STAT 583 Lecture 3

Linjun (Leon) Zhang

Department of Statistics
Rutgers University

Feb. 4th

Misc

- Midterm is scheduled at 6:50-8:50pm on March 3rd.

Last time

- Point Estimation
- Confidence intervals
- Confidence interval for a population mean (μ)

Recap

- When $X_1, ..., X_n$ is an *i.i.d.* sample from a population with known σ, and n is large,

$$[\bar{X} - 2\frac{\sigma}{\sqrt{n}}, \bar{X} + 2\frac{\sigma}{\sqrt{n}}]$$

is called an (approximate) 95% confidence interval (CI) for the population mean.

Recap

- When $X_1, ..., X_n$ is an $i.i.d.$ sample from a population with known σ, and n is large,

$$[\bar{X} - 2\frac{\sigma}{\sqrt{n}}, \bar{X} + 2\frac{\sigma}{\sqrt{n}}]$$

 is called an (approximate) 95% confidence interval (CI) for the population mean.

- For large n and unknown σ, we can safely substitute s for σ to obtain a 95% confidence interval

$$[\bar{X} - 2\frac{s}{\sqrt{n}}, \bar{X} + 2\frac{s}{\sqrt{n}}],$$

 where $s = \sqrt{\frac{(X_1-\hat{\mu})^2+...+(X_n-\hat{\mu})^2}{n-1}}$.

Recap

- When $X_1, ..., X_n$ is an $i.i.d.$ sample from a population with known σ, and n is large,

$$[\bar{X} - 2\frac{\sigma}{\sqrt{n}}, \bar{X} + 2\frac{\sigma}{\sqrt{n}}]$$

  is called an (approximate) 95% confidence interval (CI) for the population mean.
- For large n and unknown σ, we can safely substitute s for σ to obtain a 95% confidence interval

$$[\bar{X} - 2\frac{s}{\sqrt{n}}, \bar{X} + 2\frac{s}{\sqrt{n}}],$$

  where $s = \sqrt{\frac{(X_1-\hat{\mu})^2+\ldots+(X_n-\hat{\mu})^2}{n-1}}$.
- The 95% is a property of the procedure, not a specific interval. *In approximately 95% of all samples the confidence interval created according to this procedure will contain μ.*

Recap

A sample of 100 summer days measuring the number of ship passing near a power-plant location showed a mean of 7.2 ships per day with a sample variance of 8.8. Show a 95% CI for the true summer mean.

Recap

A sample of 100 summer days measuring the number of ship passing near a power-plant location showed a mean of 7.2 ships per day with a sample variance of 8.8. Show a 95% CI for the true summer mean.

Answer

- $[\bar{X} - 1.96\frac{s}{\sqrt{n}}, \bar{X} + 1.96\frac{s}{\sqrt{n}}]$

Recap

A sample of 100 summer days measuring the number of ship passing near a power-plant location showed a mean of 7.2 ships per day with a sample variance of 8.8. Show a 95% CI for the true summer mean.

Answer

- $[\bar{X} - 1.96\frac{s}{\sqrt{n}}, \bar{X} + 1.96\frac{s}{\sqrt{n}}]$
- $7.2 \pm 1.96 \cdot \sqrt{8.8}/\sqrt{100} = [6.62, 7.78]$.

Interpretation of the 95% Confidence Interval

- The 95% confidence interval for the summer mean is [6.62,7.78].
- What does it MEAN?
- Which of the following interpretations is correct?
  - A With 95% percent chance, the mean μ lies inside the interval [6.62,7.78].
  - B Given the observed data, with 95% percent chance, the mean μ lies inside the interval [6.62,7.78].
  - C 95% of the values of μ lie inside the interval [6.62,7.78].
  - D The interval [6.62,7.78] captures the true value μ 95% of the time.
  - E None of the above.

Interpretation of the 95% Confidence Interval

- The correct interpretation:

  "In an infinitely long series of trials in which repeated samples of size n are drawn from the same distribution and 95% CI's for μ are calculated using the same procedure, the proportion of intervals that actually include μ will be 95%. "

Interpretation of the 95% Confidence Interval

- The correct interpretation:
  "In an infinitely long series of trials in which repeated samples of size n are drawn from the same distribution and 95% CI's for μ are calculated using the same procedure, the proportion of intervals that actually include μ will be 95%. "
- The 95% confidence level is about the procedure, not about any particular interval obtained by applying the method to the observed data

Interpretation of the 95% Confidence Interval

- The correct interpretation:
  "In an infinitely long series of trials in which repeated samples of size n are drawn from the same distribution and 95% CI's for μ are calculated using the same procedure, the proportion of intervals that actually include μ will be 95%. "
- The 95% confidence level is about the procedure, not about any particular interval obtained by applying the method to the observed data
- For any particular CI obtained from the observed data, we do not know whether or not it contains μ

Overview

- More on general CIs: sample size determination
- One sample inference
  - CI for a binomial parameter (p) when n is large
  - CI for a population mean (μ) when σ is unknown and n is small
- Two-sample inference
  - CI for the difference between two binomial parameters $p_1 - p_2$
  - CI for the difference between two means $\mu_X - \mu_Y$
    - Large sample size
    - Small sample size
    - Paired samples

Sample size determination

- How large a sample size do I need?

Sample size determination

- How large a sample size do I need?
- Think about three things:

Sample size determination

- How large a sample size do I need?
- Think about three things:
    1. What confidence level you want (say 95%).

Sample size determination

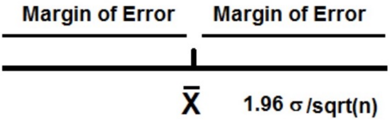
- How large a sample size do I need?
- Think about three things:
    1. What confidence level you want (say 95%).
    2. What Margin of Error (MoE) you want. ($2\sigma/\sqrt{n}$)

Sample size determination

- How large a sample size do I need?
- Think about three things:
    1. What confidence level you want (say 95%).
    2. What Margin of Error (MoE) you want. $(2\sigma/\sqrt{n})$
    3. If σ is unknown, then you need an estimate for σ.

Sample size determination

- How large a sample size do I need?
- Think about three things:
    1. What confidence level you want (say 95%).
    2. What Margin of Error (MoE) you want. $(2\sigma/\sqrt{n})$
    3. If σ is unknown, then you need an estimate for σ.
- Recall that the MoE is the distance from the center to the edge of the interval.



The sample size formula for a population mean, .

- As the $MoE = 1.96\sigma/\sqrt{n}$ in the 95% confidence interval for μ, it follows that:

$$n = (\frac{1.96\sigma}{MoE})^2 \approx (\frac{2\sigma}{MoE})^2 \approx (\frac{2s}{MoE})^2.$$

- There are different formulas for different statistical questions (not just the mean), but sample size is always a legitimate question.

Example

- An insurance company is being sued because it has paid bills late and then failed to pay interest on the late payments.
- Lawyers need to estimate the average amount of unpaid interest on the late bills.
- The population of bills is 300,000, far too many to review individually.
- How large a sample size do I need to make a 95% confidence interval for the mean amount of unpaid interest on bills paid late?
- We decide on a MoE of $\pm\$2.00$ and from a previous study we estimate σ by $s = \$25.25$.

Example

- An insurance company is being sued because it has paid bills late and then failed to pay interest on the late payments.
- Lawyers need to estimate the average amount of unpaid interest on the late bills.
- The population of bills is 300,000, far too many to review individually.
- How large a sample size do I need to make a 95% confidence interval for the mean amount of unpaid interest on bills paid late?
- We decide on a MoE of $\pm\$2.00$ and from a previous study we estimate σ by $s = \$25.25$.
- Answer:

$$n = \left(\frac{2 \times 25.25}{2}\right)^2 = 638.$$

Confidence Intervals for a binomial parameter

Suppose we have a random variable $Y \sim Bin(n, p)$, and we are interested in the CI for p. Let X_i be the binary outcome and p the sucess probability.

- Recall that $Y = \sum_{i=1}^{n} X_i$, where $X_1, ..., X_n$ are i.i.d. Bernoulli distribution with $\mathbb{P}(X_i = 1) = p, \mathbb{E}(X_i) = p, Var(X_i) = p(1-p)$.

Confidence Intervals for a binomial parameter

Suppose we have a random variable $Y \sim Bin(n, p)$, and we are interested in the CI for p. Let X_i be the binary outcome and p the sucess probability.

- Recall that $Y = \sum_{i=1}^{n} X_i$, where $X_1, ..., X_n$ are i.i.d. Bernoulli distribution with $\mathbb{P}(X_i = 1) = p, \mathbb{E}(X_i) = p, Var(X_i) = p(1-p)$.
- We estimate p by $\hat{p} = \frac{Y}{n}$.

Confidence Intervals for a binomial parameter

Suppose we have a random variable $Y \sim Bin(n, p)$, and we are interested in the CI for p. Let X_i be the binary outcome and p the sucess probability.

- Recall that $Y = \sum_{i=1}^{n} X_i$, where $X_1, ..., X_n$ are i.i.d. Bernoulli distribution with $\mathbb{P}(X_i = 1) = p, \mathbb{E}(X_i) = p, Var(X_i) = p(1-p)$.
- We estimate p by $\hat{p} = \frac{Y}{n}$.
- When $n > 30$ and both $n\hat{p}, n(1-\hat{p})$ are larger than 10, the sampling distribution of $\hat{p}$ is approximately:

$$\hat{p} \sim N(p, \frac{p(1-p)}{n}).$$

Confidence Intervals for a binomial parameter

- An approximate 95% CI for the binomial parameter is given by

$$[\hat{p} - 2sd(\hat{p}), \hat{p} + 2sd(\hat{p})] = [\hat{p} - 2\sqrt{\frac{p(1-p)}{n}}, \hat{p} + 2\sqrt{\frac{p(1-p)}{n}}]$$

Confidence Intervals for a binomial parameter

- An approximate 95% CI for the binomial parameter is given by

$$[\hat{p} - 2sd(\hat{p}), \hat{p} + 2sd(\hat{p})] = [\hat{p} - 2\sqrt{\frac{p(1-p)}{n}}, \hat{p} + 2\sqrt{\frac{p(1-p)}{n}}]$$

- Unfortunately we do not know p so we replace it with $\hat{p}$ in the standard deviation calculation (just like replacing σ with s).

Confidence Intervals for a binomial parameter

- An approximate 95% CI for the binomial parameter is given by

$$[\hat{p} - 2sd(\hat{p}), \hat{p} + 2sd(\hat{p})] = [\hat{p} - 2\sqrt{\frac{p(1-p)}{n}}, \hat{p} + 2\sqrt{\frac{p(1-p)}{n}}]$$

- Unfortunately we do not know p so we replace it with $\hat{p}$ in the standard deviation calculation (just like replacing σ with s).
- So the 95% confidence interval for p is

$$[\hat{p} - 2\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + 2\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}].$$

Confidence Intervals for a binomial parameter

- An approximate 95% CI for the binomial parameter is given by

$$[\hat{p} - 2sd(\hat{p}), \hat{p} + 2sd(\hat{p})] = [\hat{p} - 2\sqrt{\frac{p(1-p)}{n}}, \hat{p} + 2\sqrt{\frac{p(1-p)}{n}}]$$

- Unfortunately we do not know p so we replace it with $\hat{p}$ in the standard deviation calculation (just like replacing σ with s).
- So the 95% confidence interval for p is

$$[\hat{p} - 2\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + 2\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}].$$

- In general, the $(1-\alpha)$ confidence interval for p is

$$[\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}].$$

Example

- A pharmaceutical company needs to compare the performance of its clinical trials to an industry benchmark.
- One way of measuring this is to look at the proportion of trials that move from Phase I to Phase II. The *attrition rate* measures the proportion of trails that fail to make it to Phase II.
- The industry benchmark is 45%.
- The company conducted $n = 48$ trails, and 30 of them failed.
- Provide a 95% CI for the attrition rate and comment on whether there appears to be a concern with this particular companys trials.

The attrition data

- There were $n = 48$ trails.
- Of these, 30 failed, so $\hat{p} = 0.625$.
- $n = 48 > 30, n\hat{p}, n(1 - \hat{p}) > 10$.

The attrition data

- There were $n = 48$ trails.
- Of these, 30 failed, so $\hat{p} = 0.625$.
- $n = 48 > 30, n\hat{p}, n(1 - \hat{p}) > 10$.
- The approximate 95% CI is given by

$$0.625 \pm 2\sqrt{\frac{0.625 \times (1 - 0.625)}{48}} = 0.625 \pm 2 \times 0.07 = (0.485, 0.765).$$

The attrition data

- There were $n = 48$ trails.
- Of these, 30 failed, so $\hat{p} = 0.625$.
- $n = 48 > 30, n\hat{p}, n(1 - \hat{p}) > 10$.
- The approximate 95% CI is given by

$$0.625 \pm 2\sqrt{\frac{0.625 \times (1 - 0.625)}{48}} = 0.625 \pm 2 \times 0.07 = (0.485, 0.765).$$

- Reporting the interval on a percentage basis gives (48.5%, 76.5%).
- Notice that 45% is not in this interval.
- There is clear evidence that this company is not keeping up with the industry benchmark. Its attrition rate is significantly higher than the benchmark.

Manipulating confidence intervals

- In many medical settings clinicians like to report results in terms of odds ratios rather than probabilities.

Manipulating confidence intervals

- In many medical settings clinicians like to report results in terms of odds ratios rather than probabilities.
- If the probability is given by p then the odds ratio is $p/(1-p)$. A study has found the 95% confidence interval for p to be (0.485, 0.765).

Manipulating confidence intervals

- In many medical settings clinicians like to report results in terms of odds ratios rather than probabilities.
- If the probability is given by p then the odds ratio is $p/(1-p)$. A study has found the 95% confidence interval for p to be (0.485, 0.765).
- Find the 95% CI for the odds ratio.

Manipulating confidence intervals

- In many medical settings clinicians like to report results in terms of odds ratios rather than probabilities.
- If the probability is given by p then the odds ratio is $p/(1-p)$. A study has found the 95% confidence interval for p to be (0.485, 0.765).
- Find the 95% CI for the odds ratio.
- Answer: $(0.485/0.515, 0.765/0.235) = (0.942, 3.255)$.

Manipulating confidence intervals

- In many medical settings clinicians like to report results in terms of odds ratios rather than probabilities.
- If the probability is given by p then the odds ratio is $p/(1-p)$. A study has found the 95% confidence interval for p to be (0.485, 0.765).
- Find the 95% CI for the odds ratio.
- Answer: $(0.485/0.515, 0.765/0.235) = (0.942, 3.255)$.
- Remark: the point estimation for the odds ratio is $\hat{p}/(1-\hat{p}) = 0.625/0.375 = 1.67$. CI is not necessarily symmetric around the point estimator.

Manipulating confidence intervals

- The key idea: you are allowed to transform the ends of the confidence interval to obtain a new confidence interval for the transformed parameter (the transformation must be monotone though).

Manipulating confidence intervals

- The key idea: you are allowed to transform the ends of the confidence interval to obtain a new confidence interval for the transformed parameter (the transformation must be monotone though).
- Example: Let X be the temperature at 9am tomorrow in degree Celsius. You have a 95% CI for $\mu = \mathbb{E}(X)$: (12°C, 14°C).

Manipulating confidence intervals

- The key idea: you are allowed to transform the ends of the confidence interval to obtain a new confidence interval for the transformed parameter (the transformation must be monotone though).
- Example: Let X be the temperature at 9am tomorrow in degree Celsius. You have a 95% CI for $\mu = \mathbb{E}(X)$: (12°C, 14°C).
- Recall the formula that converts Celsius to Fahrenheit, $F = 32 + 9C$.

Manipulating confidence intervals

- The key idea: you are allowed to transform the ends of the confidence interval to obtain a new confidence interval for the transformed parameter (the transformation must be monotone though).
- Example: Let X be the temperature at 9am tomorrow in degree Celsius. You have a 95% CI for $\mu = \mathbb{E}(X)$: (12°C, 14°C).
- Recall the formula that converts Celsius to Fahrenheit, $F = 32 + 9C$.
- Find a 95% CI for μ when it is measured in degree Fahrenheit.

Manipulating confidence intervals

- The key idea: you are allowed to transform the ends of the confidence interval to obtain a new confidence interval for the transformed parameter (the transformation must be monotone though).
- Example: Let X be the temperature at 9am tomorrow in degree Celsius. You have a 95% CI for $\mu = \mathbb{E}(X)$: (12°C, 14°C).
- Recall the formula that converts Celsius to Fahrenheit, $F = 32 + 9C$.
- Find a 95% CI for μ when it is measured in degree Fahrenheit.
- The interval on the transformed scale is: $(12 \times 9 + 32, 14 \times 9 + 32)°\text{F} = (140°\text{F}, 158°\text{F})$.

Example

Example

A novel coronavirus (2019-nCov) outbreak spreads around the world, especially in Wuhan, China. As of last Saturday, Japan confirmed 8 among 565 citizens evacuated from Wuhan test positive for coronavirus, and Singapore confirmed 2 among 92 evacuees test positive. Assuming the cases are *i.i.d.*, what is the 95% CI for the infection rate of this virus?

Example

Example

A novel coronavirus (2019-nCov) outbreak spreads around the world, especially in Wuhan, China. As of last Saturday, Japan confirmed 8 among 565 citizens evacuated from Wuhan test positive for coronavirus, and Singapore confirmed 2 among 92 evacuees test positive. Assuming the cases are *i.i.d.*, what is the 95% CI for the infection rate of this virus?

Answer

- $\hat{p} = (8 + 2)/(565 + 92) = 1.52\%$

Example

Example

A novel coronavirus (2019-nCov) outbreak spreads around the world, especially in Wuhan, China. As of last Saturday, Japan confirmed 8 among 565 citizens evacuated from Wuhan test positive for coronavirus, and Singapore confirmed 2 among 92 evacuees test positive. Assuming the cases are *i.i.d.*, what is the 95% CI for the infection rate of this virus?

Answer

- $\hat{p} = (8 + 2)/(565 + 92) = 1.52\%$
- $n = 657 > 30, n\hat{p}, n(1 - \hat{p}) \geq 10.$

Example

Example

A novel coronavirus (2019-nCov) outbreak spreads around the world, especially in Wuhan, China. As of last Saturday, Japan confirmed 8 among 565 citizens evacuated from Wuhan test positive for coronavirus, and Singapore confirmed 2 among 92 evacuees test positive. Assuming the cases are *i.i.d.*, what is the 95% CI for the infection rate of this virus?

Answer

- $\hat{p} = (8+2)/(565+92) = 1.52\%$
- $n = 657 > 30, n\hat{p}, n(1-\hat{p}) \geq 10$.
- The approximate 95% CI is given by

$$1.52\% \pm 2\sqrt{\frac{1.52\% \times (1 - 1.52\%)}{657}} = (0.57\%, 2.47\%).$$

Example

Example

A novel coronavirus (2019-nCov) outbreak spreads around the world, especially in Wuhan, China. As of last Saturday, Japan confirmed 8 among 565 citizens evacuated from Wuhan test positive for coronavirus, and Singapore confirmed 2 among 92 evacuees test positive. Assuming the cases are *i.i.d.*, what is the 95% CI for the infection rate of this virus?

Answer

- $\hat{p} = (8+2)/(565+92) = 1.52\%$
- $n = 657 > 30, n\hat{p}, n(1-\hat{p}) \geq 10.$
- The approximate 95% CI is given by

$$1.52\% \pm 2\sqrt{\frac{1.52\% \times (1 - 1.52\%)}{657}} = (0.57\%, 2.47\%).$$

- Remark: i.i.d. assumption

A conservative 95% CI for a binomial parameter

- Recall that the 95% confidence interval for p is

$$[\hat{p} - 2\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + 2\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}]$$

A conservative 95% CI for a binomial parameter

- Recall that the 95% confidence interval for p is

$$[\hat{p} - 2\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + 2\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}]$$

- Since $0 \leq \hat{p} \leq 1$, we then have $\hat{p}(1-\hat{p}) \leq \frac{1}{4}$, and the equality holds if and only if $\hat{p} = \frac{1}{2}$.

A conservative 95% CI for a binomial parameter

- Recall that the 95% confidence interval for p is

$$[\hat{p} - 2\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + 2\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}]$$

- Since $0 \leq \hat{p} \leq 1$, we then have $\hat{p}(1-\hat{p}) \leq \frac{1}{4}$, and the equality holds if and only if $\hat{p} = \frac{1}{2}$.
- This means $2\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq 2\sqrt{\frac{1}{4n}} = \sqrt{\frac{1}{n}}$.

A conservative 95% CI for a binomial parameter

- Recall that the 95% confidence interval for p is

$$[\hat{p} - 2\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + 2\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}]$$

- Since $0 \leq \hat{p} \leq 1$, we then have $\hat{p}(1-\hat{p}) \leq \frac{1}{4}$, and the equality holds if and only if $\hat{p} = \frac{1}{2}$.
- This means $2\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq 2\sqrt{\frac{1}{4n}} = \sqrt{\frac{1}{n}}$.
- Therefore, a conservative 95% confidence interval for p is

$$[\hat{p} - \sqrt{\frac{1}{n}}, \hat{p} + \sqrt{\frac{1}{n}}].$$

Sample size for estimating a binomial parameter

- In the conservative 95% confidence interval for p, $MoE = \sqrt{\frac{1}{n}}$.

Sample size for estimating a binomial parameter

- In the conservative 95% confidence interval for p, $MoE = \sqrt{\frac{1}{n}}$.
- Therefore, a conservative approximation to the sample size formula becomes

$$n = (\frac{1}{MoE})^2.$$

Sample size for estimating a binomial parameter

- In the conservative 95% confidence interval for p, $MoE = \sqrt{\frac{1}{n}}$.
- Therefore, a conservative approximation to the sample size formula becomes

$$n = (\frac{1}{MoE})^2.$$

- You can rely on this approximation if:

Sample size for estimating a binomial parameter

- In the conservative 95% confidence interval for p, $MoE = \sqrt{\frac{1}{n}}$.
- Therefore, a conservative approximation to the sample size formula becomes

$$n = (\frac{1}{MoE})^2.$$

- You can rely on this approximation if:

  1. Your data is $i.i.d$.

Sample size for estimating a binomial parameter

- In the conservative 95% confidence interval for p, $MoE = \sqrt{\frac{1}{n}}$.
- Therefore, a conservative approximation to the sample size formula becomes

$$n = (\frac{1}{MoE})^2.$$

- You can rely on this approximation if:

1. Your data is $i.i.d.$
2. You want a 95% CI.

Sample size for estimating a binomial parameter

- In the conservative 95% confidence interval for p, $MoE = \sqrt{\frac{1}{n}}$.
- Therefore, a conservative approximation to the sample size formula becomes

$$n = (\frac{1}{MoE})^2.$$

- You can rely on this approximation if:
  1. Your data is $i.i.d$.
  2. You want a 95% CI.
  3. p lies between 0.25 and 0.75.

Example

A pollster wants to estimate the proportion of voters who will vote for a given presidential candidate. Assume that the behaviors of voters are $i.i.d.$. How large a sample is required to produce a 95% confidence interval with a margin of error of 3%?

Example

A pollster wants to estimate the proportion of voters who will vote for a given presidential candidate. Assume that the behaviors of voters are $i.i.d.$. How large a sample is required to produce a 95% confidence interval with a margin of error of 3%?

Answer

$n = (\frac{1}{MoE})^2 = (\frac{1}{3\%})^2 = 1111.$

Example

A market research company has estimated the proportion of physicians who say they will prescribe a new drug as 30%. Unfortunately they did not provide a confidence interval. You have since learned that they used a sample size of 100. What was the Margin of Error for a conservative 95% CI? What's the 95% CI?

Example

A market research company has estimated the proportion of physicians who say they will prescribe a new drug as 30%. Unfortunately they did not provide a confidence interval. You have since learned that they used a sample size of 100. What was the Margin of Error for a conservative 95% CI? What's the 95% CI?

Answer

The $MoE = 1/\sqrt{100} = 0.1$, so the confidence interval is approximately (30% - 10%, 30% + 10%) = (20%, 40%).

Confidence Intervals with unknown σ

For the CI of μ: when the sample size n is large

- If $X_1, ..., X_n$ is an $i.i.d.$ sample from a population, the 95% confidence interval (CI) for the population mean μ is

$$[\bar{X} - 2\frac{\sigma}{\sqrt{n}}, \bar{X} + 2\frac{\sigma}{\sqrt{n}}]$$

Confidence Intervals with unknown σ

For the CI of μ: when the sample size n is large

- If $X_1, ..., X_n$ is an $i.i.d.$ sample from a population, the 95% confidence interval (CI) for the population mean μ is

$$[\bar{X} - 2\frac{\sigma}{\sqrt{n}}, \bar{X} + 2\frac{\sigma}{\sqrt{n}}]$$

- For large n, you can substitute s for σ to obtain a 95% confidence interval

$$[\bar{X} - 2\frac{s}{\sqrt{n}}, \bar{X} + 2\frac{s}{\sqrt{n}}],$$

where $s = \sqrt{\frac{(X_1-\hat{\mu})^2+\ldots+(X_n-\hat{\mu})^2}{n-1}}$.

Small sample size: confidence interval for the mean.

- What if you are out of luck and your sample size in not large?

Small sample size: confidence interval for the mean.

- What if you are out of luck and your sample size in not large?
- Answer: If $n \leq 30$ then the estimation of σ is not accurate, and the normal approximation is not really appropriate since the CLT speaks only about large sample sizes.

Small sample size: confidence interval for the mean.

- What if you are out of luck and your sample size in not large?
- Answer: If $n \leq 30$ then the estimation of σ is not accurate, and the normal approximation is not really appropriate since the CLT speaks only about large sample sizes.
- For large sample size, the Central Limit Theorem told us that

$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n}).$$

Small sample size: confidence interval for the mean.

- What if you are out of luck and your sample size in not large?
- Answer: If $n \leq 30$ then the estimation of σ is not accurate, and the normal approximation is not really appropriate since the CLT speaks only about large sample sizes.
- For large sample size, the Central Limit Theorem told us that

$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n}).$$

- Equivalently,

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1).$$

Small sample size: confidence interval for the mean.

- For large sample size,
$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1).$$

Small sample size: confidence interval for the mean.

- For large sample size,

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1).$$

- When we substitute s instead of σ, we get

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

Small sample size: confidence interval for the mean.

- For large sample size,
$$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1).$$
- When we substitute s instead of σ, we get
$$\frac{\bar{X}-\mu}{s/\sqrt{n}}$$
- For large sample sizes, this is close to the standard normal random variable, but not for small n.

Small sample size: confidence interval for the mean.

- For large sample size,
$$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1).$$
- When we substitute s instead of σ, we get
$$\frac{\bar{X}-\mu}{s/\sqrt{n}}$$
- For large sample sizes, this is close to the standard normal random variable, but not for small n.
- This is because s is a random variable too and for small n, s is a bad estimate of σ.

Small sample size: confidence interval for the mean.

- Consider $\frac{\bar{X}-\mu}{s/\sqrt{n}}$, where $X_1, ..., X_n$ $i.i.d. \sim P$.
- For large sample size n, it's approximately normal distributed
- For small sample size n,
  - If the population P is not a normal distribution, there is nothing we can do at the moment for small sample sizes.
  - We only consider the case when P is normal.

Student's t-distribution

- Fact: if $X_1, ..., X_n$ *i.i.d.* $\sim N(\mu, \sigma^2)$, then

$$T_{n-1} = \frac{\bar{X} - \mu}{s/\sqrt{n}} \sim t_{n-1}$$

has a t-distribution (also called Student's t-distribution).

Student's t-distribution

- Fact: if $X_1, ..., X_n$ *i.i.d.* $\sim N(\mu, \sigma^2)$, then

$$T_{n-1} = \frac{\bar{X} - \mu}{s/\sqrt{n}} \sim t_{n-1}$$

  has a t-distribution (also called Student's t-distribution).

- Note that t_{n-1} depends on n. The subscript $n-1$ denotes the $n-1$ degrees of freedom (d.f.), which controls the shape of t_{n-1}.

Student's t-distribution

- Fact: if $X_1, ..., X_n$ *i.i.d.* $\sim N(\mu, \sigma^2)$, then
$$T_{n-1} = \frac{\bar{X} - \mu}{s/\sqrt{n}} \sim t_{n-1}$$
has a t-distribution (also called Student's t-distribution).
- Note that t_{n-1} depends on n. The subscript $n-1$ denotes the $n-1$ degrees of freedom (d.f.), which controls the shape of t_{n-1}.
- The reason why the T_{n-1} is different from a standard normal Z is that the s in the denominator will vary from sample to sample, whereas the σ in the Z is just a fixed number.

Student's t-distribution

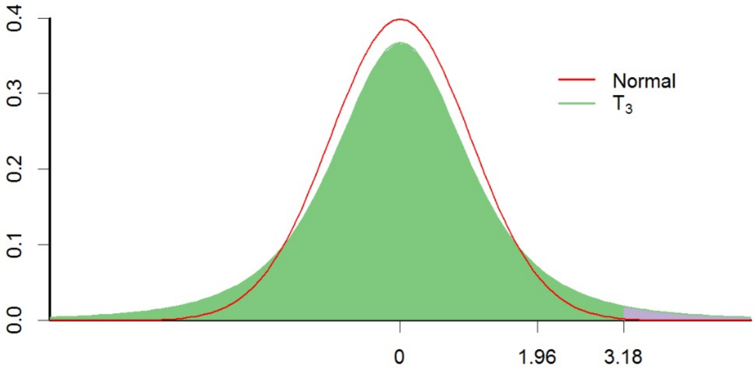
- Fact: if $X_1, ..., X_n$ *i.i.d.* $\sim N(\mu, \sigma^2)$, then

$$T_{n-1} = \frac{\bar{X} - \mu}{s/\sqrt{n}} \sim t_{n-1}$$

  has a t-distribution (also called Student's t-distribution).
- Note that t_{n-1} depends on n. The subscript $n-1$ denotes the $n-1$ degrees of freedom (d.f.), which controls the shape of t_{n-1}.
- The reason why the T_{n-1} is different from a standard normal Z is that the s in the denominator will vary from sample to sample, whereas the σ in the Z is just a fixed number.
- In fact, as n becomes large, t_{n-1} converges to a $N(0, 1)$.

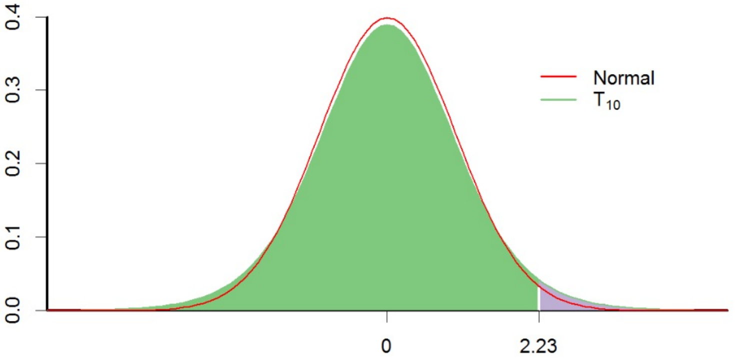
Pictures of the Students t-distribution.

Figure 1: A t-distribution with 3 degrees of freedom and the standard normal compared in red. The 97.5 percentile is identified



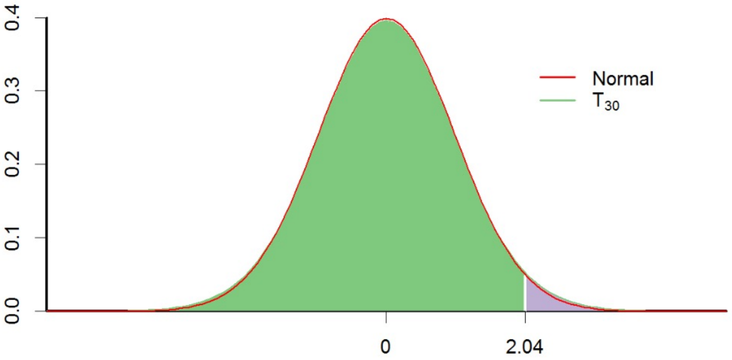
Pictures of the Students t-distribution.

Figure 2: A t-distribution with 10 degrees of freedom and the standard normal compared in red. The 97.5 percentile is identified



Pictures of the Students t-distribution.

Figure 3: A t-distribution with 30 degrees of freedom and the standard normal compared in red. The 97.5 percentile is identified



Summary of the t-distribution

- t_{n-1} looks like $N(0,1)$, but with fatter tails.

Summary of the t-distribution

- t_{n-1} looks like $N(0, 1)$, but with fatter tails.
- Thus, to get 95% on this fatter distribution, you need to go out further than on the normal distribution.

Summary of the t-distribution

- t_{n-1} looks like $N(0,1)$, but with fatter tails.
- Thus, to get 95% on this fatter distribution, you need to go out further than on the normal distribution.
- The 97.5 percentile cut-off for the t_{n-1} is greater than 1.96 (that for $N(0,1)$).

Summary of the t-distribution

- t_{n-1} looks like $N(0,1)$, but with fatter tails.
- Thus, to get 95% on this fatter distribution, you need to go out further than on the normal distribution.
- The 97.5 percentile cut-off for the t_{n-1} is greater than 1.96 (that for $N(0,1)$).
- t-tables to find these cut-offs are in the back of the textbook.

Summary of the t-distribution

- t_{n-1} looks like $N(0,1)$, but with fatter tails.
- Thus, to get 95% on this fatter distribution, you need to go out further than on the normal distribution.
- The 97.5 percentile cut-off for the t_{n-1} is greater than 1.96 (that for $N(0,1)$).
- t-tables to find these cut-offs are in the back of the textbook.
- For small degrees of freedom, e.g. 3, there can be a large difference between the cutoffs (3.18 v.s. 1.96).

Summary of the t-distribution

- t_{n-1} looks like $N(0, 1)$, but with fatter tails.
- Thus, to get 95% on this fatter distribution, you need to go out further than on the normal distribution.
- The 97.5 percentile cut-off for the t_{n-1} is greater than 1.96 (that for $N(0, 1)$).
- t-tables to find these cut-offs are in the back of the textbook.
- For small degrees of freedom, e.g. 3, there can be a large difference between the cutoffs (3.18 v.s. 1.96).
- By the time the degrees of freedom reach 30, the difference between cutoffs is very small (2.04 v.s. 1.96).

Summary of the t-distribution

- t_{n-1} looks like $N(0,1)$, but with fatter tails.
- Thus, to get 95% on this fatter distribution, you need to go out further than on the normal distribution.
- The 97.5 percentile cut-off for the t_{n-1} is greater than 1.96 (that for $N(0,1)$).
- t-tables to find these cut-offs are in the back of the textbook.
- For small degrees of freedom, e.g. 3, there can be a large difference between the cutoffs (3.18 v.s. 1.96).
- By the time the degrees of freedom reach 30, the difference between cutoffs is very small (2.04 v.s. 1.96).
- In most practical problems with large sample size, this difference is not important and simply rounding the cut-off to 2 is a reasonable rule of thumb.

Summary of the t-distribution

- t_{n-1} looks like $N(0,1)$, but with fatter tails.
- Thus, to get 95% on this fatter distribution, you need to go out further than on the normal distribution.
- The 97.5 percentile cut-off for the t_{n-1} is greater than 1.96 (that for $N(0,1)$).
- t-tables to find these cut-offs are in the back of the textbook.
- For small degrees of freedom, e.g. 3, there can be a large difference between the cutoffs (3.18 v.s. 1.96).
- By the time the degrees of freedom reach 30, the difference between cutoffs is very small (2.04 v.s. 1.96).
- In most practical problems with large sample size, this difference is not important and simply rounding the cut-off to 2 is a reasonable rule of thumb.
- When software is used to do the calculations, exact cut-off values from the relevant t-distribution are used.

Summary of the t-distribution

- Important: only use the t-interval if the population distribution (of X_i) is normal or approximately normal.

Summary of the t-distribution

- Important: only use the t-interval if the population distribution (of X_i) is normal or approximately normal.
- Key fact: for $n > 30$, t-distribution is very close to the standard normal. For smaller sample size, there is difference, which becomes more and more noticeable for smaller n.

Summary of the t-distribution

- Important: only use the t-interval if the population distribution (of X_i) is normal or approximately normal.
- Key fact: for $n > 30$, t-distribution is very close to the standard normal. For smaller sample size, there is difference, which becomes more and more noticeable for smaller n.
- Rule: use t-values whenever $n < 30$ when you substitute s for σ, if the population distribution is normal.

General confidence interval based on the t-distribution

- t-distribution is bell-shaped but has fatter tails.

General confidence interval based on the t-distribution

- t-distribution is bell-shaped but has fatter tails.
- Recall the $(1-\alpha)$ confidence interval for known σ:

$$[\bar{X} - z_{\alpha/2} \cdot \sigma/\sqrt{n}, \bar{X} + z_{\alpha/2} \cdot \sigma/\sqrt{n}].$$

  We will call this a z-interval because we rely on the normal distribution.

General confidence interval based on the t-distribution

- t-distribution is bell-shaped but has fatter tails.
- Recall the $(1-\alpha)$ confidence interval for known σ:

$$[\bar{X} - z_{\alpha/2} \cdot \sigma/\sqrt{n}, \bar{X} + z_{\alpha/2} \cdot \sigma/\sqrt{n}].$$

  We will call this a z-interval because we rely on the normal distribution.
- When we replace σ with s, the $(1-\alpha)$ confidence interval (which we now call t-interval) becomes

$$[\bar{X} - t_{\alpha/2,n-1} \cdot s/\sqrt{n}, \bar{X} + t_{\alpha/2,n-1} \cdot s/\sqrt{n}],$$

  where $t_{\alpha/2,n-1}$ is the value such that $\mathbb{P}(T_{n-1} > t_{\alpha/2,n-1}) = \alpha/2$ for the t-random variable T_{n-1} with $n-1$ degrees of freedom.

General confidence interval based on the t-distribution

- t-distribution is bell-shaped but has fatter tails.
- Recall the $(1-\alpha)$ confidence interval for known σ:

$$[\bar{X} - z_{\alpha/2} \cdot \sigma/\sqrt{n}, \bar{X} + z_{\alpha/2} \cdot \sigma/\sqrt{n}].$$

  We will call this a z-interval because we rely on the normal distribution.
- When we replace σ with s, the $(1-\alpha)$ confidence interval (which we now call t-interval) becomes

$$[\bar{X} - t_{\alpha/2,n-1} \cdot s/\sqrt{n}, \bar{X} + t_{\alpha/2,n-1} \cdot s/\sqrt{n}],$$

  where $t_{\alpha/2,n-1}$ is the value such that $\mathbb{P}(T_{n-1} > t_{\alpha/2,n-1}) = \alpha/2$ for the t-random variable T_{n-1} with $n-1$ degrees of freedom.
- Looks scary, but it's not. The only thing that's different is $t_{\alpha/2,n-1}$ instead of $z_{\alpha/2}$. I will provide you with $t_{\alpha/2,n-1}$, so no worries!

Key Take-away

- For large n, $t_{\alpha/2,n-1} \approx z_{\alpha/2}$, so we are justified in using z-values for the case of unknown σ.
- For small n, however, the t-value is larger than the corresponding z-value. Indeed, the tails of the t-distribution are fatter and the bump at the center is smaller, so you need to go further away from the center to get the same probability.
- Keep in mind that t-distribution can only be used if the original population (of X_i's) is roughly normal.

Practice Problem

A list of t-values are given below: $t_{0.05,9} = 1.833$, $t_{0.025,9} = 2.262$, $t_{0.01,9} = 2.821$.

A sample of 10 observations is drawn from a normal distribution with unknown mean and variance. The sample mean is 12.8 and sample variance s^2 is 4.1. What is a 95% CI for the true mean?

Practice Problem

A list of t-values are given below: $t_{0.05,9} = 1.833$, $t_{0.025,9} = 2.262$, $t_{0.01,9} = 2.821$.

A sample of 10 observations is drawn from a normal distribution with unknown mean and variance. The sample mean is 12.8 and sample variance s^2 is 4.1. What is a 95% CI for the true mean?

Answer

- $[\bar{X} - t_{\alpha/2,n-1} \cdot s/\sqrt{n}, \bar{X} + t_{\alpha/2,n-1} \cdot s/\sqrt{n}]$

Practice Problem

A list of t-values are given below: $t_{0.05,9} = 1.833$, $t_{0.025,9} = 2.262$, $t_{0.01,9} = 2.821$.

A sample of 10 observations is drawn from a normal distribution with unknown mean and variance. The sample mean is 12.8 and sample variance s^2 is 4.1. What is a 95% CI for the true mean?

Answer

- $[\bar{X} - t_{\alpha/2,n-1} \cdot s/\sqrt{n}, \bar{X} + t_{\alpha/2,n-1} \cdot s/\sqrt{n}]$
- $12.8 \pm 2.262 \cdot \sqrt{4.1}/\sqrt{10} = [11.35, 14.25]$.

Summary of Confidence intervals

- A general formula for CI of mean-based parameters. For a parameter θ, suppose we estimate it by $\hat{\theta}$ and $\hat{\theta}$ is approximated normal. Then a $(1-\alpha)$ confidence interval for θ is

$$[\hat{\theta} - z_{\alpha/2} \cdot \widehat{\mathrm{sd}(\hat{\theta})}, \quad \hat{\theta} + z_{\alpha/2} \cdot \widehat{\mathrm{sd}(\hat{\theta})}]$$

  where $\mathrm{sd}(\hat{\theta})$ is the standard deviation of $\hat{\theta}$, and $\widehat{\mathrm{sd}(\hat{\theta})}$ is the estimate of $\mathrm{sd}(\hat{\theta})$.

- In the setting where the sample size is small and the observations are normally distributed, we replace $z_{\alpha/2}$ with $t_{\alpha/2,n-1}$.

Confidence intervals

- When $n \geq 30$, and $n\hat{p}, n(1-\hat{p}) > 10$, if X has a binomial distribution $Bin(n, p)$, then the estimate of the binomial parameter p is $\hat{p} = \frac{X}{n}$

Confidence intervals

- When $n \geq 30$, and $n\hat{p}, n(1-\hat{p}) > 10$, if X has a binomial distribution $Bin(n,p)$, then the estimate of the binomial parameter p is $\hat{p} = \frac{X}{n}$
- $sd(\hat{p}) = \frac{p(1-p)}{n}$

Confidence intervals

- When $n \geq 30$, and $n\hat{p}, n(1-\hat{p}) > 10$, if X has a binomial distribution $Bin(n,p)$, then the estimate of the binomial parameter p is $\hat{p} = \frac{X}{n}$
- $sd(\hat{p}) = \frac{p(1-p)}{n}$
- $\widehat{sd(\hat{p})} = \frac{\hat{p}(1-\hat{p})}{n}$

Confidence intervals

- When $n \geq 30$, and $n\hat{p}, n(1-\hat{p}) > 10$, if X has a binomial distribution $Bin(n,p)$, then the estimate of the binomial parameter p is $\hat{p} = \frac{X}{n}$
- $sd(\hat{p}) = \frac{p(1-p)}{n}$
- $\widehat{sd(\hat{p})} = \frac{\hat{p}(1-\hat{p})}{n}$
- An approximate 95% confidence interval for p is

$$[\hat{p} - 2\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \ \ \hat{p} + 2\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}]$$

Confidence intervals

- When $n \geq 30$, and $n\hat{p}, n(1-\hat{p}) > 10$, if X has a binomial distribution $Bin(n, p)$, then the estimate of the binomial parameter p is $\hat{p} = \frac{X}{n}$
- $sd(\hat{p}) = \frac{p(1-p)}{n}$
- $\widehat{sd(\hat{p})} = \frac{\hat{p}(1-\hat{p})}{n}$
- An approximate 95% confidence interval for p is

$$[\hat{p} - 2\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \ \ \hat{p} + 2\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}]$$

- A conservative 95% confidence interval for p is

$$[\hat{p} - \sqrt{\frac{1}{n}}, \ \ \hat{p} + \sqrt{\frac{1}{n}}].$$

Confidence intervals

- When the sample size $n > 30$, if $X_1, X_2, ..., X_n$ are i.i.d. with mean μ and variance σ^2, then an estimate of μ is $\hat{\mu} = \bar{X} = \frac{X_1 + ... + X_n}{n}$.
- $sd(\bar{X}) = \frac{\sigma}{\sqrt{n}}$
- $\widehat{sd(\bar{X})} = \frac{s}{\sqrt{n}}$, where $s^2 = \frac{(X_1 - \bar{X})^2 + ... + (X_n - \bar{X})^2}{n-1}$
- An approximate 95% confidence interval for μ is

$$[\bar{X} - 2\frac{s}{\sqrt{n}}, \quad \bar{X} + 2\frac{s}{\sqrt{n}}].$$

Confidence intervals

- When the sample size is small, if $X_1, X_2, ..., X_n$ are i.i.d. $N(\mu, \sigma^2)$, then an estimate of μ is still $\hat{\mu} = \bar{X} = \frac{X_1 + ... + X_n}{n}$.

Confidence intervals

- When the sample size is small, if $X_1, X_2, ..., X_n$ are i.i.d. $N(\mu, \sigma^2)$, then an estimate of μ is still $\hat{\mu} = \bar{X} = \frac{X_1 + ... + X_n}{n}$.
- $sd(\bar{X}) = \frac{\sigma}{\sqrt{n}}$

Confidence intervals

- When the sample size is small, if $X_1, X_2, ..., X_n$ are i.i.d. $N(\mu, \sigma^2)$, then an estimate of μ is still $\hat{\mu} = \bar{X} = \frac{X_1 + ... + X_n}{n}$.
- $sd(\bar{X}) = \frac{\sigma}{\sqrt{n}}$
- $\widehat{sd(\bar{X})} = \frac{s}{\sqrt{n}}$, where $s^2 = \frac{(X_1 - \bar{X})^2 + ... + (X_1 - \bar{X})^2}{n-1}$

Confidence intervals

- When the sample size is small, if $X_1, X_2, ..., X_n$ are i.i.d. $N(\mu, \sigma^2)$, then an estimate of μ is still $\hat{\mu} = \bar{X} = \frac{X_1 + ... + X_n}{n}$.
- $sd(\bar{X}) = \frac{\sigma}{\sqrt{n}}$
- $\widehat{sd(\bar{X})} = \frac{s}{\sqrt{n}}$, where $s^2 = \frac{(X_1 - \bar{X})^2 + ... + (X_1 - \bar{X})^2}{n-1}$
- A 95% confidence interval for μ is

$$[\bar{X} - t_{0.025,n-1} \cdot \frac{s}{\sqrt{n}}, \quad \bar{X} + t_{0.025,n-1} \cdot \frac{s}{\sqrt{n}}].$$

CI for the difference between two binomial parameters

- When n, m are sufficiently large, if X has a binomial distribution $Bi(n, p_1)$, and Y has a binomial distribution $Bi(m, p_2)$. X and Y are independent.

CI for the difference between two binomial parameters

- When n, m are sufficiently large, if X has a binomial distribution $Bi(n, p_1)$, and Y has a binomial distribution $Bi(m, p_2)$. X and Y are independent.
- An estimate of $p_1 - p_2$ is $\hat{p}_1 - \hat{p}_2 = \frac{X}{n} - \frac{Y}{m}$.

CI for the difference between two binomial parameters

- When n, m are sufficiently large, if X has a binomial distribution $Bi(n, p_1)$, and Y has a binomial distribution $Bi(m, p_2)$. X and Y are independent.
- An estimate of $p_1 - p_2$ is $\hat{p}_1 - \hat{p}_2 = \frac{X}{n} - \frac{Y}{m}$.
- $sd(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{p_1(1-p_1)}{n} + \frac{p_2(1-p_2)}{m}}$

CI for the difference between two binomial parameters

- When n, m are sufficiently large, if X has a binomial distribution $Bi(n, p_1)$, and Y has a binomial distribution $Bi(m, p_2)$. X and Y are independent.
- An estimate of $p_1 - p_2$ is $\hat{p}_1 - \hat{p}_2 = \frac{X}{n} - \frac{Y}{m}$.
- $sd(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{p_1(1-p_1)}{n} + \frac{p_2(1-p_2)}{m}}$
- $\widehat{sd(\hat{p}_1 - \hat{p}_2)} = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n} + \frac{\hat{p}_2(1-\hat{p}_2)}{m}}$

CI for the difference between two binomial parameters

- When n, m are sufficiently large, if X has a binomial distribution $Bi(n, p_1)$, and Y has a binomial distribution $Bi(m, p_2)$. X and Y are independent.
- An estimate of $p_1 - p_2$ is $\hat{p}_1 - \hat{p}_2 = \frac{X}{n} - \frac{Y}{m}$.
- $sd(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{p_1(1-p_1)}{n} + \frac{p_2(1-p_2)}{m}}$
- $\widehat{sd(\hat{p}_1 - \hat{p}_2)} = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n} + \frac{\hat{p}_2(1-\hat{p}_2)}{m}}$
- An approximate 95% confidence interval for $p_1 - p_2$ is

$$[\hat{p}_1 - \hat{p}_2 - 2\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n} + \frac{\hat{p}_2(1-\hat{p}_2)}{m}}, \quad \hat{p}_1 - \hat{p}_2 + 2\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n} + \frac{\hat{p}_2(1-\hat{p}_2)}{m}}].$$

CI for the difference between two binomial parameters

- When n, m are sufficiently large, if X has a binomial distribution $Bi(n, p_1)$, and Y has a binomial distribution $Bi(m, p_2)$. X and Y are independent.
- An estimate of $p_1 - p_2$ is $\hat{p}_1 - \hat{p}_2 = \frac{X}{n} - \frac{Y}{m}$.
- $sd(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{p_1(1-p_1)}{n} + \frac{p_2(1-p_2)}{m}}$
- $\widehat{sd(\hat{p}_1 - \hat{p}_2)} = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n} + \frac{\hat{p}_2(1-\hat{p}_2)}{m}}$
- An approximate 95% confidence interval for $p_1 - p_2$ is

$$[\hat{p}_1 - \hat{p}_2 - 2\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n} + \frac{\hat{p}_2(1-\hat{p}_2)}{m}}, \quad \hat{p}_1 - \hat{p}_2 + 2\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n} + \frac{\hat{p}_2(1-\hat{p}_2)}{m}}].$$

- A conservative 95% confidence interval for $p_1 - p_2$ is

$$[\hat{p}_1 - \hat{p}_2 - \sqrt{\frac{1}{n} + \frac{1}{m}}, \quad \hat{p}_1 - \hat{p}_2 + \sqrt{\frac{1}{n} + \frac{1}{m}}].$$

Footnote

- $X \sim Bi(n, p_1), \; Y \sim Bi(m, p_2)$

$$
\begin{aligned}
sd(\hat{p}_1 - \hat{p}_2) =& \sqrt{Var(\hat{p}_1 - \hat{p}_2)} = \sqrt{Var(\frac{X}{n} - \frac{Y}{m})} \\
=& \sqrt{Var(\frac{X}{n}) + Var(\frac{Y}{m})} \\
=& \sqrt{\frac{1}{n^2} \cdot np_1(1 - p_1) + \frac{1}{m^2} \cdot mp_2(1 - p_2)} \\
=& \sqrt{\frac{p_1(1 - p_1)}{n} + \frac{p_2(1 - p_2)}{m}}
\end{aligned}
$$

- $2\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n} + \frac{\hat{p}_2(1-\hat{p}_2)}{m}} \leq 2\sqrt{\frac{1}{4n} + \frac{1}{4m}} = \sqrt{\frac{1}{n} + \frac{1}{m}}.$

Practice problem

Question

We have two drugs, drug A and drug B, both aimed at curing a certain illness. We wish to compare the two drugs. Define p_1 as the probability that a person with this illness will be cured if he/she takes drug A, and p_2 as the probability that a person with this illness will be cured if he/she takes drug B. We give drug A to $1,000$ people and it cures 840 of them. We give drug B to $1,200$ people and it cures 970 of them. Find the 95% CI for $p_1 - p_2$.

Practice problem

Question

We have two drugs, drug A and drug B, both aimed at curing a certain illness. We wish to compare the two drugs. Define p_1 as the probability that a person with this illness will be cured if he/she takes drug A, and p_2 as the probability that a person with this illness will be cured if he/she takes drug B. We give drug A to $1,000$ people and it cures 840 of them. We give drug B to $1,200$ people and it cures 970 of them. Find the 95% CI for $p_1 - p_2$.

Solution

$\hat{p}_1 = 0.84$, $\hat{p}_2 = 0.81$, $n = 1000$, $m = 1200$. Hence the 95% confidence interval is $0.03 \pm \sqrt{\frac{1}{1200} + \frac{1}{1000}}$, which is -0.01 to 0.07.

CI for $\mu_X - \mu_Y$ when n, m are large

If $X_1, X_2, ..., X_n$ are i.i.d. with mean μ_X and variance σ_X^2, $Y_1, Y_2, ..., Y_m$ are i.i.d. with mean μ_Y and variance σ_Y^2. $\{X_1, ..., X_n\}$ and $\{Y_1, ..., Y_m\}$ are independent.

- We estimate $\mu_X - \mu_Y$ by $\bar{X} - \bar{Y}$, where $\bar{X} = \frac{X_1 + \ldots + X_n}{n}$, $\bar{Y} = \frac{Y_1 + \ldots + Y_m}{m}$.

CI for $\mu_X - \mu_Y$ when n, m are large

If $X_1, X_2, ..., X_n$ are i.i.d. with mean μ_X and variance σ_X^2, $Y_1, Y_2, ..., Y_m$ are i.i.d. with mean μ_Y and variance σ_Y^2. $\{X_1, ..., X_n\}$ and $\{Y_1, ..., Y_m\}$ are independent.

- We estimate $\mu_X - \mu_Y$ by $\bar{X} - \bar{Y}$, where $\bar{X} = \frac{X_1 + \ldots + X_n}{n}$, $\bar{Y} = \frac{Y_1 + \ldots + Y_m}{m}$.
- The variance of $\bar{X} - \bar{Y}$ is $\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}$

CI for $\mu_X - \mu_Y$ when n, m are large

If $X_1, X_2, ..., X_n$ are i.i.d. with mean μ_X and variance σ_X^2, $Y_1, Y_2, ..., Y_m$ are i.i.d. with mean μ_Y and variance σ_Y^2. $\{X_1, ..., X_n\}$ and $\{Y_1, ..., Y_m\}$ are independent.

- We estimate $\mu_X - \mu_Y$ by $\bar{X} - \bar{Y}$, where $\bar{X} = \frac{X_1 + \ldots + X_n}{n}$, $\bar{Y} = \frac{Y_1 + \ldots + Y_m}{m}$.
- The variance of $\bar{X} - \bar{Y}$ is $\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}$
- We estimate σ_X^2 and σ_Y^2 by $s_X^2 = \frac{(X_1 - \bar{X})^2 + \ldots + (X_n - \bar{X})^2}{n-1}$, $s_Y^2 = \frac{(Y_1 - \bar{Y})^2 + \ldots + (Y_m - \bar{Y})^2}{m-1}$.

CI for $\mu_X - \mu_Y$ when n, m are large

If $X_1, X_2, ..., X_n$ are i.i.d. with mean μ_X and variance σ_X^2, $Y_1, Y_2, ..., Y_m$ are i.i.d. with mean μ_Y and variance σ_Y^2. $\{X_1, ..., X_n\}$ and $\{Y_1, ..., Y_m\}$ are independent.

- We estimate $\mu_X - \mu_Y$ by $\bar{X} - \bar{Y}$, where $\bar{X} = \frac{X_1 + ... + X_n}{n}$, $\bar{Y} = \frac{Y_1 + ... + Y_m}{m}$.
- The variance of $\bar{X} - \bar{Y}$ is $\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}$
- We estimate σ_X^2 and σ_Y^2 by $s_X^2 = \frac{(X_1 - \bar{X})^2 + ... + (X_n - \bar{X})^2}{n-1}$, $s_Y^2 = \frac{(Y_1 - \bar{Y})^2 + ... + (Y_m - \bar{Y})^2}{m-1}$.
- A $(1-\alpha)$ confidence interval for $\mu_X - \mu_Y$ is

$$[\bar{X} - \bar{Y} - z_{\alpha/2}\sqrt{\frac{s_X^2}{n} + \frac{s_Y^2}{m}}, \quad \bar{X} - \bar{Y} + z_{\alpha/2}\sqrt{\frac{s_X^2}{n} + \frac{s_Y^2}{m}}].$$

Footnote

$$
\begin{aligned}
sd(\hat{\mu}_X - \hat{\mu}_Y) =& \sqrt{Var(\hat{\mu}_X - \hat{\mu}_Y)} = \sqrt{Var(\bar{X} - \bar{Y})} \\
=& \sqrt{Var(\bar{X}) + Var(\bar{Y})} \\
=& \sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}
\end{aligned}
$$

Practice problem

A psychologist was interested in exploring whether or not male and female college students have different driving behaviors. She focused on the fastest speed ever driven by an individual. She conducted a survey of a random $n = 34$ male college students and a random $m = 30$ female college students. Here is a descriptive summary of the results of her survey: for male students, the average of their fastest speeds is 105 miles per hour (mph), and the standard deviation is 20.1 mph; for female students, the average is 90.9 mph and the standard deviation is 12.2 mph. Construct a 95% CI for the difference of the fastest speed ever driven by male and female college students.

Practice problem

A psychologist was interested in exploring whether or not male and female college students have different driving behaviors. She focused on the fastest speed ever driven by an individual. She conducted a survey of a random $n = 34$ male college students and a random $m = 30$ female college students. Here is a descriptive summary of the results of her survey: for male students, the average of their fastest speeds is 105 miles per hour (mph), and the standard deviation is 20.1 mph; for female students, the average is 90.9 mph and the standard deviation is 12.2 mph. Construct a 95% CI for the difference of the fastest speed ever driven by male and female college students.

Answer

- $[\bar{X} - \bar{Y} - z_{\alpha/2}\sqrt{\frac{s_X^2}{n} + \frac{s_Y^2}{m}}, \quad \bar{X} - \bar{Y} + z_{\alpha/2}\sqrt{\frac{s_X^2}{n} + \frac{s_Y^2}{m}}]$

Practice problem

A psychologist was interested in exploring whether or not male and female college students have different driving behaviors. She focused on the fastest speed ever driven by an individual. She conducted a survey of a random $n = 34$ male college students and a random $m = 30$ female college students. Here is a descriptive summary of the results of her survey: for male students, the average of their fastest speeds is 105 miles per hour (mph), and the standard deviation is 20.1 mph; for female students, the average is 90.9 mph and the standard deviation is 12.2 mph. Construct a 95% CI for the difference of the fastest speed ever driven by male and female college students.

Answer

- $[\bar{X} - \bar{Y} - z_{\alpha/2}\sqrt{\frac{s_X^2}{n} + \frac{s_Y^2}{m}}, \quad \bar{X} - \bar{Y} + z_{\alpha/2}\sqrt{\frac{s_X^2}{n} + \frac{s_Y^2}{m}}]$
- $(105 - 90.9) \pm 2\sqrt{\frac{20.1^2}{34} + \frac{12.2^2}{30}} = [5.89, 22.31]$ mph

CI for $\mu_X - \mu_Y$ when n, m are small

When n, m are small. If $X_1, X_2, ..., X_n$ i.i.d. $\sim N(\mu_X, \sigma_X^2)$, $Y_1, Y_2, ..., Y_m$ i.i.d. $\sim N(\mu_Y, \sigma_Y^2)$. $\{X_1, ..., X_n\}$ and $\{Y_1, ..., Y_m\}$ are independent.

- We estimate $\mu_X - \mu_Y$ by $\bar{X} - \bar{Y}$, where $\bar{X} = \frac{X_1 + ... + X_n}{n}$, $\bar{Y} = \frac{Y_1 + ... + Y_m}{m}$.

CI for $\mu_X - \mu_Y$ when n, m are small

When n, m are small. If $X_1, X_2, ..., X_n$ i.i.d. $\sim N(\mu_X, \sigma_X^2)$, $Y_1, Y_2, ..., Y_m$ i.i.d. $\sim N(\mu_Y, \sigma_Y^2)$. $\{X_1, ..., X_n\}$ and $\{Y_1, ..., Y_m\}$ are independent.

- We estimate $\mu_X - \mu_Y$ by $\bar{X} - \bar{Y}$, where $\bar{X} = \frac{X_1 + \ldots + X_n}{n}$, $\bar{Y} = \frac{Y_1 + \ldots + Y_m}{m}$.
- The variance of $\bar{X} - \bar{Y}$ is $\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}$, and we estimate σ_X^2 and σ_Y^2 by $s_X^2 = \frac{(X_1 - \bar{X})^2 + \ldots + (X_n - \bar{X})^2}{n-1}$, $s_Y^2 = \frac{(Y_1 - \bar{Y})^2 + \ldots + (Y_m - \bar{Y})^2}{m-1}$.

CI for $\mu_X - \mu_Y$ when n, m are small

When n, m are small. If $X_1, X_2, ..., X_n$ i.i.d. $\sim N(\mu_X, \sigma_X^2)$, $Y_1, Y_2, ..., Y_m$ i.i.d. $\sim N(\mu_Y, \sigma_Y^2)$. $\{X_1, ..., X_n\}$ and $\{Y_1, ..., Y_m\}$ are independent.

- We estimate $\mu_X - \mu_Y$ by $\bar{X} - \bar{Y}$, where $\bar{X} = \frac{X_1+...+X_n}{n}$, $\bar{Y} = \frac{Y_1+...+Y_m}{m}$.
- The variance of $\bar{X} - \bar{Y}$ is $\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}$, and we estimate σ_X^2 and σ_Y^2 by $s_X^2 = \frac{(X_1-\bar{X})^2+...+(X_n-\bar{X})^2}{n-1}$, $s_Y^2 = \frac{(Y_1-\bar{Y})^2+...+(Y_m-\bar{Y})^2}{m-1}$.
- We then have $\frac{\bar{X}-\bar{Y}-(\mu_X-\mu_Y)}{\sqrt{\frac{s_X^2}{n}+\frac{s_Y^2}{m}}} \approx T_k$, where $k = \min\{n, m\} - 1$

CI for $\mu_X - \mu_Y$ when n, m are small

When n, m are small. If $X_1, X_2, ..., X_n$ i.i.d. $\sim N(\mu_X, \sigma_X^2)$, $Y_1, Y_2, ..., Y_m$ i.i.d. $\sim N(\mu_Y, \sigma_Y^2)$. $\{X_1, ..., X_n\}$ and $\{Y_1, ..., Y_m\}$ are independent.

- We estimate $\mu_X - \mu_Y$ by $\bar{X} - \bar{Y}$, where $\bar{X} = \frac{X_1+...+X_n}{n}$, $\bar{Y} = \frac{Y_1+...+Y_m}{m}$.
- The variance of $\bar{X} - \bar{Y}$ is $\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}$, and we estimate σ_X^2 and σ_Y^2 by $s_X^2 = \frac{(X_1-\bar{X})^2+...+(X_n-\bar{X})^2}{n-1}$, $s_Y^2 = \frac{(Y_1-\bar{Y})^2+...+(Y_m-\bar{Y})^2}{m-1}$.
- We then have $\frac{\bar{X}-\bar{Y}-(\mu_X-\mu_Y)}{\sqrt{\frac{s_X^2}{n}+\frac{s_Y^2}{m}}} \approx T_k$, where $k = \min\{n, m\} - 1$
- A $(1-\alpha)$ confidence interval for $\mu_X - \mu_Y$ is

$$[\bar{X} - \bar{Y} - t_{\alpha/2,k}\sqrt{\frac{s_X^2}{n} + \frac{s_Y^2}{m}}, \quad \bar{X} - \bar{Y} + t_{\alpha/2,k}\sqrt{\frac{s_X^2}{n} + \frac{s_Y^2}{m}}].$$

Practice problem

We are interested in investigating any potential difference between the mean blood sugar level of diabetics (μ_X) and that of non-diabetics (μ_Y). To do this we took a sample of six diabetics and found the following blood sugar levels: 127, 144, 140, 136, 118, 138. We also took a sample of eight non-diabetics and found the following blood sugar levels: 125, 128, 133, 141, 109, 125, 126, 122. (a) Estimate $\mu_X - \mu_Y$. (b) Find the 95% CI for $\mu_X - \mu_Y$.
($s_X^2 = 93.24, s_Y^2 = 83.55, t_{0.05,5} = 2.015, t_{0.025,5} = 2.571, t_{0.05,7} = 1.895, t_{0.025,7} = 2.365$)

Practice problem

We are interested in investigating any potential difference between the mean blood sugar level of diabetics (μ_X) and that of non-diabetics (μ_Y). To do this we took a sample of six diabetics and found the following blood sugar levels: 127, 144, 140, 136, 118, 138. We also took a sample of eight non-diabetics and found the following blood sugar levels: 125, 128, 133, 141, 109, 125, 126, 122. (a) Estimate $\mu_X - \mu_Y$. (b) Find the 95% CI for $\mu_X - \mu_Y$.

($s_X^2 = 93.24$, $s_Y^2 = 83.55$, $t_{0.05,5} = 2.015$, $t_{0.025,5} = 2.571$, $t_{0.05,7} = 1.895$, $t_{0.025,7} = 2.365$)

Solution

$\hat{\mu}_X = \bar{X} = \frac{1}{6}(127 + 144 + 140 + 136 + 118 + 138) = 133.83.$

$\hat{\sigma}_X^2 = s_X^2 = \frac{1}{6-1}(127^2 + 144^2 + 140^2 + 136^2 + 118^2 + 138^2 - 6 \times 133.83^2) = 93.24.$

$\hat{\mu}_Y = \bar{Y} = \frac{1}{8}(125 + 128 + 133 + 141 + 109 + 125 + 126 + 122) = 126.13.$

$\hat{\sigma}_Y^2 = s_Y^2 = \frac{1}{8-1}(125^2 + 128^2 + 133^2 + 141^2 + 109^2 + 125^2 + 126^2 + 122^2 - 8 \times 126.125^2) = 83.55.$

Practice problem

We are interested in investigating any potential difference between the mean blood sugar level of diabetics (μ_X) and that of non-diabetics (μ_Y). To do this we took a sample of six diabetics and found the following blood sugar levels: 127, 144, 140, 136, 118, 138. We also took a sample of eight non-diabetics and found the following blood sugar levels: 125, 128, 133, 141, 109, 125, 126, 122. (a) Estimate $\mu_X - \mu_Y$. (b) Find the 95% CI for $\mu_X - \mu_Y$.

($s_X^2 = 93.24, s_Y^2 = 83.55, t_{0.05,5} = 2.015, t_{0.025,5} = 2.571, t_{0.05,7} = 1.895, t_{0.025,7} = 2.365$)

Solution

$\hat{\mu}_X = \bar{X} = \frac{1}{6}(127 + 144 + 140 + 136 + 118 + 138) = 133.83.$

$\hat{\sigma}_X^2 = s_X^2 = \frac{1}{6-1}(127^2 + 144^2 + 140^2 + 136^2 + 118^2 + 138^2 - 6 \times 133.83^2) = 93.24.$

$\hat{\mu}_Y = \bar{Y} = \frac{1}{8}(125 + 128 + 133 + 141 + 109 + 125 + 126 + 122) = 126.13.$

$\hat{\sigma}_Y^2 = s_Y^2 = \frac{1}{8-1}(125^2 + 128^2 + 133^2 + 141^2 + 109^2 + 125^2 + 126^2 + 122^2 - 8 \times 126.125^2) = 83.55.$

The 95% confidence interval is given as

$$\bar{X} - \bar{Y} \pm t_{0.025,5}\sqrt{\frac{s_X^2}{n} + \frac{s_Y^2}{m}} = (133.83 - 126.13) \pm 2.571\sqrt{\frac{93.24}{6} + \frac{83.55}{8}} = [-5.41, 20.81]$$

Footnote

$$s^2 = \frac{(X_1 - \bar{X})^2 + \ldots + (X_n - \bar{X})^2}{n-1} = \frac{X_1^2 + \ldots + X_n^2 - n(\bar{X})^2}{n-1}$$

Proof:

$$\begin{aligned}
s^2 &= \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1} = \frac{\sum_{i=1}^n (X_i^2 + (\bar{X})^2 - 2X_i \cdot \bar{X})}{n-1} \\
&= \frac{\sum_{i=1}^n X_i^2 + \sum_{i=1}^n (\bar{X})^2 - \sum_{i=1}^n 2X_i \cdot \bar{X}}{n-1} \\
&= \frac{\sum_{i=1}^n X_i^2 + n(\bar{X})^2 - 2n(\bar{X})^2}{n-1} \\
&= \frac{X_1^2 + \ldots + X_n^2 - n(\bar{X})^2}{n-1}.
\end{aligned}$$

When the variances are the same

If $X_1, X_2, ..., X_n$ i.i.d. $\sim N(\mu_X, \sigma^2)$, $Y_1, Y_2, ..., Y_n$ i.i.d. $\sim N(\mu_Y, \sigma^2)$

- If it is known that both populations have the same variance, then we can leverage this information to get a more accurate estimate of by combing both s_X and s_Y to create what is called a pooled estimate of the variance:

$$s_{pool}^2 = \frac{(n-1)s_X^2 + (m-1)s_Y^2}{n+m-2},$$

where $s_X^2 = \frac{(X_1-\bar{X})^2+...+(X_n-\bar{X})^2}{n-1}$, $s_Y^2 = \frac{(Y_1-\bar{Y})^2+...+(Y_m-\bar{Y})^2}{m-1}$.

When the variances are the same

If $X_1, X_2, ..., X_n$ i.i.d. $\sim N(\mu_X, \sigma^2)$, $Y_1, Y_2, ..., Y_n$ i.i.d. $\sim N(\mu_Y, \sigma^2)$

- If it is known that both populations have the same variance, then we can leverage this information to get a more accurate estimate of by combing both s_X and s_Y to create what is called a pooled estimate of the variance:

$$s_{pool}^2 = \frac{(n-1)s_X^2 + (m-1)s_Y^2}{n+m-2},$$

where $s_X^2 = \frac{(X_1-\bar{X})^2+...+(X_n-\bar{X})^2}{n-1}$, $s_Y^2 = \frac{(Y_1-\bar{Y})^2+...+(Y_m-\bar{Y})^2}{m-1}$.

- The pooled estimate is a weighted average of the individual sample estimates. This increases the sample size, which is almost always better!

When the variances are the same

If $X_1, X_2, \ldots, X_n$ i.i.d. $\sim N(\mu_X, \sigma^2)$, $Y_1, Y_2, \ldots, Y_n$ i.i.d. $\sim N(\mu_Y, \sigma^2)$

- If it is known that both populations have the same variance, then we can leverage this information to get a more accurate estimate of by combing both s_X and s_Y to create what is called a pooled estimate of the variance:

$$s_{pool}^2 = \frac{(n-1)s_X^2 + (m-1)s_Y^2}{n+m-2},$$

where $s_X^2 = \frac{(X_1-\bar{X})^2+\ldots+(X_n-\bar{X})^2}{n-1}$, $s_Y^2 = \frac{(Y_1-\bar{Y})^2+\ldots+(Y_m-\bar{Y})^2}{m-1}$.

- The pooled estimate is a weighted average of the individual sample estimates. This increases the sample size, which is almost always better!
- The formula for confidence intervals now becomes: $\bar{X} - \bar{Y} \pm t_{\alpha/2,k}\sqrt{\frac{s_p^2}{n} + \frac{s_p^2}{m}}$. where $k = n+m-2$.

Practice problem

We are interested in investigating any potential difference between the mean blood sugar level of diabetics (μ_X) and that of non-diabetics (μ_Y). To do this we took a sample of six diabetics and found the following blood sugar levels: 127, 144, 140, 136, 118, 138. We also took a sample of eight non-diabetics and found the following blood sugar levels: 125, 128, 133, 141, 109, 125, 126, 122. Assuming the variance for these two populations are the same. (a) Estimate $\mu_X - \mu_Y$. (b) Find the 95% CI for $\mu_X - \mu_Y$.
($s_X^2 = 93.24, s_Y^2 = 83.55, t_{0.05,12} = 1.782, t_{0.025,12} = 2.179, t_{0.05,14} = 1.761, t_{0.025,14} = 2.145$)

Practice problem

We are interested in investigating any potential difference between the mean blood sugar level of diabetics (μ_X) and that of non-diabetics (μ_Y). To do this we took a sample of six diabetics and found the following blood sugar levels: 127, 144, 140, 136, 118, 138. We also took a sample of eight non-diabetics and found the following blood sugar levels: 125, 128, 133, 141, 109, 125, 126, 122. Assuming the variance for these two populations are the same. (a) Estimate $\mu_X - \mu_Y$. (b) Find the 95% CI for $\mu_X - \mu_Y$.
($s_X^2 = 93.24, s_Y^2 = 83.55, t_{0.05,12} = 1.782, t_{0.025,12} = 2.179, t_{0.05,14} = 1.761, t_{0.025,14} = 2.145$)

Solution

$\hat{\mu}_X = \bar{X} = \frac{1}{6}(127 + 144 + 140 + 136 + 118 + 138) = 133.83.$
$\hat{\mu}_Y = \bar{Y} = \frac{1}{8}(125 + 128 + 133 + 141 + 109 + 125 + 126 + 122) = 126.13.$
$s_{pool}^2 = \frac{1}{6+8-2}((6-1) \times 93.24 + (8-1) \times 83.55) = 87.59$

Practice problem

We are interested in investigating any potential difference between the mean blood sugar level of diabetics (μ_X) and that of non-diabetics (μ_Y). To do this we took a sample of six diabetics and found the following blood sugar levels: 127, 144, 140, 136, 118, 138. We also took a sample of eight non-diabetics and found the following blood sugar levels: 125, 128, 133, 141, 109, 125, 126, 122. Assuming the variance for these two populations are the same. (a) Estimate $\mu_X - \mu_Y$. (b) Find the 95% CI for $\mu_X - \mu_Y$.
($s_X^2 = 93.24, s_Y^2 = 83.55, t_{0.05,12} = 1.782, t_{0.025,12} = 2.179, t_{0.05,14} = 1.761, t_{0.025,14} = 2.145$)

Solution

$\hat{\mu}_X = \bar{X} = \frac{1}{6}(127 + 144 + 140 + 136 + 118 + 138) = 133.83.$
$\hat{\mu}_Y = \bar{Y} = \frac{1}{8}(125 + 128 + 133 + 141 + 109 + 125 + 126 + 122) = 126.13.$
$s_{pool}^2 = \frac{1}{6+8-2}((6-1) \times 93.24 + (8-1) \times 83.55) = 87.59$
The 95% confidence interval is given as
$\bar{X} - \bar{Y} \pm t_{0.025,12}\sqrt{s_{pool}^2} = (133.83 - 126.13) \pm 2.776\sqrt{87.59} = [-6.45, 21.85]$

Paired Sample

- Sometimes it appears we have data from two samples with the further feature that there is a natural "pairing" of the data between the two samples.

Paired Sample

- Sometimes it appears we have data from two samples with the further feature that there is a natural "pairing" of the data between the two samples.
- For example, suppose that the data consists on n brother-sister pairs, with blood pressures $X_1, \ldots, X_n$ for the n sisters and blood pressures $Y_1, \ldots, Y_n$ for their respective brothers.

Paired Sample

- Sometimes it appears we have data from two samples with the further feature that there is a natural "pairing" of the data between the two samples.
- For example, suppose that the data consists on n brother-sister pairs, with blood pressures $X_1, \dots, X_n$ for the n sisters and blood pressures $Y_1, \dots, Y_n$ for their respective brothers.
- Thus X_1 and Y_1 are the blood pressures from the the sister and brother in family 1, X_2 and Y_2 are the blood pressures from the the sister and brother in family 2, and so on. The natural pairing is between the sister and the brother in the same family.

Paired Sample

- Sometimes it appears we have data from two samples with the further feature that there is a natural "pairing" of the data between the two samples.
- For example, suppose that the data consists on n brother-sister pairs, with blood pressures $X_1, \ldots, X_n$ for the n sisters and blood pressures $Y_1, \ldots, Y_n$ for their respective brothers.
- Thus X_1 and Y_1 are the blood pressures from the the sister and brother in family 1, X_2 and Y_2 are the blood pressures from the the sister and brother in family 2, and so on. The natural pairing is between the sister and the brother in the same family.
- We could not use the previous CI formula for $\mu_X - \mu_Y$, because $\{X_i\}$ and $\{Y_i\}$ are not independent.

CI for the difference between two means

Let's first consider the case where n is small.

- If $X_1, X_2, ..., X_n$ i.i.d. $\sim N(\mu_X, \sigma_X^2)$, $Y_1, Y_2, ..., Y_n$ i.i.d. $\sim N(\mu_Y, \sigma_Y^2)$. In the paired samples' case, it's not realistic to assume that $\{X_i\}$ and $\{Y_i\}$ are independent.

CI for the difference between two means

Let's first consider the case where n is small.

- If $X_1, X_2, ..., X_n$ i.i.d. $\sim N(\mu_X, \sigma_X^2)$, $Y_1, Y_2, ..., Y_n$ i.i.d. $\sim N(\mu_Y, \sigma_Y^2)$. In the paired samples' case, it's not realistic to assume that $\{X_i\}$ and $\{Y_i\}$ are independent.
- However, it's natural to assume $(X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n)$ are n independent pairs.

CI for the difference between two means

Let's first consider the case where n is small.

- If $X_1, X_2, ..., X_n$ i.i.d. $\sim N(\mu_X, \sigma_X^2)$, $Y_1, Y_2, ..., Y_n$ i.i.d. $\sim N(\mu_Y, \sigma_Y^2)$. In the paired samples' case, it's not realistic to assume that $\{X_i\}$ and $\{Y_i\}$ are independent.
- However, it's natural to assume $(X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n)$ are n independent pairs.
- We denote $D_i = X_i - Y_i$, then $D_i \ i.i.d. \sim N(\mu_X - \mu_Y, \sigma_D^2)$

CI for the difference between two means

Let's first consider the case where n is small.

- If $X_1, X_2, ..., X_n$ i.i.d. $\sim N(\mu_X, \sigma_X^2)$, $Y_1, Y_2, ..., Y_n$ i.i.d. $\sim N(\mu_Y, \sigma_Y^2)$. In the paired samples' case, it's not realistic to assume that $\{X_i\}$ and $\{Y_i\}$ are independent.
- However, it's natural to assume $(X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n)$ are n independent pairs.
- We denote $D_i = X_i - Y_i$, then $D_i \; i.i.d. \sim N(\mu_X - \mu_Y, \sigma_D^2)$
- We estimate $\mu_X - \mu_Y$ by $\bar{D} = \bar{X} - \bar{Y}$

CI for the difference between two means

Let's first consider the case where n is small.

- If $X_1, X_2, ..., X_n$ i.i.d. $\sim N(\mu_X, \sigma_X^2)$, $Y_1, Y_2, ..., Y_n$ i.i.d. $\sim N(\mu_Y, \sigma_Y^2)$. In the paired samples' case, it's not realistic to assume that $\{X_i\}$ and $\{Y_i\}$ are independent.
- However, it's natural to assume $(X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n)$ are n independent pairs.
- We denote $D_i = X_i - Y_i$, then $D_i \; i.i.d. \sim N(\mu_X - \mu_Y, \sigma_D^2)$
- We estimate $\mu_X - \mu_Y$ by $\bar{D} = \bar{X} - \bar{Y}$
- A $(1-\alpha)$ confidence interval for $\mu_X - \mu_Y$ is

$$[\bar{D} - t_{\alpha/2,n-1}\frac{s_D}{\sqrt{n}}, \quad \bar{D} + t_{\alpha/2,n-1}\frac{s_D}{\sqrt{n}}],$$

where $s_D = \sqrt{\frac{(D_1-\bar{D})^2+\ldots+(D_n-\bar{D})^2}{n}}$.

CI for the difference between two means

Let's first consider the case where n is small.

- If $X_1, X_2, ..., X_n$ i.i.d. $\sim N(\mu_X, \sigma_X^2)$, $Y_1, Y_2, ..., Y_n$ i.i.d. $\sim N(\mu_Y, \sigma_Y^2)$. In the paired samples' case, it's not realistic to assume that $\{X_i\}$ and $\{Y_i\}$ are independent.
- However, it's natural to assume $(X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n)$ are n independent pairs.
- We denote $D_i = X_i - Y_i$, then D_i $i.i.d. \sim N(\mu_X - \mu_Y, \sigma_D^2)$
- We estimate $\mu_X - \mu_Y$ by $\bar{D} = \bar{X} - \bar{Y}$
- A $(1-\alpha)$ confidence interval for $\mu_X - \mu_Y$ is

$$[\bar{D} - t_{\alpha/2,n-1}\frac{s_D}{\sqrt{n}}, \quad \bar{D} + t_{\alpha/2,n-1}\frac{s_D}{\sqrt{n}}],$$

  where $s_D = \sqrt{\frac{(D_1-\bar{D})^2+\ldots+(D_n-\bar{D})^2}{n}}$.
- If n is large, we can simply replace $t_{\alpha/2,n-1}$ with $z_{\alpha/2}$ and remove the normality assumptions.

Example

Suppose that we take the blood pressures of $n = 12$ women and their brothers, and get the following blood pressure reading:

Family	1	2	3	4	5	6	7	8	9	10	11	12
Sister	107	134	111	141	121	118	145	110	164	126	148	132
Brother	110	136	115	140	124	119	148	113	168	129	148	137

Construct the 95% CI for the difference of the mean blood pressure of men and women. ($s^2_{sister} = 307$, $s^2_{brother} = 299$, $s^2_{diff} = 3$, $t_{0.025,11} = 2.306$, $t_{0.05,11} = 1.860$, $t_{0.025,12} = 2.262$, $t_{0.05,12} = 1.833$)

Example

Suppose that we take the blood pressures of $n = 12$ women and their brothers, and get the following blood pressure reading:

Family	1	2	3	4	5	6	7	8	9	10	11	12
Sister	107	134	111	141	121	118	145	110	164	126	148	132
Brother	110	136	115	140	124	119	148	113	168	129	148	137

Construct the 95% CI for the difference of the mean blood pressure of men and women. ($s^2_{sister} = 307$, $s^2_{brother} = 299$, $s^2_{diff} = 3$, $t_{0.025,11} = 2.306$, $t_{0.05,11} = 1.860$, $t_{0.025,12} = 2.262$, $t_{0.05,12} = 1.833$)

Answer

Family	1	2	3	4	5	6	7	8	9	10	11	12
Difference	3	2	4	-1	3	1	3	3	4	3	0	5

- $\bar{D} = 2.5,\ s^2_D = 3,\ t_{0.025,11} = 2.306$

Example

Suppose that we take the blood pressures of $n = 12$ women and their brothers, and get the following blood pressure reading:

Family	1	2	3	4	5	6	7	8	9	10	11	12
Sister	107	134	111	141	121	118	145	110	164	126	148	132
Brother	110	136	115	140	124	119	148	113	168	129	148	137

Construct the 95% CI for the difference of the mean blood pressure of men and women. ($s^2_{sister} = 307$, $s^2_{brother} = 299$, $s^2_{diff} = 3$, $t_{0.025,11} = 2.306$, $t_{0.05,11} = 1.860$, $t_{0.025,12} = 2.262$, $t_{0.05,12} = 1.833$)

Answer

Family	1	2	3	4	5	6	7	8	9	10	11	12
Difference	3	2	4	-1	3	1	3	3	4	3	0	5

- $\bar{D} = 2.5,\ s^2_D = 3,\ t_{0.025,11} = 2.306$
- $2.5 \pm 2.306 \cdot \frac{\sqrt{3}}{\sqrt{12}} = [1.35, 3.65]$.

Summary

- CI for the binomial parameter p when the sample size n is large
- CI for the mean μ when the sample size n is small and samples are normal
- CI for the difference between two binomial parameters $p_1 - p_2$
- CI for the difference between two means $\mu_X - \mu_Y$
  - Large sample size
  - Small sample size
  - When we have additional information that the variances are the same
  - Paired samples

Homework

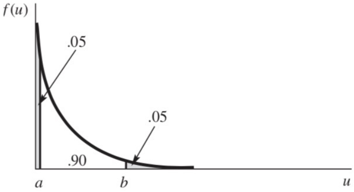
8.56, 8.60, 8.66, 8.70, 8.82, 8.85

Other confidence intervals

- A useful method for deriving confidence intervals is to use a pivotal quantity
- A pivotal quantity
  - is a function of the sample data, the unknown target parameter, and no other quantities
  - has a distribution that does not depend on the target parameter
- Example:
- We randomly sample an observation from an exponential distribution with unknown mean θ. Find a formula for a 90% CI for θ.
- If $Y \sim Exp(\theta)$, then $f_Y(y) = \frac{1}{\theta} e^{-y/\theta}$ for $y \geq 0$.
- Let $U = \frac{Y}{\theta}$, we have $f_U(u) = f_Y(u\theta) \cdot \frac{dy}{du} = \frac{1}{\theta} e^{-u} \cdot \theta = e^{-u}$ for $u > 0$, that is, $U \sim Exp(1)$
- Thus, we can use $U = \frac{Y}{\theta}$ as a pivotal quantity

Example

- We need to find two numbers a and b such that
$$\mathbb{P}(a \leq U \leq b) = 90\%$$
- One way to do this is to choose a and b to satisfy $\mathbb{P}(U < a) = \mathbb{P}(U > b) = 5\%$



- $1 - e^{-a} = .05$ and $e^{-b} = .05$, equivalently, $a = 0.051$, $b = 2.996$
- $0.9 = \mathbb{P}(0.051 \leq \frac{Y}{\theta} \leq 2.996) = \mathbb{P}(\frac{Y}{2.996} \leq \theta \leq \frac{Y}{0.051})$

Example

Example

We randomly sample 10 observations $X_1, ..., X_{10}$ from an exponential distribution with unknown mean θ. Find a formula for a 90% CI for θ.

Answer

- Hint: If $X_i \sim Exp(\theta_i)$ then $\min\{X_1, ..., X_n\} \sim Exp(\frac{1}{\frac{1}{\theta_1}+...+\frac{1}{\theta_n}})$.
- $\frac{X_1}{\theta}, ... \frac{X_{10}}{\theta} \overset{i.i.d.}{\sim} Exp(1)$
- $U = \min\{\frac{X_1}{\theta}, ... \frac{X_{10}}{\theta}\} \sim Exp(\frac{1}{10})$
- Find a, b such that $\mathbb{P}(U < a) = 1 - e^{-10a} = .05$ and $\mathbb{P}(U > b) = e^{-10b} = .05$, equivalently, $a = 0.005,\ b = 0.300$
- $0.9 = \mathbb{P}(0.005 < U < 0.300) = \mathbb{P}(0.005 < \frac{\min\{X_1,...,X_{10}\}}{\theta} < 0.300) = \mathbb{P}(\frac{\min\{X_1,...,X_{10}\}}{0.300} < \theta < \frac{\min\{X_1,...,X_{10}\}}{0.005})$

Example

Example

Suppose that we take a sample Y from a uniform distribution defined on the interval $[0, \theta]$, where θ is unknown. Find a 95% lower confidence bound for θ, that is, find $L(Y)$, such that $\mathbb{P}(\theta \geq L(Y)) = 0.95$.

Answer

- Consider $U = \frac{Y}{\theta}$
- $U \sim U(0, 1)$
- $\mathbb{P}(U < 0.95) = 0.95$
- $0.95 = \mathbb{P}(U < 0.95) = \mathbb{P}(\frac{Y}{\theta} < 0.95) = \mathbb{P}(\frac{Y}{0.95} \leq \theta)$

Confidence interval for the variance

- Suppose $Y_1, ..., Y_n \overset{i.i.d.}{\sim} N(\mu, \sigma^2)$ with μ and σ^2 unknown. We seek a $100(1-\alpha)\%$ CI for σ^2
- Recall that, for $S^2 = \frac{1}{n-1}\sum_{i=1}^n (Y_i - \bar{Y})^2$, we have
$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}.$$
- As a result,
$$\mathbb{P}(\chi^2_{1-(\alpha/2),n-1} \leq \frac{(n-1)S^2}{\sigma^2} \leq \chi^2_{\alpha/2,n-1}) = 1-\alpha$$
- A $100(1-\alpha)\%$ Confidence Interval for σ^2:
$$\mathbb{P}(\frac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi^2_{1-(\alpha/2),n-1}}) = 1-\alpha.$$

Example

Example

Suppose the maturation times for a flower species are $N(\mu, \sigma^2)$. If a random sample of $n = 13$ seeds yielded $s^2 = 10.7$, then what is a 90% CI for σ^2? ($\chi^2_{0.05,12} = 21.03, \chi^2_{0.95,12} = 5.23$)

Answer

- $$\mathbb{P}\left(\frac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}} \le \sigma^2 \le \frac{(n-1)S^2}{\chi^2_{1-(\alpha/2),n-1}}\right) = 1 - \alpha.$$
- $\frac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}} = \frac{12\times 10.7}{21.03} = 6.11$
- $\frac{(n-1)S^2}{\chi^2_{1-\alpha/2,n-1}} = \frac{12\times 10.7}{5.23} = 24.55$
- The 90% CI is $[6.11, 24.55]$.

Homework