Statistics

Marco Caserta
marco.caserta@ie.edu

IE University

Hypothesis Testing: A Summary

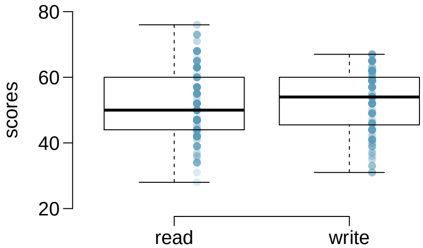
T Tests for Population Means

- Is the average salary of IEU graduates above 50k?
- Is salary of MBA graduates from school A higher than that of graduates from school B?
- Is the salary after getting an MBA higher than the salary before getting an MBA?
- Is there a reduction in stress level after taking a certain medicament?
- Is the Safe Driving Program reducing the number of accidents?
- Are males more optimistic than females?

Hypothesis	**Statistical Procedure**
A sample coming from a population with a particular mean	One-sample t test
Two related population means are equal	Paired-sample t test
Two independent population means are equal	Two-independent-sample t test

Hypothesis	**Statistical Procedure**
Two or more independent population means are equal	One-way ANOVA
Two categorical variables are independent	Chi-square test
Two or more proportions are equal	Chi-square test

200 observations were randomly sampled from the High School and Beyond survey. The same students took a reading and writing test and their scores are shown below. At a first glance, does there appear to be a difference between the average reading and writing test score?



The same students took a reading and writing test and their scores are shown below. Are the reading and writing scores of each student independent of each other?

	id	read	write
1	70	57	52
2	86	44	33
3	141	63	44
4	172	47	52
⋮	⋮	⋮	⋮
200	137	63	65

(A) Yes

(B) No

The same students took a reading and writing test and their scores are shown below. Are the reading and writing scores of each student independent of each other?

	id	read	write
1	70	57	52
2	86	44	33
3	141	63	44
4	172	47	52
⋮	⋮	⋮	⋮
200	137	63	65

(A) Yes

(B) *No*

Analyzing paired data

- When two sets of observations have this special correspondence (not independent), they are said to be paired.

Analyzing paired data

- When two sets of observations have this special correspondence (NOT INDEPENDENT), they are said to be PAIRED.
- To analyze paired data, it is often useful to look at the difference in outcomes of each pair of observations.

diff = read – write

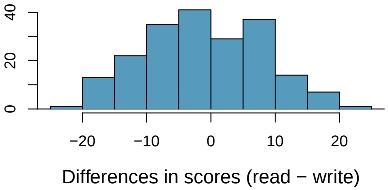
Analyzing paired data

- When two sets of observations have this special correspondence (NOT INDEPENDENT), they are said to be PAIRED.
- To analyze paired data, it is often useful to look at the difference in outcomes of each pair of observations.

$$\text{diff} = \text{read} - \text{write}$$

- It is important that we always subtract using a consistent order.

	id	read	write	diff
1	70	57	52	5
2	86	44	33	11
3	141	63	44	19
4	172	47	52	-5
⋮	⋮	⋮	⋮	⋮
200	137	63	65	-2



Parameter and point estimate

- Parameter of interest: Average difference between the reading and writing scores of all high school students.

$$\mu_{diff}$$

Parameter and point estimate

- Parameter of interest: Average difference between the reading and writing scores of all high school students.

$$\mu_{diff}$$

- Point estimate: Average difference between the reading and writing scores of sampled high school students.

$$\bar{x}_{diff}$$

Setting the hypotheses

If in fact there was no difference between the scores on the reading and writing exams, what would you expect the average difference to be?

Setting the hypotheses

If in fact there was no difference between the scores on the reading and writing exams, what would you expect the average difference to be?

0

Setting the hypotheses

If in fact there was no difference between the scores on the reading and writing exams, what would you expect the average difference to be?

0

What are the hypotheses for testing if there is a difference between the average reading and writing scores?

Setting the hypotheses

If in fact there was no difference between the scores on the reading and writing exams, what would you expect the average difference to be?

0

What are the hypotheses for testing if there is a difference between the average reading and writing scores?

H_0: There is no difference between the average reading and writing score.

$$\mu_{diff} = 0$$

H_A: There is a difference between the average reading and writing score.

$$\mu_{diff} \neq 0$$

Nothing new here

- The analysis is no different than what we have done before.
- We have data from one sample: differences.
- We are testing to see if the average difference is different than 0.

Checking assumptions & conditions

Which of the following is true?

(A) Since students are sampled randomly and are less than 10% of all high school students, we can assume that the difference between the reading and writing scores of one student in the sample is independent of another.

(B) The distribution of differences is bimodal, therefore we cannot continue with the hypothesis test.

(C) In order for differences to be random we should have sampled with replacement.

(D) Since students are sampled randomly and are less than 10% all students, we can assume that the sampling distribution of the average difference will be nearly normal.

Checking assumptions & conditions

Which of the following is true?

(a) *Since students are sampled randomly and are less than 10% of all high school students, we can assume that the difference between the reading and writing scores of one student in the sample is independent of another.*

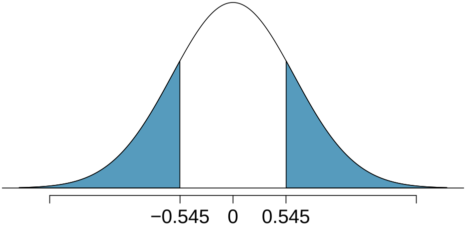
(b) The distribution of differences is bimodal, therefore we cannot continue with the hypothesis test.

(c) In order for differences to be random we should have sampled with replacement.

(d) Since students are sampled randomly and are less than 10% all students, we can assume that the sampling distribution of the average difference will be nearly normal.

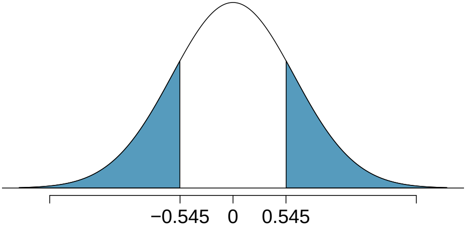
Calculating the test-statistic and the p-value

The observed average difference between the two scores is -0.545 points and the standard deviation of the difference is 8.887 points. Do these data provide convincing evidence of a difference between the average scores on the two exams? Use $\alpha = 0.05$.



Calculating the test-statistic and the p-value

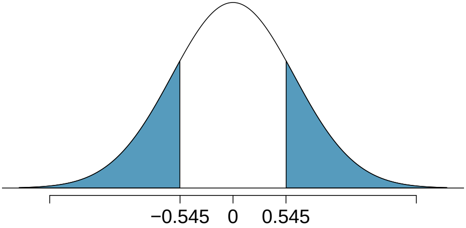
The observed average difference between the two scores is -0.545 points and the standard deviation of the difference is 8.887 points. Do these data provide convincing evidence of a difference between the average scores on the two exams? Use $\alpha = 0.05$.



$$Z = \frac{-0.545 - 0}{\frac{8.887}{\sqrt{200}}} = \frac{-0.545}{0.628} = -0.87$$

Calculating the test-statistic and the p-value

The observed average difference between the two scores is -0.545 points and the standard deviation of the difference is 8.887 points. Do these data provide convincing evidence of a difference between the average scores on the two exams? Use $\alpha = 0.05$.

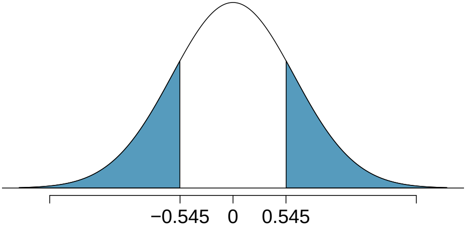


$$Z = \frac{-0.545 - 0}{\frac{8.887}{\sqrt{200}}} = \frac{-0.545}{0.628} = -0.87$$

$$p - value = 0.1949 \times 2 = 0.3898$$

Calculating the test-statistic and the p-value

The observed average difference between the two scores is -0.545 points and the standard deviation of the difference is 8.887 points. Do these data provide convincing evidence of a difference between the average scores on the two exams? Use $\alpha = 0.05$.



$$
\begin{aligned}
Z &= \frac{-0.545 - 0}{\frac{8.887}{\sqrt{200}}} \\
&= \frac{-0.545}{0.628} = -0.87 \\
p - value &= 0.1949 \times 2 = 0.3898
\end{aligned}
$$

Since p-value $>$ 0.05, fail to reject, the data do <u>not</u> provide convincing evidence of a difference between the average reading and writing scores.

Interpretation of p-value

Which of the following is the correct interpretation of the p-value?

(A) Probability that the average scores on the reading and writing exams are equal.

(B) Probability that the average scores on the reading and writing exams are different.

(C) Probability of obtaining a random sample of 200 students where the average difference between the reading and writing scores is at least 0.545 (in either direction), if in fact the true average difference between the scores is 0.

(D) Probability of incorrectly rejecting the null hypothesis if in fact the null hypothesis is true.

Interpretation of p-value

Which of the following is the correct interpretation of the p-value?

(A) Probability that the average scores on the reading and writing exams are equal.

(B) Probability that the average scores on the reading and writing exams are different.

(C) *Probability of obtaining a random sample of 200 students where the average difference between the reading and writing scores is at least 0.545 (in either direction), if in fact the true average difference between the scores is 0.*

(D) Probability of incorrectly rejecting the null hypothesis if in fact the null hypothesis is true.

HT $\leftrightarrow$ CI

Suppose we were to construct a 95% confidence interval for the average difference between the reading and writing scores. Would you expect this interval to include 0?

(A) yes

(B) no

(C) cannot tell from the information given

HT $\leftrightarrow$ CI

Suppose we were to construct a 95% confidence interval for the average difference between the reading and writing scores. Would you expect this interval to include 0?

(A) *yes*

(B) no

(C) cannot tell from the information given

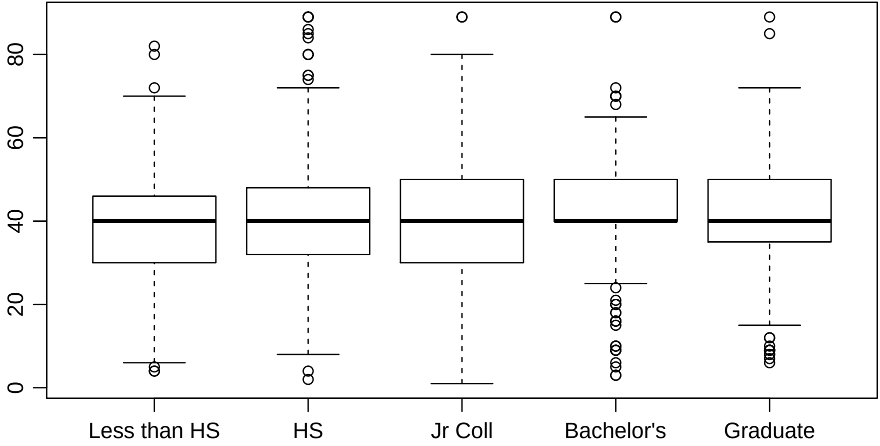
$$
\begin{aligned}
-0.545 \pm 1.96 \frac{8.887}{\sqrt{200}} &= -0.545 \pm 1.96 \times 0.628 \\
&= -0.545 \pm 1.23 \\
&= (-1.775, 0.685)
\end{aligned}
$$

The General Social Survey (GSS) conducted by the Census Bureau contains a standard ‘core’ of demographic, behavioral, and attitudinal questions, plus topics of special interest. Many of the core questions have remained unchanged since 1972 to facilitate time-trend studies as well as replication of earlier findings. Below is an excerpt from the 2010 data set. The variables are number of hours worked per week and highest educational attainment.

	degree	hrs1
1	BACHELOR	55
2	BACHELOR	45
3	JUNIOR COLLEGE	45
⋮		
1172	HIGH SCHOOL	40

Exploratory analysis

What can you say about the relationship between educational attainment and hours worked per week?



Collapsing levels into two

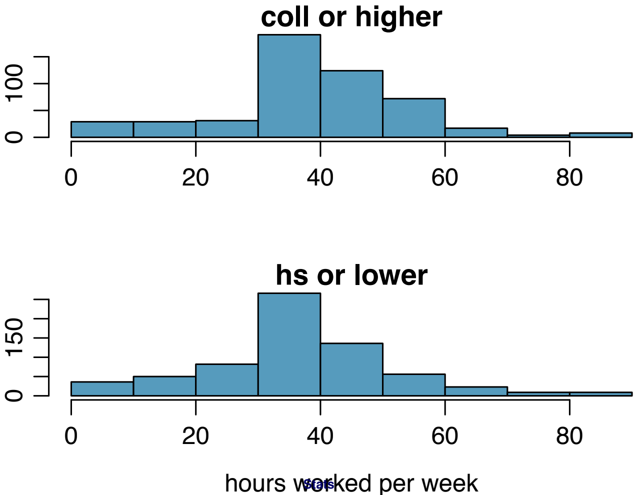
- Say we are only interested the difference between the number of hours worked per week by college and non-college graduates.

Collapsing levels into two

- Say we are only interested the difference between the number of hours worked per week by college and non-college graduates.
- Then we combine the levels of education into two:
  - `hs or lower` ← less than high school or high school
  - `coll or higher` ← junior college, bachelor's, and graduate

Exploratory analysis - another look

	$\bar{x}$	s	n
coll or higher	41.8	15.14	505
hs or lower	39.4	15.12	667



Parameter and point estimate

We want to construct a 95% confidence interval for the average difference between the number of hours worked per week by Americans with a college degree and those with a high school degree or lower. What are the parameter of interest and the point estimate?

Parameter and point estimate

We want to construct a 95% confidence interval for the average difference between the number of hours worked per week by Americans with a college degree and those with a high school degree or lower. What are the parameter of interest and the point estimate?

- Parameter of interest: Average difference between the number of hours worked per week by all Americans with a college degree and those with a high school degree or lower.

$$\mu_{coll} - \mu_{hs}$$

Parameter and point estimate

We want to construct a 95% confidence interval for the average difference between the number of hours worked per week by Americans with a college degree and those with a high school degree or lower. What are the parameter of interest and the point estimate?

- Parameter of interest: Average difference between the number of hours worked per week by all Americans with a college degree and those with a high school degree or lower.

$$\mu_{coll} - \mu_{hs}$$

- Point estimate: Average difference between the number of hours worked per week by sampled Americans with a college degree and those with a high school degree or lower.

$$\bar{x}_{coll} - \bar{x}_{hs}$$

Checking assumptions & conditions

1. **Independence within groups:**
   - Both the college graduates and those with HS degree or lower are sampled randomly.

Checking assumptions & conditions

1. Independence within groups:
    - Both the college graduates and those with HS degree or lower are sampled randomly.
    - 505 $<$ 10% of all college graduates and 667 $<$ 10% of all students with a high school degree or lower.

Checking assumptions & conditions

1. Independence within groups:
   - Both the college graduates and those with HS degree or lower are sampled randomly.
   - 505 $<$ 10% of all college graduates and 667 $<$ 10% of all students with a high school degree or lower.

We can assume that the number of hours worked per week by one college graduate in the sample is independent of another, and the number of hours worked per week by someone with a HS degree or lower in the sample is independent of another as well.

Checking assumptions & conditions

1. Independence within groups:
    - Both the college graduates and those with HS degree or lower are sampled randomly.
    - 505 $<$ 10% of all college graduates and 667 $<$ 10% of all students with a high school degree or lower.

   We can assume that the number of hours worked per week by one college graduate in the sample is independent of another, and the number of hours worked per week by someone with a HS degree or lower in the sample is independent of another as well.
2. Independence between groups: $\leftarrow$ new!

   Since the sample is random, the college graduates in the sample are independent of those with a HS degree or lower. In other words, we are dealing with two *independent* samples.

Checking assumptions & conditions

1. Independence within groups:
   - Both the college graduates and those with HS degree or lower are sampled randomly.
   - 505 $<$ 10% of all college graduates and 667 $<$ 10% of all students with a high school degree or lower.

   We can assume that the number of hours worked per week by one college graduate in the sample is independent of another, and the number of hours worked per week by someone with a HS degree or lower in the sample is independent of another as well.
2. Independence between groups: $\leftarrow$ new!

   Since the sample is random, the college graduates in the sample are independent of those with a HS degree or lower. In other words, we are dealing with two *independent* samples.
3. Sample size / skew:

   Both distributions look reasonably symmetric, and the sample sizes are at least 30, therefore we can assume that the sampling distribution of number of hours worked per week by college graduates and those with HS degree or lower are nearly normal. Hence the sampling distribution of the average difference will be nearly normal as well.

Confidence interval for difference between two means

- All confidence intervals have the same form:

$$point\ estimate \pm ME$$

Confidence interval for difference between two means

- All confidence intervals have the same form:

$$point\ estimate \pm ME$$

- And all $ME = critical\ value \times SE\ of\ point\ estimate$

Confidence interval for difference between two means

- All confidence intervals have the same form:

$$point\ estimate \pm ME$$

- And all $ME = critical\ value \times SE\ of\ point\ estimate$
- In this case the point estimate is $\bar{x}_1 - \bar{x}_2$

Confidence interval for difference between two means

- All confidence intervals have the same form:

$$point\ estimate \pm ME$$

- And all $ME = critical\ value \times SE\ of\ point\ estimate$
- In this case the point estimate is $\bar{x}_1 - \bar{x}_2$
- Since the sample sizes are large enough, the critical value is $z^\star$

Confidence interval for difference between two means

- All confidence intervals have the same form:

$$point\ estimate \pm ME$$

- And all $ME = critical\ value \times SE\ of\ point\ estimate$
- In this case the point estimate is $\bar{x}_1 - \bar{x}_2$
- Since the sample sizes are large enough, the critical value is $z^\star$
- So the only new concept is the standard error of the difference between two means...

Confidence interval for difference between two means

- All confidence intervals have the same form:

$$point\ estimate \pm ME$$

- And all $ME = critical\ value \times SE\ of\ point\ estimate$
- In this case the point estimate is $\bar{x}_1 - \bar{x}_2$
- Since the sample sizes are large enough, the critical value is $z^\star$
- So the only new concept is the standard error of the difference between two means...

Standard error of the difference between two sample means

$$SE_{(\bar{x}_1 - \bar{x}_2)} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Let's put things in context

Calculate the standard error of the average difference between the number of hours worked per week by college graduates and those with a HS degree or lower.

	$\bar{x}$	s	n
coll or higher	41.8	15.14	505
hs or lower	39.4	15.12	667

Let's put things in context

Calculate the standard error of the average difference between the number of hours worked per week by college graduates and those with a HS degree or lower.

	$\bar{x}$	s	n
coll or higher	41.8	15.14	505
hs or lower	39.4	15.12	667

$$SE_{(\bar{x}_{coll}-\bar{x}_{hs})} = \sqrt{\frac{s_{coll}^2}{n_{coll}} + \frac{s_{hs}^2}{n_{hs}}}$$

Let's put things in context

Calculate the standard error of the average difference between the number of hours worked per week by college graduates and those with a HS degree or lower.

	$\bar{x}$	s	n
coll or higher	41.8	15.14	505
hs or lower	39.4	15.12	667

$$SE_{(\bar{x}_{coll}-\bar{x}_{hs})} = \sqrt{\frac{s_{coll}^2}{n_{coll}} + \frac{s_{hs}^2}{n_{hs}}} = \sqrt{\frac{15.14^2}{505} + \frac{15.12^2}{667}}$$

Let's put things in context

Calculate the standard error of the average difference between the number of hours worked per week by college graduates and those with a HS degree or lower.

	$\bar{x}$	s	n
coll or higher	41.8	15.14	505
hs or lower	39.4	15.12	667

$$
\begin{aligned}
SE_{(\bar{x}_{coll} - \bar{x}_{hs})} &= \sqrt{\frac{s_{coll}^2}{n_{coll}} + \frac{s_{hs}^2}{n_{hs}}} \\
&= \sqrt{\frac{15.14^2}{505} + \frac{15.12^2}{667}} \\
&= 0.89
\end{aligned}
$$

Confidence interval for the difference (cont.)

Estimate (using a 95% confidence interval) the average difference between the number of hours worked per week by Americans with a college degree and those with a high school degree or lower.

$$\bar{x}_{coll} = 41.8 \qquad \bar{x}_{hs} = 39.4 \qquad SE_{(\bar{x}_{coll} - \bar{x}_{hs})} = 0.89$$

Confidence interval for the difference (cont.)

Estimate (using a 95% confidence interval) the average difference between the number of hours worked per week by Americans with a college degree and those with a high school degree or lower.

$$\bar{x}_{coll} = 41.8 \qquad \bar{x}_{hs} = 39.4 \qquad SE_{(\bar{x}_{coll} - \bar{x}_{hs})} = 0.89$$

$$(\bar{x}_{coll} - \bar{x}_{hs}) \pm z^{\star} \times SE_{(\bar{x}_{coll} - \bar{x}_{hs})} \quad = \quad (41.8 - 39.4) \pm 1.96 \times 0.89$$

Confidence interval for the difference (cont.)

Estimate (using a 95% confidence interval) the average difference between the number of hours worked per week by Americans with a college degree and those with a high school degree or lower.

$$\bar{x}_{coll} = 41.8 \qquad \bar{x}_{hs} = 39.4 \qquad SE_{(\bar{x}_{coll} - \bar{x}_{hs})} = 0.89$$

$$\begin{aligned} (\bar{x}_{coll} - \bar{x}_{hs}) \pm z^{\star} \times SE_{(\bar{x}_{coll} - \bar{x}_{hs})} &= (41.8 - 39.4) \pm 1.96 \times 0.89 \\ &= 2.4 \pm 1.74 \end{aligned}$$

Confidence interval for the difference (cont.)

Estimate (using a 95% confidence interval) the average difference between the number of hours worked per week by Americans with a college degree and those with a high school degree or lower.

$$\bar{x}_{coll} = 41.8 \qquad \bar{x}_{hs} = 39.4 \qquad SE_{(\bar{x}_{coll} - \bar{x}_{hs})} = 0.89$$

$$\begin{aligned}(\bar{x}_{coll} - \bar{x}_{hs}) \pm z^{\star} \times SE_{(\bar{x}_{coll} - \bar{x}_{hs})} &= (41.8 - 39.4) \pm 1.96 \times 0.89 \\ &= 2.4 \pm 1.74 \\ &= (0.66, 4.14)\end{aligned}$$

Interpretation of a confidence interval for the difference

Which of the following is the best interpretation of the confidence interval we just calculated?

(A) The difference between the average number of hours worked per week by college grads and those with a HS degree or lower is between 0.66 and 4.14 hours.

(B) College grads work on average of 0.66 to 4.14 hours more per week than those with a HS degree or lower.

(C) College grads work on average 0.66 hours less to 4.14 hours more per week than those with a HS degree or lower.

(D) College grads work on average 0.66 to 4.14 hours less per week than those with a HS degree or lower.

Interpretation of a confidence interval for the difference

Which of the following is the best interpretation of the confidence interval we just calculated?

(A) The difference between the average number of hours worked per week by college grads and those with a HS degree or lower is between 0.66 and 4.14 hours.

(B) *College grads work on average of 0.66 to 4.14 hours more per week than those with a HS degree or lower.*

(C) College grads work on average 0.66 hours less to 4.14 hours more per week than those with a HS degree or lower.

(D) College grads work on average 0.66 to 4.14 hours less per week than those with a HS degree or lower.

Reality check

Do these results sound reasonable? Why or why not?

Setting the hypotheses

What are the hypotheses for testing if there is a difference between the average number of hours worked per week by college graduates and those with a HS degree or lower?

Setting the hypotheses

What are the hypotheses for testing if there is a difference between the average number of hours worked per week by college graduates and those with a HS degree or lower?

H_0: $\mu_{coll} = \mu_{hs}$

There is no difference in the average number of hours worked per week by college graduates and those with a HS degree or lower. Any observed difference between the sample means is due to natural sampling variation (chance).

Setting the hypotheses

What are the hypotheses for testing if there is a difference between the average number of hours worked per week by college graduates and those with a HS degree or lower?

H_0: $\mu_{coll} = \mu_{hs}$
There is no difference in the average number of hours worked per week by college graduates and those with a HS degree or lower. Any observed difference between the sample means is due to natural sampling variation (chance).

H_A: $\mu_{coll} \neq \mu_{hs}$
There is a difference in the average number of hours worked per week by college graduates and those with a HS degree or lower.

Calculating the test-statistic and the p-value

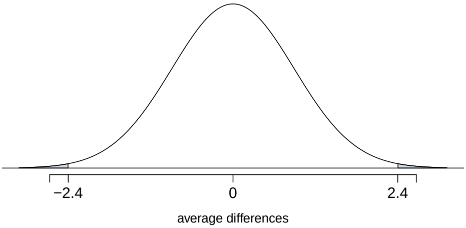
H_0: $\mu_{coll} = \mu_{hs} \rightarrow \mu_{coll} - \mu_{hs} = 0$
H_A: $\mu_{coll} \neq \mu_{hs} \rightarrow \mu_{coll} - \mu_{hs} \neq 0$

$\bar{x}_{coll} - \bar{x}_{hs} = 2.4$, $SE_{(}\bar{x}_{coll} - \bar{x}_{hs}) = 0.89$

Calculating the test-statistic and the p-value

H_0: $\mu_{coll} = \mu_{hs} \rightarrow \mu_{coll} - \mu_{hs} = 0$
H_A: $\mu_{coll} \neq \mu_{hs} \rightarrow \mu_{coll} - \mu_{hs} \neq 0$

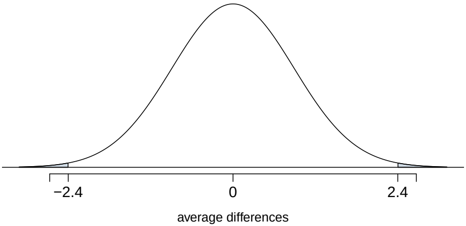
$\bar{x}_{coll} - \bar{x}_{hs} = 2.4$, $SE_{(}\bar{x}_{coll} - \bar{x}_{hs}) = 0.89$



Calculating the test-statistic and the p-value

H_0: $\mu_{coll} = \mu_{hs} \rightarrow \mu_{coll} - \mu_{hs} = 0$
H_A: $\mu_{coll} \neq \mu_{hs} \rightarrow \mu_{coll} - \mu_{hs} \neq 0$

$\bar{x}_{coll} - \bar{x}_{hs} = 2.4$, $SE_{(}\bar{x}_{coll} - \bar{x}_{hs}) = 0.89$

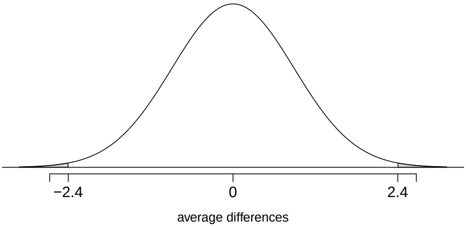


$$Z = \frac{(\bar{x}_{coll} - \bar{x}_{hs}) - 0}{SE_{(\bar{x}_{coll} - \bar{x}_{hs})}}$$

Calculating the test-statistic and the p-value

H_0: $\mu_{coll} = \mu_{hs} \rightarrow \mu_{coll} - \mu_{hs} = 0$
H_A: $\mu_{coll} \neq \mu_{hs} \rightarrow \mu_{coll} - \mu_{hs} \neq 0$

$\bar{x}_{coll} - \bar{x}_{hs} = 2.4$, $SE_{(\bar{x}_{coll} - \bar{x}_{hs})} = 0.89$

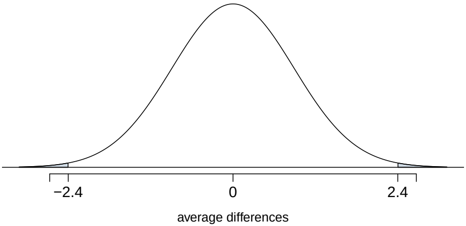


$$Z = \frac{(\bar{x}_{coll} - \bar{x}_{hs}) - 0}{SE_{(\bar{x}_{coll} - \bar{x}_{hs})}} = \frac{2.4}{0.89} = 2.70$$

Calculating the test-statistic and the p-value

H_0: $\mu_{coll} = \mu_{hs} \rightarrow \mu_{coll} - \mu_{hs} = 0$
H_A: $\mu_{coll} \neq \mu_{hs} \rightarrow \mu_{coll} - \mu_{hs} \neq 0$

$\bar{x}_{coll} - \bar{x}_{hs} = 2.4$, $SE_{(}\bar{x}_{coll} - \bar{x}_{hs}) = 0.89$

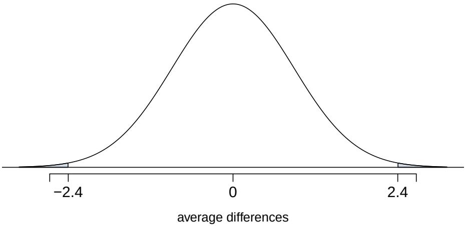


$$
\begin{aligned}
Z &= \frac{(\bar{x}_{coll} - \bar{x}_{hs}) - 0}{SE_{(\bar{x}_{coll} - \bar{x}_{hs})}} \\
&= \frac{2.4}{0.89} = 2.70 \\
upper\ tail &= 1 - 0.9965 = 0.0035
\end{aligned}
$$

Calculating the test-statistic and the p-value

H_0: $\mu_{coll} = \mu_{hs} \rightarrow \mu_{coll} - \mu_{hs} = 0$
H_A: $\mu_{coll} \neq \mu_{hs} \rightarrow \mu_{coll} - \mu_{hs} \neq 0$

$\bar{x}_{coll} - \bar{x}_{hs} = 2.4$, $SE_{(\bar{x}_{coll} - \bar{x}_{hs})} = 0.89$



$$
\begin{aligned}
Z &= \frac{(\bar{x}_{coll} - \bar{x}_{hs}) - 0}{SE_{(\bar{x}_{coll} - \bar{x}_{hs})}} \\
&= \frac{2.4}{0.89} = 2.70 \\
upper\ tail &= 1 - 0.9965 = 0.0035 \\
p - value &= 2 \times 0.0035 = 0.007
\end{aligned}
$$

Conclusion of the test

Which of the following is correct based on the results of the hypothesis test we just conducted?

(A) There is a 0.7% chance that there is no difference between the average number of hours worked per week by college graduates and those with a HS degree or lower.

(B) Since the p-value is low, we reject H_0. The data provide convincing evidence of a difference between the average number of hours worked per week by college graduates and those with a HS degree or lower.

(C) Since we rejected H_0, we may have made a Type 2 error.

(D) Since the p-value is low, we fail to reject H_0. The data do not provide convincing evidence of a difference between the average number of hours worked per week by college graduates and those with a HS degree or lower.

Conclusion of the test

Which of the following is correct based on the results of the hypothesis test we just conducted?

(A) There is a 0.7% chance that there is no difference between the average number of hours worked per week by college graduates and those with a HS degree or lower.

(B) *Since the p-value is low, we reject H_0. The data provide convincing evidence of a difference between the average number of hours worked per week by college graduates and those with a HS degree or lower.*

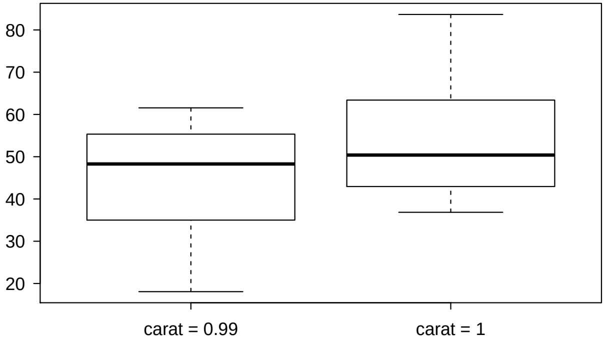
(C) Since we rejected H_0, we may have made a Type 2 error.

(D) Since the p-value is low, we fail to reject H_0. The data do not provide convincing evidence of a difference between the average number of hours worked per week by college graduates and those with a HS degree or lower.

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

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)

$$\begin{cases} H_0: & \mu_{pt99} = \mu_{pt100} \\ H_A: & \mu_{pt99} \neq \mu_{pt100} \end{cases}$$

(B)

$$\begin{cases} H_0: & \mu_{pt99} = \mu_{pt100} \\ H_A: & \mu_{pt99} > \mu_{pt100} \end{cases}$$

(C)

$$\begin{cases} H_0: & \mu_{pt99} = \mu_{pt100} \\ H_A: & \mu_{pt99} < \mu_{pt100} \end{cases}$$

(D)

$$\begin{cases} H_0: & \bar{x}_{pt99} = \bar{x}_{pt100} \\ H_A: & \bar{x}_{pt99} < \bar{x}_{pt100} \end{cases}$$

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 statistic

Test statistic for inference on the difference of two small sample means
The test statistic for inference on the difference of two small sample means ($n_1 < 30$ and/or $n_2 < 30$) mean 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 statistic (cont.)

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

in context...

Test statistic (cont.)

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

IN CONTEXT...

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

Test statistic (cont.)

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

in context...

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

$$= \frac{(44.50 - 53.43) - 0}{\sqrt{\frac{13.32^2}{23} + \frac{12.22^2}{30}}}$$

Test statistic (cont.)

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

IN CONTEXT...

$$T = \frac{\text{point estimate} - \text{null value}}{SE} = \frac{(44.50 - 53.43) - 0}{\sqrt{\frac{13.32^2}{23} + \frac{12.22^2}{30}}} = \frac{-8.93}{3.56}$$

Test statistic (cont.)

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

in context...

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

Test statistic (cont.)

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

(A) 22

(B) 23

(C) 30

(D) 29

(E) 52

Test statistic (cont.)

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

(a) *22*

(b) 23

(c) 30

(d) 29

(e) 52

$$\rightarrow df = \min(n_{pt99} - 1, n_{pt100} - 1)$$
$$= \min(23 - 1, 30 - 1)$$
$$= \min(22, 29) = 22$$

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	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

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 two tails	0.100 0.200	0.050 0.100	0.025 0.050	0.010 0.020	0.005 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

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 small so 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

What is the equivalent confidence level for a one-sided hypothesis test at $\alpha = 0.05$?

(A) 90%

(B) 92.5%

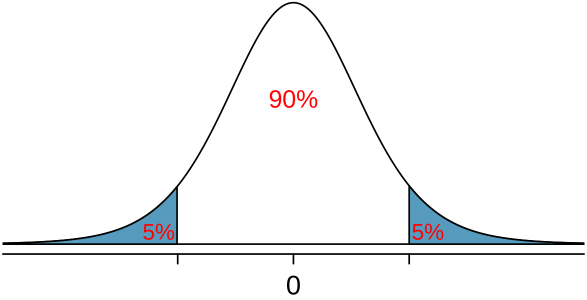
(C) 95%

(D) 97.5%

Equivalent confidence level

What is the equivalent confidence level for a one-sided hypothesis test at $\alpha = 0.05$?

(A) *90%*
(B) 92.5%
(C) 95%
(D) 97.5%



Critical value

What is the appropriate $t^\star$ 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 $t^\star$ 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 interval, and interpret it in context.

Confidence interval

Calculate the interval, and interpret it in context.

point estimate $\pm\ ME$

Confidence interval

Calculate the interval, and interpret it in context.

$$\textit{point estimate} \pm ME$$

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

Confidence interval

Calculate the interval, and interpret it in context.

$$\textit{point estimate} \pm ME$$

$$\begin{aligned} (\bar{x}_{pt99} - \bar{x}_{pt1}) \pm t^{\star}_{df} \times SE &= (44.50 - 53.43) \pm 1.72 \times 3.56 \\ &= -8.93 \pm 6.12 \end{aligned}$$

Confidence interval

Calculate the interval, and interpret it in context.

$$\textit{point estimate} \pm ME$$

$$\begin{aligned}(\bar{x}_{pt99} - \bar{x}_{pt1}) \pm t^{\star}_{df} \times SE &= (44.50 - 53.43) \pm 1.72 \times 3.56 \\ &= -8.93 \pm 6.12 \\ &= (-15.05, -2.81)\end{aligned}$$

Confidence interval

Calculate the interval, and interpret it in context.

$$\textit{point estimate} \pm ME$$

$$\begin{aligned}(\bar{x}_{pt99} - \bar{x}_{pt1}) \pm t^\star_{df} \times SE &= (44.50 - 53.43) \pm 1.72 \times 3.56 \\ &= -8.93 \pm 6.12 \\ &= (-15.05, -2.81)\end{aligned}$$

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 $n_1 < 30$ and/or $n_2 < 30$, difference between the sample means follow a t distribution with $SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_1}}$.

Recap: Inference using difference of two small sample means

- If $n_1 < 30$ and/or $n_2 < 30$, difference between the sample means follow a t distribution with $SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_1}}$.
- Conditions:
  - independence within groups (often verified by a random sample, and if sampling without replacement, $n <$ 10% of population)
  - independence between groups
  - $n_1 < 30$ and/or $n_2 < 30$ and no extreme skew in either group

Recap: Inference using difference of two small sample means

- If $n_1 < 30$ and/or $n_2 < 30$, difference between the sample means follow a t distribution with $SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_1}}$.
- Conditions:
  - independence within groups (often verified by a random sample, and if sampling without replacement, $n <$ 10% of population)
  - independence between groups
  - $n_1 < 30$ and/or $n_2 < 30$ and no extreme skew in either group
- Hypothesis testing:

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

Recap: Inference using difference of two small sample means

- If $n_1 < 30$ and/or $n_2 < 30$, difference between the sample means follow a t distribution with $SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_1}}$.
- Conditions:
  - independence within groups (often verified by a random sample, and if sampling without replacement, $n <$ 10% of population)
  - independence between groups
  - $n_1 < 30$ and/or $n_2 < 30$ and no extreme skew in either group
- Hypothesis testing:

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

- Confidence interval:

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

Melting ice cap

Scientists predict that global warming may have big effects on the polar regions within the next 100 years. One of the possible effects is that the northern ice cap may completely melt. Would this bother you a great deal, some, a little, or not at all if it actually happened?

(a) A great deal

(b) Some

(c) A little

(d) Not at all

Results from the GSS

The GSS asks the same question, below are the distributions of responses from the 2010 GSS as well as from a group of introductory statistics students at IE University:

	GSS	IE
A great deal	454	69
Some	124	30
A little	52	4
Not at all	50	2
Total	680	105

Parameter and point estimate

- **Parameter of interest:** Difference between the proportions of all IE students and all Europeans who would be bothered a great deal by the northern ice cap completely melting.

$$p_{IE} - p_{EU}$$

Parameter and point estimate

- **Parameter of interest:** Difference between the proportions of all IE students and all Europeans who would be bothered a great deal by the northern ice cap completely melting.

$$p_{IE} - p_{EU}$$

- **Point estimate:** Difference between the proportions of sampled IE students and sampled Europeans who would be bothered a great deal by the northern ice cap completely melting.

$$\hat{p}_{IE} - \hat{p}_{EU}$$

Inference for comparing proportions

- The details are the same as before...

Inference for comparing proportions

- The details are the same as before...
- CI: $point\ estimate \pm margin\ of\ error$

Inference for comparing proportions

- The details are the same as before...
- CI: $point\ estimate \pm margin\ of\ error$
- HT: Use $Z = \frac{point\ estimate - null\ value}{SE}$ to find appropriate p-value.

Inference for comparing proportions

- The details are the same as before...
- CI: $point\ estimate \pm margin\ of\ error$
- HT: Use $Z = \frac{point\ estimate - null\ value}{SE}$ to find appropriate p-value.
- We just need the appropriate standard error of the point estimate ($SE_{\hat{p}_{IE} - \hat{p}_{EU}}$), which is the only new concept.

Inference for comparing proportions

- The details are the same as before...
- CI: $point\ estimate \pm margin\ of\ error$
- HT: Use $Z = \frac{point\ estimate - null\ value}{SE}$ to find appropriate p-value.
- We just need the appropriate standard error of the point estimate ($SE_{\hat{p}_{IE} - \hat{p}_{EU}}$), which is the only new concept.

Standard error of the difference between two sample proportions

$$SE_{(\hat{p}_1 - \hat{p}_2)} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

Conditions for CI for difference of proportions

1. Independence within groups:
   - The EU group is sampled randomly and we're assuming that the IE group represents a random sample as well.

Conditions for CI for difference of proportions

1. Independence within groups:
    - The EU group is sampled randomly and we're assuming that the IE group represents a random sample as well.
    - $n_{IE} <$ 10% of all IE students and 680 $<$ 10% of all Europeans.

Conditions for CI for difference of proportions

1. Independence within groups:
   - The EU group is sampled randomly and we're assuming that the IE group represents a random sample as well.
   - $n_{IE} <$ 10% of all IE students and 680 $<$ 10% of all Europeans.

   We can assume that the attitudes of IE students in the sample are independent of each other, and attitudes of EU residents in the sample are independent of each other as well.

Conditions for CI for difference of proportions

1. Independence within groups:
   - The EU group is sampled randomly and we're assuming that the IE group represents a random sample as well.
   - $n_{IE} <$ 10% of all IE students and 680 $<$ 10% of all Europeans.

   We can assume that the attitudes of IE students in the sample are independent of each other, and attitudes of EU residents in the sample are independent of each other as well.

2. Independence between groups: The sampled IE students and the EU residents are independent of each other.

Conditions for CI for difference of proportions

1. Independence within groups:
   - The EU group is sampled randomly and we're assuming that the IE group represents a random sample as well.
   - $n_{IE} <$ 10% of all IE students and 680 $<$ 10% of all Europeans.

   We can assume that the attitudes of IE students in the sample are independent of each other, and attitudes of EU residents in the sample are independent of each other as well.
2. Independence between groups: The sampled IE students and the EU residents are independent of each other.
3. Success-failure:
   At least 15 observed successes and 15 observed failures in the two groups.

Construct a 95% confidence interval for the difference between the proportions of IE students and Europeans who would be bothered a great deal by the melting of the northern ice cap $(p_{IE} - p_{EU})$.

Data	IE	EU
A great deal	69	454
Not a great deal	36	226
Total	105	680

Construct a 95% confidence interval for the difference between the proportions of IE students and Europeans who would be bothered a great deal by the melting of the northern ice cap $(p_{IE} - p_{EU})$.

Data	IE	EU
A great deal	69	454
Not a great deal	36	226
Total	105	680
$\hat{p}$	0.657	0.668

Construct a 95% confidence interval for the difference between the proportions of IE students and Europeans who would be bothered a great deal by the melting of the northern ice cap $(p_{IE} - p_{EU})$.

Data	IE	EU
A great deal	69	454
Not a great deal	36	226
Total	105	680
$\hat{p}$	0.657	0.668

$$(\hat{p}_{IE} - \hat{p}_{EU}) \pm z^{\star} \times \sqrt{\frac{\hat{p}_{IE}(1-\hat{p}_{IE})}{n_{IE}} + \frac{\hat{p}_{EU}(1-\hat{p}_{EU})}{n_{EU}}}$$

Construct a 95% confidence interval for the difference between the proportions of IE students and Europeans who would be bothered a great deal by the melting of the northern ice cap $(p_{IE} - p_{EU})$.

Data	IE	EU
A great deal	69	454
Not a great deal	36	226
Total	105	680
$\hat{p}$	0.657	0.668

$$(\hat{p}_{IE} - \hat{p}_{EU}) \pm z^\star \times \sqrt{\frac{\hat{p}_{IE}(1 - \hat{p}_{IE})}{n_{IE}} + \frac{\hat{p}_{EU}(1 - \hat{p}_{EU})}{n_{EU}}}$$

$$= \quad (0.657 - 0.668)$$

Construct a 95% confidence interval for the difference between the proportions of IE students and Europeans who would be bothered a great deal by the melting of the northern ice cap ($p_{IE} - p_{EU}$).

Data	IE	EU
A great deal	69	454
Not a great deal	36	226
Total	105	680
$\hat{p}$	0.657	0.668

$$(\hat{p}_{IE} - \hat{p}_{EU}) \pm z^{\star} \times \sqrt{\frac{\hat{p}_{IE}(1-\hat{p}_{IE})}{n_{IE}} + \frac{\hat{p}_{EU}(1-\hat{p}_{EU})}{n_{EU}}}$$

$$= \quad (0.657 - 0.668) \pm 1.96$$

Construct a 95% confidence interval for the difference between the proportions of IE students and Europeans who would be bothered a great deal by the melting of the northern ice cap $(p_{IE} - p_{EU})$.

Data	IE	EU
A great deal	69	454
Not a great deal	36	226
Total	105	680
$\hat{p}$	0.657	0.668

$$(\hat{p}_{IE} - \hat{p}_{EU}) \pm z^{\star} \times \sqrt{\frac{\hat{p}_{IE}(1-\hat{p}_{IE})}{n_{IE}} + \frac{\hat{p}_{EU}(1-\hat{p}_{EU})}{n_{EU}}}$$

$$= \quad (0.657 - 0.668) \pm 1.96 \times \sqrt{\frac{0.657 \times 0.343}{105} + \frac{0.668 \times 0.332}{680}}$$

Construct a 95% confidence interval for the difference between the proportions of IE students and Europeans who would be bothered a great deal by the melting of the northern ice cap ($p_{IE} - p_{EU}$).

Data	IE	EU
A great deal	69	454
Not a great deal	36	226
Total	105	680
$\hat{p}$	0.657	0.668

$$
\begin{aligned}
& (\hat{p}_{IE} - \hat{p}_{EU}) \pm z^{\star} \times \sqrt{\frac{\hat{p}_{IE}(1-\hat{p}_{IE})}{n_{IE}} + \frac{\hat{p}_{EU}(1-\hat{p}_{EU})}{n_{EU}}} \\
= \quad & (0.657 - 0.668) \pm 1.96 \times \sqrt{\frac{0.657 \times 0.343}{105} + \frac{0.668 \times 0.332}{680}} \\
= \quad & -0.011 \pm
\end{aligned}
$$

Construct a 95% confidence interval for the difference between the proportions of IE students and Europeans who would be bothered a great deal by the melting of the northern ice cap ($p_{IE} - p_{EU}$).

Data	IE	EU
A great deal	69	454
Not a great deal	36	226
Total	105	680
$\hat{p}$	0.657	0.668

$$(\hat{p}_{IE} - \hat{p}_{EU}) \pm z^{\star} \times \sqrt{\frac{\hat{p}_{IE}(1-\hat{p}_{IE})}{n_{IE}} + \frac{\hat{p}_{EU}(1-\hat{p}_{EU})}{n_{EU}}}$$

$$= (0.657 - 0.668) \pm 1.96 \times \sqrt{\frac{0.657 \times 0.343}{105} + \frac{0.668 \times 0.332}{680}}$$

$$= -0.011 \pm 1.96 \times 0.0497$$

Construct a 95% confidence interval for the difference between the proportions of IE students and Europeans who would be bothered a great deal by the melting of the northern ice cap ($p_{IE} - p_{EU}$).

Data	IE	EU
A great deal	69	454
Not a great deal	36	226
Total	105	680
$\hat{p}$	0.657	0.668

$$
\begin{aligned}
& (\hat{p}_{IE} - \hat{p}_{EU}) \pm z^{\star} \times \sqrt{\frac{\hat{p}_{IE}(1-\hat{p}_{IE})}{n_{IE}} + \frac{\hat{p}_{EU}(1-\hat{p}_{EU})}{n_{EU}}} \\
= \quad & (0.657 - 0.668) \pm 1.96 \times \sqrt{\frac{0.657 \times 0.343}{105} + \frac{0.668 \times 0.332}{680}} \\
= \quad & -0.011 \pm 1.96 \times 0.0497 \\
= \quad & -0.011 \pm 0.097
\end{aligned}
$$

Construct a 95% confidence interval for the difference between the proportions of IE students and Europeans who would be bothered a great deal by the melting of the northern ice cap ($p_{IE} - p_{EU}$).

Data	IE	EU
A great deal	69	454
Not a great deal	36	226
Total	105	680
$\hat{p}$	0.657	0.668

$$
\begin{aligned}
& (\hat{p}_{IE} - \hat{p}_{EU}) \pm z^{\star} \times \sqrt{\frac{\hat{p}_{IE}(1-\hat{p}_{IE})}{n_{IE}} + \frac{\hat{p}_{EU}(1-\hat{p}_{EU})}{n_{EU}}} \\
= \quad & (0.657 - 0.668) \pm 1.96 \times \sqrt{\frac{0.657 \times 0.343}{105} + \frac{0.668 \times 0.332}{680}} \\
= \quad & -0.011 \pm 1.96 \times 0.0497 \\
= \quad & -0.011 \pm 0.097 \\
= \quad & (-0.108, 0.086)
\end{aligned}
$$

Which of the following is the correct set of hypotheses for testing if the proportion of all IE students who would be bothered a great deal by the melting of the northern ice cap differs from the proportion of all Europeans who do?

(A) $H_0 : p_{IE} = p_{EU}$
$H_A : p_{IE} \neq p_{EU}$

(B) $H_0 : \hat{p}_{IE} = \hat{p}_{EU}$
$H_A : \hat{p}_{IE} \neq \hat{p}_{EU}$

(C) $H_0 : p_{IE} - p_{EU} = 0$
$H_A : p_{IE} - p_{EU} \neq 0$

(D) $H_0 : p_{IE} = p_{EU}$
$H_A : p_{IE} < p_{EU}$

Which of the following is the correct set of hypotheses for testing if the proportion of all IE students who would be bothered a great deal by the melting of the northern ice cap differs from the proportion of all Europeans who do?

(A) $H_0 : p_{IE} = p_{EU}$
$H_A : p_{IE} \neq p_{EU}$

(B) $H_0 : \hat{p}_{IE} = \hat{p}_{EU}$
$H_A : \hat{p}_{IE} \neq \hat{p}_{EU}$

(C) $H_0 : p_{IE} - p_{EU} = 0$
$H_A : p_{IE} - p_{EU} \neq 0$

(D) $H_0 : p_{IE} = p_{EU}$
$H_A : p_{IE} < p_{EU}$

Both (a) and (c) are correct.

Flashback to working with one proportion

- When constructing a confidence interval for a population proportion, we check if the observed number of successes and failures are at least 10.

$$n\hat{p} \geq 15 \qquad n(1 - \hat{p}) \geq 15$$

Flashback to working with one proportion

- When constructing a confidence interval for a population proportion, we check if the observed number of successes and failures are at least 10.

$$n\hat{p} \geq 15 \qquad n(1-\hat{p}) \geq 15$$

- When conducting a hypothesis test for a population proportion, we check if the expected number of successes and failures are at least 10.

$$np_0 \geq 15 \qquad n(1-p_0) \geq 15$$

Pooled estimate of a proportion

- In the case of comparing two proportions where $H_0 : p_1 = p_2$, there isn't a given null value we can use to calculated the expected number of successes and failures in each sample.

Pooled estimate of a proportion

- In the case of comparing two proportions where $H_0 : p_1 = p_2$, there isn't a given null value we can use to calculated the expected number of successes and failures in each sample.
- Therefore, we need to first find a common (POOLED) proportion for the two groups, and use that in our analysis.

Pooled estimate of a proportion

- In the case of comparing two proportions where $H_0 : p_1 = p_2$, there isn't a given null value we can use to calculated the expected number of successes and failures in each sample.
- Therefore, we need to first find a common (POOLED) proportion for the two groups, and use that in our analysis.
- This simply means finding the proportion of total successes among the total number of observations.

Pooled estimate of a proportion

$$\hat{p} = \frac{\#\ of\ successes_1 + \#\ of\ successes_2}{n_1 + n_2}$$

Calculate the estimated pooled proportion of IE students and Europeans who would be bothered a great deal by the melting of the northern ice cap. Which sample proportion ($\hat{p}_{IE}$ or $\hat{p}_{EU}$) the pooled estimate is closer to? Why?

Data	IE	EU
A great deal	69	454
Not a great deal	36	226
Total	105	680
$\hat{p}$	0.657	0.668

Calculate the estimated pooled proportion of IE students and Europeans who would be bothered a great deal by the melting of the northern ice cap. Which sample proportion ($\hat{p}_{IE}$ or $\hat{p}_{EU}$) the pooled estimate is closer to? Why?

Data	IE	EU
A great deal	69	454
Not a great deal	36	226
Total	105	680
$\hat{p}$	0.657	0.668

$$\hat{p} = \frac{\#\, of\ successes_1 + \#\, of\ successes_2}{n_1 + n_2}$$

Calculate the estimated pooled proportion of IE students and Europeans who would be bothered a great deal by the melting of the northern ice cap. Which sample proportion ($\hat{p}_{IE}$ or $\hat{p}_{EU}$) the pooled estimate is closer to? Why?

Data	IE	EU
A great deal	69	454
Not a great deal	36	226
Total	105	680
$\hat{p}$	0.657	0.668

$$\hat{p} = \frac{\#\ of\ successes_1 + \#\ of\ successes_2}{n_1 + n_2}$$

$$= \frac{69 + 454}{105 + 680}$$

Calculate the estimated pooled proportion of IE students and Europeans who would be bothered a great deal by the melting of the northern ice cap. Which sample proportion ($\hat{p}_{IE}$ or $\hat{p}_{EU}$) the pooled estimate is closer to? Why?

Data	IE	EU
A great deal	69	454
Not a great deal	36	226
Total	105	680
$\hat{p}$	0.657	0.668

$$\hat{p} = \frac{\# of\ successes_1 + \# of\ successes_2}{n_1 + n_2}$$

$$= \frac{69 + 454}{105 + 680} = \frac{523}{785}$$

Calculate the estimated pooled proportion of IE students and Europeans who would be bothered a great deal by the melting of the northern ice cap. Which sample proportion ($\hat{p}_{IE}$ or $\hat{p}_{EU}$) the pooled estimate is closer to? Why?

Data	IE	EU
A great deal	69	454
Not a great deal	36	226
Total	105	680
$\hat{p}$	0.657	0.668

$$
\begin{aligned}
\hat{p} &= \frac{\#\,of\ successes_1 + \#\,of\ successes_2}{n_1 + n_2} \\
&= \frac{69 + 454}{105 + 680} = \frac{523}{785} = 0.666
\end{aligned}
$$

Do these data suggest that the proportion of all IE students who would be bothered a great deal by the melting of the northern ice cap differs from the proportion of all Europeans who do? Calculate the test statistic, the p-value, and interpret your conclusion in context of the data.

Data	IE	EU
A great deal	69	454
Not a great deal	36	226
Total	105	680
$\hat{p}$	0.657	0.668

Do these data suggest that the proportion of all IE students who would be bothered a great deal by the melting of the northern ice cap differs from the proportion of all Europeans who do? Calculate the test statistic, the p-value, and interpret your conclusion in context of the data.

Data	IE	EU
A great deal	69	454
Not a great deal	36	226
Total	105	680
$\hat{p}$	0.657	0.668

$$Z = \frac{(\hat{p}_{IE} - \hat{p}_{EU})}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n_{IE}} + \frac{\hat{p}(1-\hat{p})}{n_{EU}}}}$$

Do these data suggest that the proportion of all IE students who would be bothered a great deal by the melting of the northern ice cap differs from the proportion of all Europeans who do? Calculate the test statistic, the p-value, and interpret your conclusion in context of the data.

Data	IE	EU
A great deal	69	454
Not a great deal	36	226
Total	105	680
$\hat{p}$	0.657	0.668

$$Z = \frac{(\hat{p}_{IE} - \hat{p}_{EU})}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n_{IE}} + \frac{\hat{p}(1-\hat{p})}{n_{EU}}}}$$

$$= \frac{(0.657 - 0.668)}{\sqrt{\frac{0.666 \times 0.334}{105} + \frac{0.666 \times 0.334}{680}}} =$$

Do these data suggest that the proportion of all IE students who would be bothered a great deal by the melting of the northern ice cap differs from the proportion of all Europeans who do? Calculate the test statistic, the p-value, and interpret your conclusion in context of the data.

Data	IE	EU
A great deal	69	454
Not a great deal	36	226
Total	105	680
$\hat{p}$	0.657	0.668

$$Z = \frac{(\hat{p}_{IE} - \hat{p}_{EU})}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n_{IE}} + \frac{\hat{p}(1-\hat{p})}{n_{EU}}}}$$

$$= \frac{(0.657 - 0.668)}{\sqrt{\frac{0.666\times0.334}{105} + \frac{0.666\times0.334}{680}}} = \frac{-0.011}{0.0495}$$

Do these data suggest that the proportion of all IE students who would be bothered a great deal by the melting of the northern ice cap differs from the proportion of all Europeans who do? Calculate the test statistic, the p-value, and interpret your conclusion in context of the data.

Data	IE	EU
A great deal	69	454
Not a great deal	36	226
Total	105	680
$\hat{p}$	0.657	0.668

$$Z = \frac{(\hat{p}_{IE} - \hat{p}_{EU})}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n_{IE}} + \frac{\hat{p}(1-\hat{p})}{n_{EU}}}}$$

$$= \frac{(0.657 - 0.668)}{\sqrt{\frac{0.666\times0.334}{105} + \frac{0.666\times0.334}{680}}} = \frac{-0.011}{0.0495} = -0.22$$

Do these data suggest that the proportion of all IE students who would be bothered a great deal by the melting of the northern ice cap differs from the proportion of all Europeans who do? Calculate the test statistic, the p-value, and interpret your conclusion in context of the data.

Data	IE	EU
A great deal	69	454
Not a great deal	36	226
Total	105	680
$\hat{p}$	0.657	0.668

$$Z = \frac{(\hat{p}_{IE} - \hat{p}_{EU})}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n_{IE}} + \frac{\hat{p}(1-\hat{p})}{n_{EU}}}}$$

$$= \frac{(0.657 - 0.668)}{\sqrt{\frac{0.666\times0.334}{105} + \frac{0.666\times0.334}{680}}} = \frac{-0.011}{0.0495} = -0.22$$

$$p - value = 2 \times P(Z < -0.22)$$

Do these data suggest that the proportion of all IE students who would be bothered a great deal by the melting of the northern ice cap differs from the proportion of all Europeans who do? Calculate the test statistic, the p-value, and interpret your conclusion in context of the data.

Data	IE	EU
A great deal	69	454
Not a great deal	36	226
Total	105	680
$\hat{p}$	0.657	0.668

$$Z = \frac{(\hat{p}_{IE} - \hat{p}_{EU})}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n_{IE}} + \frac{\hat{p}(1-\hat{p})}{n_{EU}}}}$$

$$= \frac{(0.657 - 0.668)}{\sqrt{\frac{0.666\times0.334}{105} + \frac{0.666\times0.334}{680}}} = \frac{-0.011}{0.0495} = -0.22$$

$$p - value = 2 \times P(Z < -0.22) = 2 \times 0.41 = 0.82$$

Recap - comparing two proportions

- Population parameter: $(p_1 - p_2)$, point estimate: $(\hat{p}_1 - \hat{p}_2)$

Recap - comparing two proportions

- Population parameter: $(p_1 - p_2)$, point estimate: $(\hat{p}_1 - \hat{p}_2)$
- Conditions:

Recap - comparing two proportions

- Population parameter: $(p_1 - p_2)$, point estimate: $(\hat{p}_1 - \hat{p}_2)$
- Conditions:
  - independence within groups
    - random sample and 10% condition met for both groups
  - independence between groups
  - at least 15 successes and failures in each group
    - if not $\rightarrow$ randomization (Section 6.4)

Recap - comparing two proportions

- Population parameter: $(p_1 - p_2)$, point estimate: $(\hat{p}_1 - \hat{p}_2)$
- Conditions:
  - independence within groups
    - random sample and 10% condition met for both groups
  - independence between groups
  - at least 15 successes and failures in each group
    - if not $\rightarrow$ randomization (Section 6.4)
- $SE_{(\hat{p}_1 - \hat{p}_2)} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$
  - for CI: use $\hat{p}_1$ and $\hat{p}_2$
  - for HT:
    - when $H_0 : p_1 = p_2$: use $\hat{p}_{pool} = \frac{\#\, suc_1 + \# suc_2}{n_1 + n_2}$
    - when $H_0 : p_1 - p_2 =$ *(some value other than 0)*: use $\hat{p}_1$ and $\hat{p}_2$
      - this is pretty rare

Reference - standard error calculations

	one sample	two samples
mean	$SE = \frac{s}{\sqrt{n}}$	$SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
proportion	$SE = \sqrt{\frac{p(1-p)}{n}}$	$SE = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$

Reference - standard error calculations

	one sample	two samples
mean	$SE = \frac{s}{\sqrt{n}}$	$SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
proportion	$SE = \sqrt{\frac{p(1-p)}{n}}$	$SE = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$

- When working with means, it's very rare that σ is known, so we usually use s.

Reference - standard error calculations

	one sample	two samples
mean	$SE = \frac{s}{\sqrt{n}}$	$SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
proportion	$SE = \sqrt{\frac{p(1-p)}{n}}$	$SE = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$

- When working with means, it's very rare that σ is known, so we usually use s.
- When working with proportions,
  - if doing a hypothesis test, p comes from the null hypothesis
  - if constructing a confidence interval, use $\hat{p}$ instead

Effect Size

- Statistical Tests should report EFFECT SIZE. Consider the following example:
  - look younger with Botox
  - look 10 years younger with Botox
- How can we estimate effect size? Use CONFIDENCE INTERVALS!

A Treatment for Weight Loss

$$\begin{cases} H_0: & \mu_{diff} \geq 0 \\ H_a: & \mu_{diff} < 0 \end{cases}$$

- The sample size has an effect and, if we deal with very large samples, we will (almost) always end up rejecting H_0.
- The difference from zero might be really small. E.g.:

Treatment 1: 95% CI

$$[-0.5, -0.1]$$

at most, a loss of 550gr

Treatment 2: 95% CI

$$[-10, -5]$$

effect size is quite large here

Measuring and Reporting Effect Size

- Report confidence intervals (and not only the p-value)
- Cohen's d:

$$d = \frac{\mu_1 - \mu_2}{s_p}$$

  - $d = 0.2 \rightarrow$ small
  - $d = 0.5 \rightarrow$ medium
  - $d = 0.8 \rightarrow$ large

- Compute η^2, partial η^2, etc.
- Have a look at the following paper: "The Earth is Round ($p < .05$)," Jacob Cohen, American Psychologist, Vol. 49, No. 12, 997–1003 (1994)