Inference for a Difference in Two Means

Dr. Ab Mosca (they/them)

Slides based off slides courtesy of OpenIntro and John McGreedy of Johns Hopkins University

Plan for Today

- Recap tests we've look at so far
- Introduce inference for a difference of means
- Use some statistical tools

Inference so far…

Is a sample mean ($\bar{x}$) significantly different from the population mean (μ)?

- When population variance (σ^2) is known, or sample size (N) is large (>30)
  - Z-test
- When population variance (σ^2) is unknown, or sample size (N) is small (<30)
  - T-test
- Assumptions:
  - Independence: sample observations are independent
  - Normality: sample data must be nearly normally distributed

Inference so far…

Is a sample proportion (p) significantly different from the population proportion ($\hat{p}$)?

- When population variance (σ^2) is known, or sample size (N) is large (>30)
  - Z-test
- When population variance (σ^2) is unknown, or sample size (N) is small (<30)
  - T-test
- Assumptions:
  - Independence: sample observations are independent
  - Normality: we expect at least 10 successes and 10 failures

Inference so far…

Is there a significant mean difference ($\overline{x}_{diff}$) between paired data?

- When population variance (σ^2) is known, or sample size (N) is large (>30)
  - Z-test
- When population variance (σ^2) is unknown, or sample size (N) is small (<30)
  - T-test
- Assumptions:
  - Independence: sample observations are independent
  - Normality: sample data must be nearly normally distributed

Inference so far…

Is a sample mean ($\bar{x}$) or proportion ($\hat{p}$) significantly different from the population mean (μ) or proportion (p)?

- When population variance (σ^2) is known, or sample size (N) is large (>30)
  - Z-test
- When population variance (σ^2) is unknown, or sample size (N) is small (<30)
  - T-test

Is there a significant difference between paired data?

- When population variance (σ^2) is known, or sample size (N) is large (>30)
  - Z-test
- When population variance (σ^2) is unknown, or sample size (N) is small (<30)
  - T-test

Come up with a research question and hypotheses for comparing a sample mean/proportion to the population mean/proportion, and for paired data.

Another Type of Question

What if we wanted to answer a questions like these?

- Do juniors and seniors at WSU spend different amounts of time working?
- Are pumpkins heavier than butternut squash on average?

Another Type of Question

What if we wanted to answer a questions like these?

- Do juniors and seniors at WSU spend different amounts of time working?
- Are pumpkins heavier than butternut squash on average?

Do our current tools work?

Another Type of Question

What if we wanted to answer a questions like these?

- Do juniors and seniors at WSU spend different amounts of time working?
- Are pumpkins heavier than butternut squash on average?

Do our current tools work?

No! In these cases we want to compare *unpaired* means

Another Type of Question

Is there a significant mean difference ($\overline{x}_{diff}$) between two independent groups?

T-test

- Assumptions:
  - Independence: observations are independent within and between groups
  - Normality: sample data must be nearly normally distributed within each group

Another Type of Question

Is there a significant mean difference ($\overline{x}_{diff}$) between two independent groups?

T-test

- Assumptions:
  - Independence: observations are independent within and between groups
  - Normality: sample data must be nearly normally distributed within each group

	Sample Size	Mean	Standard Deviation
Group A	n_a	$\bar{x}_a$	s_a
Group B	n_b	$\bar{x}_b$	s_b

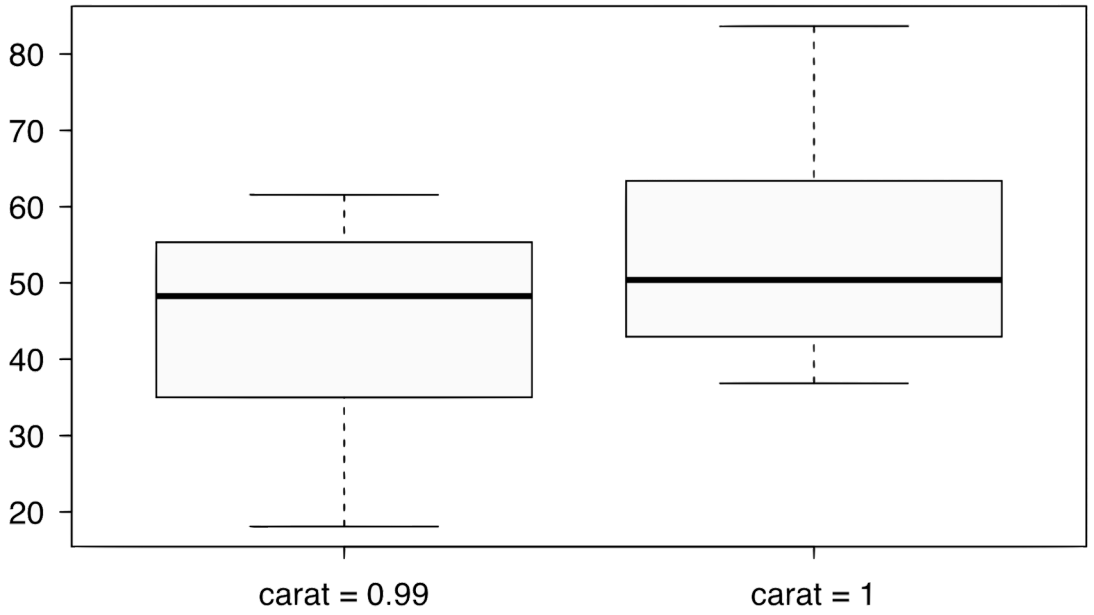
$$SE = \sqrt{\frac{s_a^2}{n_a} + \frac{s_b^2}{n_b}}$$

use the smaller of $n_a - 1$ and $n_b - 1$ for df

Ex. Diamonds

- Weights of diamonds are measured in carats
- 1 carat = 100 points, 0.99 carats = 99 points, etc.
- The difference between the size of a 0.99 carat diamond and a 1 carat diamond is undetectable to the naked human eye, but does the price of a 1 carat diamond tend to be higher than the price of a 0.99 diamond?
- We are going to test to see if there is a difference between the average prices of 0.99 and 1 carat diamonds
- In order to be able to compare equivalent units, we divide the prices of 0.99 carat diamonds by 99 and 1 carat diamonds by 100, and compare the average point prices

Data



	0.99 carat	*1 carat*
	pt99	pt100
$\bar{x}$	44.50	53.43
s	13.32	12.22
n	23	30

Note: These data are a random sample from the diamonds data set in ggplot2 R package.

Parameter and point estimate

- *Parameter of interest*: Average difference between the point prices of *all* 0.99 carat and 1 carat diamonds

$$\mu_{pt99} - \mu_{pt100}$$

Parameter and point estimate

- *Parameter of interest*: Average difference between the point prices of *all* 0.99 carat and 1 carat diamonds

$$\mu_{pt99} - \mu_{pt100}$$

- *Point estimate*: Average difference between the point prices of *sampled* 0.99 carat and 1 carat diamonds

$$\bar{x}_{pt99} - \bar{x}_{pt100}$$

Hypotheses

Which of the following is the correct set of hypotheses for testing if the average point price of 1 carat diamonds (pt100) is higher than the average point price of 0.99 carat diamonds (pt99)?

A. $H_0 : \boldsymbol{\mu}_{pt99} = \boldsymbol{\mu}_{pt100}$

$H_A : \boldsymbol{\mu}_{pt99} \neq \boldsymbol{\mu}_{pt100}$

B. $H_0 : \boldsymbol{\mu}_{pt99} = \boldsymbol{\mu}_{pt100}$

$H_A : \boldsymbol{\mu}_{pt99} > \boldsymbol{\mu}_{pt100}$

C. $H_0 : \boldsymbol{\mu}_{pt99} = \boldsymbol{\mu}_{pt100}$

$H_A : \boldsymbol{\mu}_{pt99} < \boldsymbol{\mu}_{pt100}$

D. $H_0 : \bar{x}_{pt99} = \bar{x}_{pt100}$

$H_A : \bar{x}_{pt99} < \bar{x}_{pt100}$

Hypotheses

Which of the following is the correct set of hypotheses for testing if the average point price of 1 carat diamonds (pt100) is higher than the average point price of 0.99 carat diamonds (pt99)?

A. $H_0 : \mu_{pt99} = \mu_{pt100}$

$H_A : \mu_{pt99} \neq \mu_{pt100}$

B. $H_0 : \mu_{pt99} = \mu_{pt100}$

$H_A : \mu_{pt99} > \mu_{pt100}$

C. $H_0 : \mu_{pt99} = \mu_{pt100}$

$H_A : \mu_{pt99} < \mu_{pt100}$

D. $H_0 : \bar{x}_{pt99} = \bar{x}_{pt100}$

$H_A : \bar{x}_{pt99} < \bar{x}_{pt100}$

Conditions

Which of the following does not need to be satisfied in order to conduct this hypothesis test using theoretical methods?

A. Point price of one 0.99 carat diamond in the sample should be independent of another, and the point price of one 1 carat diamond should independent of another as well
B. Point prices of 0.99 carat and 1 carat diamonds in the sample should be independent.
C. Distributions of point prices of 0.99 and 1 carat diamonds should not be extremely skewed
D. Both sample sizes should be at least 30

Conditions

Which of the following does not need to be satisfied in order to conduct this hypothesis test using theoretical methods?

A. Point price of one 0.99 carat diamond in the sample should be independent of another, and the point price of one 1 carat diamond should independent of another as well
B. Point prices of 0.99 carat and 1 carat diamonds in the sample should be independent.
C. Distributions of point prices of 0.99 and 1 carat diamonds should not be extremely skewed
D. Both sample sizes should be at least 30

Test statistics

Test statistic for inference on the difference of two small sample means

The test statistic for inference on the difference of two means where σ_1 and σ_2 are unknown is the T statistic.

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

where

$$SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \qquad \text{and} \qquad df = min(n_1 - 1, n_2 - 1)$$

Note: The calculation of the *df* is actually much more complicated. For simplicity we'll use the above formula to estimate the true df when conducting the analysis by hand

Test statistics (cont.)

	0.99 carat pt99	*1 carat* pt100
$\bar{x}$	44.50	53.43
s	13.32	12.22
n	23	30

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

$$SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \quad \text{and} \quad df = min(n_1 - 1, n_2 - 1)$$

Find T

Test statistics (cont.)

	0.99 carat pt99	*1 carat* pt100
$\bar{x}$	44.50	53.43
s	13.32	12.22
n	23	30

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

$$SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \quad \text{and} \quad df = min(n_1 - 1, n_2 - 1)$$

Find T

$$\begin{aligned} T &= \frac{(44.50 - 53.43) - 0}{\sqrt{\frac{13.32^2}{23} + \frac{12.22^2}{30}}} \\ &= \frac{-8.93}{3.56} \\ &= -2.508 \end{aligned}$$

Test statistics (cont.)

Which of the following is the correct *df* for this hypothesis test?

$$df = min(n_1 - 1, n_2 - 1)$$

A. 22
B. 23
C. 30
D. 29
E. 52

Test statistics (cont.)

Which of the following is the correct *df* for this hypothesis test?

$$df = min(n_1 - 1, n_2 - 1)$$

$$df = \min(n_{pt99} - 1, n_{pt100} - 1)$$

$$= \min(23 - 1, 30 - 1)$$

$$= \min(22, 29)$$

A. 22
B. 23
C. 30
D. 29
E. 52

p-value

Which of the following is the correct p-value for this hypothesis test?

$$T = -2.508 \qquad df = 22$$

A. between 0.005 and 0.01
B. between 0.01 and 0.025
C. between 0.02 and 0.05
D. between 0.01 and 0.02

one tail	0.100	0.050	0.025	0.010
two tails	0.200	0.100	0.050	0.020
df 21	1.32	1.72	2.08	2.52
22	1.32	1.72	2.07	2.51
23	1.32	1.71	2.07	2.50
24	1.32	1.71	2.06	2.49
25	1.32	1.71	2.06	2.49

p-value

Which of the following is the correct p-value for this hypothesis test?

$$T = -2.508 \qquad df = 22$$

A. between 0.005 and 0.01
B. between 0.01 and 0.025
C. between 0.02 and 0.05
D. between 0.01 and 0.02

one tail	0.100	0.050	*0.025*	*0.010*
two tails	0.200	0.100	0.050	0.020
df 21	1.32	1.72	2.08	2.52
22	1.32	1.72	*2.07*	*2.51*
23	1.32	1.71	2.07	2.50
24	1.32	1.71	2.06	2.49
25	1.32	1.71	2.06	2.49

Synthesis

What is the conclusion of the hypothesis test? How (if at all) would this conclusion change your behavior if you went diamond shopping?

Synthesis

What is the conclusion of the hypothesis test? How (if at all) would this conclusion change your behavior if you went diamond shopping?

- p-value is smaller than 0.05 so we reject H_0. The data provide convincing evidence to suggest that the point price of 0.99 carat diamonds is lower than the point price of 1 carat diamonds
- Maybe buy a 0.99 carat diamond? It looks like a 1 carat, but is significantly cheaper

Equivalent confidence level

If our confidence level for the one-sided test was $\alpha = 0.05$, what is the equivalent confidence level for a two-sided hypothesis test?

A. 90%
B. 92.5%
C. 95%
D. 97.5%

Equivalent confidence level

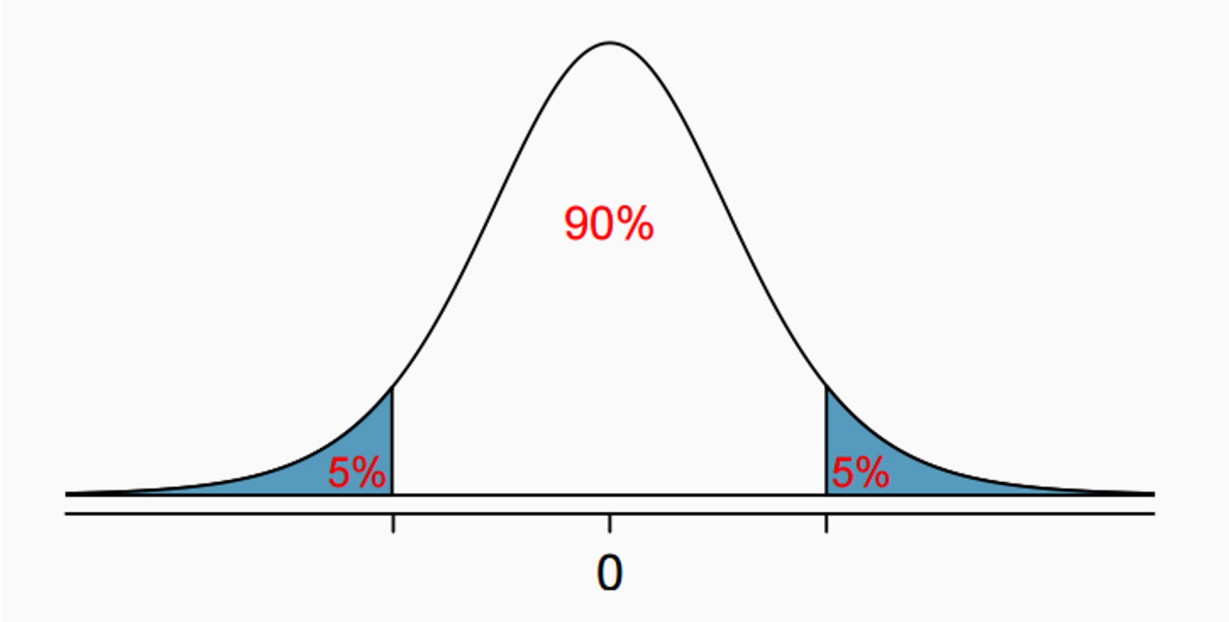
If our confidence level for the one-sided test was α = 0.05, what is the equivalent confidence level for a two-sided hypothesis test?

A. 90%

B. 92.5%

C. 95%

D. 97.5%



Critical value

What is the appropriate t^* for a confidence interval for the average difference between the point prices of 0.99 and 1 carat diamonds?

A. 1.32
B. 1.72
C. 2.07
D. 2.82

one tail		0.100	0.050	0.025	0.010	0.005
two tails		0.200	0.100	0.050	0.020	0.010
df	21	1.32	1.72	2.08	2.52	2.83
	22	1.32	1.72	2.07	2.51	2.82
	23	1.32	1.71	2.07	2.50	2.81
	24	1.32	1.71	2.06	2.49	2.80
	25	1.32	1.71	2.06	2.49	2.79

Critical value

What is the appropriate t^* for a confidence interval for the average difference between the point prices of 0.99 and 1 carat diamonds?

A. 1.32
B. 1.72
C. 2.07
D. 2.82

one tail	0.100	0.050	0.025	0.010	0.005
two tails	0.200	*0.100*	0.050	0.020	0.010
df 21	1.32	1.72	2.08	2.52	2.83
22	1.32	*1.72*	2.07	2.51	2.82
23	1.32	1.71	2.07	2.50	2.81
24	1.32	1.71	2.06	2.49	2.80
25	1.32	1.71	2.06	2.49	2.79

Confidence interval

Calculate the confidence interval for the difference in diamond prices.

$$CI = point\ estimate\ \pm\ t^*_{df}\ *SE$$

$$SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \qquad \text{and} \qquad df = min(n_1 - 1, n_2 - 1)$$

Confidence interval

Calculate the confidence interval for the difference in diamond prices.

$$CI = point\ estimate\ \pm\ t^*_{df}\ * SE$$

$$SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \qquad \text{and} \qquad df = min(n_1 - 1, n_2 - 1)$$

$$\left(\bar{x}_{pt99} - \bar{x}_{pt1}\right) \pm t^*_{df} \times SE = (44.50\ - 53.43) \pm 1.72 \times 3.56$$

$$= -8.93 \pm 6.12$$

$$= (-15.05, -2.81)$$

Confidence interval

What does this 90% CI mean?

$(-15.05, -2.81)$

Confidence interval

What does this 90% CI mean?

$$(-15.05, -2.81)$$

We are 90% confident that the average point price of a 0.99 carat diamond is $15.05 to $2.81 lower than the average point price of a 1 carat diamond

Recap: Inference using difference of two small sample means

- If σ_1 or σ_2 is unknown, difference between the sample means follow a ***t***-distribution with $SE = \sqrt{\frac{s_a^2}{n_a} + \frac{s_b^2}{n_b}}$

Recap: Inference using difference of two small sample means

- If σ_1 or σ_2 is unknown, difference between the sample means follow a ***t***-distribution with $SE = \sqrt{\frac{s_a^2}{n_a} + \frac{s_b^2}{n_b}}$
- Conditions:
  - Independence
  - Normality

Recap: Inference using difference of two small sample means

- If σ_1 or σ_2 is unknown, difference between the sample means follow a ***t***-distribution with $SE = \sqrt{\frac{s_a^2}{n_a} + \frac{s_b^2}{n_b}}$
- Conditions:
  - Independence
  - Normality
- Hypothesis testing:

$$T_{df} = \frac{point\ estimate\ - null\ value}{SE}, where\ df = \min(n_1 - 1, n_2 - 1)$$

Recap: Inference using difference of two small sample means

- If σ_1 or σ_2 is unknown, difference between the sample means follow a ***t***-distribution with $SE = \sqrt{\frac{s_a^2}{n_a} + \frac{s_b^2}{n_b}}$
- Conditions:
  - Independence
  - Normality
- Hypothesis testing:

$$T_{df} = \frac{point\ estimate\ - null\ value}{SE}, where\ df = \min(n_1 - 1, n_2 - 1)$$

- Confidence interval:

$$point\ estimate \pm t^*_{df} \times SE$$

Statistical Tools

- How do we analyze data when there is too much to calculate by hand?

Statistical Tools

- How do we analyze data when there is too much to calculate by hand?
  - Option 1: Use excel or some other coding-free software
  - Option 2: Learn to use a statistical software (like R, SAS, SPSS)

Statistical Tools

- How do we analyze data when there is too much to calculate by hand?
  - Option 1: Use excel or some other coding-free software
  - Option 2: Learn to use a statistical software (like R, SAS, SPSS)
- Ex. With R https://www.kaggle.com/datasets/neuromusic/avocado-prices