MATH 7343 Applied Statistics

Prof. (Aidong) Adam Ding

Northeastern University

Review

- Last time, finished Chapter 8. We introduced 3 theoretical results, and used the first one to derive a formula for Confidence Intervals.

- Today, we will study Confidence Intervals in more detail.

Chapter8 Sampling Distribution of the Mean

There are three theoretical results we covered last time.

- (1) $\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$
- (2) $\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}$
- (3) $\frac{\overline{X}-\mu}{s/\sqrt{n}} \sim t_{n-1}$
- Based on the result (1): a (1-α) CI for μ, when $\frac{\sigma}{\sqrt{n}}$ is known, is $(\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}})$.

Confidence Interval of the Variance

- (2) $\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}$
- For population variance parameter σ^2, the point estimator is s^2. Let $\chi^2_{n-1,\,\alpha}$ denote the upper α percentile of χ^2_{n-1}, which is contained in TableA.8. Based on the result (2):

$$P(\chi^2_{n-1,\,1-\alpha/2} \le \frac{(n-1)s^2}{\sigma^2} \le \chi^2_{n-1,\,\alpha/2}) = 1-\alpha$$

$$\Leftrightarrow P(\frac{(n-1)s^2}{\chi^2_{n-1,\,\alpha/2}} \le \sigma^2 \le \frac{(n-1)s^2}{\chi^2_{n-1,\,1-\alpha/2}}) = 1-\alpha$$

So a (1-α) CI for σ^2, is $(\frac{(n-1)s^2}{\chi^2_{n-1,\,\alpha/2}}, \frac{(n-1)s^2}{\chi^2_{n-1,\,1-\alpha/2}})$.

We will focus on CI for population mean in rest of class.

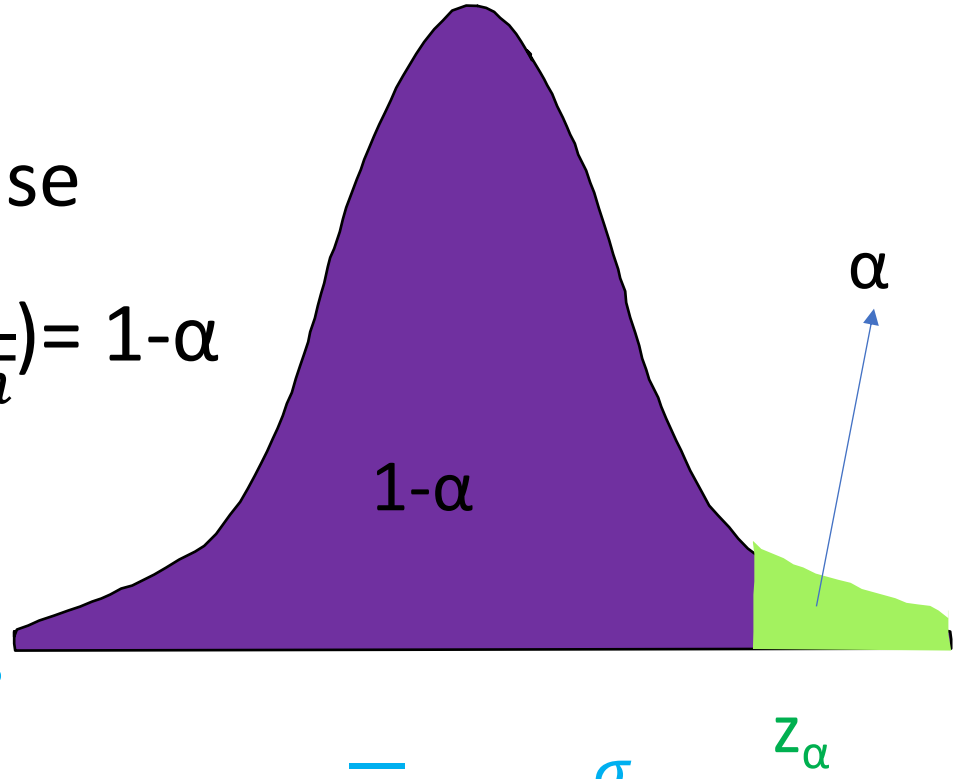
One-sided CI of the population Mean μ

- A (1-α) CI for μ, when $\frac{\sigma}{\sqrt{n}}$ is known, is $(\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}})$.
- Notice the above interval is symmetric around. We call it *two-sided confidence interval*.
- If we only want a lower bound for μ, use

$$P(\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \le z_\alpha) = 1-\alpha \iff P(\overline{X} - \mu \le z_\alpha \frac{\sigma}{\sqrt{n}}) = 1-\alpha$$

$\iff P(\overline{X} - z_\alpha \frac{\sigma}{\sqrt{n}} \le \mu) = 1-\alpha$. Hence

one-sided (1-α) CI for μ is $(\overline{X} - z_\alpha \frac{\sigma}{\sqrt{n}}, \infty)$.



- For upper bound, a 1-sided (1-α) CI for μ is $(-\infty, \overline{X} + z_\alpha \frac{\sigma}{\sqrt{n}})$.

One-sided CI of the population Mean μ

- A 2-sided (1-α) CI for μ is $(\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}})$.
- A 1-sided (1-α) CI for μ is $(\overline{X} - z_{\alpha}\frac{\sigma}{\sqrt{n}}, \infty)$ or $(-\infty, \overline{X} + z_{\alpha}\frac{\sigma}{\sqrt{n}})$.
- For example,

  a 2-sided 95% CI for μ is $(\overline{X} - 1.96\frac{\sigma}{\sqrt{n}}, \overline{X} + 1.96\frac{\sigma}{\sqrt{n}})$;

  a lower 1-sided 95% CI for μ is $(\overline{X} - 1.64\frac{\sigma}{\sqrt{n}}, \infty)$;

  an upper 1-sided 95% CI for μ is $(-\infty, \overline{X} + 1.64\frac{\sigma}{\sqrt{n}})$;

Notice a 2-sided 90% CI for μ is $(\overline{X} - 1.64\frac{\sigma}{\sqrt{n}}, \overline{X} + 1.64\frac{\sigma}{\sqrt{n}})$.

t-interval of the population Mean μ

- (1) $\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1) \Rightarrow$ A (1-α) CI for μ is $(\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}})$.
- Using same derivations on (3) $\frac{\overline{X}-\mu}{s/\sqrt{n}} \sim t_{n-1}$, we get

  A (1-α) CI for μ is $(\overline{X} - t_{n-1,\,\alpha/2}\frac{s}{\sqrt{n}}, \overline{X} + t_{n-1,\,\alpha/2}\frac{s}{\sqrt{n}})$.

Get $t_{n-1,\,\alpha/2}$ from TableA.4. Or in R, qt(1- α/2, df=n-1)

In most cases, we *do not know* the population parameter σ. So we should use the t-interval.

- A 1-sided (1-α) t-interval for μ is $(\overline{X} - t_{n-1,\,\alpha}\frac{s}{\sqrt{n}}, \infty)$

  or $(-\infty, \overline{X} + t_{n-1,\,\alpha}\frac{s}{\sqrt{n}})$.

t-interval of the population Mean μ

Example: A random sample of 24 male runners are taken. The sample mean is $\overline{X}$ =60kg. Suppose that the *sample standard deviation* of the male runners population is s=5kg.

Then, since $t_{23,\,0.025}$ =2.069, a 95% 2-sided CI for μ is

$$(60\text{-}2.069\,\frac{5}{\sqrt{24}}\,,\ 60\text{+}2.069\,\frac{5}{\sqrt{24}}) = (58.9,62.1).$$

Since $t_{23,\,0.05}$ =1.714, a 95% 1-sided CI for μ is

$$(60\text{-}1.714\,\frac{5}{\sqrt{24}}\,,\ \infty) \text{ or } (-\infty,\ 60\text{+}1.714\,\frac{5}{\sqrt{24}})\ .$$

R Commands for CI of the population Mean μ

Ex (In CalculateCI.pdf) Recall the earlier data set of grocery shoppers.

18.71	32.82	37.52	33.26	6.90
31.99	39.28	69.49	19.55	12.66
27.07	63.85	34.76	20.89	16.55
23.85	30.54	40.80	52.36	15.01
14.35	14.52	20.58	33.80	13.72
36.22	29.15	43.97	45.58	15.33
21.13	14.55	13.67	61.57	18.30
20.91	64.30	11.34	18.22	17.15
2.32	26.04	28.76	8.04	9.45
19.54	11.63	6.61	12.95	10.26

- Z-interval: we can just get the sample statistics using R and plug in formulas as before: $(\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}})$. Just need to find $\overline{X}$ and n from the data set.

R Commands for CI of the population Mean μ

Example (In CalculateCI.pdf)

```
> x <- scan(file="grocery.txt") #Read in unformatted data
Read 50 items
> (n <- length(x)) #Sample size, the extra () displays the result 'n'
[1] 50
> (xbar <- mean(x) ) #Sample mean
[1] 25.8364
```

- From the R outputs, n =50, $\overline{X}$ =25.8364.
- If σ^2=5, a 95% CI for μ (z-interval) is $(\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}})$ so

$$(25.8364 - 1.96\frac{\sqrt{5}}{\sqrt{50}}, 25.8364 + 1.96\frac{\sqrt{5}}{\sqrt{50}}) = (25.217, 26.456).$$

- However, should we use the z-interval here? Do we trust σ^2=5?

```
> var(x)
[1] 260.9305
```

- In this data set, s^2= 260.9305. So σ^2=5 **unlikely**.
- We really should be using the t-interval in practice! (unknown σ^2)

R Commands for CI of the population Mean μ

Example: The 95% t-interval is $(\overline{X}\text{-}t_{n-1,\,0.025}\frac{s}{\sqrt{n}}, \overline{X}\text{+}t_{n-1,\,0.025}\frac{s}{\sqrt{n}})$.

We can get the n=50, $\overline{X}$ and s from R outputs as before.
To get $t_{49,\,0.025}$ from Table A.4, there is no df=49 in the table. In such case, we use the closest df=50, i.e., $t_{50,\,0.025}$=2.009, and plug-in formula.
Or get $t_{49,\,0.025}$ directly from R: qt(1-0.025, df=49).

- We can let R do all the calculation:

```
> alpha <- 0.05
> xbar + c(-1,1) * qt(1-alpha/2, df=n-1) * sd(x)/sqrt(n)
[1] 21.24567 30.42713
```

- So the 95% t-interval is = (21.25, 30.43).

- An even quicker way is to use the t.test() procedure (which is mainly used for hypothesis test in next chapter, but does produce CI too.)

R Commands for CI of the population Mean μ

Example (In CalculateCI.pdf)

- An even quicker way is to use the t.test() procedure (which is mainly used for hypothesis test in next chapter, but does produce CI too.)

```
> t.test(x, mu=22, conf.level=0.95)
        One Sample t-test
data:  x
t = 1.6794, df = 49, p-value = 0.09945
alternative hypothesis: true mean is not equal to 22
95 percent confidence interval:
 21.24567 30.42713
sample estimates:
mean of x
  25.8364
```

- So the 95% 2-sided t-interval is (21.25, 30.43).

CI of the population Mean μ

- Question: Which is longer, z-interval or t-interval?

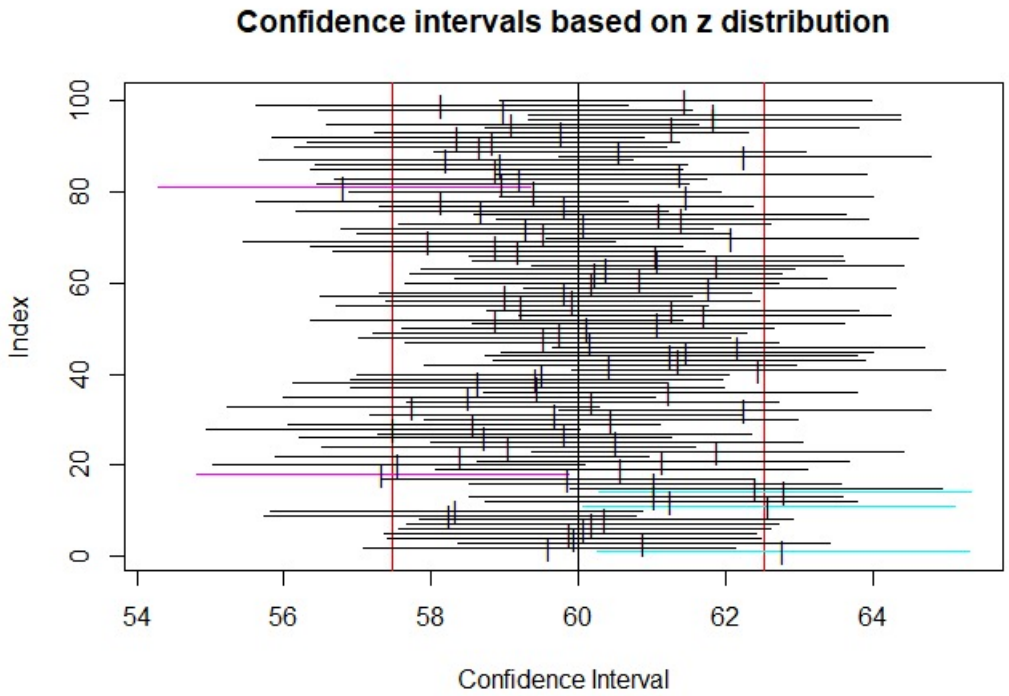
Meaning of Confidence interval

- In above example, a 95% t-interval is (21.25, 30.43). Does this mean that the true parameter μ falls within the interval (21.25, 30.43) with 95% probability?
- No! There is a 95% probability for μ to fall in the confidence interval calculated by this formula over many independent data sets. But on each data set, the interval changes, not always (21.25, 30.43).
- μ is an *unknown* parameter. It is either in (21.25, 30.43) or not. So probability is either 0 or 1. We just do not know it!

Meaning of Confidence interval

- Use the R demo to generate 100 CIs.

```
> library(TeachingDemos)
> ci.examp(mean.sim=60,sd=5,n=15,reps=100, method='z', lower.conf=0.025,upper.conf=0.975)
```



Sample size calculation

- We want to know how much coca-cola is there in a 2-liter bottle on average. Assume that we know σ =0.1. How many bottles are needed to determine the mean volume to within 0.01 liter with 95% confidence?
- Solution: Want $\overline{X}$ to be within $\pm$ 0.01 of μ.

Need the 95% CI $\overline{X}$ $\pm$L have ½-length L $\leq$0.01.

Since 95% CI is $\overline{X} \pm z_{0.025}\frac{\sigma}{\sqrt{n}}$, this is $z_{0.025}\frac{\sigma}{\sqrt{n}} \leq 0.01$.

Solve $1.96\frac{0.1}{\sqrt{n}} \leq 0.01$, we get $\sqrt{n} \geq 1.96\frac{0.1}{0.01}$.

So $n \geq \left(\frac{1.96\times 0.1}{0.01}\right)^2 = 384.16$. *Round up*, the answer is n=385 bottles.

Sample size calculation

- Generally, say, we want to estimate the parameter μ at (1-α) confidence to within ±L, that is, we want

$$P\left(\overline{X} - L \leq \mu \leq \overline{X} + L\right) \geq 1 - \alpha$$

- Then we should solve $z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq L$.
- The solution is $\mathrm{n} \geq \left(z_{\alpha/2}\frac{\sigma}{L}\right)^2$. For final answer, remember to *round up* to an integer.

- Question: Why we use the z-interval instead of t-interval in the sample size calculation?

Module 3 done

- We finished covering the confidence intervals (Chapter 9). You need to know:
- What does CI mean? (coverage over many data sets.)
- Which CI formula to use? t or z, 1-sided vs. 2-sided
- Sample size calculation.
- Using R to calculate CI's.

- We now start on hypothesis tests (Chapter 10).

Chapter 10 Hypothesis Testing

- Recall the definition of statistics:

A form of Mathematics concerned with methods for Collecting, Organizing, Presenting, and Analyzing Data, as well as drawing valid conclusions and making reasonable decisions on the basis of such analyses.

- **Hypothesis testing** is a way to formalize the "drawing conclusion" part.

Chapter 10 Hypothesis Testing

There are four elements in the hypothesis testing

- **Alternative hypothesis** H_A :

  A statement about the parameter. Typically the goal is to establish the truth of H_A. (We want to prove this!)

- **Null hypothesis** H_0 :

  A statement about the parameter, the logical opposite of H_A. The goal is *to reject* H_0.

- **Test criterion** (or test statistic):

  A statistic that is used to make decisions.

- **Rejection region**:

  A rule for using the test criterion.

Chapter 10 Hypothesis Testing

Example: A standard pain reliever is known to bring relief in 3.5 minutes on average.

A researcher wants to show that a new drug is better.

Data: Time to pain relief for persons in the trial.

$n = 120, \overline{X} = 3.2, s = 1.14$.

Parameter μ = mean time to pain relief in all humans.

What are the four elements of hypothesis testing for this study?

Pain reliever example

Data: $n = 120, \overline{X} = 3.2, s = 1.14$.

Want: a new drug is better than the standard pain reliever (3.5 minutes to pain relief).

Parameter μ = mean time to pain relief in all humans.

- $\mathbf{H_A}$: $\mu < 3.5$
- $\mathbf{H_0}$: $\mu \geq 3.5$
- **Test statistic** $\overline{X}$
- **Rejection region**: any rule you can think of.

E.g., $\overline{X}$ <3.5, or $\overline{X}$ <3.2, or $\overline{X}$ >4, etc. Which one to use?

Two types of errors in hypothesis testing

Decision	True state: H_0 is true (innocent)	True state: H_A is true (murder)
Fail to reject H_0 (Not guilty)	Correct $1-\alpha$	**Type II error** β
Reject H_0 (Guilty)	**Type I error** α	Correct $1-\beta$

α = P(reject H_0 | H_0 is true) is also called **significance level** (or size)

β = P(fail to reject H_0 | H_A is true). $1-\beta$ is called the **power**.

Pain reliever example

- $\mathbf{H_0}: \mu \geq 3.5$

Since $\overline{X} \approx \mu$ we want to reject $\mathbf{H_0}$ if $\overline{X}$ is small.

Say, we use the rule: if $\overline{X}$ <3.5, then reject H_0.

Then the significance level is

$P(\text{reject } H_0 \mid \mu = 3.5) = P(\overline{X} < 3.5 \mid \mu = 3.5)$

$= 1/2$

<u>This is not a good rule</u>, since Type I error is 50%.

Pain reliever example

- $\mathbf{H_0}: \mu \geq 3.5$

We want to use the rejection region of $\overline{X} < x_0$.

As $x_0 \searrow$, type I error $\alpha = P(\overline{X} < x_0 \mid \mu = 3.5) \searrow$

type II error $\beta = P(\overline{X} \geq x_0 \mid \mu < 3.5) \nearrow$

As $x_0 \nearrow$, type I error $\alpha = P(\overline{X} < x_0 \mid \mu = 3.5) \nearrow$

type II error $\beta = P(\overline{X} \geq x_0 \mid \mu < 3.5) \searrow$

How to balance these errors?

Pain reliever example: Formulation of test

- μ = true mean, μ_0=3.5 (standard drug). Test

$$\mathbf{H_0}: \mu \geq \mu_0 \text{ versus } \mathbf{H_A}: \mu < \mu_0$$

- Choose α=0.05 significance level (tolerable error)

Reject $\mathbf{H_0}$ if $\overline{X} < x_0$. x_0=?

Want $\alpha = P(\text{reject } H_0 \mid \mu = \mu_0) = P(\overline{X} < x_0 \mid \mu = \mu_0) = 0.05$

Now, we can solve the value of x_0 from probability theory.

Pain reliever example: Formulation of test

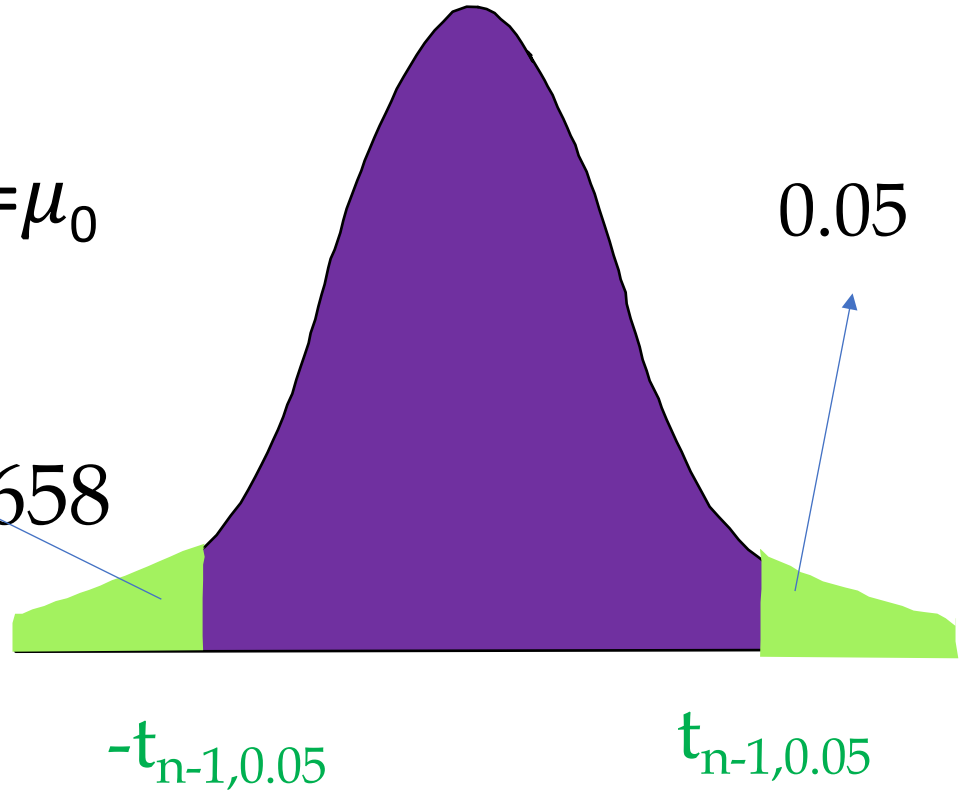
Want $P(\overline{X} < x_0 \mid \mu = \mu_0) = 0.05$

From probability theory, under H_0: $\mu = \mu_0$

$$P\left(\frac{\overline{X} - \mu_0}{s/\sqrt{n}} < -t_{n-1,0.05}\right) = 0.05$$

Use Table A.4, $-t_{119,0.05} \approx -t_{120,0.05} = -1.658$

Hence $$P\left(\frac{\overline{X} - 3.5}{1.14/\sqrt{120}} < -1.658\right) = 0.05$$



The rejection region is , $\frac{\overline{X} - 3.5}{1.14/\sqrt{120}} < -1.658$

i.e., $\overline{X} < 3.5 - 1.658\,(1.14/\sqrt{120}) \Leftrightarrow \overline{X} < 3.327$

Pain reliever example: test

- μ = true mean, μ_0=3.5 (standard drug). Test

  $\mathbf{H_0}$: $\mu \geq \mu_0$ versus $\mathbf{H_A}$: $\mu < \mu_0$

- Choose α=0.05. Then reject $\mathbf{H_0}$ if $\overline{X} < 3.327$.

Here, we observe $\bar{X}_{obs}$ =3.2 <3.327.

Hence, we do reject $\mathbf{H_0}$ at α=0.05 level.

Conclusion: the new drug is better.

Summary

Module 3 (confidence intervals) done. You should know:

- What does CI mean? (coverage over many data sets.)
- Which CI formula to use? t or z, 1-sided vs. 2-sided
- Sample size calculation.
- Using R to calculate CI's.
- Monte Carlo simulation.
- Homework 3 due in one week.

- We also started on the concept of hypothesis test (Chapter 10). We will finish the topic of hypothesis tests in next lecture.