Chapter 5: Estimation

(Ott & Longnecker Sections: 5.3, 10.2)

https://dzwang91.github.io/stat324/

Part 4

WISCONSIN

UNIVERSITY OF WISCONSIN–MADISON

Confidence interval

"Is it possible to have a 100% CI?"

Confidence interval

"Is it possible to have a 100% CI?"

YES, $(-\infty, +\infty)$ is a 100% confidence interval. But wait, is it useful?!

Confidence interval

"Is it possible to have a 100% CI?"

YES, $(-\infty, +\infty)$ is a 100% confidence interval. But wait, is it useful?!

- This kind of confidence interval is uninformative.
- We want the higher confidence level, and the narrower confidence interval, the more accurate estimate.

Confidence interval

"Is it possible to have a 100% CI?"

YES, $(-\infty, +\infty)$ is a 100% confidence interval. But wait, is it useful?!

- This kind of confidence interval is uninformative.
- We want the higher confidence level, and the narrower confidence interval, the more accurate estimate.

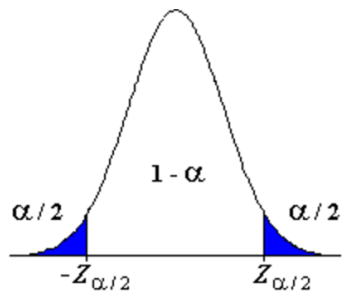
"For any given confidence level, how can we adjust the sample size to get the desired width of the confidence interval?"

- If we want to get a 95% confidence interval with width 5(U-L=5), then what's the required sample size?

Review of confidence interval: case 1

If we know the population standard deviation σ,

1. Choose a confidence level $1-\alpha$. Typically, if we require 95% confidence level, then $\alpha = 0.05$.
2. Use z table to find the $z_{\frac{\alpha}{2}}$ critical value such that $P(-z_{\frac{\alpha}{2}} \leq Z \leq z_{\frac{\alpha}{2}}) = 1-\alpha$.



3. Construct the interval: (L, U), where $L = \bar{X} - z_{\frac{\alpha}{2}} * \frac{\sigma}{\sqrt{n}}$, $U = \bar{X} + z_{\frac{\alpha}{2}} * \frac{\sigma}{\sqrt{n}}$.
4. Conclude: $P(L \leq \mu \leq U) = 1-\alpha$. We are $(1-\alpha) \times 100\%$ confident that the population mean is between (L, U).

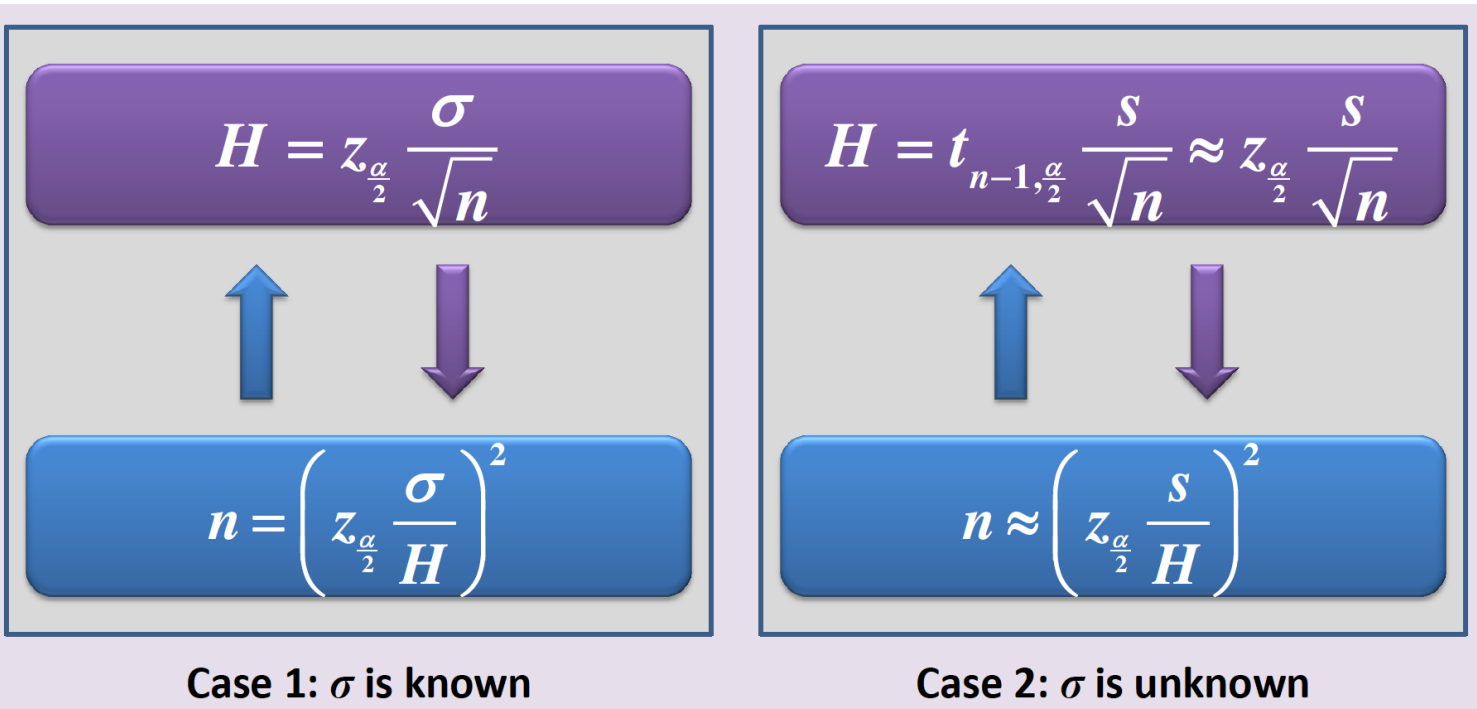
Review of confidence interval: case 2

If we don't know the population standard deviation σ,

1. Choose a confidence level $1-\alpha$. Typically, if we require 95% confidence level, then $\alpha = 0.05$.
2. Find the value t such that $P(-t \leq T_{n-1} \leq t) = 1-\alpha$. It also means $P(T_{n-1} \geq t) = \frac{\alpha}{2}$. Use t table with degrees of freedom n-1. We denote the value t as $t_{n-1,\alpha/2}$.
3. Construct the interval: (L, U), where $L = \bar{X} - t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}$, $U = \bar{X} + t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}$.
4. Conclude: $P(L \leq \mu \leq U) = 1-\alpha$. We are $(1-\alpha)\times 100\%$ confident that the population mean is between (L, U).

Determining sample size

- Let H be the half width: $H = \frac{U-L}{2}$.



Example

We want a 95% CI for μ. We desire the half-width to be no larger than 0.1mm. Then what's the required sample size?

Example

We want a 95% CI for μ. We desire the half-width to be no larger than 0.1mm. Then what's the required sample size?

1. Case 1: if σ, the true population standard deviation is known, since $z_{\alpha/2} = 1.96$, we would just need to solve the equation

$$0.1 = 1.96 * \frac{\sigma}{\sqrt{n}},$$

Example

We want a 95% CI for μ. We desire the half-width to be no larger than 0.1mm. Then what's the required sample size?

1. Case 1: if σ, the true population standard deviation is known, since $z_{\alpha/2} = 1.96$, we would just need to solve the equation

$$0.1 = 1.96 * \frac{\sigma}{\sqrt{n}},$$

2. Case 2: if σ is unknown, in this case, we solve the equation

$$0.1 = t_{(n-1,\alpha/2)} \frac{s}{\sqrt{n}} \approx z_{\alpha/2} \frac{s}{\sqrt{n}}$$

So if we are given $s = 0.3385$ mm.

$$0.1 = 1.96(0.3385/\sqrt{n}),$$

which gives:

$$n = \frac{(1.96^2)(0.3385^2)}{0.1^2} = 44.01, \text{ which we round up to } 45.$$

What we have learnt so far...

We've talked about the estimation of population mean

- Point estimator: sample mean (why do we choose this estimator?)
- Interval estimator:
  1. σ is known: case 1, use Z table
  2. σ is unknown: case 2, use T table

Next topic

Estimation of population proportion.

Example

- An accounting firm has a large list of clients (the population), and each client has a file with information about that client. The firm has noticed errors in some of these files, and has decided that it would be worthwhile to know the proportion of files that contain an error (population proportion).
- Call the population proportion of files in error π. It was decided to take a simple random sample of size $n = 50$, and use the results of the sample to estimate π. Each selected file was thoroughly reviewed, and classified as either containing an error (call this 1), or not (call this 0). The results are as follows:

Files with an error: 10; Files without any errors: 40.

Example continued

What is the file reviewing process from the statistical perspective?

Example continued

What is the file reviewing process from the statistical perspective?

- It is a Binomial Process.
  1. Let the random variable Y_i be the *indicator* that the ith file sampled had errors: that is, Y_i is 1 if the file contains an error and 0 otherwise. The pmf of Y_i for all i is:

Y_i	$p(Y_i)$
0	$1-\pi$
1	π

  2. Then the random variable

$$B = Y_1 + Y_2 + ... + Y_n = \sum_{i=1}^{n} Y_i \sim Bin(n, \pi).$$

B counts the number of files with errors. (In the example, we happened to realize $b = 10$ errors out of $n = 50$ files sampled.)

Goal 1: point estimate of π

What is a natural estimator of the true proportion of files with errors?

Goal 1: point estimate of π

What is a natural estimator of the true proportion of files with errors?

- Sample proportion is the proportion of successes in the sample, which is given by the formula:

$$\text{Sample proportion: } \hat{\pi} = P = \frac{\sum_{i=1}^{n} Y_i}{n}.$$

Goal 1: point estimate of π

What is a natural estimator of the true proportion of files with errors?

- Sample proportion is the proportion of successes in the sample, which is given by the formula:

$$\text{Sample proportion: } \hat{\pi} = P = \frac{\sum_{i=1}^{n} Y_i}{n}.$$

- Recall $\mathbb{E}(Y_i) = \pi$ and $VAR(Y_i) = \pi(1-\pi)$. Hence,

$$\mathbb{E}(P) = \pi,\ VAR(P) = \frac{\pi(1-\pi)}{n}, SE(P) = \sqrt{\frac{\pi(1-\pi)}{n}}.$$

Goal 1: point estimate of π

What properties of the sample proportion P do you find?

Goal 1: point estimate of π

What properties of the sample proportion P do you find?

- The estimator P is unbiased for π.
- If $\pi = 0$ or 1, then the standard error is 0. (Does this make sense?)
- The estimated standard error of P:

$$\text{Estimated standard error of P} = \sqrt{\frac{P(1-P)}{n}}.$$

Goal 2: confidence interval of π

How do we make a CI for π?

Goal 2: confidence interval of π

How do we make a CI for π?

- The general form of CI:

$$L = \text{estimator} - \text{critical value} \times \text{standard error},$$

$$U = \text{estimator} + \text{critical value} \times \text{standard error}.$$

- Key ingredient: figure out the sampling distribution of sample proportion P

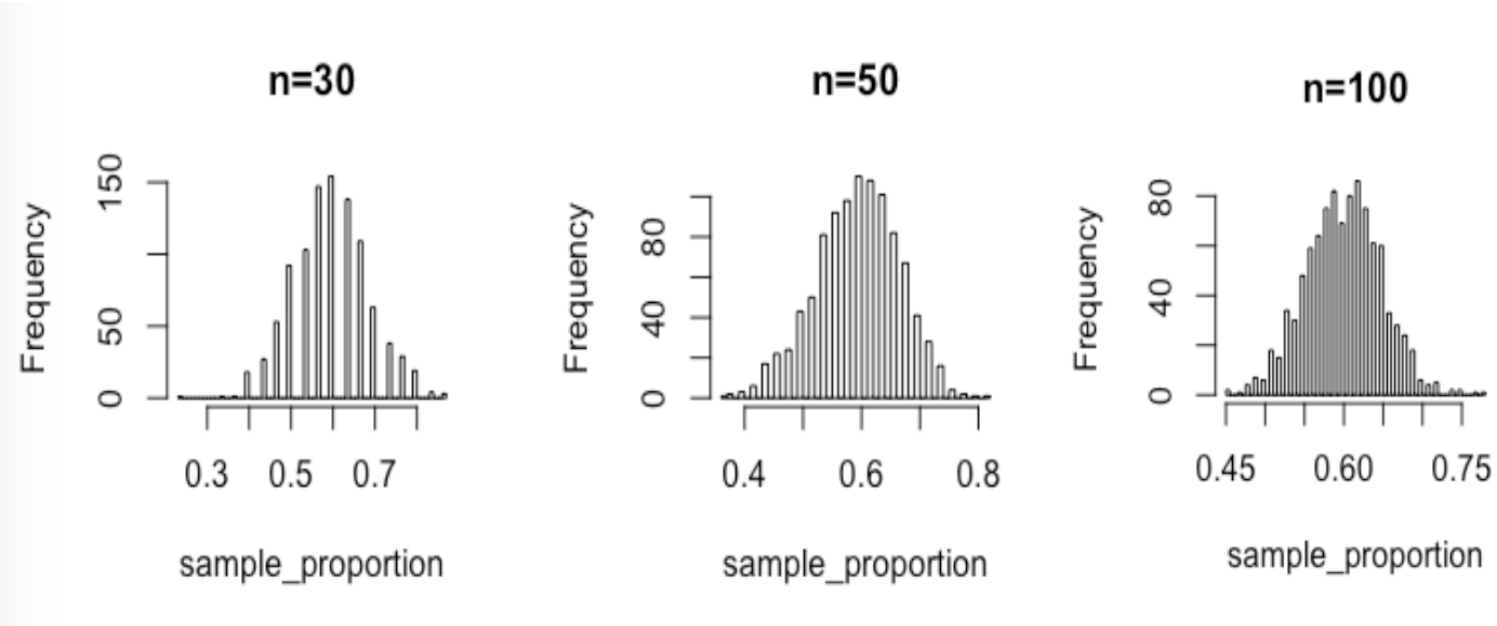
Goal 2: confidence interval of π

What is the sampling distribution of sample proportion?

Goal 2: confidence interval of π

What is the sampling distribution of sample proportion?

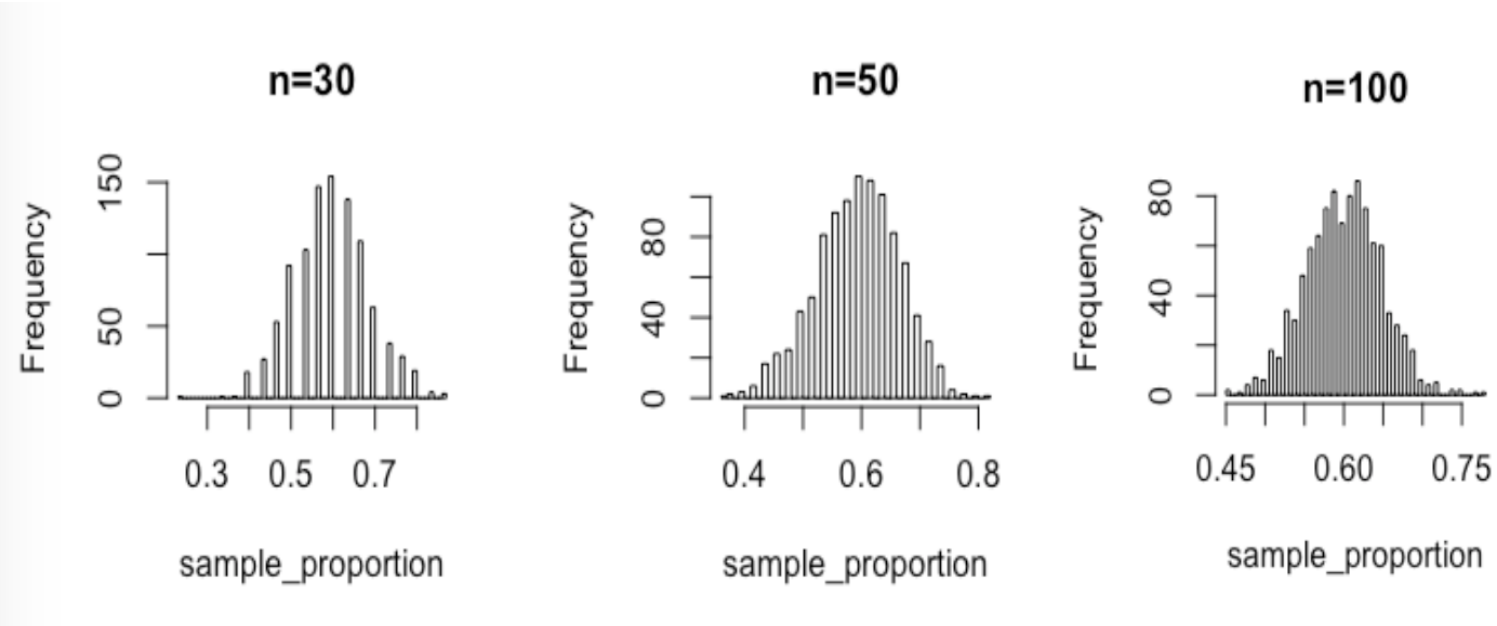
- Simulation setup: set the population proportion $\pi = 0.6$, number of trials $n = 30, 50$ and 100.



Goal 2: confidence interval of π

What is the sampling distribution of sample proportion?

- Simulation setup: set the population proportion $\pi = 0.6$, number of trials $n = 30, 50$ and 100.



- This is central limit theorem!

Goal 2: confidence interval of π

Theorem (CLT)

Let Y_1, Y_2, ..., Y_n be a collection of iid RVs with $\mathbb{E}(Y_i) = \mu$ and $VAR(Y_i) = \sigma^2$. For large enough n, the distribution of $\bar{Y} = \frac{\sum_{i=1}^{n} Y_i}{n}$ will be approximately normal with $\mathbb{E}(\bar{Y}) = \mu$ and $VAR(\bar{Y}) = \frac{\sigma^2}{n}$. That is,

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

- Sample proportion: $P = \frac{\sum_i^n Y_i}{n}$
- $\mathbb{E}(Y_i) = \pi$, $VAR(Y_i) = \pi(1-\pi)$
- For large samples, P is approximately distributed as a normal:

$$P \approx N(\pi, \frac{\pi(1-\pi)}{n}).$$

Goal 2: confidence interval of π

- An approximate $100(1-\alpha)\%$ CI for π:

$$(P - z_{\alpha/2}\sqrt{\frac{P(1-P)}{n}}, P + z_{\alpha/2}\sqrt{\frac{P(1-P)}{n}}).$$

Goal 2: confidence interval of π

- An approximate $100(1-\alpha)\%$ CI for π:

$$(P - z_{\alpha/2}\sqrt{\frac{P(1-P)}{n}}, P + z_{\alpha/2}\sqrt{\frac{P(1-P)}{n}}).$$

When is this approximation good?

- If $n\pi > 5$ and $n(1-\pi) > 5$, the approximation will be good. In this expression π can be approximated by P. The rule then becomes, you should have observed at least 5 successes and at least 5 failures.

Example

For the audit data,

- sample proportion is $P = 10/50 = 0.2$,
- estimated standard error is $\sqrt{(0.2 * 0.8)/50} = 0.057$.
- The CLT should be a good approximation since we have 10 successes and 40 failures, more than 5 each.
- Thus an approximate 95% CI for π would be $0.2 \pm 1.96 * 0.057$, or $(0.088, 0.312)$.

What's the next?

We'll talk about the bootstrap method in next lecture.