

① [7.1] Chapter 7

- Now, we will talk about estimation and confidence.
Our main targets will be population means and proportions.

Definitions

Point Estimate: Best guess for the value of a parameter

Interval Estimate: An entire interval that is likely to contain a parameter.

(point estimate)

[Ex] We think the average US height is 5 feet 10 inches

We think the average US height is between 5 and 6 feet.

Not all estimators are created equal. What makes a good estimator?

1) Unbiased: An estimator is unbiased if the mean of the estimator is the parameter being estimated

[Ex] \bar{X} estimates μ

\Rightarrow mean of \bar{X} is μ

\hat{p} estimates p

\Rightarrow mean of \hat{p} is p

2) Small standard error (standard deviation σ)

The standard deviation of an estimator is sometimes called the standard error. We would like the s.e. to be small.

② Interval Estimates

Confidence Intervals: A confidence interval is an interval with a confidence level, between 0 and 1, of how sure we are that the parameter is in the interval. That is, the confidence level states how likely it is that the interval contains the parameter.

The confidence level is determined by the procedure you use; not necessarily the interval. The confidence is the probability that your procedure will produce an interval that contains the parameter.

Elections
[Ex] Support for candidate X is $37\% \pm 5\%$ with 95% confidence. What does this mean?

It means 95% of all possible samples will produce an interval that contains the parameter value.

The interval is $(37 - 5)\%, (37 + 5)\% = (32\%, 42\%)$

The parameter value is the actual %age who voted for candidate X.

③ [7.2] C.I. for p , the Population Proportion

A 95% C.I. for p is given by (as long as $n\hat{p} \geq 15$, $n(1-\hat{p}) \geq 15$)

$$\hat{p} \pm \text{s.e.} \times 1.96, \text{ where s.e.} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \quad (\text{recall } \hat{p} = \frac{\# \text{ successes}}{\text{Sample Size}} = \frac{x}{n})$$

So the interval is $(\hat{p} - 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}})$ with 95% confidence.

This means we are 95% confident that our method will capture the true value of the parameter, p .

Where is this coming from? The Central Limit Theorem.

Notice $P(-c \leq \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \leq c) \approx .95$ if $c = 1.96$ by CLT.

~~[EX] We wish to know what percentage of U.S. identify as republican. We take a SRS of 1000 people and 400 identified as republican. Construct a 95% C.I. for the true proportion of republicans.~~

~~$$\hat{p} = \frac{400}{1000} = 0.4, \quad n\hat{p} = 1000(0.4) = 400 \geq 15 \quad \text{and} \quad n(1-\hat{p}) = 1000(0.6) = 600 \geq 15$$~~

~~So, our CI is OK to find using CLT~~~~So, with 95% confidence, p is in the interval~~

~~$$\begin{aligned} \left(\hat{p} - 1.96 \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + 1.96 \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right) &= \left(0.4 - 1.96 \sqrt{\frac{0.4(0.6)}{1000}}, 0.4 + 1.96 \sqrt{\frac{0.4(0.6)}{1000}} \right) \\ &= \left(0.4 - 1.96 \sqrt{0.00024}, 0.4 + 1.96 \sqrt{0.00024} \right) \\ &= \left(0.4 - 1.96(0.0155), 0.4 + 1.96(0.0155) \right) \\ &= (0.4 - 0.0304, 0.4 + 0.0304) = (0.3696, 0.4304) \end{aligned}$$~~

(4) Now, playing with the inequality

$$\begin{aligned} -c \leq \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \leq c &\Rightarrow -c \sqrt{\frac{p(1-p)}{n}} \leq \hat{p} - p \leq c \sqrt{\frac{p(1-p)}{n}} \\ &\Rightarrow -\hat{p} - c \sqrt{\frac{p(1-p)}{n}} \leq -p \leq -\hat{p} + c \sqrt{\frac{p(1-p)}{n}} \\ &\Rightarrow \hat{p} - c \sqrt{\frac{p(1-p)}{n}} \leq p \leq \hat{p} + c \sqrt{\frac{p(1-p)}{n}} \end{aligned}$$

Now, replace p with \hat{p} on left + right hand sides

Then

$$P\left(\hat{p} - c \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq p \leq \hat{p} + c \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) \approx 0.95 \text{ if } c = 1.96$$

So $\hat{p} \pm 1.96 \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ gives us an approximate 95% CI for p .

[9X] We wish to know what percentage of US identify as Republican. We take a SRS of 1000 people and 400 identified as Republican. Construct a 95% CI for the true proportion of Republicans.

$$\hat{p} = \frac{400}{1000} = 0.4, n\hat{p} = 1000(0.4) = 400 > 15, n(1-\hat{p}) = 1000(0.6) = 600 > 15 \checkmark$$

$$\begin{aligned} \text{So, our CI is } \hat{p} \pm 1.96 \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} &= 0.4 \pm 1.96 \sqrt{\frac{0.4(0.6)}{1000}} \\ &= 0.4 \pm 1.96 \sqrt{0.00024} \\ &= 0.4 \pm 0.0304 \end{aligned}$$

5. Other C.I.'s (not just 95%)

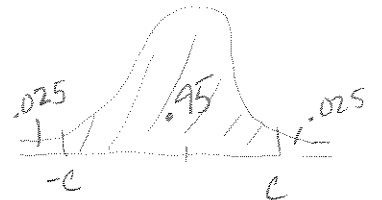
For what constant c does

1) $P(-c \leq Z \leq c) = .95$

2) $P(-c \leq Z \leq c) = .90$

3) $P(-c \leq Z \leq c) = .99$

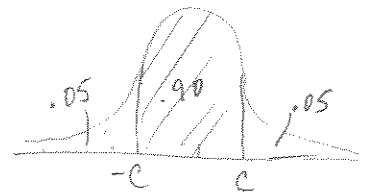
1) $P(-c \leq Z \leq c) = .95 \Rightarrow P(Z \leq -c) = 0.025$



Find this area in Appendix A, Table A or B.

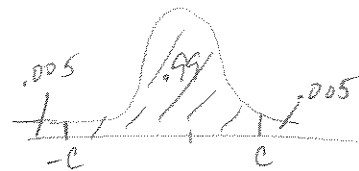
It is 0.0250 and gives Z value of 1.96

2) $P(-c \leq Z \leq c) = .90 \Rightarrow P(Z \leq -c) = .05$



In Table A, area .05 gives Z value of 1.645

3) $P(-c \leq Z \leq c) = .99 \Rightarrow P(Z \leq -c) = .005$



In Table A, area .005 gives Z value of 2.58

So a 90% CI for p is

$$\hat{p} \pm 1.645 \times \text{s.e.}$$

95% CI for p is

$$\hat{p} \pm 1.96 \times \text{s.e.}$$

99% CI for p is

$$\hat{p} \pm 2.58 \times \text{s.e.}$$

where $\text{s.e.} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

What are 2 ways to decrease the length of CI?

- ① use larger n
- ② decrease confidence

⑥ CI for μ

- Recall \bar{X} is approximately normal, by the CLT, with mean of $\bar{X} = \mu$ and s.d of $\bar{X} = \sigma/\sqrt{n}$.

We could try a similar approach to what we did for p , but σ is a problem.

We would like $\bar{X} \pm$ margin of error, but margin of error will depend on σ because the margin of error is coming from the s.d. of \bar{X} . This is a problem b/c usually σ is unknown.

So, we replace σ by its estimate, $S =$ sample s.d.

New Problem: Assume X_1, \dots, X_n are coming from a normal population. We know

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \text{ is standard normal (from Chapter 6.5)}$$

But, $\frac{\bar{X} - \mu}{S/\sqrt{n}}$ is not standard normal because S is a random variable. So, if original population is normal, $\frac{\bar{X} - \mu}{S/\sqrt{n}}$ has a t -distribution. (Student's t -distribution)

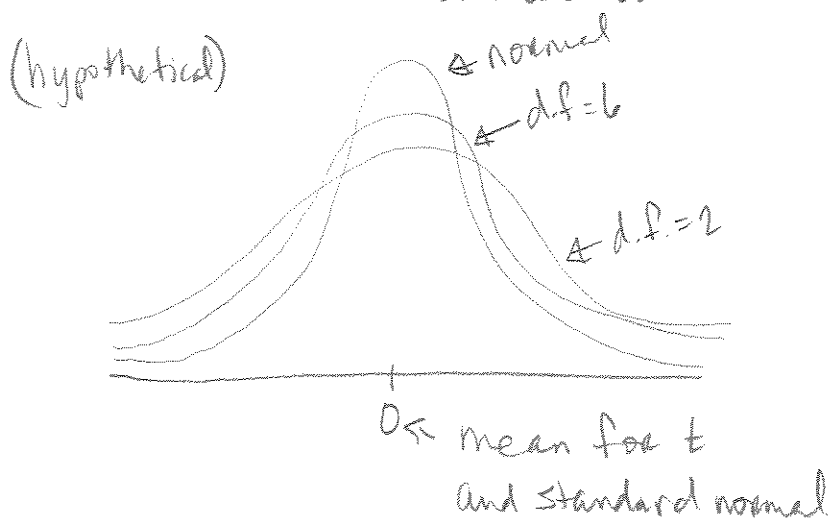
William Gosset

- From Guinness brewing in early 1900's.
- He was concerned about sample sizes in his quality control. By this time people knew the CLT only worked for large n .

③ Gosset wanted to know how to deal with small sample sizes. He discovered the t -distribution for this problem. But, he could not publish his results because Guinness worried about giving trade secrets. So Gosset published his results under the pseudonym "Student". This is why it is called the "Student's t -distribution".

New Random Variable: t -random variable

The t -random variables have one parameter called the "degrees of freedom" or d.f. The t looks very similar to a normal, but with much thicker tails controlled by the d.f. The higher the d.f., the closer the t comes to a standard normal.



⑧. A 95% CI for μ is then

$$\bar{X} \pm t_{.025} \cdot \text{s.e.}, \text{ where } \text{s.e.} = \frac{S}{\sqrt{n}}$$

$$t_{.025} \text{ comes from } P(T_{n-1} \geq t_{.025}) = .025$$

T_{n-1} is a t -random variable with $n-1$ d.f.
(Similar to Z for a normal random variable)

EX) Suppose we want to know the average selling price of a particular pda on ebay. We sample the results of 7 auctions with results

$$\bar{X} = 233.37, S = 14.64$$

Construct a 95% CI for μ assuming population is normal.

$$\bar{X} \pm t_{.025} \cdot \text{s.e.}, \text{ s.e.} = \frac{S}{\sqrt{n}}$$

Use Table B, p. A3

$$\text{d.f.} = n - 1 = 7 - 1 = 6, t_{.025} = 2.447 \quad (\text{note: in same situation, } Z = 1.96)$$

Then, we get

$$233.37 \pm 2.447 \cdot \frac{14.64}{\sqrt{7}} = 233.37 \pm 13.540$$

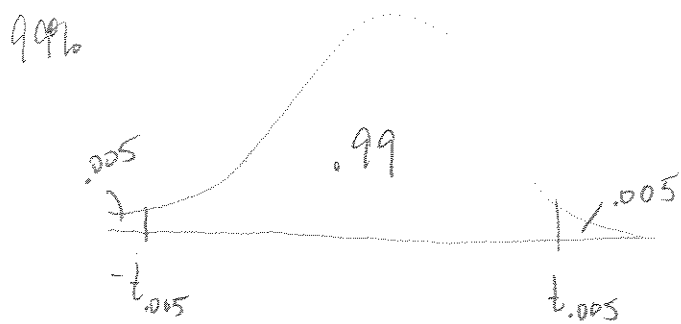
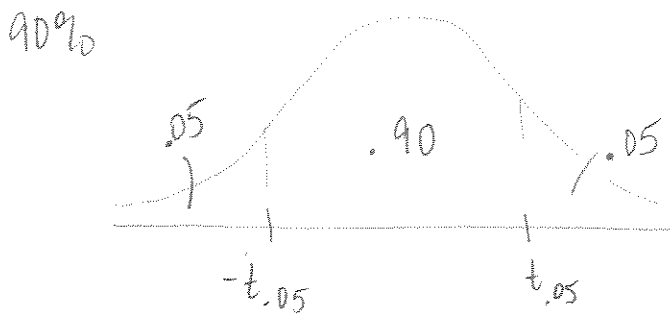
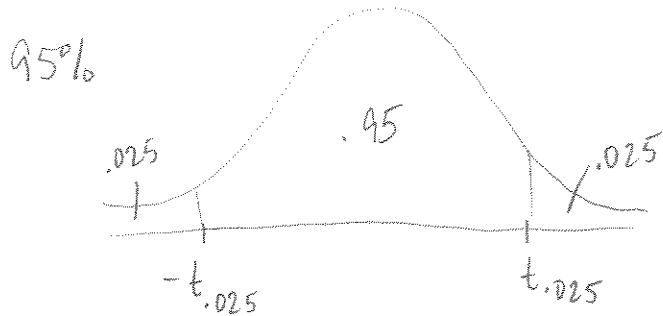
which gives the interval

$$(219.83, 246.91)$$

for which we are 95% confident μ is in.

⑨. 90, 95, and 99% CI for μ

To generalize the 95% CI, we only need replace $t_{.025}$ in the equation $\bar{X} \pm t_{.025} \cdot S.E.$



Then we have

90% CI is $\bar{X} \pm t_{.05} \cdot S.E.$

95% CI is $\bar{X} \pm t_{.025} \cdot S.E.$

99% CI is $\bar{X} \pm t_{.005} \cdot S.E.$

Where

$$P(T_{n-1} \geq t_{.05}) = .05, \quad P(T_{n-1} \geq t_{.025}) = .025, \quad \text{and} \quad P(T_{n-1} \geq t_{.005}) = .005$$

10. [EX] Math SAT. We sample 9 students and record their SAT math scores. We obtain

$$\bar{X} = 500 \text{ and } S = 110$$

a) 90% CI

$$\bar{X} \pm t_{.05} \cdot \text{s.e.}, \quad \text{s.e.} = \frac{S}{\sqrt{n}} = \frac{110}{\sqrt{9}} = \frac{110}{3} \approx 36.67$$

$$\text{d.f.} = 9 - 1 = 8, \quad t_{.05} = 1.86$$

So, the 90% CI has endpoints

$$500 \pm 1.86(36.67) = 500 \pm 68.2$$

b) 95% CI

d.f. = 8, $t_{.025} = 2.306$, so the 95% CI has endpoints

$$500 \pm 2.306(36.67) = 500 \pm 84.56$$

c) 99% CI

d.f. = 8, $t_{.005} = 3.355$, so the 99% CI has endpoints

$$500 \pm 3.355(36.67) = 500 \pm 123.03$$