1

Lecture 23

Topics: Confidence intervals

Confidence intervals

Hypothesis testing is a tool for showing some value of a population parameter θ is Not plausible.

Confidence intervals is a tool ~~for~~ for showing which θ's <u>are</u> plausible.

All confidence intervals (CIs) done in this class follow a similar structure (based on z-scores of course !! ☺)

$$\hat{\theta} \pm (\text{some number}) \cdot sd(\hat{\theta})$$

This number is determined by how confident you want your interval to wrap around the true θ.
Larger number $\Rightarrow$ more confident
Smaller number $\Rightarrow$ less confident

This will typically be approximated

2

How does this come from z-scores?

In the hypothesis testing setup the quick guess method uses z-score of the form

$$\frac{\hat{\theta} - \theta}{sd(\hat{\theta})} \approx N(0,1)$$

$\therefore$ with about 95% confidence

$\frac{\hat{\theta} - \theta}{sd(\hat{\theta})}$ is btwn -2 & 2

i.e. $\hat{\theta} - \theta$ is btwn $-2sd(\hat{\theta})$ & $2sd(\hat{\theta})$

i.e. θ is btwn $\underbrace{\hat{\theta} - 2sd(\hat{\theta}) \ \& \ \hat{\theta} + 2sd(\hat{\theta})}_{\text{an approx 95\% CI}}$

The "quick guess" Method for a 68% CI, 95% CI & 99% CI

68% CI for θ: $\hat{\theta} \pm sd(\hat{\theta})$

95% CI for θ: $\hat{\theta} \pm 2\, sd(\hat{\theta})$

99% CI for θ: $\hat{\theta} \pm 3\, sd(\hat{\theta})$

where $sd(\hat{\theta})$ approximates the denominator of the approximate z-scores used in the hypothesis testing cases (1), (2), (3) & (4). from the previous lecture.

③

Population Parameter	approximate sd($\hat{\theta}$) for quick guess CI's.
(1) Population average μ (population s.d. σ known)	$\sigma/\sqrt{n}$
(2) Population average μ (population s.d. σ unknown)	$\hat{\sigma}/\sqrt{n}$
(3) Population proportion p	$\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
(4) The difference btwn two population averages $\mu_x - \mu_y$	$\sqrt{\frac{\hat{\sigma}_x^2}{n} + \frac{\hat{\sigma}_y^2}{m}}$

The official Confidence intervals.

The quick guess method gives approximate 68%, 95% & 99% CI's. what if you want an exact 85% CI, for example?

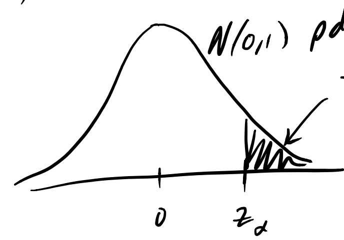
More abstractly how do you find a $100(1-2\alpha)\%$ confidence interval, where α is a number btwn 0 & 1. For example,

%	$\alpha = \frac{1}{2}(1-\frac{\%}{100})$
80	0.1
85	0.075
90	0.05
95	0.025
99	0.005

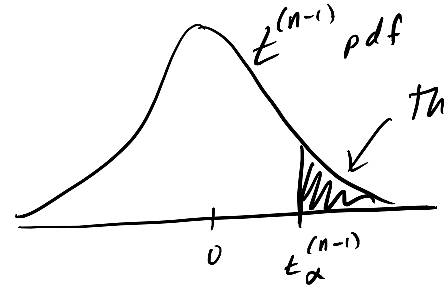
④

Population Parameter and assumptions	$100(1-2\alpha)\%$ CI's.
(1) Population average μ; population s.d. σ known; $n > 30$	$\hat{\mu} \pm z_\alpha \sigma/\sqrt{n}$
(2) Population average μ; population s.d. σ unknown; populatis is Normal	$\hat{\mu} \pm t_\alpha^{(n-1)} \hat{\sigma}/\sqrt{n}$
(3) Population proportion p; $n > 30$	$\hat{p} \pm z_\alpha \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
(4) The difference btwn two population averages $\mu_x - \mu_y$; $\sigma_x = \sigma_y$ & both populations are Normal.	$\hat{\mu}_x - \hat{\mu}_y \pm t_\alpha^{(n+m-2)} \hat{\sigma}_p \sqrt{\frac{1}{n}+\frac{1}{m}}$

where z_α is the "z-score" of $N(0,1)$ with α probability to the right



and $t_\alpha^{(n-1)}$ is the "z-score" of $t^{(n-1)}$ with α probability to the right



example:

warehouse storing light bulbs.
You want to estimate the ave lifetime (in hrs) accross the whole warehouse.
Let this average be denoted μ.
Sample $n = 25$ bulbs ... let

$$X_1, X_2, \ldots, X_n$$

denote their lifetime (in hrs).

Suppose $\hat{\mu} = \frac{1}{n}(X_1 + \cdots + X_n) = 855$

$$\hat{\sigma} = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (X_i - \hat{\mu})^2} = 42$$

Find a 95 % CI for μ assuming the histogram of all lightbulbs in the warehouse is approx normal.

Quick guess method

$$\hat{\mu} \pm 2\, sd(\hat{\mu}) \cong \hat{\mu} \pm 2 \frac{\hat{\sigma}}{\sqrt{n}}$$

$$= 855 \pm 2 \frac{42}{\sqrt{25}}$$

$$= 855 \pm 16.8$$

Official method (2)

$$\hat{\mu} \pm t_{\alpha}^{(n-1)} \frac{\hat{\sigma}}{\sqrt{n}} = 855 \pm t_{\alpha}^{(n-1)} \frac{42}{\sqrt{25}}$$

where $95\% \leftrightarrow \alpha = 0.025 = \frac{1}{2}(1 - \frac{95}{100})$

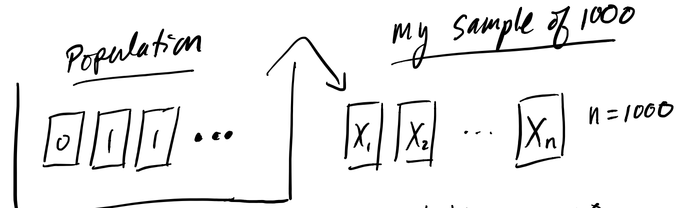
From the table $t_{0.025}^{(24)} = 2.06$ gives

$$855 \pm 2.06 \frac{42}{\sqrt{25}} = 855 \pm 17.3$$

example:

I just interviewed 1000 UC Davis students & only 34% said they liked statistics.
Make a 99% confidence interval for the proportion of all UC Davis students who like statistics.

Here is a box model for our problem



A ticket for each UCD student.
$\boxed{1}$ if they like stats
$\boxed{0}$ if Not
Let p be the proportion of "likes" in the box

Each X_i is 1 or 0
Let $\hat{p}$ be the proportion of likes in the sample

$$\underbrace{\hat{p} = 0.34}_{\text{data.}}$$

Quick guess method

$$\hat{p} \pm 3\, sd(\hat{p}) \overset{MF2}{=} \hat{p} \pm 3 \sqrt{\frac{p(1-p)}{n}}$$

$$\cong \hat{p} \pm 3 \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

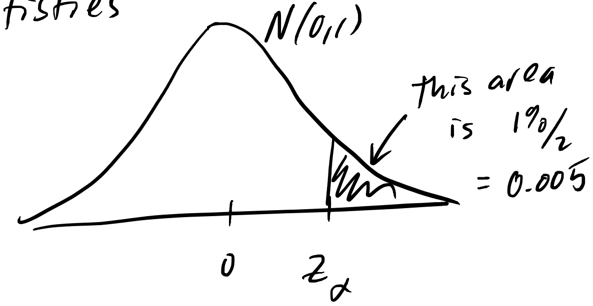
$$= 0.34 \pm 3 \sqrt{\frac{0.34(1-0.34)}{1000}}$$

$$= 0.34 \pm 0.045$$

Official Method (3)

$\hat{p} \pm z_\alpha \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ where z_α

satisfies



Since setting $99\% = 100(1-2\alpha)\%$

corresponds to $\alpha = 0.05$.

$\therefore\ z_\alpha = 2.58$ so that

$$\hat{p} \pm z_\alpha \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.34 \pm \underbrace{0.039}$$

This looks quite different from the quick guess but Not when compared to 0.34

example

Study the relationship between the number of social security recipients (per 1000) in southern vrs northern cities

Data:

Sample 20 southern cities & 10 Northern cities.

For each city measure the number of social security recipients (per 1000). Here is the data

	South	North
Samples:	$X_1, \ldots, X_{20}$	$Y_1, \ldots, Y_{10}$
Sample Ave:	$\bar{X} = 178.5$	$\bar{Y} = 156.0$
Sample s.d:	$s_x = 68.0$	$s_y = 19.5$

Make a 95% CI for the difference in population averages

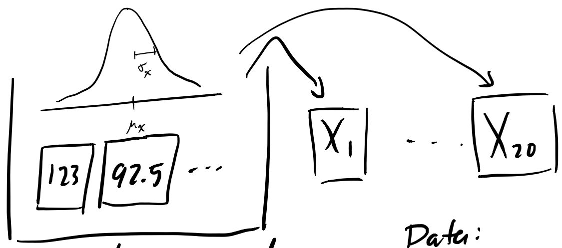
$$\mu_x - \mu_y$$

ave value over all southern cities

ave value over all Northern cities

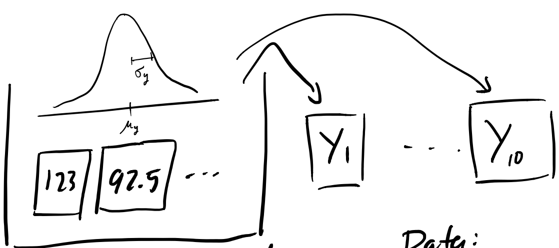
Assume both populations are normal and $\sigma_x = \sigma_y$.

Box Model picture



- tickets represent **Southern** cities.
- The number on the ticket represent # of Soc Security recipients per 1000
- μ_x = Ave ticket #
  σ_x = s.d. ticket #

Data:
$\hat{\mu}_x = 178.5$
$\hat{\sigma}_x = 68.0$



- tickets represent **Northern** cities.
- The number on the ticket represent # of Soc Security recipients per 1000
- μ_y = Ave ticket #
  σ_y = s.d. ticket #

Data:
$\hat{\mu}_y = 156.0$
$\hat{\sigma}_y = 19.5$

Quick guess method

$$(\hat{\mu}_x - \hat{\mu}_y) \pm 2 \underbrace{\sqrt{\frac{\hat{\sigma}_x^2}{20} + \frac{\hat{\sigma}_y^2}{10}}}$$

for 95% CI

approximates $sd(\hat{\mu}_x - \hat{\mu}_y) \overset{MF2}{=} \sqrt{\frac{\sigma_x^2}{20} + \frac{\sigma_y^2}{10}}$

$$= (178.5 - 156.0) \pm 2\sqrt{\frac{68^2}{20} + \frac{19.5^2}{10}}$$

$$= 22.5 \pm 32.8$$

"official" method

$$(\hat{\mu}_x - \hat{\mu}_y) \pm t_\alpha^{(20+10-2)} \hat{\sigma}_p \sqrt{\frac{1}{20} + \frac{1}{10}}$$

where $95\% = 100(1-2\alpha)\%$ corresponds to $\alpha = 0.025$ so

$$t_{0.025}^{(28)} = 2.05$$

and

$$\hat{\sigma}_p = \sqrt{\frac{20-1}{20+10-2}\hat{\sigma}_x^2 + \frac{10-1}{20+10-2}\hat{\sigma}_y^2}$$

$$= 57.1$$ sorta like an ave of $\hat{\sigma}_x = 68.0$ & $\hat{\sigma}_y = 19.5$

$\therefore$

$$(\hat{\mu}_x - \hat{\mu}_y) \pm t_\alpha^{(20+10-2)} \hat{\sigma}_p \sqrt{\frac{1}{20} + \frac{1}{10}}$$

$$= 22.5 \pm \underbrace{45.33}$$

This looks different from the quick guess method but the conclusion is the same, i.e. there is no conclusive evidence that Northern & Southern cities are different.