Chapter 5: Estimation

(Ott & Longnecker Sections: 5.3, 10.2)

Duzhe Wang

https://dzwang91.github.io/stat371/

Part 5

UNIVERSITY OF WISCONSIN–MADISON

Outline

1. Determining sample size

2. Estimation and inference for population proportions

Confidence interval

1. “Is it possible to have a 100% CI?”

Confidence interval

1. "Is it possible to have a 100% CI?"
   YES, $(-\infty, +\infty)$ is a 100% confidence interval. But wait, is it useful? !

Confidence interval

1. "Is it possible to have a 100% CI?"
   YES, $(-\infty, +\infty)$ is a 100% confidence interval. But wait, is it useful? !
2. This kind of confidence interval is uninformative.

Confidence interval

1. "Is it possible to have a 100% CI?"
   YES, $(-\infty, +\infty)$ is a 100% confidence interval. But wait, is it useful? !
2. This kind of confidence interval is uninformative.
3. We want the higher confidence level, and the narrower confidence interval, the more accurate estimate.

Confidence interval

1. "Is it possible to have a 100% CI?"
   YES, $(-\infty, +\infty)$ is a 100% confidence interval. But wait, is it useful? !
2. This kind of confidence interval is uninformative.
3. We want the higher confidence level, and the narrower confidence interval, the more accurate estimate.
4. A natural question: For any given confidence level, how can we adjust the sample size to get the desired width of the confidence interval?

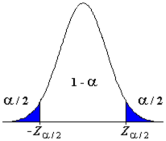
A typical problem

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

$$H = z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

$$n = \left(z_{\frac{\alpha}{2}} \frac{\sigma}{H} \right)^2$$

Case 1: σ is known

$$H = t_{n-1,\frac{\alpha}{2}} \frac{s}{\sqrt{n}} \approx z_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}}$$

$$n \approx \left(z_{\frac{\alpha}{2}} \frac{s}{H} \right)^2$$

Case 2: σ is unknown

Example

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

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

Outline

1. Determining sample size

2. Estimation and inference for population proportions

What we have learnt so far...

We've talked about the estimation of population means

1. Point estimate: sample mean
2. interval estimate: σ is given and σ is unknown.

What we have learnt so far...

We've talked about the estimation of population means

1. Point estimate: sample mean
2. interval estimate: σ is given and σ is unknown.

We'll now discuss estimation of population proportions.

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. 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.

Goal 1: find an estimate of π

- what do you observe from the file reviewing process from the statistical perspective?

Goal 1: find an estimate of π

- what do you observe from the file reviewing process from the statistical perspective?
- The procedure by which the files were selected is a binomial process. 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

Goal 1: find an estimate of π

- what do you observe from the file reviewing process from the statistical perspective?
- The procedure by which the files were selected is a binomial process. 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

- 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: find an estimate of π

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

Goal 1: find an estimate of π

- What is a natural estimator of the true proportion of files with errors?
- Recall we use sample mean to estimate population mean. Now we use sample proportion to estimate population proportion.

Goal 1: find an estimate of π

- What is a natural estimator of the true proportion of files with errors?
- Recall we use sample mean to estimate population mean. Now we use sample proportion to estimate population proportion.
- 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: find an estimate of π

- What is a natural estimator of the true proportion of files with errors?
- Recall we use sample mean to estimate population mean. Now we use sample proportion to estimate population proportion.
- 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 $E(Y_i) = \pi$ and $VAR(Y_i) = \pi(1-\pi)$. Hence:

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

Goal 1: find an estimate of π

- What properties of sample proportion P do you find?

Goal 1: find an estimate of π

- What properties of sample proportion P do you find?
- The estimator P is unbiased for π.

Goal 1: find an estimate of π

- What properties of sample proportion P do you find?
- The estimator P is unbiased for π.
- Note that if $\pi = 0$ or 1 the standard error is 0. Does this make sense?

Goal 1: find an estimate of π

- What properties of sample proportion P do you find?
- The estimator P is unbiased for π.
- Note that if $\pi = 0$ or 1 the standard error is 0. Does this make sense?
- We can get the estimated standard error of P by plugging in our estimator of π:

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

Goal 2: make a CI for π

- What do we need to know in order to make a CI?

Goal 2: make a CI for π

- What do we need to know in order to make a CI?
- We must know the distribution of sample proportion P!

Goal 2: make a CI for π

- What do we need to know in order to make a CI?
- We must know the distribution of sample proportion P!
- The exact distribution of P is related to a binomial, but it turns out that making an exact CI based on this fact is very mathematically challenging and difficult.

Goal 2: make a CI for π

- What do we need to know in order to make a CI?
- We must know the distribution of sample proportion P!
- The exact distribution of P is related to a binomial, but it turns out that making an exact CI based on this fact is very mathematically challenging and difficult.
- Do we have any other tool to overcome this challenge?

Goal 2: make a CI for π

- What do we need to know in order to make a CI?
- We must know the distribution of sample proportion P!
- The exact distribution of P is related to a binomial, but it turns out that making an exact CI based on this fact is very mathematically challenging and difficult.
- Do we have any other tool to overcome this challenge?

Don't forget the central limit theorem!!

Review of central limit theorem

- Let X_1, X_2, ..., X_n be a collection of iid RVs with $E(X_i) = \mu$ and $VAR(X_i) = \sigma^2$. For large enough n, the distribution of $\bar{X}$ will be approximately normal with $E(\bar{X}) = \mu$ and $VAR(\bar{X}) = \frac{\sigma^2}{n}$. That is, $\bar{X} \approx N(\mu, \frac{\sigma^2}{n})$.
- Sample proportion $P = \frac{\sum_i^n Y_i}{n}$, very similar to $\bar{X}$ in the above theorem.
- As long as the sample size is large enough, all the conditions of the CLT are met, because the Y_i are iid, and P is just a sample mean of a bunch of zeros and ones. Thus, for large samples, P is approximately distributed as a normal:

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

Goal 2: make a CI for π

- What is the $100(1-\alpha)\%$ CI for π?
- Recall the general form of CI: estimate $\pm$ multiplier $\times$ estimated SE of the estimator.
- This means that an approximate $100(1-\alpha)\%$ CI for π would be of the form:

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

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

Goal 2: make a CI for π

- What is the $100(1-\alpha)\%$ CI for π?
- Recall the general form of CI: estimate $\pm$ multiplier $\times$ estimated SE of the estimator.
- This means that an approximate $100(1-\alpha)\%$ CI for π would be of the form:

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

- When is this approximation good?
- Generally, if $n\pi > 5$ and $n(1-\pi) > 5$, the approximation will be good. In this expression π can be approximated by P as estimated by the sample. The rule then becomes, you should have observed at least 5 successes and at least 5 failures.
- For the audit data, our estimate would be $P = 10/50 = 0.2$, with estimated standard error $\sqrt{(0.2 * 0.8)/50} = 0.057$. 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 $(0.088, 0.312)$.

One issue

- Can the CI for a proportion go below 0 or above 1 using the CLT method?

One issue

- Can the CI for a proportion go below 0 or above 1 using the CLT method?
- Yes, it will happen because the interval is approximate. Practically, you would probably use a lower or upper bound of 0 or 1, rather then extending the interval into a range that is physically impossible.

What's the next?

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