
Lecture 07: Estimates & Sample Size

— Applied Statistics - STAN - —
5.37 & 5.38

Lecturer: Erika Siregar

What to Learn Today

- Point Estimate
- Margin of Error
- Confidence Interval
- Calculating the sample size required to estimate population parameter.
 - **Populasi** : semua objek yang ingin kita teliti → all → **parameter**
 - Sample: sebagian → statistik
 - μ > < \bar{x}
 -

Activities in Statistics

1. Descriptive → mean, median, modus, varians, frequency table, etc.
2. **Inferential**
 - a. Use sample data to estimate population parameters → proportion, mean, variance
 - b. Use sample data to test hypotheses → **next week!**

How to Estimate population parameters?

- Point Estimate
 - Single value used to estimate the real population parameter value
 - μ vs \bar{x} , σ vs s , σ^2 vs s^2 .
- Dalam menghitung estimator kita perlu menentukan peluang error yang bisa kita tolerir (α) dan tingkat kepercayaan atau confidence levelnya ($1 - \alpha$).
 - α = peluang error \rightarrow how much error we are willing to afford? \rightarrow 1%, 5%, 10%
 - **Confidence level (CL) = $1 - \alpha$** \rightarrow the probability that shows how confident we are that the CI will actually contain the true population parameter value.
 - Misal: $\alpha = 0.05 \rightarrow$ CL = 0.95 atau 95%.
 - α and CL are given.
- Sebuah Estimator pasti memiliki selisih dengan nilai yang sebenarnya \rightarrow **margin of error (E).**

Most Common Confidence Levels	Corresponding Values of α
90% (or 0.90) confidence level:	$\alpha = 0.10$
95% (or 0.95) confidence level:	$\alpha = 0.05$
99% (or 0.99) confidence level:	$\alpha = 0.01$

Margin of Error (E)

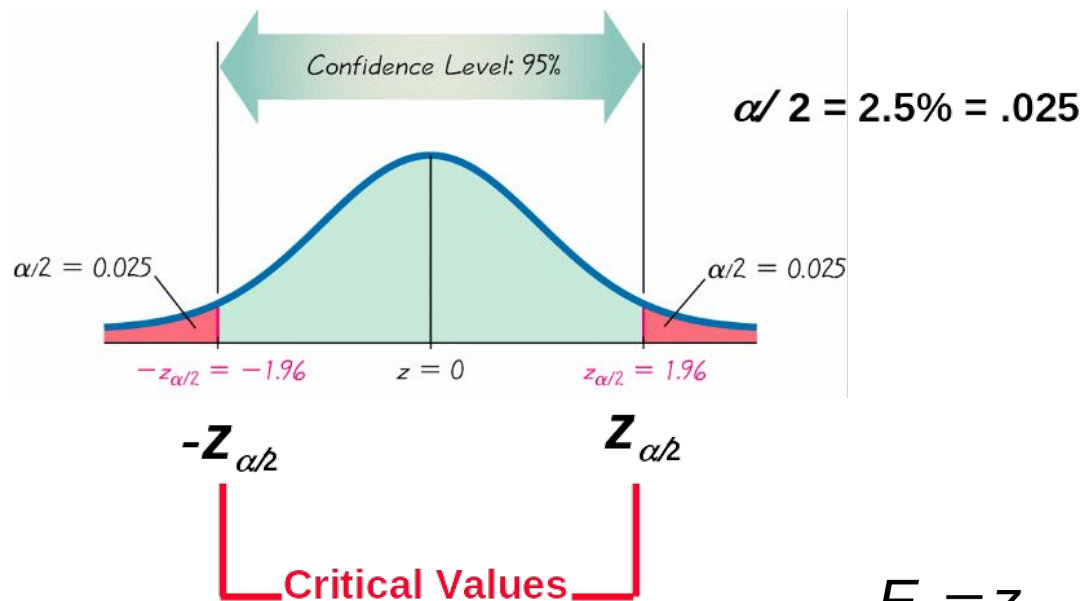
- maximum error of the estimate
- |estimator - true population value|
- **Critical value** * **sd** of sample proportions

- $$E = z_{\alpha/2} \sqrt{\frac{\hat{p}\hat{q}}{n}}$$

- p = population proportion
- \hat{p} = sample proportion
- n = number of sample values
- E = margin of error
- $z_{\alpha/2}$ = z score separating an area of $\alpha/2$ in the right tail of the standard normal distribution

Critical Value

- **critical values** = values that separate the likely and the unlikely → a pair (left tail & right tail)



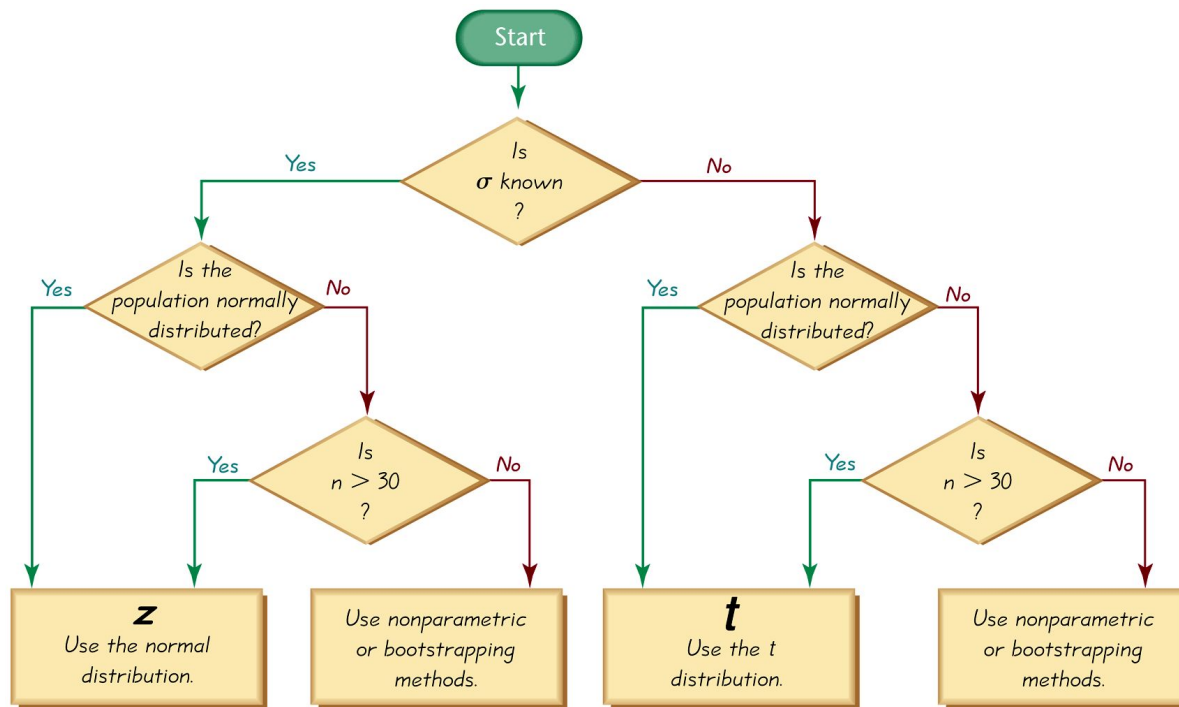
In this picture, the critical value is calculated using **normal distribution (z)**. But this is not always the case.

In other cases, the probability distribution **t and chi-square** are used.

$$E = z_{\alpha/2} \sqrt{\frac{\hat{p}\hat{q}}{n}}$$

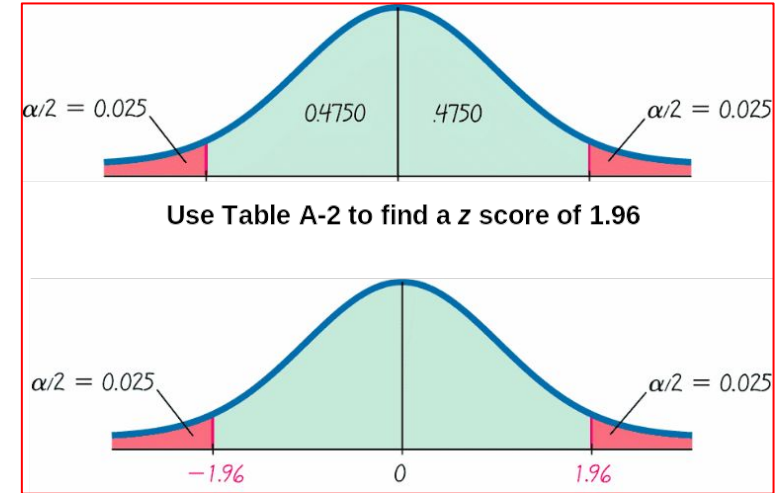
Critical Value (2)

- Calculating critical value depends on the probability distribution being used.



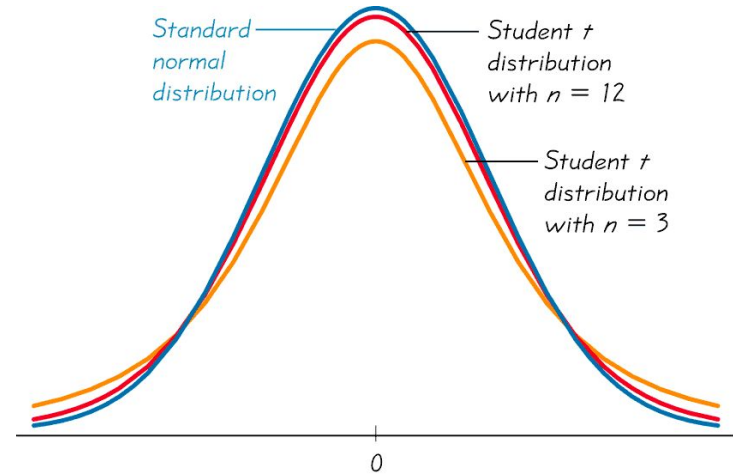
Critical Value (3)

- Z score is also called The z score separating the **right-tail** region is commonly **denoted** by $z_{\alpha/2}$
 - associated with a sample proportion has a probability of $\alpha/2$ of falling in the right tail.
- The z score separating the **left-tail** region is commonly denoted by $z_{-\alpha/2}$.
 - associated with a sample proportion has a probability of $\alpha/2$ of falling in the left tail.



Student t Distribution

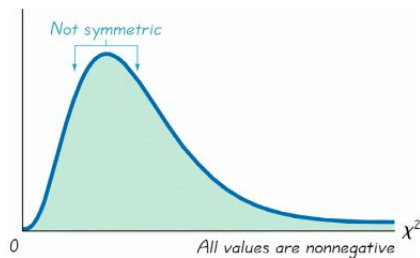
1. The Student t distribution is different for different sample sizes (see the following slide, for the cases $n = 3$ and $n = 12$).
2. The Student t distribution has the same general **symmetric bell shape** as the standard normal distribution but it reflects the greater variability (with **wider distributions**) that is expected with small samples.
3. The Student t distribution has a **mean of $t = 0$** (just as the standard normal distribution has a mean of $z = 0$).
4. The **standard deviation** of the Student t distribution **varies with the sample size and is greater than 1** (unlike the standard normal distribution, which has a = 1).
5. As the sample size **n gets larger**, the **Student t distribution gets closer to the normal distribution**.
6. To obtain value:
 - a. Baca table
 - b. Atau gunakan teknologi: R, excel, dll



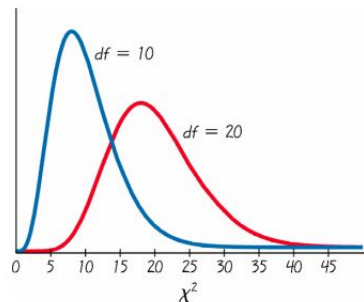
Degree of freedom (d.o.f)
= jumlah sample yang
valuenya bisa bervariasi
→ $n = 5 \rightarrow \mathbf{x\ bar = 3} \rightarrow x_1$
 $= 10, x_2 = 8, x_3 = 15, x_4 = 6,$
 $x_5 \rightarrow \text{d.o.f} = 5 - 1 = 4$

Chi Square Distribution (χ^2)

- The chi-square distribution is asymmetric, unlike the normal and Student t distributions. → thus, we need to find **2 critical values (left and right)**.
- The values of chi-square can be zero or positive, but they cannot be negative.
- The chi-square distribution is different for each number of dof, which is **df = n - 1**. As the **number of degrees of freedom increases**, the **chi-square distribution approaches a normal distribution**.
- Value of χ^2 can be obtained from **table or R 'qchisq(prob, df)'**.



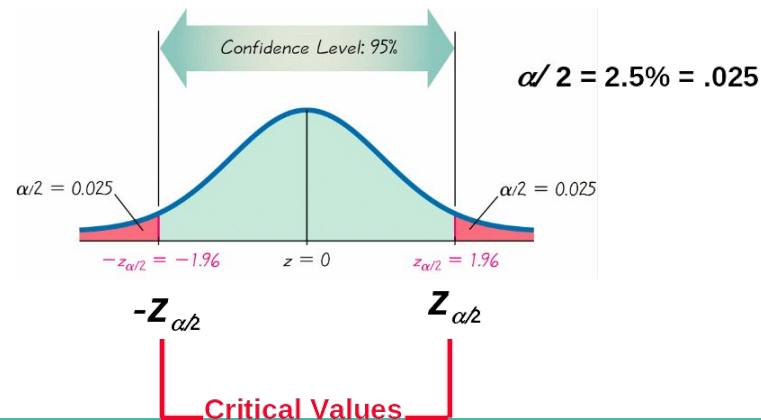
Chi-Square Distribution



Chi-Square Distribution for
df = 10 and df = 20

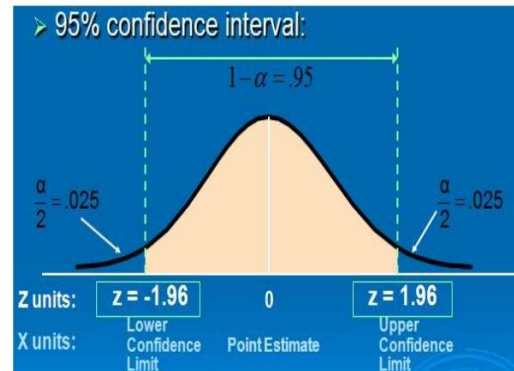
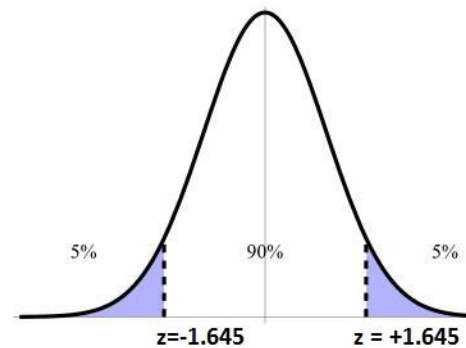
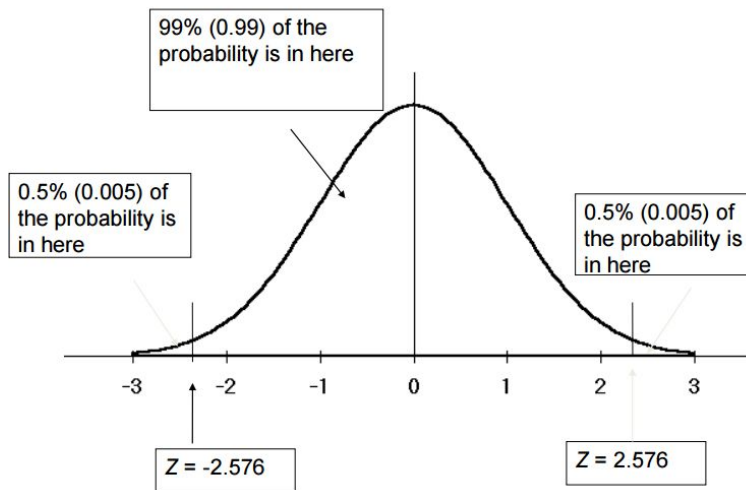
Confidence Interval

- With E & confidence level, we can compute CI
- Confidence Interval (CI)
 - CI = interval yang kita percayai mengandung nilai parameter populasi yang sebenarnya → perlu confidence level
 - Format CI: **Batas bawah** < p < **batas atas**
 - **(estimator - E) < p < (estimator + E)**
 - **Contoh: $0.405 < \mu < 0.455$** → confidence level = 95%
 - **we are 95% confident that the interval from 0.405 to 0.455 actually does contain the true value of the population proportion p.**




Several Options for Confidence Level

Common Values	
Confidence Level	Value of $Z_{\alpha/2}$
80%	1.28
90%	1.645
95%	1.96
98%	2.33
99%	2.58
99.8%	3.08
99.9%	3.27



Calculating the sample size required to estimate population parameter.

- The more samples we have, the more confident we are in estimating the population parameter.
- But how many samples that we must, at least, have to be confident about our estimation?

$$E = z_{\alpha/2} \sqrt{\frac{\hat{p}\hat{q}}{n}}$$


$$n = \frac{(Z_{\alpha/2})^2 \hat{p} \hat{q}}{E^2}$$

When \hat{p} is unknown \longrightarrow

$$n = \frac{(Z_{\alpha/2})^2 0.25}{E^2}$$

Finding the Point Estimate and E from a CI

Point estimate of p :

$$p = \frac{(\text{upper confidence limit}) + (\text{lower confidence limit})}{2}$$

$$\circ \quad 0.405 < p < 0.455$$

Margin of Error:

$$E = \frac{(\text{upper confidence limit}) - (\text{lower confidence limit})}{2}$$

Round-off Rule

- **CI limits:** 3 significant digits (**3 angka di belakang koma**)
- Sample size (n) → dibulatkan ke atas → $n = 35.7 \rightarrow 38$, $100.03 \rightarrow 101$ → $999.17 \rightarrow 1000$

\bar{x}

Summary of CI and Point Estimate

No	Jenis	Point estimate	E	CI	n
1	Proporsi (p)	\hat{p}	$E = z_{\alpha/2} \sqrt{\frac{\hat{p}\hat{q}}{n}}$ critical_value * sigma/akar n	$\hat{p} - E < p < \hat{p} + E$ Or $\hat{p} \pm E$ or $(\hat{p} - E, \hat{p} + E)$	$n = \frac{(z_{\alpha/2})^2 \hat{p} \hat{q}}{E^2}$
					If \hat{p} is unknown $n = \frac{(z_{\alpha/2})^2 0.25}{E^2}$
2	Mean (μ) → When σ is known	\bar{X}	$E = z_{\alpha/2} \sigma / \sqrt{n}$	$\bar{X} - E < \mu < \bar{X} + E$ Or $(\bar{X} - E, \bar{X} + E)$ Or $\bar{X} \pm E$	$n = \left[\frac{(z_{\alpha/2}) \cdot \sigma}{E} \right]^2$

Summary of CI and Point Estimate

No	Jenis	Point Estimate	E	CI	n
3	Mean (μ) → when σ is unknown	\bar{X}	$E = t_{\alpha/2} \frac{s}{\sqrt{n}}$ <p>where $t_{\alpha/2}$ has $n - 1$ dof</p>	<div style="border: 1px solid red; padding: 5px; display: inline-block;"> $\bar{X} - E < \mu < \bar{X} + E$ </div> <p>Use student-t distribution</p> $t = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}}$ <p>With critical value: $t_{\alpha/2}$</p> <p>The value for t can be found in t table (just like that of z). We must know about the degree of freedom (dof). dof is the number of samples whose values can vary after imposing certain restrictions on all data values. → Thus, dof = n - 1.</p>	

Summary of CI and Point Estimate

No	Jenis	Point estimate	E	CI	n
4	Variance (σ^2)	S^2	$E = t_{\alpha/2} \frac{s}{\sqrt{n}}$ <p>where $t_{\alpha/2}$ has $n - 1$ dof</p>	$\frac{(n-1)s^2}{\chi_R^2} < \sigma^2 < \frac{(n-1)s^2}{\chi_L^2}$ <p>Use Chi-Square distribution (χ^2)</p> $\chi^2 = \frac{(n-1)s^2}{\sigma^2}$ <p>dof = n - 1.</p>	

Summary of CI and Point Estimate

No	Jenis	Point estimate	E	CI	n
5	Standard Deviation (σ)	S	$E = t_{\alpha/2} \frac{s}{\sqrt{n}}$ <p>where $t_{\alpha/2}$ has $n - 1$ dof</p>	$\sqrt{\frac{(n-1)s^2}{\chi_R^2}} < \sigma < \sqrt{\frac{(n-1)s^2}{\chi_L^2}}$ <p>Use Chi-Square distribution (χ^2)</p> $\chi^2 = \frac{(n-1)s^2}{\sigma^2}$ <p>dof = n - 1.</p>	Use the table

Table for finding sample size to calculate σ^2

Sample Size for σ^2		Sample Size for σ	
To be 95% confident that s^2 is within	of the value of σ^2 , the sample size n should be at least	To be 95% confident that s is within	of the value of σ , the sample size n should be at least
1%	77,208	1%	19,205
5%	3,149	5%	768
10%	806	10%	192
20%	211	20%	48
30%	98	30%	21
40%	57	40%	12
50%	38	50%	8
To be 99% confident that s^2 is within	of the value of σ^2 , the sample size n should be at least	To be 99% confident that s is within	of the value of σ , the sample size n should be at least
1%	133,449	1%	33,218
5%	5,458	5%	1,336
10%	1,402	10%	336
20%	369	20%	85
30%	172	30%	38
40%	101	40%	22
50%	68	50%	14

Example:

A Pew Research Center poll of **1501** randomly selected U.S. adults showed that **70%** of the respondents believe in global warming. The sample results are **n = 1501**, and $\hat{p} = 0.70$.

Find the **95%** confidence interval estimate of the population proportion p .

Answers:

$$E = z_{\alpha/2} \sqrt{\frac{\hat{p}\hat{q}}{n}} = 1.96 \sqrt{\frac{(0.70)(0.30)}{1501}}$$

$$E = 0.023183$$

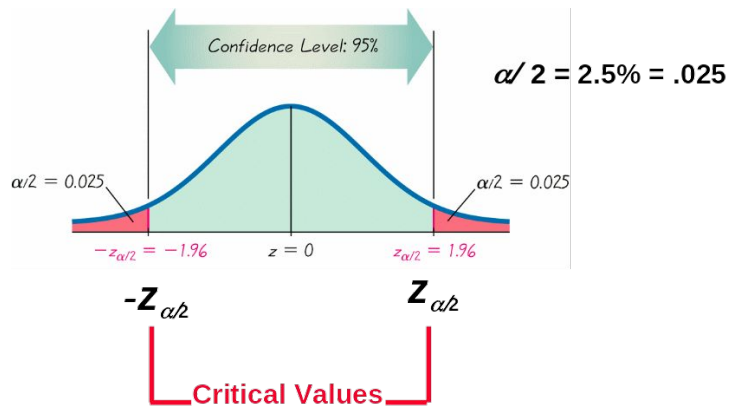
CI = estimator - E < parameter < estimator + E

Calculate the CI

$$\hat{p} - E < p < \hat{p} + E$$

$$0.70 - 0.023183 < p < 0.70 + 0.023183$$

$$0.677 < p < 0.723 \rightarrow \text{Round off to 3 significant digits}$$



Example

The government wants to determine the current percentage of Indonesian adults who now use the Internet. **How many adults** must be surveyed in order to be **95%** confident that the sample percentage is in error by **no more than three percentage points**?

- a. In 2006, **73%** of adults used the internet
- b. No known possible value of the proportion.

$$E = z_{\alpha/2} \sqrt{\frac{\hat{p}\hat{q}}{n}} = 1.96 \sqrt{\frac{(0.70)(0.30)}{1501}}$$

$$E = 0.023183$$

$$\begin{aligned} n &= \frac{(z_{\alpha/2})^2 \hat{p}\hat{q}}{E^2} \\ &= \frac{(1.96)^2 (0.73)(0.27)}{(0.03)^2} \\ &= 841.3104 \\ &= 842 \end{aligned}$$

Answer for a:

$$\hat{p} = 0.73 \text{ and } \hat{q} = 1 - \hat{p} = 0.27$$

$$\alpha = 0.05 \text{ so } z_{\alpha/2} = 1.96$$

$$E = 0.03$$

$$n = \frac{(z_{\alpha/2})^2 \hat{p}\hat{q}}{E^2}$$
$$= \frac{(1.96)^2 (0.73)(0.27)}{(0.03)^2}$$

$$= 841.3104$$

$$= 842$$

To be 95% confident that our sample percentage is within three percentage points of the true percentage for all adults, we should obtain a simple random sample of 842 adults.

Do the same thing for CL 90% and 99%.

Answer for b:

$$\alpha = 0.05 \quad \text{so} \quad z_{\alpha/2} = 1.96$$

$$E = 0.03$$

$$n = \frac{(z_{\alpha/2})^2 \cdot 0.25}{E^2}$$

$$n = \frac{(z_{\alpha/2})^2 \cdot 0.25}{E^2}$$

$$= \frac{(1.96)^2 \cdot 0.25}{(0.03)^2}$$

$$= 1067.1111$$

$$= 1068$$

To be 95% confident that our sample percentage is within three percentage points of the true percentage for all adults, we should obtain a simple random sample of 1068 adults.

Do the same thing for confidence level of 90% and 99%

Example: estimator of mean

People have died in boat and aircraft accidents because an obsolete estimate of the mean weight of men was used. In recent decades, the mean weight of men has increased considerably, so **we need to update our estimate of that mean** so that boats, aircraft, elevators, and other such devices do not become dangerously overloaded. We obtain these sample statistics for the simple random sample: **$n = 40$** and **$\bar{X} = 172.55 \text{ lb}$** . Research from several other sources suggests that the **population** of weights of men has a **standard deviation given by $= 26 \text{ lb}$** .

- Find the **best point estimate of the mean** weight of the population of all men.
- Construct a **95% confidence interval** estimate of the mean weight of all men.
- What do the results suggest about the mean weight of **166.3 lb** that was used to determine the safe passenger capacity of water vessels in 1960 (as given in the National Transportation and Safety Board safety recommendation M-04-04)?

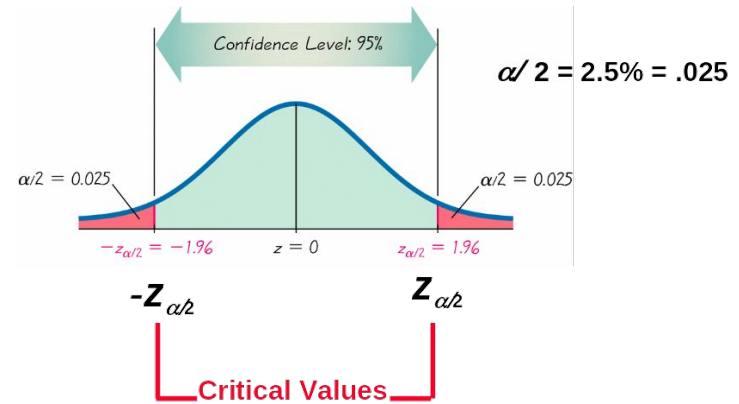
Answer:

a. The sample mean of 172.55 lb is the best point estimate of the mean weight of the population of all men.

b. Menentukan CI

- $CL = 95\% \rightarrow \alpha = 0.05 \rightarrow \alpha/2 = 0.025$
- $z_{\alpha/2} = 1.96 \rightarrow$ cari nilai z yang peluangnya = 0.975
- $\sigma = 26, n = 40$
- $E = z_{\alpha/2} \sigma / \sqrt{n} = 1.96 * (26/\sqrt{40}) = 8.0574835.$
- Maka CI:

$$\begin{aligned}\bar{X} - E &< \mu < \bar{X} + E \\ 172.55 - 8.0574835 &< \mu < 172.55 + 8.0574835 \\ 164.49 &< \mu < 180.61\end{aligned}$$



Interpretasi:

Dengan kepercayaan sebesar 95%, kita bisa mengatakan bahwa the real μ populasi berada di antara 164.49 dan 180.61

Answer (2)

$$\bar{X} - E < \mu < \bar{X} + E$$

$$172.55 - 8.0574835 < \mu < 172.55 + 8.0574835$$

$$164.49 < \mu < 180.61$$

Based on the confidence interval, it is possible that the mean weight of **166.3 lb** used in 1960 **could be** the mean weight of men today. However, the best point estimate of **172.55 lb** suggests that the mean weight of men is now **considerably greater than 166.3 lb**. Considering that an underestimate of the mean weight of men could result in lives lost through overloaded boats and aircraft, these results strongly suggest that **additional data should be collected**. (Additional data have been collected, and the assumed mean weight of men has been increased.)

Example

Assume that we want to estimate the mean IQ score for the population of statistics students. How many statistics students must be randomly selected for IQ tests if we want **95%** confidence that the sample mean is within **3 IQ points of the population mean**? $\sigma = 15$

$$\alpha = 0.05$$

$$\alpha / 2 = 0.025$$

$$z_{\alpha/2} = 1.96$$

$$E = 3$$

$$\sigma = 15$$

$$n = \left[\frac{1.96 \cdot 15}{3} \right]^2 = 96.04 = 97$$

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

With a simple random sample of only 97 statistics students, we will be 95% confident that the sample mean is within 3 IQ points of the true population mean .

Example of CI for μ when σ is unknown

A common claim is that garlic lowers cholesterol levels. In a test of the effectiveness of garlic, **49 subjects** were treated with doses of raw garlic, and their cholesterol levels were measured before and after the treatment. The changes in their levels of LDL cholesterol (in mg/dL) have a **mean of 0.4** and a standard deviation of 21.0. Use the sample statistics of **$n = 49$, $\bar{X} = 0.4$ and $s = 21.0$** to construct a **95%** confidence interval estimate of the mean net change in LDL cholesterol after the garlic treatment. What does the confidence interval suggest about the effectiveness of garlic in reducing LDL cholesterol?

Requirements are satisfied: simple random sample and **$n = 49$** (i.e., $n > 30$).

95% implies $\alpha = 0.05$. \rightarrow With $n = 49$, the **$df = n - 1 = 49 - 1 = 48$** .

From the table, closest df is 50, two tails, so $t/2 = 2.009 \rightarrow$ Using $t/2 = 2.009$, $s = 21.0$ and $n = 49$ the margin of error is:

$$E = t_{\alpha/2} \frac{s}{\sqrt{n}} = 2.009 \frac{21.0}{\sqrt{49}} = 6.027$$

```
> qt(0.025, 48, lower.tail = FALSE)
[1] 2.010635
```

Construct the CI

$$\bar{x} = 0.4, E = 6.027$$

$$\bar{x} - E < \mu < \bar{x} + E$$

$$0.4 - 6.027 < \mu < 0.4 + 6.027$$

$$-5.6 < \mu < 6.4$$

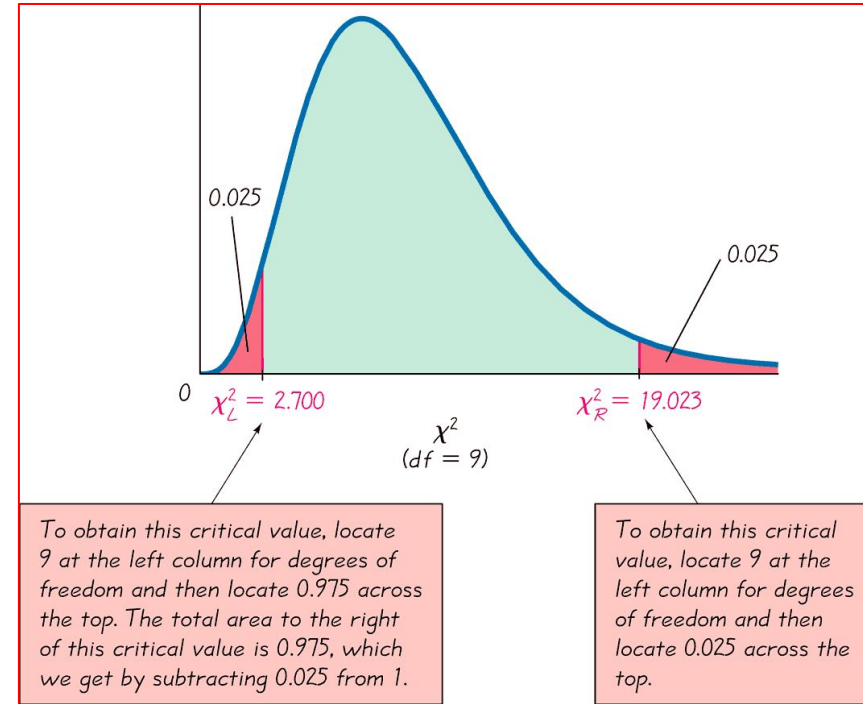
We are 95% confident that the limits of -5.6 and 6.4 actually do contain the value of μ , the mean of the changes in LDL cholesterol for the population. Because the confidence interval limits contain the value of 0, it is very possible that the mean of the changes in LDL cholesterol is equal to 0, suggesting that the garlic treatment did not affect the LDL cholesterol levels. It does not appear that the garlic treatment is effective in lowering LDL cholesterol.

Example

A simple random sample of **ten** voltage levels is obtained. Construction of a confidence interval for the population standard deviation requires the **left** and **right** critical values of χ^2 corresponding to a confidence level of **95%** and a sample size of **n = 10**. Find the critical value of χ^2 separating an area of 0.025 in the left tail, and find the critical value of χ^2 separating an area of 0.025 in the right tail.

Answer:

For a sample of 10 values taken from a normally distributed population, the chi-square statistic $2 = (n - 1)s^2/2$ has a 0.95 probability of falling between the chi-square critical values of 2.700 and 19.023.



Example

Listed below are **ten** voltage levels (in volts) recorded in the author's home on ten different days. These ten values have a standard deviation of **s = 0.15** volt. Use the sample data to construct a **95%** confidence interval estimate of the standard deviation of all voltage levels.

123.3 123.5 123.7 123.4 123.6 123.5 123.5
123.4 123.6 123.8

$n = 10$ so $df = 10 - 1 = 9$

Use table A-4 to find:

$$\chi_L^2 = 2.700 \text{ and } \chi_R^2 = 19.023$$

Construct the confidence interval: $n = 10$, $s = 0.15$

$$\frac{(n-1)s^2}{\chi_R^2} < \sigma^2 < \frac{(n-1)s^2}{\chi_L^2}$$

$$\frac{(10-1)(0.15)^2}{19.023} < \sigma^2 < \frac{(10-1)(0.15)^2}{2.700}$$

$$0.010645 < \sigma^2 < 0.075000$$

$$0.10 \text{ volt} < \sigma < 0.27 \text{ volt.}$$

Example : menghitung sample size σ

We want to estimate the **standard deviation** of **all voltage levels** in a home. We want to be **95%** confident that our estimate is within **20%** of the true value of . How large should the **sample** be? Assume that the population is normally distributed

Answer:

From Table 7-2, we can see that 95% confidence and an error of 20% for correspond to a sample of size 48. We should obtain a simple random sample of 48 voltage levels from the population of voltage levels.

Example 2

We want to estimate the σ of all IQ scores of people with exposure to lead. We want to be **99%** confident that our estimate is within **5%** of the true value of σ . How large should the sample be?

Answer:

Take a look at the table on the next slide

Table for finding sample size to calculate σ^2

Sample Size for σ^2		Sample Size for σ	
To be 95% confident that s^2 is within	of the value of σ^2 , the sample size n should be at least	To be 95% confident that s is within	of the value of σ , the sample size n should be at least
1%	77,208	1%	19,205
5%	3,149	5%	768
10%	806	10%	192
20%	211	20%	48
30%	98	30%	21
40%	57	40%	12
50%	38	50%	8
To be 99% confident that s^2 is within	of the value of σ^2 , the sample size n should be at least	To be 99% confident that s is within	of the value of σ , the sample size n should be at least
1%	133,449	1%	33,218
5%	5,458	5%	1,336
10%	1,402	10%	336
20%	369	20%	85
30%	172	30%	38
40%	101	40%	22
50%	68	50%	14

We want to estimate the σ of all IQ scores of people with exposure to lead. We want to be **99%** confident that our estimate is within **5%** of the true value of σ . How large should the sample be?

Answer:

From the table, we can see 1336

Question

$$E = z_{\alpha/2} \sqrt{\frac{\hat{p}\hat{q}}{n}}$$

$$\hat{p} - E < p < \hat{p} + E$$

1. Dari survey yang dilakukan terhadap 1000 orang diketahui bahwa 14% responden merupakan perokok dengan margin of error (E) ± 4 percentage points (0.04). Apakah informasi ini sudah cukup untuk melakukan penghitungan dan interpretasi CI? \rightarrow Cukup, tapi masih perlu informasi untuk CL atau α agar bisa melakukan interpretasi.
2. Jika confidence level = 90%, berapakan nilai α ? Berapakah critical value-nya (right tail), jika diasumsikan populasi berdistribusi normal?

$$n = \left[\frac{(z_{\alpha/2}) \cdot \sigma}{E} \right]^2 \quad \rightarrow z_{0.05}$$

Question (2)

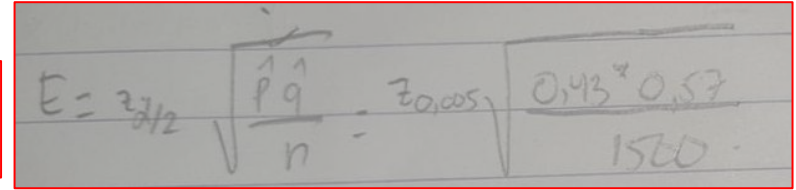
$$E = z_{\alpha/2} \sqrt{\frac{\hat{p}\hat{q}}{n}}$$

Alpha = 0.01 \rightarrow 0.005

3. Dari sebuah survey yang dilakukan terhadap 1500 orang pegawai swasta, diketahui bahwa **43%** diantaranya suka mengonsumsi kopi kekinian. Tentukan **99% CI** untuk estimator p.

a. Step 1: Hitung E

```
> qnorm(1-0.005)*sqrt(0.43*0.57/1500)  
[1] 0.03292631
```



Handwritten formula for E: $E = z_{\alpha/2} \sqrt{\frac{\hat{p}\hat{q}}{n}} = z_{0.005} \sqrt{\frac{0.43 \times 0.57}{1500}}$

b. Step 2: Bentuk confidence intervalnya

$$\hat{p} - E < p < \hat{p} + E$$

$$(0.43 - 0.03292631) < p < (0.43 + 0.03292631)$$

$$0.3970737 < p < 0.4629263$$

$$0.397 < p < 0.463$$

Questions (3)

4. Diketahui sebuah CI $0.58 < p < 0.81$ dengan confidence level 95%. Hitunglah \hat{p} dan E.

Point estimate of p :

$$\hat{p} = \frac{(\text{upper confidence limit}) + (\text{lower confidence limit})}{2}$$

$$p = 0.695$$

Margin of Error:

$$E = \frac{(\text{upper confidence limit}) - (\text{lower confidence limit})}{2}$$

$$p = 0.115$$

Thank You