Prelim 2 Review

October 29 - Nov 5, 2019

Topics

Topics covered after Prelim 1

- Sampling Distribution and Central Limit Theorem, bootstrap
- t-distribution
- Confidence Intervals (CI), Hypothesis Testing (HT)
- Five Scenarios for CI and HT:
  1. one population mean (μ)
  2. One population proportion (p)
  3. Difference of two related means (μ_d) – essentially case 1
  4. Difference of two unrelated means ($\mu_1 - \mu_2$)
  5. Difference of two population proportions ($p_1 - p_2$)
- Categorical data: χ^2 tests of independence and goodness-of-fit

Five Scenarios for CI and HT

1. one population mean (μ)
   - average prelim 1 score in a class
2. One population proportion (p)
   - proportion of students scoring A in prelim 1
3. Difference of two related means (μ_d) – essentially case 1
   - difference of average scores in prelim 1 and prelim 2
4. Difference of two unrelated means ($\mu_1 - \mu_2$)
   - difference of average prelim 1 scores in labs 402 and 405
5. Difference of two population proportions ($p_1 - p_2$)
   - difference of proportion of students scoring A in labs 402 and 405

Checkpoints

- Calculate quantiles using R commands

  `qnorm`, `qt`, `qbinom`, `qunif`

- Distinguish between population and sample, paramater and statistics
- Interpret standard deviation, standard error, confidence level, margin of error, p-value, significance level, type-I error, type-II error, power **in the context of a problem**
- Explain in words the meaning of confidence interval, conclusion of a hypothesis test **in the context of a problem**
- Check assumptions for CI/HT in each of the five scenarios

On CI and HT

CI, margin of error, sample size

> Based on a poll of 100 Cornell undergrads, a 95% confidence interval was constructed to estimate the proportion of students who want to major in STEM fields. The interval was $(0.49, 0.61)$.

- What proportion of undergrads in the survey wanted to major in STEM fields?
- What is the margin of error in this study?
- If we want to reduce the margin of error by half, how many more students do we need to survey?

CI, margin of error, sample size

> Based on a poll of 100 Cornell undergrads, a 95% confidence interval was constructed to estimate the proportion of students who want to major in STEM fields. The interval was $(0.49, 0.61)$.

What proportion of undergrads in the survey wanted to major in STEM fields?

1. 0.49
2. 0.61
3. (0.49+0.61)/2
4. 0.95

CI, margin of error, sample size

> Based on a poll of 100 Cornell undergrads, a 95% confidence interval was constructed to estimate the proportion of students who want to major in STEM fields. The interval was $(0.49, 0.61)$.

What is the margin of error in this study?

1. 0.05
2. 0.61 - 0.49
3. (0.61 - 0.49)/2
4. (0.61 + 0.49)/2

CI, margin of error, sample size

> Based on a poll of 100 Cornell undergrads, a 95% confidence interval was constructed to estimate the proportion of students who want to major in STEM fields. The interval was $(0.49, 0.61)$.

If we want to reduce the margin of error by half, how many students do we need to survey?

1. 50
2. 200
3. 300
4. 400

Type-I and Type-II errors

> A clinical trial ended with a report: “At 1% level of significance, there is not sufficient evidence that the new drug can reduce cholesterol level in middle aged men.”

- What type of error may have been committed?

1. Type-I
2. Type-II

- If we change the level of significance to 5%, would the chance of this type of error increase?

1. Yes
2. No

- What should be the next step from here?

Type-I and Type-II errors

> A clinical trial ended with a report: “At 1% level of significance, there **was** sufficient evidence that the new drug can reduce cholesterol level in middle aged men.”

- What type of error may have been committed?

1. Type-I
2. Type-II

- If we change the level of significance to 5%, would the chance of this type of error increase?

1. Yes
2. No

- What should be the next step from here?

Five Scenarios for CI and HT

One population Mean

Recommended Exercise: 5.1, 5.3, 5.4, 5.5, 5.7, 5.12

1. one population mean (μ)

> Historically, the ~~average~~ scores of students in BTRY6010 prelim1 are normally distributed with a standard deviation of 7. This year, a randomly selected sample of 16 students has an average score of 82 with a standard deviation of 5.

Q: What is the population parameter of interest?

A: average score of **all** students in BTRY6010 prelim1 (μ)

or,

A: **population** mean score of students in BTRY6010 prelim1 (μ)

1. one population mean (μ)

> Historically, the ~~average~~ scores of students in BTRY6010 prelim1 are normally distributed with a standard deviation of 7. This year, a randomly selected sample of 16 students has an average score of 82 with a standard deviation of 5.

Q: Construct a 95% confidence interval for the average prelim1 score of all students.

A: Step-by-step:

- CI: (point estimate - margin of error, point estimate + margin of error)
- margin of error = multiplier ($z_{\alpha/2}$ or $t_{n-1,\alpha/2}$) $\times$ SE(point estimate)
- point estimate: sample statistic $\bar{x}_n = 82$, sample size $n = 16$
- SE($\bar{x}$) = $\sigma/\sqrt{n}$ = $7/\sqrt{16}$ = 7/4 = 1.75
- multiplier: data is normally distributed, so use $z_{\alpha/2}$ with $\alpha = 0.05$

1. one population mean (μ)

> Historically, the ~~average~~ scores of students in BTRY6010 prelim1 are normally distributed with a standard deviation of 7. This year, a randomly selected sample of 16 students has an average score of 82 with a standard deviation of 5.

- (Draw picture on board)

```
alpha = 0.05
z = -qnorm(alpha/2, mean=0, sd=1)
z
```

```
## [1] 1.959964
```

- A 95% CI for μ is given by

$$(82 - 1.96 \times 1.75, 82 + 1.96 \times 1.75)$$

1. one population mean (μ)

> Historically, the ~~average~~ scores of students in BTRY6010 prelim1 are normally distributed with a standard deviation of 7. This year, a randomly selected sample of 16 students has an average score of 82 with a standard deviation of 5.

Q: Interpret the 95% CI **in the context of the problem**

- We are **95% confident** that the average prelim1 score of **all** BTRY6010 students is somewhere between $82 - 1.96 \times 1.75$ and $82 + 1.96 \times 1.75$.

Q: Explain the term 95% level of confidence **in the context of this problem**

- If we draw many simple random samples of 16 students and create 95% CIs using this procedure, we expect **95% of those intervals** to capture the **average prelim1 score of all BTRY6010 students**.

1. one population mean (μ)

> Historically, the ~~average~~ scores of students in BTRY6010 prelim1 are normally distributed with a standard deviation of 7. This year, a randomly selected sample of 16 students has an average score of 82 with a standard deviation of 5.

Q: How would our calculations change if

1. we did not know that the population standard deviation was 7?
   - SE($\bar{x}$) = $S/\sqrt{n}$ = $5/\sqrt{16}$ = 1.25
   - multiplier: $t_{15,0.025}$ = `-qt(.025, df = 15)`
2. we did not know that the scores are normally distributed?
   - we would not have enough information to calculate the 95% CI

1. one population mean (μ)

1. we did not know that the scores were normal, but $n = 35$?

- If we knew $\sigma = 7$, use $z_{\alpha/2}$, $n = 35$ and $\sigma = 7$
- If not, use $t_{34,0.025}$, $n = 35$ and $S = 5$ ($z_{0.025}$ **is fine, too!**)

1. one population mean (μ)

> Historically, the ~~average~~ scores of students in BTRY6010 prelim1 are normally distributed with a standard deviation of 7. This year, a randomly selected sample of 16 students has an average score of 82 with a standard deviation of 5. **Is the average prelim1 score of all BTRY6010 students different from 85**?

- Define parameter of interest

  Let μ be the average prelim1 score of **all** BTRY6010 students

- Define null and alternative hypotheses

$$H_0 : \mu = 85 \quad H_A : \mu \neq 85$$

- Identify null value μ_0, sample statistic (or point estimate) and its distribution under null

  $\bar{x} = 82$, $\mu_0 = 85$, under null $\bar{x} \sim N(\mu_0, \sigma/\sqrt{n})$

1. one population mean (μ)

> Historically, the ~~average~~ scores of students in BTRY6010 prelim1 are normally distributed with a standard deviation of 7. This year, a randomly selected sample of 16 students has an average score of 82 with a standard deviation of 5. **Is the average prelim1 score of all BTRY6010 students different from 85**?

- Calculate test statistic, note its distribution under null

$$Z = \frac{\bar{x} - \mu_0}{SE(\bar{x})} = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{82 - 85}{7/\sqrt{16}}$$

Under null, $Z \sim N(0, 1)$

```
xbar = 82; mu0 = 85; sigma = 7; n = 16
se_xbar = sigma/sqrt(n)
teststat = (xbar - mu0)/se_xbar
teststat
```

```
## [1] -1.714286
```

1. one population mean (μ)

> Historically, the ~~average~~ scores of students in BTRY6010 prelim1 are normally distributed with a standard deviation of 7. This year, a randomly selected sample of 16 students has an average score of 82 with a standard deviation of 5. **Is the average prelim1 score of all BTRY6010 students different from 85**?

- Calculate p-value

  $pval = P(Z < -1.714) + P(Z > 1.714) = 2P(Z > 1.714)$ (Draw picture on board)

```
pval = 2*pnorm(abs(teststat), lower.tail = FALSE)
pval
```

```
## [1] 0.08647627
```

1. one population mean (μ)

> Historically, the ~~average~~ scores of students in BTRY6010 prelim1 are normally distributed with a standard deviation of 7. This year, a randomly selected sample of 16 students has an average score of 82 with a standard deviation of 5. **Is the average prelim1 score of all BTRY6010 students different from 85**?

- Make a decision for a specific level of significance α
  - **Remember: $pvalue < \alpha$ means Reject H_0 in favor of H_A**
  - Our p-value is 8.6%
  - we reject H_0 in favor of H_A at 10% level of significance
  - we fail to reject H_0 in favor of H_A at 5% level of significance
  - we fail to reject H_0 in favor of H_A at 1% level of significance

1. one population mean (μ)

> Historically, the ~~average~~ scores of students in BTRY6010 prelim1 are normally distributed with a standard deviation of 7. This year, a randomly selected sample of 16 students has an average score of 82 with a standard deviation of 5.

Q: Is the average prelim1 score of all BTRY6010 students different from 85?

- Summarize this conclusion in your report:

> "At 5% level of significance, we do not have sufficient evidence ($p = 0.086$) to say that the average prelim1 score of all BTRY6010 students is different from 85."

1. one population mean (μ)

> Historically, the ~~average~~ scores of students in BTRY6010 prelim1 are normally distributed with a standard deviation of 7. This year, a randomly selected sample of 16 students has an average score of 82 with a standard deviation of 5.

Q: How will the result change if knew that the scores are normal, but did not know that the standard deviation is 7?

- We would have used a t-test with $S = 5$, $df = 15$, i.e.,
- Our test statistic would have been

$$T = \frac{\bar{x} - \mu_0}{SE(\bar{x})} = \frac{\bar{x} - \mu_0}{S/\sqrt{n}} = \frac{82 - 85}{5/\sqrt{16}}$$

and we would have used t-distribution with $df = 15$ to calculate p-value

1. one population mean (μ)

> Historically, the ~~average~~ scores of students in BTRY6010 prelim1 are normally distributed with a standard deviation of 7. This year, a randomly selected sample of 16 students has an average score of 82 with a standard deviation of 5.

Q: How will the result change if knew that the scores are normal, but did not know that the standard deviation is 7?

```
xbar = 82; mu0 = 85; S = 5; n = 16
se_xbar = S/sqrt(n)
teststat = (xbar - mu0)/se_xbar
teststat
```

```
## [1] -2.4
```

```
pval = 2*pt(abs(teststat), df = n-1, lower.tail = FALSE)
pval
```

```
## [1] 0.02982493
```

1. one population mean (μ)

> Historically, the ~~average~~ scores of students in BTRY6010 prelim1 are normally distributed with a standard deviation of 7. This year, a randomly selected sample of 16 students has an average score of 82 with a standard deviation of 5. Is the average prelim1 score of all BTRY6010 students **less than** 85?

- Definition of alternative hypothesis changes

$$H_0 : \mu = 85 \quad H_A : \mu < 85$$

- sample statistic, test statistic, their null distributions are the same
- However, the p-value calculation changes (**"extreme" is only one-sided now**)

$pval = P(Z < -1.714)$ (Draw picture on board)

```
pval = pnorm(-abs(teststat))
pval
```

```
## [1] 0.04323813
```

1. one population mean (μ)

> Historically, the ~~average~~ scores of students in BTRY6010 prelim1 are normally distributed with a standard deviation of 7. This year, a randomly selected sample of 16 students has an average score of 82 with a standard deviation of 5. Is the average prelim1 score of all BTRY6010 students **less than** 85?

- Q: How does this p-value ($\sim 4\%$) compare with the one from two-sided hypothesis ($\sim 8\%$)?
- The p-value from one-sided hypothesis is **exactly half** of the p-value from two-sided hypothesis
- Q: Is it always the case?
- No, but you can find it with a small calculation

1. one population mean (μ)

> Historically, the ~~average~~ scores of students in BTRY6010 prelim1 are normally distributed with a standard deviation of 7. This year, a randomly selected sample of 16 students has an average score of 82 with a standard deviation of 5. Is the average prelim1 score of all BTRY6010 students **more than** 85?

$$H_0 : \mu = 85 \quad H_A : \mu > 85$$

- sample statistic, test statistic, their null distributions are the same

- $pval = P(Z > -1.714)$ (Draw picture on board)

```
pval = pnorm(teststat, lower.tail=FALSE)
pval
```

```
## [1] 0.9567619
```

- p-value is **large**, so fail to reject H_0 in favor or $H_A : \mu > 85$. Since $\bar{x} = 82$, **was that really a surprise**?

1. one population mean (μ)

To sum up …

- We saw how to calculate/interpret CI and HT
- Did not check all the assumptions carefully (more on this in next two lectures)
- **Most importantly**, did almost all our calculations with 3 things:

  1. point estimate ($\bar{x}$);
  2. its standard error $SE(\bar{x})$
  3. A quantile from Normal/t-distribution ($z_{\alpha/2}$ or $t_{n-1,\alpha/2}$ for CI),

  or

  iii') Probability to the left or right of the observed test statistic (Z or T) for HT

- **For the other 4 scenarios: formulas of i) and ii) change, but everything else remains (mostly) same!!**

One population proportion

Recommended Exercise: 6.1, 6.3, 6.5, 6.9, 6.12, 6.16,

6.17

One population proportion (p)

> Historically, 50% of students in BTRY6010 get A in prelim1. This year, 25 out of a randomly selected sample of 40 students received A in prelim1.

Q: Construct a 90% confidence interval of the proportion of BTRY6010 students who received A in prelim1.

- population parameter p: proportion of all BTRY6010 students who received A in prelim1 this year
- sample statistic: $X = 25, n = 40, \hat{p} = X/n = 25/40 = 0.625$
- Its standard error $SE(\hat{p}) = \sqrt{\frac{0.625\times(1-0.625)}{40}}$
- $\alpha = 1 - 0.9$, multiplier $z_{\alpha/2}$

```
alpha = 1-0.9; z = -qnorm(alpha/2); z
```

```
## [1] 1.644854
```

One population proportion (p)

> Historically, 50% of students in BTRY6010 get A in prelim1. This year, 25 out of a randomly selected sample of 40 students received A in prelim1.

Q: Construct a 90% confidence interval of the proportion of all BTRY6010 students who received A in prelim1.

```
x = 25; n=40; phat = x/n
se = sqrt(phat*(1-phat)/n)
alpha = 1-0.9; z = -qnorm(alpha/2)
phat
phat-z*se
phat+z*se
```

```
## [1] 0.625
```

```
## [1] 0.4990921
```

```
## [1] 0.7509079
```

One population proportion (p)

> Historically, 50% of students in BTRY6010 get A in prelim1. This year, 25 out of a randomly selected sample of 40 students received A in prelim1.

Q: Construct a 90% confidence interval of the proportion of all BTRY6010 students who received A in prelim1.

- Interpret this CI?
- Check Assumptions

  $n\hat{p} \geq 10, n(1 - \hat{p}) \geq 10$

One population proportion (p)

> Historically, 50% of students in BTRY6010 get A in prelim1. This year, 25 out of a randomly selected sample of 40 students received A in prelim1.

Q: Is the proportion of A students this year significantly higher than the historical average?

- $H_0 : p = 0.5$ and $H_A : p > 0.5$
- null value $p_0 = 0.5$, sample statistic $\hat{p} = 0.625$
- Test Statistic (follows N(0,1) under null) $Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$
- Calculate p-value (Draw picture)

One population proportion (p)

> Historically, 50% of students in BTRY6010 get A in prelim1. This year, 25 out of a randomly selected sample of 40 students received A in prelim1.

Q: Is the proportion of A students this year significantly higher than the historical proportion of A students?

```
x=25; n=40; phat = x/n; p0 = 0.5
se = sqrt(p0*(1-p0)/n)
teststat = (phat-p0)/se
pval = pnorm(teststat, lower.tail = FALSE)
pval
```

```
## [1] 0.05692315
```

- Conclusion (Reject/Fail to reject H_0), report your findings
- Check assumptions (more on this in next lecture)

$$np_0 \geq 10, n(1-p_0) \geq 10$$

Paired Mean

Recommended Exercise: 5.17, 5.18, 5.19, 5.21

1. paired mean (μ_D)

> Prelim 1 and Prelim 2 scores of $n = 10$ randomly selected students are recorded. Find a 95% confidence interval for the mean difference of the two prelim scores. Clearly state any assumption you are making.

```
prelim1 = c(50, 61, 43, 73, 75, 68, 59, 32, 47, 52)
prelim2 = c(55, 62, 49, 71, 68, 72, 43, 38, 45, 45)
```

Calculate the difference of scores, and construct CI for one population mean

1. paired mean (μ_D)

> Prelim 1 and Prelim 2 scores of $n = 10$ randomly selected students are recorded. Find a 95% confidence interval for the mean difference of the two prelim scores. Clearly state any assumption you are making.

```
diff = prelim2 - prelim1
diff
```

```
##  [1]   5   1   6  -2  -7   4 -16   6  -2  -7
```

```
mean(diff)
```

```
## [1] -1.2
```

```
sd(diff)
```

```
## [1] 7.161626
```

1. paired mean (μ_D)

> Prelim 1 and Prelim 2 scores of $n = 10$ randomly selected students are recorded. Find a 95% confidence interval for the mean difference of the two prelim scores. Clearly state any assumption you are making.

1. paired mean (μ_D)

> Prelim 1 and Prelim 2 scores of $n = 10$ randomly selected students are recorded. Based on this data, is it reasonable to conclude that students performed differently on average in the two prelims? Clearly state any assumption you are making.