Statistical Literacy — MINT

Lecture 8: Confidence intervals for means

Rémi Viné

The Graduate Institute | Geneva

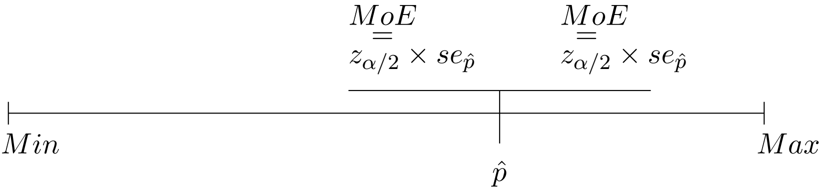
November 13th, 2023

Outline

Recap'

Confidence intervals for proportions

- The point estimate, $\hat{p}$, is the proportion in the sample



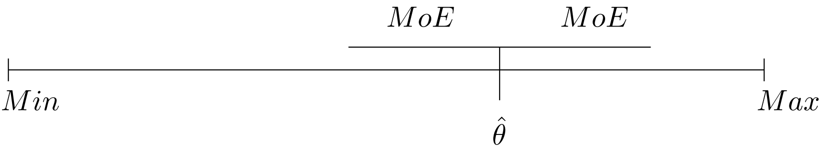
- A 99% CI for $n = 81$, and $\hat{p} = 0.23$:
  - $n\hat{p}\hat{q} = 81 \times 0.23 \times 0.77 = 14.3451 >> 1$
  - $z_{0.005} \approx 2.58$
  - $se_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{0.23 \times 0.77}{81}} \approx 0.0468$
  - $MoE = z_{0.005} \times se_{\hat{p}} = 2.58 \times 0.0468 = 0.1206\ (no\ rounding\ in\ the\ process)$

  $\Rightarrow$ $(0.1094; 0.3506) \rightarrow$ and interpret!

Housekeeping

- Problem set 8 is now available

Generalities on confidence intervals



- A confidence interval is $PointEstimate \pm MoE$
- The $Margin\ of\ Error$ is constructed using two elements
  1. The standard error (standard deviation of the sampling distribution, $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$)
  2. The **critical value** (cv)

$$\Rightarrow MoE = \sigma_{\bar{x}} \times cv$$

Confidence intervals in this class

1. Confidence intervals for proportions
   - The critical value is always taken from a Normal distribution
   - The standard error is even easier to compute
     - $\rightarrow$ One need only the proportions and the sample size
2. Confidence intervals for means
   - The critical value is taken from a Normal distribution **only** if the population variance (σ^2) is available; otherwise, need to use another distribution (*Student distribution* - to encounter today)

Confidence intervals: step by step

1. Verify the conditions of validity (*i.e.*, can the sampling distribution be considered as Normally distributed?)
2. Select a level of confidence
3. Find the point estimate
4. Compute the standard error
5. Find the critical value from the appropriate distribution
6. Compute the Margin of Error and build the confidence interval
7. Conclude

Confidence interval for means

Same process with a couple of differences

- The point estimate is now an average extracted from the sample: $\bar{x}$ (this is the $\hat{\theta}$)
- The **condition of validity** is now simply the application of the Central Limit Theorem: $n > 30$ *($\rightarrow$ usual convention)*
- The **standard error** is back to the usual definition: $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$
  - In the case where the standard deviation in the population is not available, then we compute the standard error using the sample standard deviation
    - $s_{\bar{x}} = \frac{s}{\sqrt{n}}$
    - *E.g.*, if $n = 50$, $d.o.f = 49$ for CI

Confidence interval for means

Find the critical value

- There are 2 cases:

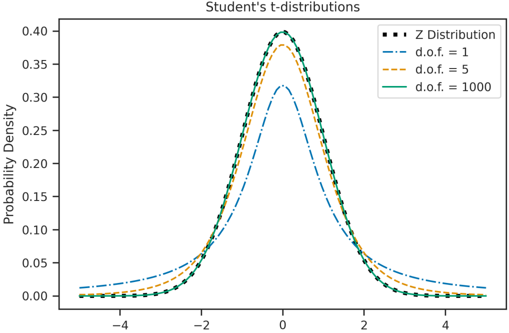
  1. If the population variance, σ^2, is available, then the critical value is extracted from a Normal distribution
  2. If the variance of the population is **NOT** available (most of the time), then the critical value is extracted from a **Student distribution**

     - Similar procedure as for the Normal distribution
     - But with another distribution, find the critical value, $t_{\alpha/2}^{d.o.f.}$
     - $d.o.f.$ stands for **degrees of freedom** (number of values free to vary). Here $d.o.f. = n - 1$

Confidence interval for means

The Student distribution

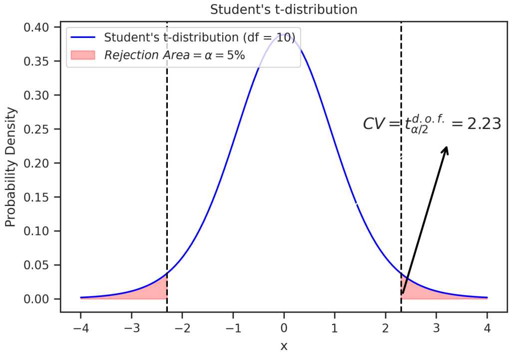
- The sample variance can differ across samples. Need to adjust for this extra variability
- This is what the Student distribution is about (**more scattered** than a Normal distribution **but symmetric** as well)
- Less degree of freedom $\Rightarrow$ more dispersion and thus larger critical values for a given level of significance (α)



Confidence interval for means (no information on population variance)

Find the critical value in a Student distribution

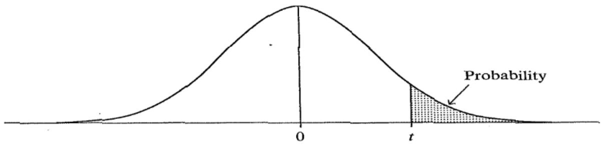
- The procedure is the same, and one can read a Student table to find the critical value that corresponds to a given probability
- As before, the rejection area is split into two equal sides



Confidence interval for means (no information on population variance)

Find the critical value for $\alpha = 5\%$ and $d.o.f. = 10$

TABLE B: *t* **Distribution Critical Values**



	Confidence Level					
	80%	90%	95%	98%	99%	99.8%
	Right-Tail Probability					
df	$t_{.100}$	$t_{.050}$	$t_{.025}$	$t_{.010}$	$t_{.005}$	$t_{.001}$
1	3.078	6.314	12.706	31.821	63.656	318.289
2	1.886	2.920	4.303	6.965	9.925	22.328
3	1.638	2.353	3.182	4.541	5.841	10.214
4	1.533	2.132	2.776	3.747	4.604	7.173
5	1.476	2.015	2.571	3.365	4.032	5.894
6	1.440	1.943	2.447	3.143	3.707	5.208
7	1.415	1.895	2.365	2.998	3.499	4.785
8	1.397	1.860	2.306	2.896	3.355	4.501
9	1.383	1.833	2.262	2.821	3.250	4.297
10	1.372	1.812	2.228	2.764	3.169	4.144

Confidence interval for means

Find the critical value for $\alpha = 5\%$ and $d.o.f. = n-1$

	With population variance	Without population variance
Standard error	$\sigma_{\bar{x}}$	$s_{\bar{x}}$
Critical value	$z_{\alpha/2}$	$t_{\alpha/2}^{d.o.f.}$ with $d.o.f. = n-1$
Margin of Error	$z_{\alpha/2} \times \sigma_{\bar{x}}$	$t_{\alpha/2}^{d.o.f.} \times s_{\bar{x}}$
$n \uparrow$	$\sigma_{\bar{x}} \downarrow$	$s_{\bar{x}} \downarrow$ & $t_{\alpha/2}^{d.o.f.} \downarrow$

- Define the interval

$\rightarrow$ $\bar{x} \pm$ MoE; or: $\bar{x} - \text{MoE} < \mu < \bar{x} + \text{MoE}$; or: $[\bar{x} - \text{MoE}; \bar{x} + \text{MoE}]$

- Interpret

$\rightarrow$ We can be $(1-\alpha)\%$ confident that the mean in the population is located in the interval $[\bar{x} - \text{MoE}; \bar{x} + \text{MoE}]$

$\rightarrow$ Constructing 100 intervals, the population mean would belong to the constructed interval $[\bar{x} - \text{MoE}; \bar{x} + \text{MoE}]$ about $(1-\alpha)$ times

Example

- Based on a (random) sample of 51 refugee camps, a manager of UNHCR observed that the average population in refugee camps is 250 people and the standard deviation is 15 people. What would be the 99% confidence interval to capture the average size of all refugee camps?

Example

- Based on a (random) sample of 51 refugee camps, a manager of UNHCR observed that the average population in refugee camps is 250 people and the standard deviation is 15 people. What would be the 99% confidence interval to capture the average size of all refugee camps?

1. Validity condition: $n > 30$ so the sampling distribution is approximately Normal

Example

- Based on a (random) sample of 51 refugee camps, a manager of UNHCR observed that the average population in refugee camps is 250 people and the standard deviation is 15 people. What would be the 99% confidence interval to capture the average size of all refugee camps?

2. Level of confidence: as it is a 99% confidence interval, the level of confidence is 99% (and $\alpha = 1\%$)

Example

- Based on a (random) sample of 51 refugee camps, a manager of UNHCR observed that the average population in refugee camps is 250 people and the standard deviation is 15 people. What would be the 99% confidence interval to capture the average size of all refugee camps?

3. Standard Error: there is no information on the population variance, so the standard error is the following: $s_{\bar{x}} = \frac{s}{\sqrt{n}} = \frac{15}{\sqrt{51}} \approx 2.1$ people

Example

- Based on a (random) sample of 51 refugee camps, a manager of UNHCR observed that the average population in refugee camps is 250 people and the standard deviation is 15 people. What would be the 99% confidence interval to capture the average size of all refugee camps?

4. Critical value: there is no information on the population variance, so the critical value is to be found in a Student distribution: $t_{\alpha/2}^{51-1} = t_{0.005}^{50} = 2.678$

Example

- Based on a (random) sample of 51 refugee camps, a manager of UNHCR observed that the average population in refugee camps is 250 people and the standard deviation is 15 people. What would be the 99% confidence interval to capture the average size of all refugee camps?

5. Margin of error: $MoE \approx 2.1 \times 2.678 \approx 5.625$ people (the last calculation is done using the non-rounded standard error)

$\Rightarrow$ Hence the 99% confidence interval is 250 ± 5.625 or $(244.375; 255.625)$

We can be 99% confident that the average number of people living in refugee camps is comprised between 244.375 and 255.625.

Example

- Based on a (random) sample of 51 refugee camps, a manager of UNHCR observed that the average population in refugee camps is 250 people and the standard deviation is 15 people. **Now assume that, for some reason, this standard deviation is the population standard deviation.** What would be the new 99% confidence interval to capture the average size of all refugee camps?

Example

- Based on a (random) sample of 51 refugee camps, a manager of UNHCR observed that the average population in refugee camps is 250 people and the standard deviation is 15 people. **Now assume that, for some reason, this standard deviation is the population standard deviation.** What would be the new 99% confidence interval to capture the average size of all refugee camps?

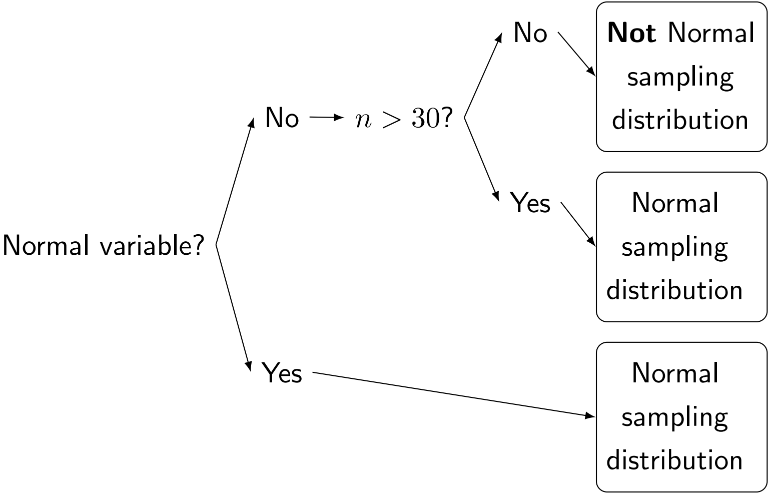
- Nothing but the critical value changes. Now the critical value is $z_{\alpha/2} = z_{0.005} \approx 2.58$

- Hence the new margin of error is smaller, it is now $MoE = 5.419$
- The confidence interval, for the same confidence level, will be narrower

  $\rightarrow$ $(244.581; 255.419)$

Wrap-up on condition for inference

Applicability of the Central Limit Theorem



Wrap-up on confidence intervals

Assuming respective validity conditions are met

- Type
  - Means? → σ known?
    - No → $MoE \Leftrightarrow t^{n-1}_{\alpha/2} \times \frac{s}{\sqrt{n}}$
    - Yes → $MoE \Leftrightarrow z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}}$
  - Proportion? → $MoE \Leftrightarrow z_{\alpha/2} \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

Next session

- Next session is on hypotheses testing (a second inference tool). We will focus on hypothesis testing for means.