Math 170S

Lecture Notes Section 7.4 [*†]

Sample size

Instructor: Swee Hong Chan

NOTE: Materials that appear in the textbook but do not appear in the lecture notes might still be tested. Please send me an email if you find typos.

[*]Version date: Saturday 31st October, 2020, 23:59.

[†]This notes is based on Hanbaek Lyu's and Liza Rebrova's notes from the previous quarter, and I would like to thank them for their generosity. "*Nanos gigantum humeris insidentes* (I am but a dwarf standing on the shoulders of giants)".

1 Example: Case I

The friendly instructor wants to know the average annual income of citizens in Atlantis.

He knows from experience that this kind of survey has standard deviation of approximately 15 cowrie shells.

How many people should he interview so that he can be 95% confident that his estimate is off by at most 1 cowrie shell?

2 Setting: Case I

- **Object:** X is a random variable with **unknown mean** μ but **known variance** σ^2.

- **Input:**

  - The **estimate error** ε and the **confidence constant** $1-\alpha$

- **Output:** The **number of samples** n that allows us to say

  "μ is contained in the interval $\left[\bar{x}-\varepsilon, \bar{x}+\varepsilon\right]$ with confidence (approximately) 1-α."

  Important: We don't know the value of $\bar{x}$ and n yet since the survey has not been conducted!

3 Answer to Example: Case I

We have from the question that

$$\sigma = 15; \qquad 1 - \alpha = 0.95; \qquad \varepsilon = 1.$$

From Section 7.1, the following equation holds:

$$\varepsilon \approx z_{\alpha/2} \frac{\sigma}{\sqrt{n}}.$$

Substituting all the values that we know already, we get

$$1 \approx (1.96) \frac{15}{\sqrt{n}} \qquad \Rightarrow \qquad \sqrt{n} \approx (1.96)(15)$$

$$\Rightarrow \qquad n \approx 864.36.$$

Since n **must be an integer**, we round up and take $n = 865$.

4 The value ε: Case I

Theorem 1. *In this case, n is equal to*

$$n = \left\lceil \left(\frac{z_{\alpha/2}\sigma}{\varepsilon} \right)^2 \right\rceil ,$$

where $\lceil x \rceil$ is the value of x rounded-up. □

5 Example: Case II

Your friendly instructor wants to know the average annual income of citizens in Shangrila.

Unlike in Atlantis, he **does not know the variance** of this distribution anymore, but he knows that this distribution is **normal**.

He has also interviewed 15 people, with sample mean 189 cowrie shells and sample variance 74.8 cowrie shells.

How many more people should he interview so that he can be 90% confident that his estimate is off by at most 3 cowrie shells?

6 Setting: Case II

- **Object:** X is a **normal** random variable with **unknown mean** μ and **unknown variance**.
- **Input:**
  - The **estimate error** ε and the **confidence constant** $1 - \alpha$
- **Output:** The **number of samples** n that allows us to say

  "μ is contained in the interval $\left[\overline{\mathrm{x}} - \varepsilon, \overline{\mathrm{x}} + \varepsilon\right]$ with confidence (approximately) 1-α."

7 The value ε: Case II

Theorem 2. *In this case, we should find n by following these steps:*

- *First, let n be given by*

$$\left\lceil \left(\frac{z_{\alpha/2}\, s}{\varepsilon}\right)^2 \right\rceil.$$

- *Check if n satisfies*

$$\varepsilon \geq t_{\alpha/2}(n-1)\frac{s}{\sqrt{n}}.$$

  - *If the inequality above is satisfied, then use that value of n as our answer;*
  - *If the inequality above is NOT satisfied, then increase the value of n by 1, and repeat until the inequality is satisfied.* □

8 Answer to Example: Case II

We have from the question that

$$1 - \alpha = 0.9, \qquad \varepsilon = 3.$$

From the preliminary data, for $n = 15$, we have

$$\bar{\mathrm{x}} = 189; \qquad s^2 = 74.8.$$

Our preliminary value for n is

$$n \approx \left(\frac{z_{\alpha/2}\sigma}{\varepsilon}\right)^2 = \frac{z^2_{\alpha/2}\sigma^2}{\varepsilon^2} = \frac{z^2_{\alpha/2}s^2}{\varepsilon^2} = \frac{(1.645)^2(74.8)}{(3)^2} = 22.41.$$

So we get that n should be greater than 23. Let's check if $\varepsilon \geq t_{\alpha/2}(n-1)\frac{s}{\sqrt{n}}$.

$$t_{\alpha/2}(n-1)\frac{s}{\sqrt{n}} = t_{0.05}(22)\frac{\sqrt{74.8}}{\sqrt{23}} = (1.717)\frac{\sqrt{74.8}}{\sqrt{23}} \approx 3.096.$$

This is greater than $\varepsilon = 3$, so $n = 23$ is not good enough.

Let's try $n = 24$:

$$t_{\alpha/2}(n-1)\frac{s}{\sqrt{n}} = t_{0.05}(23)\frac{\sqrt{74.8}}{\sqrt{24}} = (1.714)\frac{\sqrt{74.8}}{\sqrt{24}} \approx 3.0259.$$

This is still bigger than 3.

So let's try $n = 25$:

$$t_{\alpha/2}(n-1)\frac{s}{\sqrt{n}} = t_{0.05}(24)\frac{\sqrt{74.8}}{\sqrt{25}} = (1.711)\frac{\sqrt{74.8}}{\sqrt{25}} \approx 2.9595.$$

This is less than $\varepsilon = 3$, so $n = 25$ is our answer. Thus FI needs to interview 10 more people.

Remark 3. This approach only works if we assume that **sample variance** s^2 is relatively close to the **real variance** σ^2.

9 Example: Case III

A presidential candidate wants to know the percentage of voters that support their presidency.

How many people do they need to interview so that their prediction is off by at most 0.03 with 95% confidence?

10 The value ε: Case III

- **Object:** Y is a **Bernoulli** random variable with **unknown parameter** p.
- **Input:**
  - The **estimate error** ε and the **confidence constant** $1 - \alpha$
- **Output:** The **number of samples** n that allows us to say

  "p is contained in the interval $\left[\, \bar{\mathrm{y}} - \varepsilon, \bar{\mathrm{y}} + \varepsilon \right]$ with confidence (approximately) 1-α."

11 The value ε: Case III

Theorem 4. *In this case, n is equal to*

$$n = \left\lceil \left(\frac{z_{\alpha/2}}{2\varepsilon} \right)^2 \right\rceil .$$

In the case when we also know the approximate value of $\overline{\mathrm{y}}$, we can use the smaller value of n given by

$$n = \left\lceil \left(\frac{z_{\alpha/2}}{\varepsilon} \right)^2 (\overline{\mathrm{y}})(1 - \overline{\mathrm{y}}) \right\rceil . \qquad \square$$

12 Answer to Example: Case III

We have from the question that

$$1 - \alpha = 0.95, \qquad \varepsilon = 0.03.$$

So we have

$$n = \left\lceil \left(\frac{z_{0.025}}{2(0.03)} \right)^2 \right\rceil = \lceil 1067.11 \rceil = 1068.$$

Now, let's change the question a little bit.

Suppose that the candidate already suspects that $\bar{\text{y}}$ **is approximately** 0.2. They just want to conduct another poll to verify their suspicion.

So we have

$$n = \left\lceil \left(\frac{z_{0.025}}{0.03}\right)^2 (0.2)(0.8) \right\rceil = \lceil 682.9 \rceil = 683.$$

This requires less sample than before, but requires knowing $\bar{\text{y}}$ **is approximately** 0.2.