Class 25

DATA1220-55, Fall 2024

Sarah E. Grabinski

2024-11-01

In-Class Quiz

- Friday, November 15th, in-class (closed-note, open-R)

In-Class Quiz

- Friday, November 15th, in-class (closed-note, open-R)
- Covers Chapters 2.1-2.2, 5.1-5.3, 6.1-6.2/6.4, and/or 7.1/7.3

In-Class Quiz

- Friday, November 15th, in-class (closed-note, open-R)
- Covers Chapters 2.1-2.2, 5.1-5.3, 6.1-6.2/6.4, and/or 7.1/7.3
- Quiz will take $\sim$30 minutes, review of answers will follow

In-Class Quiz

- Friday, November 15th, in-class (closed-note, open-R)
- Covers Chapters 2.1-2.2, 5.1-5.3, 6.1-6.2/6.4, and/or 7.1/7.3
- Quiz will take ~30 minutes, review of answers will follow
- 5 extra credit points available (+0-5% to final grade)

Take-Home Quiz

- Due Monday, November 18th by 6:00pm (open-note, open-R)

Take-Home Quiz

- Due Monday, November 18th by 6:00pm (open-note, open-R)
- Covers Chapters 2.1-2.2, 5.1-5.3, 6.1-6.2/6.4, and/or 7.1/7.3

Take-Home Quiz

- Due Monday, November 18th by 6:00pm (open-note, open-R)
- Covers Chapters 2.1-2.2, 5.1-5.3, 6.1-6.2/6.4, and/or 7.1/7.3
- Will require use of R, but will be in Google Forms or Canvas

Take-Home Quiz

- Due Monday, November 18th by 6:00pm (open-note, open-R)
- Covers Chapters 2.1-2.2, 5.1-5.3, 6.1-6.2/6.4, and/or 7.1/7.3
- Will require use of R, but will be in Google Forms or Canvas
- Worth 10% of final grade, will be bonus points available

Statistical Analysis Workflow

1. Develop research question.

Statistical Analysis Workflow

1. Develop research question.
2. Identify target, study, and sample populations.

Statistical Analysis Workflow

1. Develop research question.
2. Identify target, study, and sample populations.
3. Collect data / take sample.

Statistical Analysis Workflow

1. Develop research question.
2. Identify target, study, and sample populations.
3. Collect data / take sample.
4. Check assumptions.

Statistical Analysis Workflow

1. Develop research question.
2. Identify target, study, and sample populations.
3. Collect data / take sample.
4. Check assumptions.
5. Calculate sample statistics for sample.

Statistical Analysis Workflow

1. Develop research question.
2. Identify target, study, and sample populations.
3. Collect data / take sample.
4. Check assumptions.
5. Calculate sample statistics for sample.
6. Infer the sampling distribution of the study population from the sample statistics.

Statistical Analysis Workflow

1. Develop research question.
2. Identify target, study, and sample populations.
3. Collect data / take sample.
4. Check assumptions.
5. Calculate sample statistics for sample.
6. Infer the sampling distribution of the study population from the sample statistics.
7. Test a hypothesis.

Statistical Analysis Workflow

1. Develop research question.
2. Identify target, study, and sample populations.
3. Collect data / take sample.
4. Check assumptions.
5. Calculate sample statistics for sample.
6. Infer the sampling distribution of the study population from the sample statistics.
7. Test a hypothesis.
8. Apply results to target population.

Example: One-Sample Proportion

Research question: Did it rain more than usual in April of 2024?

Example: One-Sample Proportion

Research question: Did it rain more than usual in April of 2024?

Based on historical data you pulled for 1980-2020 from the National Oceanic and Atmospheric Administration (NOAA), it has rained on 594 of the 1200 days in April.

Example: One-Sample Proportion

Research question: Did it rain more than usual in April of 2024?

Based on historical data you pulled for 1980-2020 from the National Oceanic and Atmospheric Administration (NOAA), it has rained on 594 of the 1200 days in April.

In 2024, it rained on 18 days and didn't rain on 12 days in April.

Sample Statistics

The sample statistic for a population proportion p is the sample proportion $\hat{p}$.

In 2024, it rained on 18 days and didn't rain on 12 days in April, and April has 30 days total.

Sample Statistics

The sample statistic for a population proportion p is the sample proportion $\hat{p}$.

In 2024, it rained on 18 days and didn't rain on 12 days in April, and April has 30 days total.

```
p_hat <- 18 / 30

p_hat
```

```
[1] 0.6
```

Inference: Rain Days April 2024

With 95% confidence, what percentage of days were rainy in April 2024?

Inference: Rain Days April 2024

With 95% confidence, what percentage of days were rainy in April 2024?

- $\hat{p} \pm Z^* \times SE_{\hat{p}}$
- $Z^* = Z_{1-\alpha/2}$
- $\text{Confidence} = 1 - \alpha$

Finding Z^*

If our confidence level is 95% (0.95), then our α is 0.05. We need the Z-score that corresponds to the probability $p = 1 - \alpha/2 = 0.975$. Use the `qnorm()` function to find Z^*.

```
z_star <- qnorm(0.975)

z_star
```

```
[1] 1.959964
```

Standard Error for a Single Proportion

```r
se_phat <- sqrt((0.6 * (1 - 0.6)) / 30)

se_phat
```

```
[1] 0.08944272
```

Confidence Interval for $\hat{p}$

Lower bound:

```
p_hat - z_star * se_phat
```

```
[1] 0.4246955
```

Upper bound:

```
p_hat + z_star * se_phat
```

```
[1] 0.7753045
```

Test Statistic

```r
z_test <- (p_hat - 0.5) / 0.09

z_test
```

```
[1] 1.111111
```

P-Value

```
pnorm(z_test, lower.tail = F) * 2
```

```
[1] 0.2665205
```

The Central Limit Theorem

The distribution of the sample statistic $\bar{x}$ or $\hat{p}$ approximates the normal distribution $N\left(\text{populationparameter}, \text{standarderror}\right)$ as $n \to \infty$.

The Central Limit Theorem

The distribution of the sample statistic $\bar{x}$ or $\hat{p}$ approximates the normal distribution $N\left(\mathrm{population parameter}, \mathrm{standard error}\right)$ as $n \to \infty$.

- $\bar{x} \sim N\left(\mu, \frac{\sigma}{n}\right))$
- $\hat{p} \sim N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$

The Central Limit Theorem

The distribution of the sample statistic $\bar{x}$ or $\hat{p}$ approximates the normal distribution $N\left(\mathrm{population parameter}, \mathrm{standard error}\right)$ as $n \to \infty$.

- $\bar{x} \sim N\left(\mu, \frac{\sigma}{n}\right))$
- $\hat{p} \sim N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$

The ***sampling distribution*** is normal with $\mu = \mathrm{sample statistic}$ and $\sigma = \mathrm{standard error}$.

Inference & Hypothesis Testing with Means

- The distribution of sample means $\bar{x}$ calculated from samples of size n from the same population approximates a normal distribution (i.e. the *sampling distribution*)

Inference & Hypothesis Testing with Means

- The distribution of sample means $\bar{x}$ calculated from samples of size n from the same population approximates a normal distribution (i.e. the *sampling distribution*)
- Observations in sample assumed to be ***independent and identically distributed*** (***i.i.d.***)

Inference & Hypothesis Testing with Means

- The distribution of sample means $\bar{x}$ calculated from samples of size n from the same population approximates a normal distribution (i.e. the *sampling distribution*)
- Observations in sample assumed to be ***independent and identically distributed*** (***i.i.d.***)
- Need $n \geq 30$ observations in sample

Inference & Hypothesis Testing with Means

- The distribution of sample means $\bar{x}$ calculated from samples of size n from the same population approximates a normal distribution (i.e. the *sampling distribution*)
- Observations in sample assumed to be ***independent and identically distributed*** (***i.i.d.***)
- Need $n \geq 30$ observations in sample
- Underlying population distribution is normal (less strict as sample n increases)

Sample Means & The Standard Normal (z) Distribution

- As n increases, the sampling distribution of $\bar{x}$ approximates the distribution $N\left(\mu, \frac{\sigma}{\sqrt{n}}\right))$

Sample Means & The Standard Normal (z) Distribution

- As n increases, the sampling distribution of $\bar{x}$ approximates the distribution $N\left(\mu, \frac{\sigma}{\sqrt{n}}\right))$
- When assumptions met, $\bar{x} \approx \mu$ and $s \approx \sigma$

Sample Means & The Standard Normal (z) Distribution

- As n increases, the sampling distribution of $\bar{x}$ approximates the distribution $N\left(\mu, \frac{\sigma}{\sqrt{n}}\right))$
- When assumptions met, $\bar{x} \approx \mu$ and $s \approx \sigma$
- $s \approx \sigma$ is a strong assumption!

The t distribution

- Better than z when the population standard deviation σ is unknown (almost always)

The t distribution

- Better than z when the population standard deviation σ is unknown (almost always)
- Appears normal, but is flatter to allow more uncertainty about $SE = \frac{s}{\sqrt{n}}$ of μ

The t distribution

- Better than z when the population standard deviation σ is unknown (almost always)
- Appears normal, but is flatter to allow more uncertainty about $SE = \frac{s}{\sqrt{n}}$ of μ
- Centered at 0 with the single parameter ***degrees of freedom*** $(df = n - 1)$

The t distribution

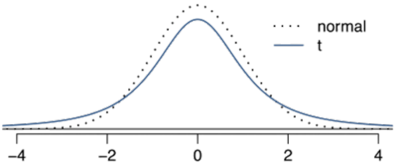


Figure 1: The t distribution versus the standard normal (z) distribution

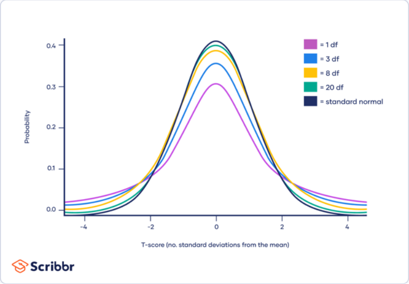


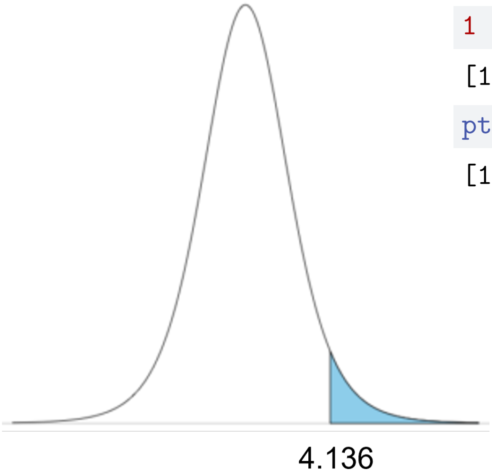
Figure 2: The t distribution is centered at 0 and has the parameter *degrees of freedom* (df)

t Distribution Test Statistic

$$T_{df} = \frac{\text{pointestimate} - \text{nullvalue}}{SE}$$

$$T_{n-1} = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

P-Values in R




```r
1 - pt(4.135, 205)
```

```
[1] 2.587688e-05
```

```r
pt(4.135, 205, lower.tail = F)
```

```
[1] 2.587688e-05
```

Confidence Intervals

- When $s \approx \sigma$, the confidence interval is $\text{point estimate} \pm Z^* \times SE$

Confidence Intervals

- When $s \approx \sigma$, the confidence interval is $\text{point estimate} \pm Z^* \times SE$
- When σ is unknown, we use $\text{point estimate} \pm T^* \times SE$

Confidence Intervals

- When $s \approx \sigma$, the confidence interval is $\text{point estimate} \pm Z^* \times SE$
- When σ is unknown, we use $\text{point estimate} \pm T^* \times SE$
- $T^* = T_{1-\alpha/2}$