$ confidence intervals. As more assumptions are introduced, confidence coefficients for the intervals become more exact (i.e. closer to $100(1-\alpha)\%$ level).

  1. (Least best scenario) If X is badly skewed or has outliers, then prefer to have even larger sample sizes like $n \geq 50$, and even that may not produce good results.
  2. (Most likely in practice) If starting from Normal or like Normal (unimodal, symmetric, and continuous), then need $n \geq 30$ for the CLT to work with the unknown σ^2.
  3. (Best case scenario) If assume Normal and known variance σ^2, then this procedure even works for $n << 30$.

- Examples:

1. Example: Let X equal the life of a a 60-watt light bulb marketed by a certain manufacturer with $X \sim \text{Normal}(\mu, \sigma^2 = 1296)$. Suppose a random sample of size 27 from this distribution yields $\bar{x} = 1478$.

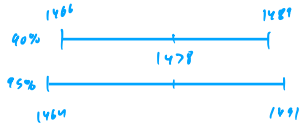
   Construct 90% and 95% confidence intervals for $E(X) = \mu$.

   $n = 27$, $X \sim$ normal, σ^2 ✓ ⇒ Exact z interval

   → 90% CI $= \bar{x} - z_{0.05}\frac{\sigma}{\sqrt{n}} = 1478 \pm 1.645\frac{\sqrt{1296}}{\sqrt{27}}$

   ↓

   $= [1466, 1489]$

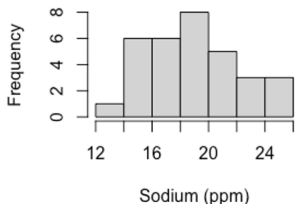
   → 95% CI $= \bar{x} - z_{0.025}\frac{\sigma}{\sqrt{n}} = 1478 \pm 1.96\frac{\sqrt{1296}}{\sqrt{27}}$

   ↓

   $= [1464, 1491]$

2. Lake Macatawa, an inlet lake on the east side of Lake Michigan, is divided into an east basin and a west basin. To measure the effect on the lake of salting city streets in the winter, students took 32 random samples of water from the west basin and measured the amount of sodium in parts per million in order to make a statistical inference about the unknown mean μ. They obtained the following data:

13.0	18.5	16.4	14.8	19.4	17.3	23.2	24.9
20.8	19.3	18.8	23.1	15.2	19.9	19.1	18.1
25.1	16.8	20.4	17.4	25.2	23.1	15.3	19.4
16.0	21.7	15.2	21.3	21.5	16.8	15.6	17.6



→ like Normal ⇒ good approximation

   Construct a 95% confidence interval for μ the mean amount of sodium in the west basin.

   $n = 32$, $X \sim$??, σ^2 ✗ ⇒ Approximate z interval

   → 95% CI $\approx \bar{x} \pm z_{0.025}\frac{s}{\sqrt{n}} = 19.07 \pm 1.96\left(\frac{3.26}{\sqrt{32}}\right)$ → calculate in Excel /R

   ↓

   $= [17.94, 20.20]$

   → Graphing calculator → 2-interval & use sample st dev as σ because of large sample

3. Example: Let X be the amount of orange juice (in grams per day) consumed by an American. Suppose $V(X) = \sigma^2 = 96$. To estimate μ, an orange growers' association took a random sample of $n = 576$ and found $\bar{x} = 133$.

   Construct a 98% confidence interval for μ.

   $n = 576$, $X \sim$??, σ^2 ✓ ⇒ Approximate z interval

   → 98% CI $\approx \bar{x} \pm z_{0.01}\frac{\sigma}{\sqrt{n}} = 133 \pm 2.33\frac{\sqrt{96}}{\sqrt{576}}$

   ↓

   $= [132.05, 133.95]$

Small sample confidence intervals

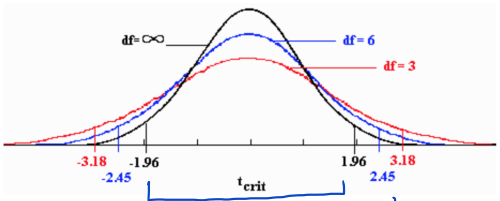
- Suppose $X_1, \ldots, X_n$ are a random sample with **"small" n** from **Normal**(μ, σ^2), with **unknown variance σ^2**.

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} = T \sim t_{n-1} \Rightarrow$$

$$1-\alpha = P\left(\bar{X} - t_{\alpha/2}\frac{S}{\sqrt{n}} \le \mu \le \bar{X} + t_{\alpha/2}\frac{S}{\sqrt{n}}\right)$$

$$\Rightarrow \bar{x} \pm t_{\alpha/2}\frac{s}{\sqrt{n}}$$

- Effect of converting to a t-interval
  - All else equal, t-intervals are wider than the corresponding Z-intervals because we are approximating σ with s (estimating another parameter in addition to μ).



  - However, the length of t-intervals are very much dependent on the value of the observed sample standard deviation s.

    If the observed s is smaller than σ, we can get a narrower using a t-interval compared to a Z-interval. But on average, $\bar{x} \pm z_{\alpha/2}\,(\sigma/\sqrt{n})$ is the shorter of the two confidence intervals.

  - When n gets larger ($n \ge 30$), then $t_{n-1} \approx Z$, which is why we can just use the Z critical values and the approximate Z-interval.

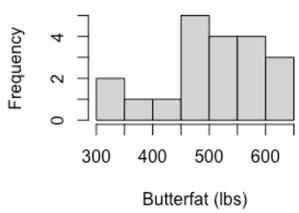
- What if data is not Normal?
  - Generally, this procedue works well when underlying distribution is symmetric, unimodal, and continuous and is still quite good (i.e. it is robust) for many non-normal distributions.
  - However it is not good (i.e. dangerous to use) if the distribution is highly skewed. If this is the case, safer to use certain nonparametric methods for finding a confidence interval for the median of the distribution (we will not cover this).

- Examples:

1. Let X equal the amount of butterfat in pounds produced by a typical cow during a 305-day milk production period between her first and second calves. Assume that the distribution of $X \sim \text{Normal}(\mu, \sigma^2)$. To estimate μ, a farmer measured the butterfat production for $n = 20$ cows and obtained the following data:

$n = 20$
$X \sim$ Normal
σ^2 X
$\Rightarrow$ Exact t interval

481	537	513	583	453	510	570	500	457	555
618	327	350	643	499	421	505	637	599	392



→ Assuming Normal, so don't need to check

Construct a 97% confidence interval for μ.

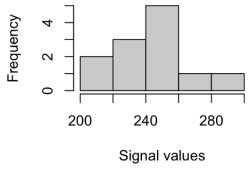
97% CI $= \bar{x} \pm t_{0.015,19} \frac{s}{\sqrt{n}} = 507.5 \pm 2.346 \left(\frac{89.75}{\sqrt{20}} \right) = [460.41, 554.58]$

↳ invT(area = 0.015, df = 19)

2. In nuclear physics, detectors are often used to measure the energy of a particle. To calibrate a detector, particles of known energy are directed into it. The values of signals from 12 different detectors, for the same energy, are shown below.

$n = 12$
$X \sim$??
σ^2 X
$\Rightarrow$ Approximate t interval

260	216	259	206	265	284
232	250	225	242	240	252

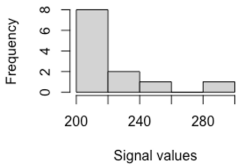


unimodal & not skewed
$\Rightarrow$ good enough

Construct a 95% confidence interval for μ.

95% CI $= \bar{x} \pm t_{0.025,11} \frac{s}{\sqrt{n}} = 244.25 \pm 2.20 \left(\frac{22.10}{\sqrt{12}} \right)$
$= [230.21, 258.29]$

3. Continuing example: Suppose the data looked like this (still $n = 12$):



Badly skewed $\Rightarrow$ don't use t-internal

Question

Suppose $n < 30$

$X \sim ?$, not skewed

σ^2 ✓

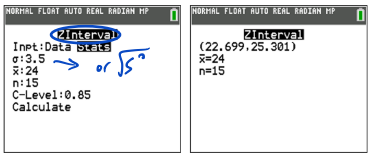
which interval?

$\Rightarrow$ Approximate Z-interval

Calculator session

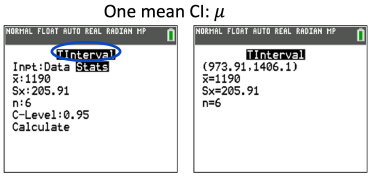
STAT > TESTS >

One mean CI: μ



large n

One mean CI: μ



Small n

One-sided confidence intervals

- Can also create one-sided confidence intervals if interested in the probability θ is larger or smaller than a certain number.
- By the same arguments as shown above, we can determine that **$100(1-\alpha)\%$ one-sided confidence interval** are given by

Lower bound $\Rightarrow$ "At least"

$P(\hat{\theta} - z_\alpha \sigma_{\hat{\theta}} \le \theta) = 1-\alpha$

$\Rightarrow [\hat{\theta} - z_\alpha \sigma_{\hat{\theta}}, \infty)$

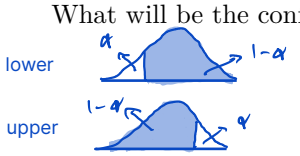
Upper bound $\Rightarrow$ "At most"

$P(\theta \le \hat{\theta} + z_\alpha \sigma_{\hat{\theta}}) = 1-\alpha$

$\Rightarrow (-\infty, \hat{\theta} + z_\alpha \sigma_{\hat{\theta}}]$

- Suppose that we compute both a $100(1-\alpha)\%$ lower bound and a $100(1-\alpha)\%$ upper bound for θ. We then decide to use both of these bounds to form a CI for θ.

  What will be the confidence coefficient of this interval?



Combined $\rightarrow$ $1-2\alpha = P(\hat{\theta} - z_\alpha \sigma_{\hat{\theta}} \le \theta \le \hat{\theta} + z_\alpha \sigma_{\hat{\theta}})$

$< 1-\alpha \Rightarrow$ lose confidence

- For one mean $\theta = \mu$, we will use the same criteria discussed above to determine Z vs t (knowing if exact vs approximate), then just use either the lower or upper bound interval if interested in a one-sided CI.
- Example: Using the Lake Macatawa data (unknown distribution and variance with $n = 32$), construct a 95% lower CI for μ and a 95% upper CI for μ.

  Then combine to form a two-sided interval and compare to the 95% two-sided interval found earlier.

Lower bound

$[\bar{x} - z_{0.05} \frac{s}{\sqrt{n}}, \infty)$ ($z_{0.05} = 1.645$)

$= [18.12, \infty)$

Upper bound

$(-\infty, \bar{x} + z_{0.05} \frac{s}{\sqrt{n}}]$

$= (0, 20.07]$

↓ Natural lower bound

Combined

$[18.12, 20.07] =$ 90 % CI two-sided

95 % Two-sided $= \bar{x} \pm z_{0.025} \frac{s}{\sqrt{n}} = [17.94, 20.20]$ ($z_{0.025} = 1.96$)

Confidence intervals for the difference of two means

Introduction

- Often we want to compare means of two different populations. For example, compare: average heights of male vs females for a species, average GPA of students in different school districts, mean response for two different treatments in an experiment, etc.
- We can compute confidence intervals for difference in means.

Confidence intervals for difference in means

- Now we are interested in the parameter $\theta =$ $\mu_1 - \mu_2$

  The unbiased estimator is $\bar{X}_1 - \bar{X}_2$

  Again, we just need to us the correct standard error $\sigma_{\bar{X}_1 - \bar{X}_2}$
- Variables that affect the formation of our confidence intervals:
  - Independent or dependent samples
  - Sample sizes n_1 and n_2 (large or small)
  - Population distributions X_1 and X_2 (normal or not normal)
  - Population variances σ_1^2 and σ_2^2 (known or unknown and ratio of variances)

  (Same as before)
- Similar logic (large vs small sample) can be used for two samples with regards to the form of the interval once we decide on independent vs dependent samples.

Independent, large sample confidence intervals

- Suppose we have **independent, large** random samples from some distributions X_1 and X_2 with sizes n_1 and n_2, respectively.

$$\rightarrow \frac{\hat{\theta} - \theta}{\sigma_{\hat{\theta}}} = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sigma_{\bar{X}_1 - \bar{X}_2}} \overset{\text{approx}}{\sim} Z \sim \text{normal}(0,1)$$

$$\hookrightarrow \bar{X}_1 - \bar{X}_2 \overset{\text{approx}}{\sim} \mathcal{N}\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

$$\rightarrow 1 - \alpha \approx P\left(\bar{X}_1 - \bar{X}_2 - z_{\alpha/2}\,\sigma_{\bar{X}_1 - \bar{X}_2} \le \mu_1 - \mu_2 \le \quad + \quad\right) \Longrightarrow \quad (\bar{X}_1 - \bar{X}_2) \pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

$$\hookrightarrow = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \quad \text{(If known)} \qquad \approx \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \quad \text{(If unknown)}$$

- Again, we need larger sample sizes with more unknowns in order to still have good approximations. Additionally, if we are starting from Normal or the variances are assumed known, then the approximations are better. (or exact)

- Examples:

1. A ecological study was conducted to compare rates of growth of trees at two sites by measuring leaf lengths of trees planted the previous year. It is known that the lengths of these leaves are normally distributed regardless of the conditions in which they grow and the variance of the lengths is $\sigma_1^2 = 1.69$ cm^2 at site 1 and $\sigma_1^2 = 2$ cm^2 at site 2. Two independent random samples of leaf lengths from the two sites are observed as below:

Independent samples

$n_1 = 8, n_2 = 7$

$X_1, X_2 \sim$ normal

σ_1^2, σ_2^2 ✓

⟹ Exact z-interval

	Site Leaf Length (cm)							
Site 1	5.18	1.48	1.82	2.35	3.04	5.49	1.03	4.04
Site 2	7.45	7.27	4.06	5.75	3.31	8.19	6.4	

$\bar{x}_1 = 3.054$, $\bar{x}_2 = 6.061$

Construct an 80% CI for $\mu_1 - \mu_2$ where μ_1, μ_2 are the mean leaf lengths of sites 1 and sites 2, respectively.

$\bar{x}_1 - \bar{x}_2 = -3.007$

$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{1.69}{8} + \frac{2}{7}} \approx 0.705$

80% CI $= \bar{x}_1 - \bar{x}_2 \pm z_{0.10}\, \sigma_{\bar{x}_1 - \bar{x}_2}$ ($z_{0.10} = 1.282$)

$= [-3.911, -2.104]$

– When interpreting confidence intervals for the difference in parameters, there are three scenarios for intervals:

Below zero	Contains zero	Above zero
$\mu_1 - \mu_2 < 0$	$\mu_1 - \mu_2 \overset{?}{=} 0$	$\mu_1 - \mu_2 > 0$
$\Rightarrow \mu_2 > \mu_1$	$\Rightarrow \mu_1 \overset{??}{=} \mu_2$	$\Rightarrow \mu_1 > \mu_2$

– For example, suppose interval is $[-1, 3]$. This contains zero and we would conclude there is no difference in θ_1 and θ_2; however, keep in mind that 3 is also a "believable" value for the difference.

2. A comparison of the durability of two types of automobile tires was obtained by road testing samples of $n_1 = n_2 = 100$ tires of each type. The number of miles until wear-out was recorded, where wear-out was defined as the number of miles until the amount of remaining tread reached a pre-specified small value. The measurements for the two types of tires were obtained independently, and the following means and variances were computed (in miles):

Independent samples

$n_1 = n_2 = 100$

$X_1, X_2 \sim ??$

σ_1^2, σ_2^2 ✗

⟹ Approximate Z-interval

$$\bar{x}_1 = 26,400, \quad s_1^2 = 1,440,000 \qquad \text{and} \qquad \bar{x}_2 = 25,100, \quad s_2^2 = 1,960,000$$

Construct a 90% CI for $\mu_1 - \mu_2$.

90% CI $\approx (\bar{x}_1 - \bar{x}_2) \pm z_{0.05}\, \sigma_{\bar{x}_1 - \bar{x}_2} = (26,400 - 25,100) \pm 1.645 \sqrt{\frac{1,440,000}{100} + \frac{1,960,000}{100}}$

$= [916.66, 1023.38]$

⟹ We are 90% confident that the true mean difference in number of miles until wear out for tire 1 and tire 2 is between 916.66 and 1023.38 miles.

OR

We are 90% confident that the true mean number of miles until wear out for tire 1 is between 916.66 and 1023.38 miles greater than that of tire 2.

Independent, small sample confidence intervals

- Suppose we have **independent, small** random samples from $X_1 \sim$ **Normal** (μ_1, σ_1^2) and $X_2 \sim$ **Normal** (μ_2, σ_2^2) with n_1 and n_2, and with **unknown common variance** $\boldsymbol{\sigma_1^2 = \sigma_2^2 = \sigma^2}$.

$$\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sigma_{\bar{X}_1 - \bar{X}_2}} = T \sim t_{n_1+n_2-2}$$

$$\Rightarrow 1-\alpha = P\left(\bar{x}_1 - \bar{x}_2 - t_{\alpha/2,\, n_1+n_2-2}\, s_p\sqrt{\tfrac{1}{n_1}+\tfrac{1}{n_2}} \le \mu_1 - \mu_2 \le \;+\right)$$

$$\Rightarrow \bar{x}_1 - \bar{x}_2 \pm t_{\alpha/2,\, n_1+n_2-2}\, s_p\sqrt{\tfrac{1}{n_1}+\tfrac{1}{n_2}}$$

$$\hookrightarrow s_p\sqrt{\tfrac{1}{n_1}+\tfrac{1}{n_2}}$$

$$\hookrightarrow \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}}$$

  - The standard error for this comes from the usual unbiased estimator of the common variance σ^2, which is obtained by pooling the sample data to obtain the pooled estimator S_p^2. This is just a weighted average of S_1^2 and S_2^2 with larger weight given to the sample variance associated with the larger sample size. $E(s_p^2) = \sigma^2$
  - Proof:

    From above, we know

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}}} \sim \text{Normal}\,(0,1)$$

$$\hookrightarrow = \sigma\sqrt{\tfrac{1}{n_1}+\tfrac{1}{n_2}}$$

    and from earlier theorems

$$\frac{(n_1-1)}{\sigma^2}S_1^2 \sim \chi^2_{\,n_1-1} \qquad \text{and} \qquad \frac{(n_2-1)}{\sigma^2}S_2^2 \sim \chi^2_{\,n_2-1}$$

    Thus,

$$U = \frac{(n_1-1)}{\sigma^2}S_1^2 + \frac{(n_2-1)}{\sigma^2}S_2^2 \sim \chi^2_{n_1+n_2-2} \qquad \left(= \sum \chi^2_r = \chi^2_{\sum r}\right)$$

    Combining all of this, we can form the following

$$T = \frac{Z}{\sqrt{U/(n_1+n_2-2)}} \sim t_{n_1+n_2-2}$$

$$= \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\left[(n_1-1)s_1^2 + (n_2-1)s_2^2\right]/(n_1+n_2-2)}} \cdot \sqrt{\tfrac{1}{n_1}+\tfrac{1}{n_2}} \quad *$$

$$= \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}}}\sqrt{\tfrac{1}{n_1}+\tfrac{1}{n_2}}$$

$$\hookrightarrow = s_p^2$$

need common σ^2

Interpretation → we are 95% confident that the true mean difference in standardized test scores in math for large high schools in between 3.65 points less than to 9.054 points above that of small high schools.

Independent samples

$n_1 = 9$, $n_2 = 15$

$X_1, X_2 \sim$ Normal

$\sigma_1^2 = \sigma_2^2 = \sigma^2$ ✗

⇒ Exact t-interval

– Example: Suppose that scores on a standardized test in mathematics taken by students from large and small high schools are $N(\mu_1, \sigma^2)$ and $N(\mu_2, \sigma^2)$, respectively, where σ^2 is unknown. If a random sample of $n_1 = 9$ students from large high schools and a random sample from $n_2 = 15$ small high school yielded

$$\bar{x}_1 = 81.31, \quad s_1^2 = 60.76 \qquad \text{and} \qquad \bar{x}_2 = 78.61, \quad s_2^2 = 48.24$$

Construct a 95% CI for $\mu_1 - \mu_2$.

$95\%\ CI = (\bar{x}_1 - \bar{x}_2) \pm t_{0.025, 22}\, s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$

$= (81.31 - 78.61) \pm 2.074\,(7.266) \sqrt{\frac{1}{9} + \frac{1}{15}} = [-3.65, 9.054]$

$= 2.7$

$\sqrt{\frac{(9-1)\,60.76 + (15-1)\,48.24}{9+15-2}}$

Demo

unweighted avg of variances → $\frac{s_1^2 + s_2^2}{2} = \frac{60.76 + 48.24}{2} = 54.5$

weighted avg → $s_p^2 = (7.266)^2 = 52.79$

⇒ "pulled" towards variance of X_2 which is smaller (but larger n)

- If we don't assume a common variance, we can still do the above procedure if the sample variances are close enough.

☆ As a rule of thumb, if the ratio of S_1^2/S_2^2 is between 0.5 and 2 (i.e., if one variance is no more than double the other), then we can use the pooled formula. → ex) $\frac{60.76}{48.24} \approx 1.26$ ✓

- Now we have all the assumptions from above scenario, (independent, small samples, both Normal with unknown variances), except we **cannot assume a common variance** AND they are drastically different. There are two cases

  1. We **do know the ratio of variances $\sigma_1^2/\sigma_2^2 = d$**.

     Can still construct a $100(1-\alpha)\%$ confidence interval for $\mu_1 - \mu_2$ using a modified s_p (will not cover this one).

  2. We **do not know the ratio of the variances** and yet suspect that the unknown σ_1^2 and σ_2^2 **differ by a a lot**.

     Can still construct a $100(1-\alpha)\%$ confidence interval for $\mu_1 - \mu_2$ using a similar interval to one with large samples and [redacted] unknown variances, except we use t-critical values (due to the unknown variances) with adjusted degrees of freedom to provide a larger MOE (will not cover this one either).

Dependent samples confidence intervals

- All of the previous intervals required independent samples as one of the assumptions. This is often not the case in practice.
- Independent vs dependent samples
  - Independent: Groups are unrelated, no connection, no relationship

    This is often not the case in practice, sometimes by design.
  - Dependent: Groups have some relationship between one another, can link the two; PAIRS

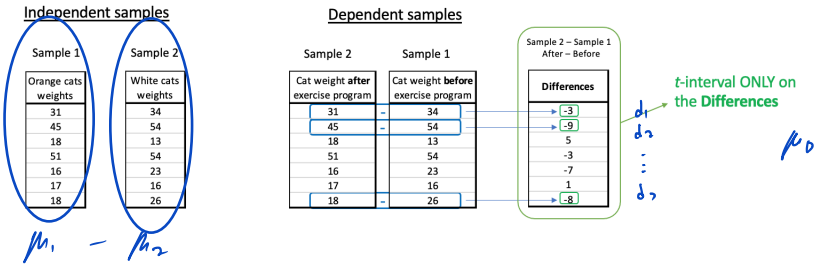
- If samples are dependent, they can be dependent in one of two ways.

  The interval that we construct is that same for both, but nonetheless it is important to know the structure of our data and how it was obtained.

  - **Paired**: Two values from the SAME subject.
  - **Matched**: Two values from DIFFERENT subjects connected in some way.
  - Examples) Determine if the following samples are independent or dependent (and matched or paired).

    1. Comparing the blood pressure of MATH 321 students before the final exam and after completing the final exam. — paired
    2. Are brothers or sisters smarter? A researcher studied ACT scores of 8 brother and sister pairs. — Matched
    3. A study is conducted to see what effect a new drug has on dexterity. A random sample of 30 students is chosen. They are given a series of tasks to perform and a score reflecting their performance. A dose of the drug is given to the 30 students and they again perform similar tasks and are scored again. — paired
    4. Looking to see if there is a difference in the price of the same Video Game Consoles at Target or Walmart. — Matched
    5. Seeing if the height of Faculty is shorter than the undergraduate population. — Independent

- Independent vs dependent strategy



- If X_1 and X_2 may be **dependent random variables**, then we cannot use the t-statistics and confidence intervals that we just developed, because they were based on the assumption of independence.

- **Matched-pair (dependent) t-interval for $\mu_1 - \mu_2$**

  Suppose we have **dependent** random samples from of size n from X_1 and X_2 (which can think of as ordered pairs $(X_{1,i}, X_{2,i})$).

  Let $D = X_1 - X_2$. This can be thought of as a random sample from $D \sim \text{Normal}(\mu_D, \sigma_D^2)$, where μ_D and σ_D^2 are the mean and variance of the difference in each pair.

  Then we can construct a $100(1-\alpha)\%$ for D in the same way as the as we did for one mean:

small sample → $\frac{\bar{D} - \mu_0}{\sigma_{\bar{D}}} = \frac{\bar{D} - \mu_D}{S_D/\sqrt{n}} = T \sim t_{n-1} \Rightarrow$

$1 - \alpha = P\left(\bar{d} - t_{\alpha/2, n-1}\frac{S_D}{\sqrt{n}} \le \mu_D \le \bar{d} + t_{\alpha/2, n-1}\frac{S_D}{\sqrt{n}}\right)$

$\Rightarrow \bar{d} \pm t_{\alpha/2, n-1}\frac{S_D}{\sqrt{n}}$

large sample → $\approx z$ $\approx N(0,1) \Rightarrow$ Z-interval

↳ can have unknown distribution

Dependent samples
$n = 8$
$D \sim$ normal
σ_D X
⟹ Exact t-interval

- Example: To compare the wearing of two types of automobile tires, A and B, a tire of type A and of type B are randomly mounted on the wheels of each of 8 automobiles. The automobiles are operated for a certain number of miles, and the amount of wear recorded for each tire below. Assuming the difference in wears of the tires are normally distributed, construct a 95% CI and a 95% lower-bounded CI for the difference in the mean wear of the two type of tires.

Car	1	2	3	4	5	6	7	8
Tire A	10.5	9.8	12.3	9.7	13.2	8.8	11	11.3
Tire B	10.4	9.6	12	9.3	12.8	8.3	10.4	10.6
$d = A-B$	0.1	0.2	...					0.7

95% CI $= \bar{d} \pm t_{0.025,7} \frac{s_d}{\sqrt{8}}$
$= 0.4 \pm 2.365\left(\frac{0.7}{\sqrt{8}}\right)$
$= [0.233, 0.567]$

< interpret same way as independent samples difference of two means >

95% LB CI $= \left[\bar{d} - t_{0.05,7} \frac{s_d}{\sqrt{8}}, \infty\right)$
$= \left[0.4 - 1.895 \frac{0.7}{\sqrt{8}}, \infty\right)$
$= [0.266, \infty)$

We are 95% confident that the true mean difference in wear of tires between brand A and brand B is at least 0.266

Finding the minimum sample size

Motivation

- In statistical consulting, the first question frequently asked is, "How large should the sample size be to estimate a mean?"

  Determining the sample size is an important step when planning a study because of the following considerations:

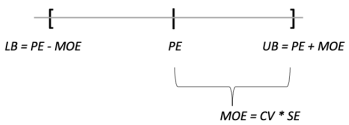
  - If n is too large, it is a waste of resources (studies are expensive, time and $$$).
  - If n is too small, they are less confident in the results (i.e. too imprecise); no real insight is gained.

- In the context of estimation, researchers want to figure out how large their sample needs to be to yield a confidence interval with a predetermined width.

  In doing so, they are controlling the precision!

Margin of error (MOE) revisited

- Recall: MOE is what you add and subtract from your point estimate to get your upper bound (UB) and lower bound (LB) of your confidence interval.



- If you are given an interval, your margin of error is the following:

$$MOE = \frac{UB - LB}{2} = \frac{Width}{2} \qquad \rightarrow \qquad Width = 2 \times MOE$$

- This is what we are controlling in the process of selecting the minimum sample size!

  For example, suppose a mathematics department wishes to evaluate a new method of teaching calculus with a computer. At the end of the course, the evaluation will be made on the basis of standard test, in which they would like to estimate μ, the mean score for students in the new class.

  In planning this course, they wish to determine how many students should take the course in order to be fairly confident that $\bar{x} \pm 1$ contains the unknown test mean μ.

- More formal definition: The **error in estimation ϵ** is the distance between an estimator and its target parameter. That is

  $\left[\hat{\theta} - \varepsilon, \hat{\theta} + \varepsilon\right] \Rightarrow |\hat{\theta} - \theta| = \varepsilon$

  Typically, we are given a **maximum error in estimation**, which means we want the margin of error to be no more than ϵ (or "within" ϵ), less than is okay.

$MoE \leq \varepsilon$

Finding minimum sample size

- The process for finding the minimum sample size for a given a maximum error in estimation ϵ is the same regardless of what type of interval we are using.

  Just start with the formula for Margin of Error and rearrange to solve for n.

$MOE = CV * SE$

$\downarrow = z_{\alpha/2}\, \sigma_{\hat{\theta}}$

- Here are the derivations / calculations for some of the different intervals that we have discussed. For each situation, we want the $100(1-\alpha)\%$ confidence interval for θ, $\hat{\theta} \pm z_{\alpha/2}\sigma_{\hat{\theta}}$, to be no longer than that given by $\hat{\theta} \pm \epsilon$.

- One proportion

$\varepsilon \geq z_{\alpha/2}\, \sigma_{\hat{p}}$

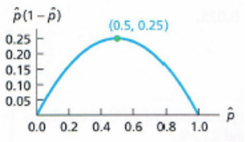
$\downarrow = z_{\alpha/2} \sqrt{\frac{p(1-p)}{n}}$

$\Rightarrow \quad n \geq \frac{z_{\alpha/2}^2\, p(1-p)}{\varepsilon^2}$

  - Again, we do not know the value p obviously and we cannot substitute $\hat{p}$ like before because we haven't collected data yet. So we have two options for specifying $p = p^*$:

    1. Set p^* based on previous research or experience.
    2. If no prior information is available, set $p^* = 0.5$.

       This results in the largest n for a specific MOE, so it is a safe (conservative) estimate. So to achieve a maximum error of estimate of at most ϵ, use the following:



$p^* = 0.5 \Rightarrow n \geq \frac{z_{\alpha/2}^2\,(0.5)(0.5)}{\varepsilon^2}$

$= \frac{z_{\alpha/2}^2}{4\varepsilon^2}$

$\Rightarrow$ all other values of p^* result in n less than that with $p^* = 0.5$

– Example: The unemployment rate in a certain country has been about 8%. This rate has changed by small amount and economists wish to update their estimate of the unemployment rate p in order to make decisions about national policy. Find the sample size needed to achieve a maximum error of the estimate of

1. $\epsilon = 0.001$ for a 95% CI for p

$$n \geq \frac{1.96^2 (0.08)(0.92)}{0.001^2} = 282{,}741.76 \approx 282{,}742$$

round up

2. $\epsilon = 0.01$ for a 99% CI for p

$$n \geq \frac{2.575^2 (0.08)(0.92)}{0.01^2} \approx 4881$$

3. $\epsilon = 0.01$ for a 95% CI for p

$$n \geq \frac{1.96^2 (0.08)(0.92)}{0.01^2} \approx 2828$$

much smaller with prior knowledge

4. $\epsilon = 0.01$ for a 95% CI for p, except assume now assume that we have no prior information about p.

$$n \geq \frac{1.96^2 (0.5)(0.5)}{0.01^2} \approx 9604$$

(worst case n)

- One mean

$$E \geq z_{\alpha/2}\, \sigma_{\bar{x}} \quad \downarrow = z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \quad \Rightarrow \quad n \geq \frac{z_{\alpha/2}^2\, \sigma^2}{E^2}$$

– Researchers have to make an assumption about the value of σ in order to do sample size calculations, which can be tricky. And resulting estimates for minimum sample sizes can change drastically based on how much variability is in the process they are studying. So there are a few options:

  * Assume a value for σ^2.
  * Use the best approximation available such as an estimate s obtained from a previous sample.
  * Use an upper bound on σ^2 if available.
  * Use knowledge of the range of the measurements in the population.

– Examples

1. (Continuing the math department example) Given past experience it is believed scores on such a common final are normally distributed with standard deviation of 15. Using $\bar{x}$ as an estimate, find the sample size needed to achieve a maximum error of the estimate of

   (a) $\epsilon = 1$ for a 95% CI for μ

   $$n \geq \frac{1.96^2 (15^2)}{1^2} \approx 865$$

   (b) $\epsilon = 2$ for a 95% CI for μ

   $$n \geq \frac{1.96^2 (15^2)}{2^2} \approx 217$$

   (c) $\epsilon = 2$ for a 90% CI for μ

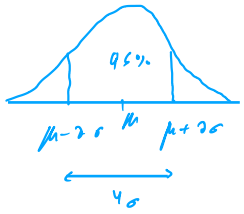
   $$n \geq \frac{1.645^2 (15^2)}{2^2} \approx 152$$

   (d) $\epsilon = 2$ for a 90% CI for μ, except with $\sigma = 22.5$ → 1.5 * original σ

   more than double with new σ

   $$n \geq \frac{1.645 (22.5^2)}{2^2} \approx 343$$

2. Continuing math example: Suppose test grades typically range between 60 and 95. Based on the empirical rule, 95% of data is between 2σ of the population mean μ. We can use this fact to get an approximate sample size, for say $\epsilon = 1$.



$4\sigma = 95 - 60$

$\sigma = \frac{35}{4} \approx 8.75$

$n \geq \frac{2^2 (8.75^2)}{1^2} \approx 307$

use # of $\pm\sigma$ to approximate $z_{\alpha/2}$

- Two means

  – If we make two simplifying assumptions, then we can get sample size estimates for this scenario as well (else it becomes like solving a system).

  Equal sample sizes: $n_1 = n_2 = n$ and equal variances $\sigma_1^2 = \sigma_2^2 = \sigma^2$.

$$\epsilon \geq z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

$$= z_{\alpha/2} \sqrt{\frac{\sigma^2}{n} + \frac{\sigma^2}{n}} \quad \text{< by assumption >}$$

$$= z_{\alpha/2} \sqrt{\frac{2\sigma^2}{n}}$$

$$= z_{\alpha/2} \frac{(\sqrt{2})\sigma}{\sqrt{n}} \implies n \geq \frac{z_{\alpha/2}^2 (2) \sigma^2}{\epsilon^2}$$

– Example: An experimenter wishes to compare the effectiveness of two methods of training. The selected participants are to be divided into two groups of equal size, the first receiving training method 1 and the second receiving training method 2. After training, each participant will a task and have their time recorded.

The goal is to estimate the mean difference in times within 1 minute with 95% confidence, assume $\sigma_1^2 = \sigma_2^2 = 2$.

$$n \geq \frac{1.96^2(2)(2)}{1^2} \approx 16$$

Note we could use the range strategy to get an estimate of common σ if we did not have the assumption it equaled 2.

- Observations: when estimating μ with $\bar{x}$, all else equal:

larger margin of error ϵ → smaller sample size n

more confident (smaller α) → larger sample size n

Larger variance σ^2 → larser sample size n

fixed width

MOE = CV * SE

increase decrease ⇒ ↑ n