Class 26

DATA1220-55, Fall 2024

Sarah E. Grabinski

2024-11-04

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 ~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

The Central Limit Theorem

The distribution of the sample statistic $\bar{x}$ or $\hat{p}$ approximates the normal distribution $N\left(\text{population parameter}, \text{standard error}\right)$ 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 $N\left(\text{populationparameter}, \text{standarderror}\right)$ 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 $\mu = \text{samplestatistic}$ and $\sigma = \text{standarderror}$.

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)

Population Parameters versus Sample Statistics

Table 1: Sample statistics are used to estimate unknowable population parameters

Measure	Sample Statistic	Population Parameter
Mean	$\bar{x}$	μ
Paired Difference in Means	$\bar{x}_{\text{difference}}$	$\mu_{\text{difference}}$
Difference in Means	$\bar{x}_1 - \bar{x}_2$	$\mu_1 - \mu_2$
Standard Deviation	s	σ

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 parameter ***degrees of freedom*** $(\mathrm{df} = n - 1)$

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 parameter ***degrees of freedom*** $(\mathrm{df} = n - 1)$
- $\bar{x} \sim t\,(\text{df} = n - 1)$

The t distribution

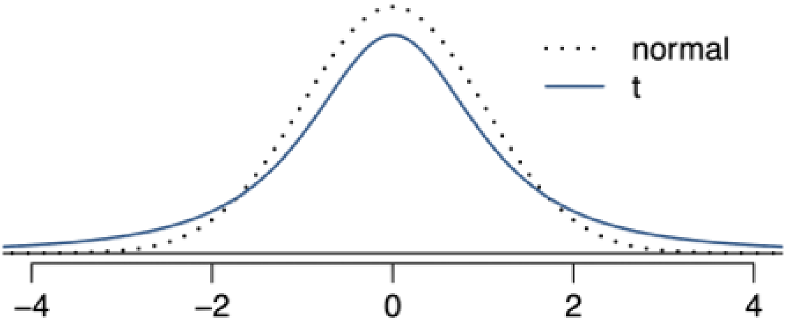

Figure 1: The t distribution versus the standard normal (z) distribution

The t distribution

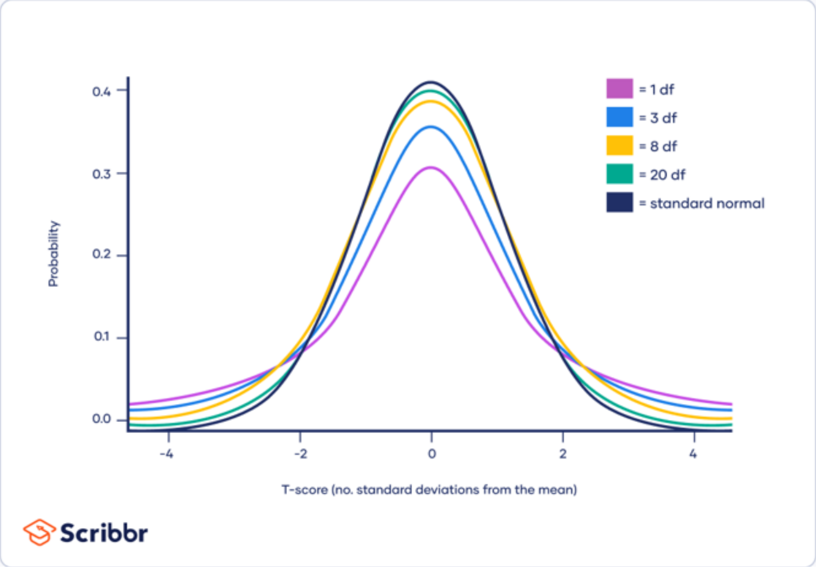

Figure 2: The t distribution is centered at 0 and has the parameter *degrees of freedom* (df)

Inference with the t distribution

Confidence intervals take the form point estimate $\pm T^* \times \mathrm{SE}$.

$$\bar{x} \pm T^* \times \frac{s}{\sqrt{n}}$$

$$\bar{x}_{\text{difference}} \pm T^* \times \frac{s_{\text{difference}}}{\sqrt{n}}$$

$$\bar{x}_1 - \bar{x}_2 \pm T^* \times \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Finding T^*

For a sample of size n, T^* is the value from a t distribution with $\mathrm{df} = n - 1$ and probability $\alpha/2$ or $1 - \alpha/2$.

$$T^* = T_{n-1,\alpha/2}$$

Finding T^*

For a sample of size n, T^* is the value from a t distribution with $\mathrm{df} = n - 1$ and probability $\alpha/2$ or $1 - \alpha/2$.

$$T^* = T_{n-1,\alpha/2}$$

Remember, to find the significance threshold α...
confidence $= 1 - \alpha$!

Getting T^* in R

For a 95% confidence interval for a sample of size $n = 100$, T^* is given by the `qt()` function. It takes the probability $p = \alpha/2$ or $p = 1 - \alpha/2$ and degrees of freedom $\mathrm{df} = n - 1$ as parameters.

```
qt(0.975, df = 100-1)
```

```
[1] 1.984217
```

Sample Size and the t distribution

When $n = 100$ and our confidence is 95%, $T^*_{100} = 1.98$. If we decrease the same size to $n = 50$, do you think T^*_{50} will be larger or smaller than T^*_{100}?

Sample Size and the t distribution

When $n = 100$ and our confidence is 95%, $T^*_{100} = 1.98$. If we decrease the same size to $n = 50$, do you think T^*_{50} will be larger or smaller than T^*_{100}?

Sampling distributions from small samples are more variable / have more error.

Sample Size and the t distribution

When $n = 100$ and our confidence is 95%, $T^*_{100} = 1.98$. If we decrease the same size to $n = 50$, do you think T^*_{50} will be larger or smaller than T^*_{100}?

Sampling distributions from small samples are more variable / have more error.

As the sample size n decreases, so does the degrees of freedom df for the t distribution.

Sample Size and the t distribution

When $n = 100$ and our confidence is 95%, $T^*_{100} = 1.98$. If we decrease the same size to $n = 50$, do you think T^*_{50} will be larger or smaller than T^*_{100}?

Sampling distributions from small samples are more variable / have more error.

As the sample size n decreases, so does the degrees of freedom df for the t distribution.

As df $= n - 1$ decreases, the tails of the distribution get fatter with more uncertainty.

Sample Size and the t distribution

When $n = 100$ and our confidence is 95%, $T^*_{100} = 1.98$. If we decrease the same size to $n = 50$, do you think T^*_{50} will be larger or smaller than T^*_{100}?

Sampling distributions from small samples are more variable / have more error.

As the sample size n decreases, so does the degrees of freedom df for the t distribution.

As $\text{df} = n - 1$ decreases, the tails of the distribution get fatter with more uncertainty.

As the sample n gets smaller, T^*_{n-1} for a given confidence level $1 - \alpha$ gets larger.

```
qt(0.975, df = 50-1)
```

```
[1] 2.009575
```

Hypothesis Tests Using Means

- One-Sample t-test: one sample mean where $\bar{x} \approx \mu$

Hypothesis Tests Using Means

- One-Sample t-test: one sample mean where $\bar{x} \approx \mu$
- Paired t-test: paired difference where $\bar{x}_{\text{difference}} \approx \mu$

Hypothesis Tests Using Means

- One-Sample t-test: one sample mean where $\bar{x} \approx \mu$
- Paired t-test: paired difference where $\bar{x}_{\text{difference}} \approx \mu$
- Two-Sample t-test: two sample means where $\bar{x}_1 - \bar{x}_2 \approx \mu_1 - \mu_2$

One-Sample t-Test Hypotheses

The null distribution for a one-sample t-test is $\bar{x} \sim N\left(\mu, \mathrm{SE}_{\bar{x}}\right)$.

$$H_0\colon\ \bar{x} = \mu$$

$$H_A\colon\ \bar{x} \neq \mu$$

One-Sample t-Test Hypotheses

The null distribution for a one-sample t-test is $\bar{x} \sim N\left(\mu, \mathrm{SE}_{\bar{x}}\right)$.

$$H_0\colon\ \bar{x} = \mu$$

$$H_A\colon\ \bar{x} \neq \mu$$

μ comes from a reference distribution (e.g. historical data, arbitrary threshold).

One-Sample Test Statistic

The degrees of freedom for a one-sample t-test is $\mathrm{df} = n - 1$.

$$T_{\mathsf{df}} = \frac{\text{point estimate} - \text{null value}}{SE}$$

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

Paired Data

- Occurs when 2 observations come from the same unit (i.e. ***not*** independent)
  - Example: Observations from the same subject at 2 time points (e.g. before/after)
  - Example: Observations from matched pairs like twins, husbands-wives, etc.

Paired Data

- Occurs when 2 observations come from the same unit (i.e. ***not*** independent)
  - Example: Observations from the same subject at 2 time points (e.g. before/after)
  - Example: Observations from matched pairs like twins, husbands-wives, etc.
- Paired data is analyzed as $\bar{x}_1 - \bar{x}_2 = \bar{x}_{\text{difference}}$

Paired Data

- Occurs when 2 observations come from the same unit (i.e. ***not*** independent)
  - Example: Observations from the same subject at 2 time points (e.g. before/after)
  - Example: Observations from matched pairs like twins, husbands-wives, etc.
- Paired data is analyzed as $\bar{x}_1 - \bar{x}_2 = \bar{x}_{\text{difference}}$
- Because both observations come from the same unit, $\bar{x}_{\text{difference}}$ is treated as a single sample

Paired Data

- Occurs when 2 observations come from the same unit (i.e. ***not*** independent)
  - Example: Observations from the same subject at 2 time points (e.g. before/after)
  - Example: Observations from matched pairs like twins, husbands-wives, etc.
- Paired data is analyzed as $\bar{x}_1 - \bar{x}_2 = \bar{x}_{\text{difference}}$
- Because both observations come from the same unit, $\bar{x}_{\text{difference}}$ is treated as a single sample
- The sample statistic $s_{\text{difference}}$ is the standard deviation of $\bar{x}_{\text{difference}}$

Paired t-Test Hypotheses

The null distribution for a paired t-test is $\bar{x}_{\text{difference}} \sim N\left(\mu_{\text{difference}}, \mathrm{SE}_{\text{difference}}\right)$.

$$H_0 : \bar{x}_{\text{difference}} = \mu_{\text{difference}}$$

$$H_A : \bar{x}_{\text{difference}} \neq \mu_{\text{difference}}$$

Paired t-Test Hypotheses

The null distribution for a paired t-test is
$\bar{x}_{\text{difference}} \sim N\left(\mu_{\text{difference}}, \mathrm{SE}_{\text{difference}}\right)$.

$$H_0 : \bar{x}_{\text{difference}} = \mu_{\text{difference}}$$

$$H_A : \bar{x}_{\text{difference}} \neq \mu_{\text{difference}}$$

$\mu_{\text{difference}} = 0$ for most paired data.

Paired Test Statistic

The degrees of freedom for a paired t-test is $\mathrm{df} = n - 1$.

When $\mu_{\text{difference}} \neq 0$...

$$T_{\text{df}} = \frac{\text{point estimate} - \text{null value}}{\mathrm{SE}}$$

$$T_{n-1} = \frac{\bar{x}_{\text{difference}} - \mu_{\text{difference}}}{\frac{s_{\text{difference}}}{\sqrt{n}}}$$

When $\mu_{\text{difference}} = 0$...

$$T_{\text{df}} = \frac{\text{point estimate}}{\mathrm{SE}}$$

$$T_{n-1} = \frac{\bar{x}_{\text{difference}}}{\frac{s_{\text{difference}}}{\sqrt{n}}}$$

Two-Sample t-Test Hypotheses

The null distribution for a two-sample t-test is $\bar{x}_1 - \bar{x}_2 \sim N\left(\mu_1 - \mu_2, \mathrm{SE}_{\bar{x}_1 - \bar{x}_2}\right)$.

$$H_0 \colon \bar{x}_1 - \bar{x}_2 = \mu_1 - \mu_2$$

$$H_A \colon \bar{x}_1 - \bar{x}_2 \neq \mu_1 - \mu_2$$

Two-Sample t-Test Hypotheses

The null distribution for a two-sample t-test is
$\bar{x}_1 - \bar{x}_2 \sim N\left(\mu_1 - \mu_2, \mathrm{SE}_{\bar{x}_1 - \bar{x}_2}\right)$.

$$H_0 : \bar{x}_1 - \bar{x}_2 = \mu_1 - \mu_2$$

$$H_A : \bar{x}_1 - \bar{x}_2 \neq \mu_1 - \mu_2$$

$\mu_1 - \mu_2 = 0$ for many two-sample tests..

Two-Sample Test Statistic

The degrees of freedom for a two-sample t-test is $\mathrm{df} = \text{min}\left(n_1 - 1, n_2 - 1\right)$.

$$T_{\text{df}} = \frac{\text{point estimate} - \text{null value}}{SE}$$

$$T_{n-1} = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$