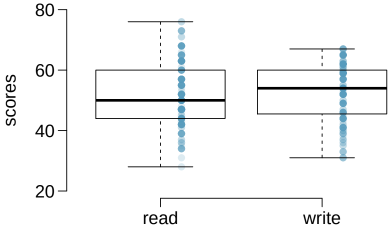
Chapter 5: Inference for numerical data

OpenIntro Statistics, 2nd Edition

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.

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

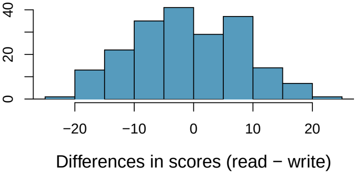
Analyzing paired data

- 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.*

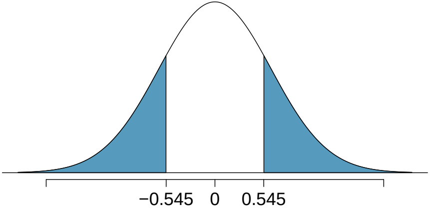
(b) The distribution of differences is bimodal, therefore we cannot continue with the hypothesis test.

(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.

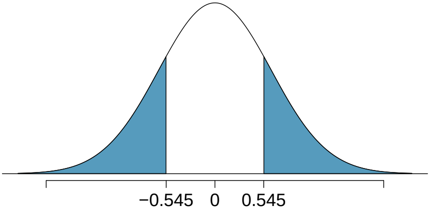
Calculating the test-statistic and the p-value

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

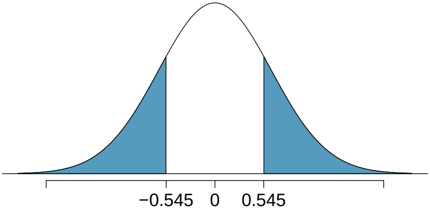
The observed average difference between the two scores is -0.545 points and the standard deviation of the difference is 8.887 points. Do these data provide convincing evidence of a difference between the average scores on the two exams? Use $\alpha = 0.05$.



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

Calculating the test-statistic and the p-value

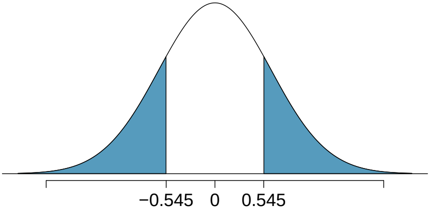
The observed average difference between the two scores is -0.545 points and the standard deviation of the difference is 8.887 points. Do these data provide convincing evidence of a difference between the average scores on the two exams? Use $\alpha = 0.05$.



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

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 not 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

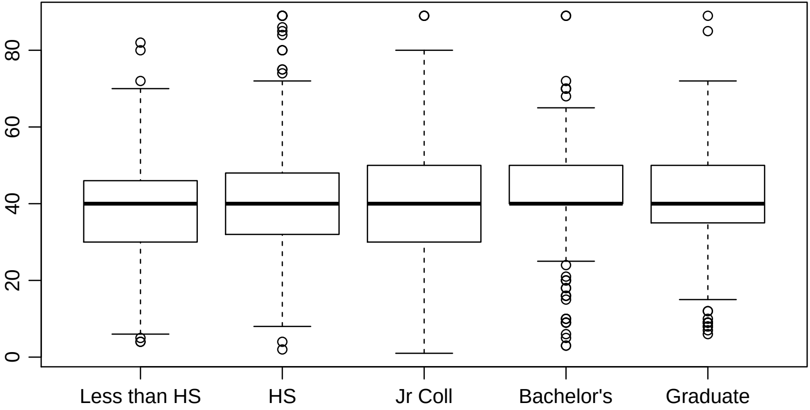
$$\begin{aligned} -0.545 \pm 1.96\frac{8.887}{\sqrt{200}} &= -0.545 \pm 1.96 \times 0.628 \\ &= -0.545 \pm 1.23 \\ &= (-1.775, 0.685) \end{aligned}$$

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

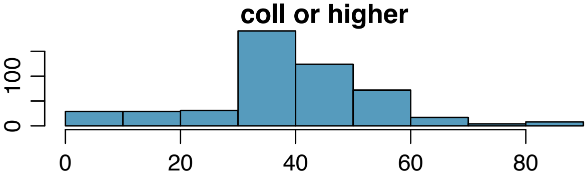
- Say we are only interested the difference between the number of hours worked per week by college and non-college graduates.

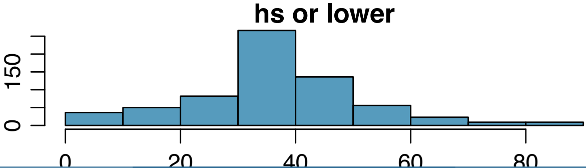
Collapsing levels into two

- 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.

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.
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:

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

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:

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

- And all $ME = \textit{critical value} \times \textit{SE of point estimate}$
- 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:

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

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

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

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

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

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.

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

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

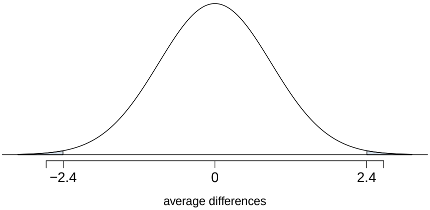
H_0: $\mu_{coll} = \mu_{hs} \rightarrow \mu_{coll} - \mu_{hs} = 0$
H_A: $\mu_{coll} \neq \mu_{hs} \rightarrow \mu_{coll} - \mu_{hs} \neq 0$

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

Calculating the test-statistic and the p-value

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

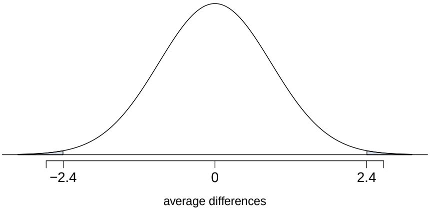
$$\bar{x}_{coll} - \bar{x}_{hs} = 2.4, \; SE_{(\bar{x}_{coll} - \bar{x}_{hs})} = 0.89$$



Calculating the test-statistic and the p-value

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

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

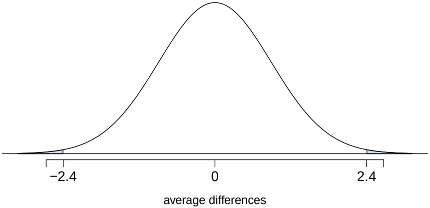


$$Z = \frac{(\bar{x}_{coll} - \bar{x}_{hs}) - 0}{SE_{(\bar{x}_{coll} - \bar{x}_{hs})}}$$

Calculating the test-statistic and the p-value

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

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

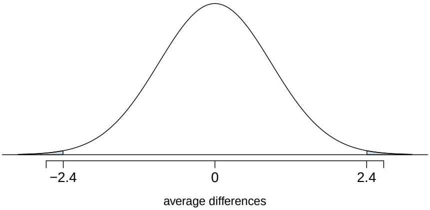


$$Z = \frac{(\bar{x}_{coll} - \bar{x}_{hs}) - 0}{SE_{(\bar{x}_{coll} - \bar{x}_{hs})}} = \frac{2.4}{0.89} = 2.70$$

Calculating the test-statistic and the p-value

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

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

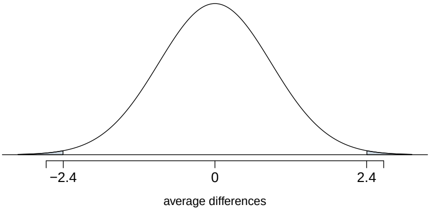


$$\begin{aligned} Z &= \frac{(\bar{x}_{coll} - \bar{x}_{hs}) - 0}{SE_{(\bar{x}_{coll} - \bar{x}_{hs})}} \\ &= \frac{2.4}{0.89} = 2.70 \\ upper\ tail &= 1 - 0.9965 = 0.0035 \end{aligned}$$

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.

Friday the 13th

Between 1990 - 1992 researchers in the UK collected data on traffic flow, accidents, and hospital admissions on Friday 13th and the previous Friday, Friday 6th. Below is an excerpt from this data set on traffic flow. We can assume that traffic flow on given day at locations 1 and 2 are independent.

	type	date	6th	13th	diff	location
1	traffic	1990, July	139246	138548	698	loc 1
2	traffic	1990, July	134012	132908	1104	loc 2
3	traffic	1991, September	137055	136018	1037	loc 1
4	traffic	1991, September	133732	131843	1889	loc 2
5	traffic	1991, December	123552	121641	1911	loc 1
6	traffic	1991, December	121139	118723	2416	loc 2
7	traffic	1992, March	128293	125532	2761	loc 1
8	traffic	1992, March	124631	120249	4382	loc 2
9	traffic	1992, November	124609	122770	1839	loc 1
10	traffic	1992, November	117584	117263	321	loc 2

Scanlon, T.J., Luben, R.N., Scanlon, F.L., Singleton, N. (1993), "Is Friday the 13th Bad For Your Health?," BMJ, 307, 1584-1586.

Friday the 13th

- We want to investigate if people's behavior is different on Friday 13th compared to Friday 6th.

Friday the 13th

- We want to investigate if people's behavior is different on Friday 13th compared to Friday 6th.
- One approach is to compare the traffic flow on these two days.

Friday the 13$^{\text{th}}$

- We want to investigate if people's behavior is different on Friday 13$^{\text{th}}$ compared to Friday 6$^{\text{th}}$.
- One approach is to compare the traffic flow on these two days.
- H_0 : Average traffic flow on Friday 6$^{\text{th}}$ and 13$^{\text{th}}$ are equal.
  H_A : Average traffic flow on Friday 6$^{\text{th}}$ and 13$^{\text{th}}$ are different.

Friday the 13th

- We want to investigate if people's behavior is different on Friday 13th compared to Friday 6th.
- One approach is to compare the traffic flow on these two days.
- H_0 : Average traffic flow on Friday 6th and 13th are equal.
  H_A : Average traffic flow on Friday 6th and 13th are different.

Each case in the data set represents traffic flow recorded at the same location in the same month of the same year: one count from Friday 6th and the other Friday 13th. Are these two counts independent?

Friday the 13th

- We want to investigate if people's behavior is different on Friday 13th compared to Friday 6th.
- One approach is to compare the traffic flow on these two days.
- H_0 : Average traffic flow on Friday 6th and 13th are equal.
  H_A : Average traffic flow on Friday 6th and 13th are different.

Each case in the data set represents traffic flow recorded at the same location in the same month of the same year: one count from Friday 6th and the other Friday 13th. Are these two counts independent?

No

Hypotheses

What are the hypotheses for testing for a difference between the average traffic flow between Friday 6th and 13th?

(a) $H_0 : \mu_{6th} = \mu_{13th}$
$H_A : \mu_{6th} \neq \mu_{13th}$

(b) $H_0 : p_{6th} = p_{13th}$
$H_A : p_{6th} \neq p_{13th}$

(c) $H_0 : \mu_{diff} = 0$
$H_A : \mu_{diff} \neq 0$

(d) $H_0 : \bar{x}_{diff} = 0$
$H_A : \bar{x}_{diff} = 0$

Hypotheses

What are the hypotheses for testing for a difference between the average traffic flow between Friday 6th and 13th?

(a) $H_0 : \mu_{6th} = \mu_{13th}$
$H_A : \mu_{6th} \neq \mu_{13th}$

(b) $H_0 : p_{6th} = p_{13th}$
$H_A : p_{6th} \neq p_{13th}$

(c) $H_0 : \mu_{diff} = 0$
$H_A : \mu_{diff} \neq 0$

(d) $H_0 : \bar{x}_{diff} = 0$
$H_A : \bar{x}_{diff} = 0$

Conditions

- *Independence:* We are told to assume that cases (rows) are independent.

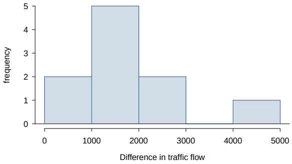
Conditions

- *Independence:* We are told to assume that cases (rows) are independent.
- *Sample size / skew:*

Conditions

- *Independence:* We are told to assume that cases (rows) are independent.
- *Sample size / skew:*

- The sample distribution does not appear to be extremely skewed, but it's very difficult to assess with such a small sample size. We might want to think about whether we would expect the population distribution to be skewed or not – probably not, it should be equally likely to have days with lower than average traffic and higher than average traffic.

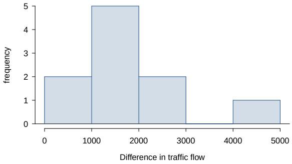


- $n < 30!$

Conditions

- *Independence:* We are told to assume that cases (rows) are independent.
- *Sample size / skew:*

- The sample distribution does not appear to be extremely skewed, but it's very difficult to assess with such a small sample size. We might want to think about whether we would expect the population distribution to be skewed or not – probably not, it should be equally likely to have days with lower than average traffic and higher than average traffic.



- $n < 30!$

So what do we do when the sample size is small?

Review: what purpose does a large sample serve?

As long as observations are independent, and the population distribution is not extremely skewed, a large sample would ensure that...

- the sampling distribution of the mean is nearly normal
- the estimate of the standard error, as $\frac{s}{\sqrt{n}}$, is reliable

The normality condition

- The CLT, which states that sampling distributions will be nearly normal, holds true for *any* sample size as long as the population distribution is nearly normal.

The normality condition

- The CLT, which states that sampling distributions will be nearly normal, holds true for *any* sample size as long as the population distribution is nearly normal.
- While this is a helpful special case, it's inherently difficult to verify normality in small data sets.

The normality condition

- The CLT, which states that sampling distributions will be nearly normal, holds true for *any* sample size as long as the population distribution is nearly normal.
- While this is a helpful special case, it's inherently difficult to verify normality in small data sets.
- We should exercise caution when verifying the normality condition for small samples. It is important to not only examine the data but also think about where the data come from.
    - For example, ask: would I expect this distribution to be symmetric, and am I confident that outliers are rare?

The t distribution

- When working with small samples, and the population standard deviation is unknown (almost always), the uncertainty of the standard error estimate is addressed by using a new distribution: the *t distribution*.

The t distribution

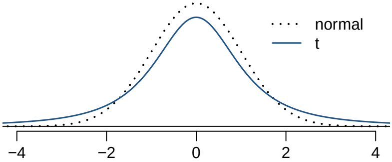
- When working with small samples, and the population standard deviation is unknown (almost always), the uncertainty of the standard error estimate is addressed by using a new distribution: the *t distribution*.
- This distribution also has a bell shape, but its tails are *thicker* than the normal model's.

The t distribution

- When working with small samples, and the population standard deviation is unknown (almost always), the uncertainty of the standard error estimate is addressed by using a new distribution: the *t distribution*.
- This distribution also has a bell shape, but its tails are *thicker* than the normal model's.
- Therefore observations are more likely to fall beyond two SDs from the mean than under the normal distribution.

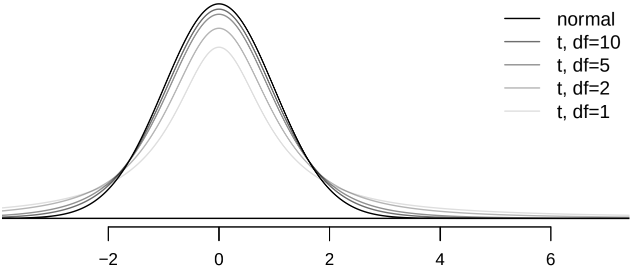
The t distribution

- When working with small samples, and the population standard deviation is unknown (almost always), the uncertainty of the standard error estimate is addressed by using a new distribution: the *t distribution*.
- This distribution also has a bell shape, but its tails are *thicker* than the normal model's.
- Therefore observations are more likely to fall beyond two SDs from the mean than under the normal distribution.
- These extra thick tails are helpful for resolving our problem with a less reliable estimate the standard error (since n is small)



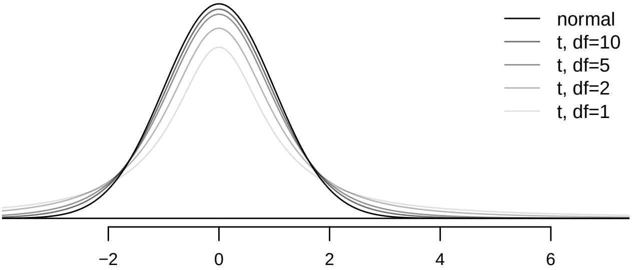
The t distribution (cont.)

- Always centered at zero, like the standard normal (z) distribution.
- Has a single parameter: *degrees of freedom* (df).



The t distribution (cont.)

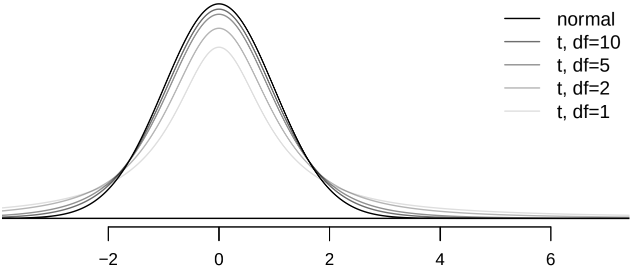
- Always centered at zero, like the standard normal (z) distribution.
- Has a single parameter: *degrees of freedom* (df).



What happens to shape of the t distribution as df increases?

The t distribution (cont.)

- Always centered at zero, like the standard normal (z) distribution.
- Has a single parameter: *degrees of freedom* (df).



What happens to shape of the t distribution as df increases?

Approaches normal.

Back to Friday the 13th

	type	date	6th	13th	diff	location
1	traffic	1990, July	139246	138548	698	loc 1
2	traffic	1990, July	134012	132908	1104	loc 2
3	traffic	1991, September	137055	136018	1037	loc 1
4	traffic	1991, September	133732	131843	1889	loc 2
5	traffic	1991, December	123552	121641	1911	loc 1
6	traffic	1991, December	121139	118723	2416	loc 2
7	traffic	1992, March	128293	125532	2761	loc 1
8	traffic	1992, March	124631	120249	4382	loc 2
9	traffic	1992, November	124609	122770	1839	loc 1
10	traffic	1992, November	117584	117263	321	loc 2

$\downarrow$

$$\bar{x}_{diff} = 1836$$

$$s_{diff} = 1176$$

$$n = 10$$

Finding the test statistic

Test statistic for inference on a small sample mean

The test statistic for inference on a small sample ($n < 50$) mean is the T statistic with $df = n - 1$.

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

Finding the test statistic

Test statistic for inference on a small sample mean

The test statistic for inference on a small sample ($n < 50$) mean is the T statistic with $df = n - 1$.

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

in context...

$$point\ estimate \quad = \quad \bar{x}_{diff} = 1836$$

Finding the test statistic

Test statistic for inference on a small sample mean

The test statistic for inference on a small sample ($n < 50$) mean is the T statistic with $df = n - 1$.

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

in context...

$$\begin{aligned} \textit{point estimate} &= \bar{x}_{diff} = 1836 \\ SE &= \frac{s_{diff}}{\sqrt{n}} = \frac{1176}{\sqrt{10}} = 372 \end{aligned}$$

Finding the test statistic

Test statistic for inference on a small sample mean

The test statistic for inference on a small sample ($n < 50$) mean is the T statistic with $df = n - 1$.

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

in context...

$$\begin{aligned}
\textit{point estimate} &= \bar{x}_{diff} = 1836 \\
SE &= \frac{s_{diff}}{\sqrt{n}} = \frac{1176}{\sqrt{10}} = 372 \\
T &= \frac{1836 - 0}{372} = 4.94
\end{aligned}$$

Finding the test statistic

Test statistic for inference on a small sample mean

The test statistic for inference on a small sample ($n < 50$) mean is the T statistic with $df = n - 1$.

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

in context...

$$\begin{aligned} point\ estimate &= \bar{x}_{diff} = 1836 \\ SE &= \frac{s_{diff}}{\sqrt{n}} = \frac{1176}{\sqrt{10}} = 372 \\ T &= \frac{1836 - 0}{372} = 4.94 \\ df &= 10 - 1 = 9 \end{aligned}$$

Note: Null value is 0 because in the null hypothesis we set $\mu_{diff} = 0$.

Finding the p-value

- The p-value is, once again, calculated as the area tail area under the t distribution.

Finding the p-value

- The p-value is, once again, calculated as the area tail area under the t distribution.
- Using R:

```
> 2 * pt(4.94, df = 9, lower.tail = FALSE)

[1] 0.0008022394
```

Finding the p-value

- The p-value is, once again, calculated as the area tail area under the t distribution.
- Using R:

```
> 2 * pt(4.94, df = 9, lower.tail = FALSE)

[1] 0.0008022394
```

- Using a web applet:
  http://www.socr.ucla.edu/htmls/SOCR_Distributions.html

Finding the p-value

- The p-value is, once again, calculated as the area tail area under the t distribution.
- Using R:

```
> 2 * pt(4.94, df = 9, lower.tail = FALSE)

[1] 0.0008022394
```

- Using a web applet:
  http://www.socr.ucla.edu/htmls/SOCR_Distributions.html
- Or when these aren't available, we can use a t table.

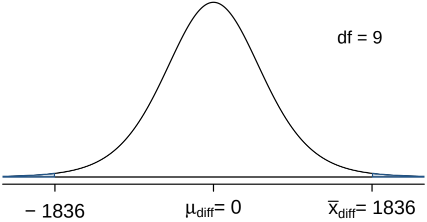
Finding the p-value

Locate the calculated T statistic on the appropriate df row, obtain the p-value from the corresponding column heading (one or two tail, depending on the alternative hypothesis).

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 1	3.08	6.31	12.71	31.82	63.66
2	1.89	2.92	4.30	6.96	9.92
3	1.64	2.35	3.18	4.54	5.84
⋮	⋮	⋮	⋮	⋮	
17	1.33	1.74	2.11	2.57	2.90
18	1.33	1.73	2.10	2.55	2.88
19	1.33	1.73	2.09	2.54	2.86
20	1.33	1.72	2.09	2.53	2.85
⋮	⋮	⋮	⋮	⋮	
400	1.28	1.65	1.97	2.34	2.59
500	1.28	1.65	1.96	2.33	2.59
∞	1.28	1.64	1.96	2.33	2.58

Finding the p-value (cont.)

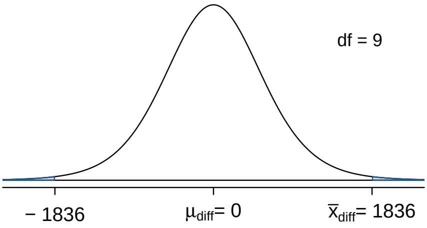
one tail		0.100	0.050	0.025	0.010	0.005
two tails		0.200	0.100	0.050	0.020	0.010
df	6	1.44	1.94	2.45	3.14	3.71
	7	1.41	1.89	2.36	3.00	3.50
	8	1.40	1.86	2.31	2.90	3.36
	9	1.38	1.83	2.26	2.82	3.25
	10	1.37	1.81	2.23	2.76	3.17



$$T = 4.94$$

Finding the p-value (cont.)

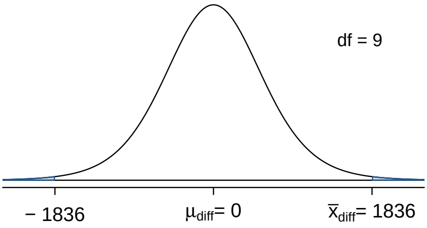
one tail		0.100	0.050	0.025	0.010	0.005
two tails		0.200	0.100	0.050	0.020	0.010
df	6	1.44	1.94	2.45	3.14	3.71
	7	1.41	1.89	2.36	3.00	3.50
	8	1.40	1.86	2.31	2.90	3.36
	9	1.38	1.83	2.26	2.82	3.25
	10	1.37	1.81	2.23	2.76	3.17



$$T = 4.94$$

Finding the p-value (cont.)

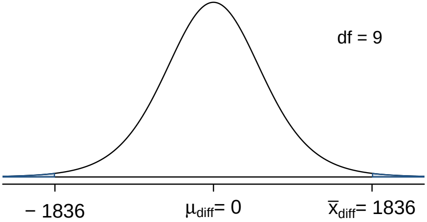
one tail		0.100	0.050	0.025	0.010	0.005	
two tails		0.200	0.100	0.050	0.020	*0.010*	→
df	6	1.44	1.94	2.45	3.14	3.71	
	7	1.41	1.89	2.36	3.00	3.50	
	8	1.40	1.86	2.31	2.90	3.36	
	9	1.38	1.83	2.26	2.82	*3.25*	→
	10	1.37	1.81	2.23	2.76	3.17	



$$T = 4.94$$

Finding the p-value (cont.)

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	6	1.44	1.94	2.45	3.14	3.71	
	7	1.41	1.89	2.36	3.00	3.50	
	8	1.40	1.86	2.31	2.90	3.36	
	9	1.38	1.83	2.26	2.82	*3.25*	→
	10	1.37	1.81	2.23	2.76	3.17	

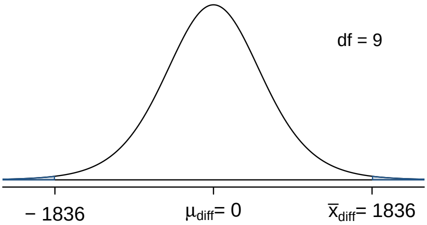


$T = 4.94$

What is the conclusion of the hypothesis test?

Finding the p-value (cont.)

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 6	1.44	1.94	2.45	3.14	3.71	
7	1.41	1.89	2.36	3.00	3.50	
8	1.40	1.86	2.31	2.90	3.36	
9	1.38	1.83	2.26	2.82	*3.25*	→
10	1.37	1.81	2.23	2.76	3.17	



$T = 4.94$

What is the conclusion of the hypothesis test?

The data provide convincing evidence of a difference between traffic flow on Friday 6th and 13th.

What is the difference?

- We concluded that there is a difference in the traffic flow between Friday 6^{th} and 13^{th}.

What is the difference?

- We concluded that there is a difference in the traffic flow between Friday 6th and 13th.
- But it would be more interesting to find out what exactly this difference is.

What is the difference?

- We concluded that there is a difference in the traffic flow between Friday 6^{th} and 13^{th}.
- But it would be more interesting to find out what exactly this difference is.
- We can use a confidence interval to estimate this difference.

Confidence interval for a small sample mean

- Confidence intervals are always of the form

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

Confidence interval for a small sample mean

- Confidence intervals are always of the form

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

- ME is always calculated as the product of a critical value and SE.

Confidence interval for a small sample mean

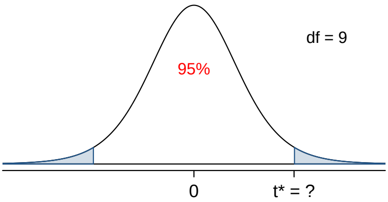
- Confidence intervals are always of the form

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

- ME is always calculated as the product of a critical value and SE.
- Since small sample means follow a t distribution (and not a z distribution), the critical value is a $t^\star$ (as opposed to a $z^\star$).

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

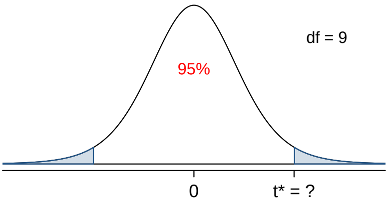
Finding the critical t ($t^\star$)



$n = 10$, $df = 10 - 1 = 9$, $t^\star$ is at the intersection of row $df = 9$ and two tail probability 0.05.

	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	6	1.44	1.94	2.45	3.14	3.71
	7	1.41	1.89	2.36	3.00	3.50
	8	1.40	1.86	2.31	2.90	3.36
	9	1.38	1.83	2.26	2.82	3.25
	10	1.37	1.81	2.23	2.76	3.17

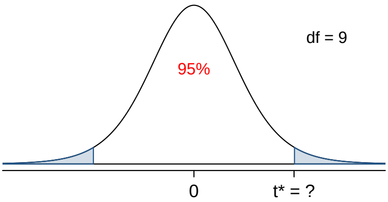
Finding the critical t ($t^\star$)



$n = 10$, $df = 10 - 1 = 9$, $t^\star$ is at the intersection of row $df = 9$ and two tail probability 0.05.

	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	6	1.44	1.94	2.45	3.14	3.71
	7	1.41	1.89	2.36	3.00	3.50
	8	1.40	1.86	2.31	2.90	3.36
	9	1.38	1.83	2.26	2.82	3.25
	10	1.37	1.81	2.23	2.76	3.17

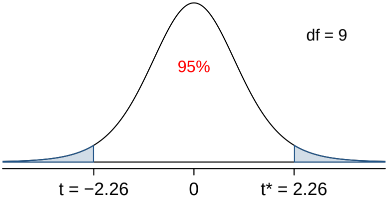
Finding the critical t ($t^\star$)



$n = 10$, $df = 10 - 1 = 9$, $t^\star$ is at the intersection of row $df = 9$ and two tail probability 0.05.

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 6	1.44	1.94	2.45	3.14	3.71
7	1.41	1.89	2.36	3.00	3.50
8	1.40	1.86	2.31	2.90	3.36
9	1.38	1.83	2.26	2.82	3.25
10	1.37	1.81	2.23	2.76	3.17

Finding the critical t ($t^\star$)



$n = 10$, $df = 10 - 1 = 9$, $t^\star$ is at the intersection of row $df = 9$ and two tail probability 0.05.

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 6	1.44	1.94	2.45	3.14	3.71
7	1.41	1.89	2.36	3.00	3.50
8	1.40	1.86	2.31	2.90	3.36
9	1.38	1.83	*2.26*	2.82	3.25
10	1.37	1.81	2.23	2.76	3.17

Constructing a CI for a small sample mean

Which of the following is the correct calculation of a 95% confidence interval for the difference between the traffic flow between Friday 6th and 13th?

$$\bar{x}_{diff} = 1836 \qquad s_{diff} = 1176 \qquad n = 10 \qquad SE = 372$$

(a) $1836 \pm 1.96 \times 372$

(b) $1836 \pm 2.26 \times 372$

(c) $1836 \pm -2.26 \times 372$

(d) $1836 \pm 2.26 \times 1176$

Constructing a CI for a small sample mean

Which of the following is the correct calculation of a 95% confidence interval for the difference between the traffic flow between Friday 6th and 13th?

$$\bar{x}_{diff} = 1836 \qquad s_{diff} = 1176 \qquad n = 10 \qquad SE = 372$$

(a) $1836 \pm 1.96 \times 372$

(b) $1836 \pm 2.26 \times 372$ $\rightarrow$ *(995, 2677)*

(c) $1836 \pm -2.26 \times 372$

(d) $1836 \pm 2.26 \times 1176$

Interpreting the CI

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

$$\mu_{diff:6th-13th} = (995, 2677)$$

We are 95% confident that ...

(a) the difference between the average number of cars on the road on Friday 6th and 13th is between 995 and 2,677.

(b) on Friday 6th there are 995 to 2,677 fewer cars on the road than on the Friday 13th, on average.

(c) on Friday 6th there are 995 fewer to 2,677 more cars on the road than on the Friday 13th, on average.

(d) on Friday 13th there are 995 to 2,677 fewer cars on the road than on the Friday 6th, on average.

Interpreting the CI

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

$$\mu_{diff:6th-13th} = (995, 2677)$$

We are 95% confident that ...

(a) the difference between the average number of cars on the road on Friday 6^{th} and 13^{th} is between 995 and 2,677.

(b) on Friday 6^{th} there are 995 to 2,677 fewer cars on the road than on the Friday 13^{th}, on average.

(c) on Friday 6^{th} there are 995 fewer to 2,677 more cars on the road than on the Friday 13^{th}, on average.

(d) *on Friday 13^{th} there are 995 to 2,677 fewer cars on the road than on the Friday 6^{th}, on average.*

Synthesis

Does the conclusion from the hypothesis test agree with the findings of the confidence interval?

Do you think the findings of this study suggests that people believe Friday 13$^{\text{th}}$ is a day of bad luck?

Synthesis

Does the conclusion from the hypothesis test agree with the findings of the confidence interval?

Yes, the hypothesis test found a significant difference, and the CI does not contain the null value of 0.

Do you think the findings of this study suggests that people believe Friday 13$^{\text{th}}$ is a day of bad luck?

Synthesis

Does the conclusion from the hypothesis test agree with the findings of the confidence interval?

Yes, the hypothesis test found a significant difference, and the CI does not contain the null value of 0.

Do you think the findings of this study suggests that people believe Friday 13th is a day of bad luck?

No, this is an observational study. We have just observed a significant difference between the number of cars on the road on these two days. We have not tested for people's beliefs.

Recap: Inference using a small sample mean

- If $n < 30$, sample means follow a t distribution with $SE = \frac{s}{\sqrt{n}}$.

Recap: Inference using a small sample mean

- If $n < 30$, sample means follow a t distribution with $SE = \frac{s}{\sqrt{n}}$.
- Conditions:
  - independence of observations (often verified by a random sample, and if sampling without replacement, $n <$ 10% of population)
  - $n < 30$ and no extreme skew

Recap: Inference using a small sample mean

- If $n < 30$, sample means follow a t distribution with $SE = \frac{s}{\sqrt{n}}$.
- Conditions:
  - independence of observations (often verified by a random sample, and if sampling without replacement, $n <$ 10% of population)
  - $n < 30$ and no extreme skew
- Hypothesis testing:

$$T_{df} = \frac{\text{point estimate} - \text{null value}}{SE}, \text{ where } df = n - 1$$

Recap: Inference using a small sample mean

- If $n < 30$, sample means follow a t distribution with $SE = \frac{s}{\sqrt{n}}$.
- Conditions:
  - independence of observations (often verified by a random sample, and if sampling without replacement, $n <$ 10% of population)
  - $n < 30$ and no extreme skew
- Hypothesis testing:

$$T_{df} = \frac{\text{point estimate} - \text{null value}}{SE}, \text{ where } df = n - 1$$

- Confidence interval:

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

Recap: Inference using a small sample mean

- If $n < 30$, sample means follow a t distribution with $SE = \frac{s}{\sqrt{n}}$.
- Conditions:
  - independence of observations (often verified by a random sample, and if sampling without replacement, $n <$ 10% of population)
  - $n < 30$ and no extreme skew
- Hypothesis testing:

$$T_{df} = \frac{\text{point estimate} - \text{null value}}{SE}, \text{ where } df = n - 1$$

- Confidence interval:

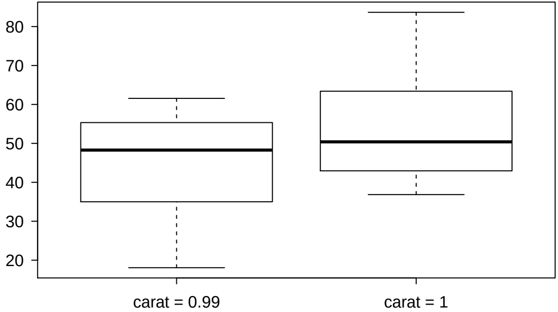
$$\text{point estimate} \pm t^{\star}_{df} \times SE$$

Note: The example we used was for paired means (difference between dependent groups). We took the difference between the observations and used only these differences (one sample) in our analysis, therefore the mechanics are the same as when we are working with just one sample.

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) $H_0 : \mu_{pt99} = \mu_{pt100}$
$H_A : \mu_{pt99} \neq \mu_{pt100}$

(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) $H_0 : \bar{x}_{pt99} = \bar{x}_{pt100}$
$H_A : \bar{x}_{pt99} < \bar{x}_{pt100}$

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 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...

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

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...

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

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
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 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%

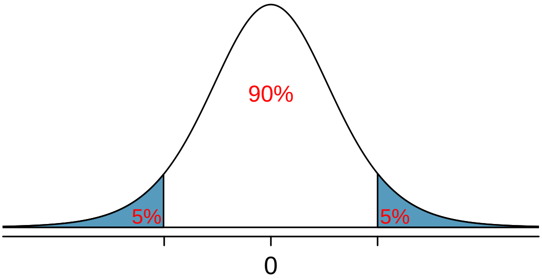
(c) 95%

(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.

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

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:

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

- The Wolf River in Tennessee flows past an abandoned site once used by the pesticide industry for dumping wastes, including chlordane (pesticide), aldrin, and dieldrin (both insecticides).

- The Wolf River in Tennessee flows past an abandoned site once used by the pesticide industry for dumping wastes, including chlordane (pesticide), aldrin, and dieldrin (both insecticides).
- These highly toxic organic compounds can cause various cancers and birth defects.

- The Wolf River in Tennessee flows past an abandoned site once used by the pesticide industry for dumping wastes, including chlordane (pesticide), aldrin, and dieldrin (both insecticides).
- These highly toxic organic compounds can cause various cancers and birth defects.
- The standard methods to test whether these substances are present in a river is to take samples at six-tenths depth.

- The Wolf River in Tennessee flows past an abandoned site once used by the pesticide industry for dumping wastes, including chlordane (pesticide), aldrin, and dieldrin (both insecticides).
- These highly toxic organic compounds can cause various cancers and birth defects.
- The standard methods to test whether these substances are present in a river is to take samples at six-tenths depth.
- But since these compounds are denser than water and their molecules tend to stick to particles of sediment, they are more likely to be found in higher concentrations near the bottom than near mid-depth.

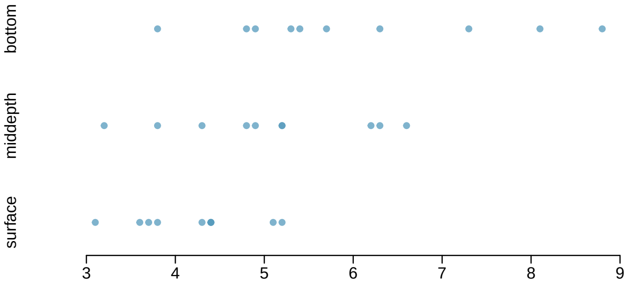
Data

Aldrin concentration (nanograms per liter) at three levels of depth.

	aldrin	depth
1	3.80	bottom
2	4.80	bottom
...		
10	8.80	bottom
11	3.20	middepth
12	3.80	middepth
...		
20	6.60	middepth
21	3.10	surface
22	3.60	surface
...		
30	5.20	surface

Exploratory analysis

Aldrin concentration (nanograms per liter) at three levels of depth.



	n	mean	sd
bottom	10	6.04	1.58
middepth	10	5.05	1.10
surface	10	4.20	0.66
overall	30	5.1 0	1.37

Research question

Is there a difference between the mean aldrin concentrations among the three levels?

Research question

Is there a difference between the mean aldrin concentrations among the three levels?

- To compare means of 2 groups we use a Z or a T statistic.

Research question

Is there a difference between the mean aldrin concentrations among the three levels?

- To compare means of 2 groups we use a Z or a T statistic.
- To compare means of 3+ groups we use a new test called *ANOVA* and a new statistic called *F*.

ANOVA

ANOVA is used to assess whether the mean of the outcome variable is different for different levels of a categorical variable.

ANOVA

ANOVA is used to assess whether the mean of the outcome variable is different for different levels of a categorical variable.

H_0 : The mean outcome is the same across all categories,

$$\mu_1 = \mu_2 = \cdots = \mu_k,$$

where μ_i represents the mean of the outcome for observations in category i.

H_A : At least one mean is different than others.

Conditions

1. The observations should be independent within and between groups
   - If the data are a simple random sample from less than 10% of the population, this condition is satisfied.
   - Carefully consider whether the data may be independent (e.g. no pairing).
   - Always important, but sometimes difficult to check.

Conditions

1. The observations should be independent within and between groups
   - If the data are a simple random sample from less than 10% of the population, this condition is satisfied.
   - Carefully consider whether the data may be independent (e.g. no pairing).
   - Always important, but sometimes difficult to check.
2. The observations within each group should be nearly normal.
   - Especially important when the sample sizes are small.

How do we check for normality?

Conditions

1. The observations should be independent within and between groups
   - If the data are a simple random sample from less than 10% of the population, this condition is satisfied.
   - Carefully consider whether the data may be independent (e.g. no pairing).
   - Always important, but sometimes difficult to check.
2. The observations within each group should be nearly normal.
   - Especially important when the sample sizes are small.

How do we check for normality?

3. The variability across the groups should be about equal.
   - Especially important when the sample sizes differ between groups.

How can we check this condition?

z/t test vs. ANOVA - Purpose

z/t test

Compare means from *two* groups to see whether they are so far apart that the observed difference cannot reasonably be attributed to sampling variability.

$$H_0 : \mu_1 = \mu_2$$

ANOVA

Compare the means from *two or more* groups to see whether they are so far apart that the observed differences cannot all reasonably be attributed to sampling variability.

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$$

z/t test vs. ANOVA - Method

z/t test

Compute a test statistic (a ratio).

$$z/t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{SE(\bar{x}_1 - \bar{x}_2)}$$

ANOVA

Compute a test statistic (a ratio).

$$F = \frac{\text{variability bet. groups}}{\text{variability w/in groups}}$$

z/t test vs. ANOVA - Method

z/t test

Compute a test statistic (a ratio).

$$z/t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{SE(\bar{x}_1 - \bar{x}_2)}$$

ANOVA

Compute a test statistic (a ratio).

$$F = \frac{\text{variability bet. groups}}{\text{variability w/in groups}}$$

- Large test statistics lead to small p-values.
- If the p-value is small enough H_0 is rejected, we conclude that the population means are not equal.

z/t test vs. ANOVA

- With only two groups t-test and ANOVA are equivalent, but only if we use a pooled standard variance in the denominator of the test statistic.

z/t test vs. ANOVA

- With only two groups t-test and ANOVA are equivalent, but only if we use a pooled standard variance in the denominator of the test statistic.
- With more than two groups, ANOVA compares the sample means to an overall *grand mean*.

Hypotheses

What are the correct hypotheses for testing for a difference between the mean aldrin concentrations among the three levels?

(a) $H_0 : \mu_B = \mu_M = \mu_S$
$H_A : \mu_B \neq \mu_M \neq \mu_S$

(b) $H_0 : \mu_B \neq \mu_M \neq \mu_S$
$H_A : \mu_B = \mu_M = \mu_S$

(c) $H_0 : \mu_B = \mu_M = \mu_S$
H_A : At least one mean is different.

(d) $H_0 : \mu_B = \mu_M = \mu_S = 0$
H_A : At least one mean is different.

(e) $H_0 : \mu_B = \mu_M = \mu_S$
$H_A : \mu_B > \mu_M > \mu_S$

Hypotheses

What are the correct hypotheses for testing for a difference between the mean aldrin concentrations among the three levels?

(a) $H_0 : \mu_B = \mu_M = \mu_S$
$H_A : \mu_B \neq \mu_M \neq \mu_S$

(b) $H_0 : \mu_B \neq \mu_M \neq \mu_S$
$H_A : \mu_B = \mu_M = \mu_S$

(c) $H_0 : \mu_B = \mu_M = \mu_S$
H_A : *At least one mean is different.*

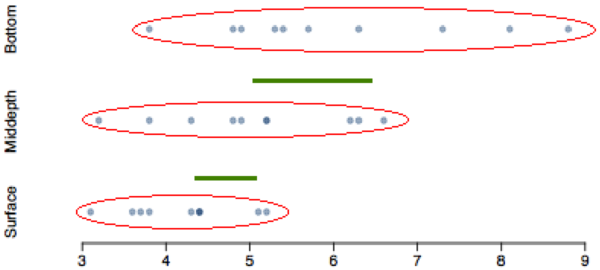
(d) $H_0 : \mu_B = \mu_M = \mu_S = 0$
H_A : At least one mean is different.

(e) $H_0 : \mu_B = \mu_M = \mu_S$
$H_A : \mu_B > \mu_M > \mu_S$

Test statistic

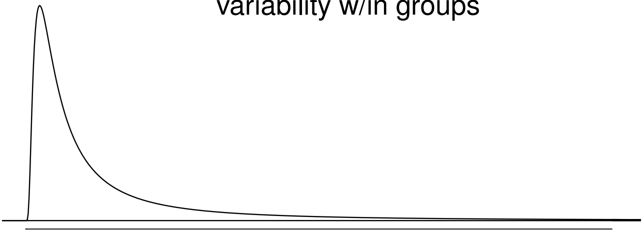
Does there appear to be a lot of variability within groups? How about between groups?

$$F = \frac{\text{variability bet. groups}}{\text{variability w/in groups}}$$



F distribution and p-value

$$F = \frac{\text{variability bet. groups}}{\text{variability w/in groups}}$$



- In order to be able to reject H_0, we need a small p-value, which requires a large F statistic.
- In order to obtain a large F statistic, variability between sample means needs to be greater than variability within sample means.

		Df	Sum Sq	Mean Sq	F value	Pr(>F)
(Group)	depth	2	16.96	8.48	6.13	0.0063
(Error)	Residuals	27	37.33	1.38		
	Total	29	54.29			

Degrees of freedom associated with ANOVA

- groups: $df_G = k - 1$, where k is the number of groups
- total: $df_T = n - 1$, where n is the total sample size
- error: $df_E = df_T - df_G$

		Df	Sum Sq	Mean Sq	F value	Pr(>F)
(Group)	depth	2	16.96	8.48	6.13	0.0063
(Error)	Residuals	27	37.33	1.38		
	Total	29	54.29			

Degrees of freedom associated with ANOVA

- groups: $df_G = k - 1$, where k is the number of groups
- total: $df_T = n - 1$, where n is the total sample size
- error: $df_E = df_T - df_G$

- $df_G = k - 1 = 3 - 1 = 2$

		Df	Sum Sq	Mean Sq	F value	Pr(>F)
(Group)	depth	2	16.96	8.48	6.13	0.0063
(Error)	Residuals	27	37.33	1.38		
	Total	29	54.29			

Degrees of freedom associated with ANOVA

- groups: $df_G = k - 1$, where k is the number of groups
- total: $df_T = n - 1$, where n is the total sample size
- error: $df_E = df_T - df_G$

- $df_G = k - 1 = 3 - 1 = 2$
- $df_T = n - 1 = 30 - 1 = 29$

		Df	Sum Sq	Mean Sq	F value	Pr(>F)
(Group)	depth	2	16.96	8.48	6.13	0.0063
(Error)	Residuals	27	37.33	1.38		
	Total	29	54.29			

Degrees of freedom associated with ANOVA

- groups: $df_G = k - 1$, where k is the number of groups
- total: $df_T = n - 1$, where n is the total sample size
- error: $df_E = df_T - df_G$

- $df_G = k - 1 = 3 - 1 = 2$
- $df_T = n - 1 = 30 - 1 = 29$
- $df_E = 29 - 2 = 27$

		Df	Sum Sq	Mean Sq	F value	Pr(>F)
(Group)	depth	2	*16.96*	8.48	6.13	0.0063
(Error)	Residuals	27	37.33	1.38		
	Total	29	54.29			

Sum of squares between groups, SSG

Measures the variability between groups

$$SSG = \sum_{i=1}^{k} n_i (\bar{x}_i - \bar{x})^2$$

where n_i is each group size, $\bar{x}_i$ is the average for each group, $\bar{x}$ is the overall (grand) mean.

		Df	Sum Sq	Mean Sq	F value	Pr(>F)
(Group)	depth	2	*16.96*	8.48	6.13	0.0063
(Error)	Residuals	27	37.33	1.38		
	Total	29	54.29			

Sum of squares between groups, SSG

Measures the variability between groups

$$SSG = \sum_{i=1}^{k} n_i(\bar{x}_i - \bar{x})^2$$

where n_i is each group size, $\bar{x}_i$ is the average for each group, $\bar{x}$ is the overall (grand) mean.

	n	mean
bottom	10	6.04
middepth	10	5.05
surface	10	4.2
overall	30	5.1

		Df	Sum Sq	Mean Sq	F value	Pr(>F)
(Group)	depth	2	*16.96*	8.48	6.13	0.0063
(Error)	Residuals	27	37.33	1.38		
	Total	29	54.29			

Sum of squares between groups, SSG

Measures the variability between groups

$$SSG = \sum_{i=1}^{k} n_i(\bar{x}_i - \bar{x})^2$$

where n_i is each group size, $\bar{x}_i$ is the average for each group, $\bar{x}$ is the overall (grand) mean.

	n	mean
bottom	10	6.04
middepth	10	5.05
surface	10	4.2
overall	30	5.1

$$SSG = \left(10 \times (6.04 - 5.1)^2\right)$$

		Df	Sum Sq	Mean Sq	F value	Pr(>F)
(Group)	depth	2	*16.96*	8.48	6.13	0.0063
(Error)	Residuals	27	37.33	1.38		
	Total	29	54.29			

Sum of squares between groups, SSG

Measures the variability between groups

$$SSG = \sum_{i=1}^{k} n_i (\bar{x}_i - \bar{x})^2$$

where n_i is each group size, $\bar{x}_i$ is the average for each group, $\bar{x}$ is the overall (grand) mean.

	n	mean
bottom	10	6.04
middepth	10	5.05
surface	10	4.2
overall	30	5.1

$$\begin{aligned} SSG &= \left(10 \times (6.04 - 5.1)^2\right) \\ &+ \left(10 \times (5.05 - 5.1)^2\right) \end{aligned}$$

		Df	Sum Sq	Mean Sq	F value	Pr(>F)
(Group)	depth	2	16.96	8.48	6.13	0.0063
(Error)	Residuals	27	37.33	1.38		
	Total	29	54.29			

Sum of squares between groups, SSG

Measures the variability between groups

$$SSG = \sum_{i=1}^{k} n_i (\bar{x}_i - \bar{x})^2$$

where n_i is each group size, $\bar{x}_i$ is the average for each group, $\bar{x}$ is the overall (grand) mean.

	n	mean
bottom	10	6.04
middepth	10	5.05
surface	10	4.2
overall	30	5.1

$$\begin{aligned} SSG &= \left(10 \times (6.04 - 5.1)^2\right) \\ &+ \left(10 \times (5.05 - 5.1)^2\right) \\ &+ \left(10 \times (4.2 - 5.1)^2\right) \end{aligned}$$

		Df	Sum Sq	Mean Sq	F value	Pr(>F)
(Group)	depth	2	*16.96*	8.48	6.13	0.0063
(Error)	Residuals	27	37.33	1.38		
	Total	29	54.29			

Sum of squares between groups, SSG

Measures the variability between groups

$$SSG = \sum_{i=1}^{k} n_i(\bar{x}_i - \bar{x})^2$$

where n_i is each group size, $\bar{x}_i$ is the average for each group, $\bar{x}$ is the overall (grand) mean.

	n	mean
bottom	10	6.04
middepth	10	5.05
surface	10	4.2
overall	30	5.1

$$\begin{aligned} SSG &= \left(10 \times (6.04 - 5.1)^2\right) \\ &+ \left(10 \times (5.05 - 5.1)^2\right) \\ &+ \left(10 \times (4.2 - 5.1)^2\right) \\ &= 16.96 \end{aligned}$$

		Df	Sum Sq	Mean Sq	F value	Pr(>F)
(Group)	depth	2	16.96	8.48	6.13	0.0063
(Error)	Residuals	27	37.33	1.38		
	Total	29	54.29			

Sum of squares total, SST

Measures the variability between groups

$$SST = \sum_{i=1}^{n} (x_i - \bar{x})$$

where x_i represent each observation in the dataset.

		Df	Sum Sq	Mean Sq	F value	Pr(>F)
(Group)	depth	2	16.96	8.48	6.13	0.0063
(Error)	Residuals	27	37.33	1.38		
	Total	29	54.29			

Sum of squares total, SST

Measures the variability between groups

$$SST = \sum_{i=1}^{n} (x_i - \bar{x})$$

where x_i represent each observation in the dataset.

$$SST = (3.8 - 5.1)^2 + (4.8 - 5.1)^2 + (4.9 - 5.1)^2 + \cdots + (5.2 - 5.1)^2$$

		Df	Sum Sq	Mean Sq	F value	Pr(>F)
(Group)	depth	2	16.96	8.48	6.13	0.0063
(Error)	Residuals	27	37.33	1.38		
	Total	29	*54.29*			

Sum of squares total, SST

Measures the variability between groups

$$SST = \sum_{i=1}^{n} (x_i - \bar{x})$$

where x_i represent each observation in the dataset.

$$\begin{aligned} SST &= (3.8 - 5.1)^2 + (4.8 - 5.1)^2 + (4.9 - 5.1)^2 + \cdots + (5.2 - 5.1)^2 \\ &= (-1.3)^2 + (-0.3)^2 + (-0.2)^2 + \cdots + (0.1)^2 \end{aligned}$$

		Df	Sum Sq	Mean Sq	F value	Pr(>F)
(Group)	depth	2	16.96	8.48	6.13	0.0063
(Error)	Residuals	27	37.33	1.38		
	Total	29	*54.29*			

Sum of squares total, SST

Measures the variability between groups

$$SST = \sum_{i=1}^{n} (x_i - \bar{x})$$

where x_i represent each observation in the dataset.

$$\begin{aligned} SST &= (3.8 - 5.1)^2 + (4.8 - 5.1)^2 + (4.9 - 5.1)^2 + \cdots + (5.2 - 5.1)^2 \\ &= (-1.3)^2 + (-0.3)^2 + (-0.2)^2 + \cdots + (0.1)^2 \\ &= 1.69 + 0.09 + 0.04 + \cdots + 0.01 \end{aligned}$$

		Df	Sum Sq	Mean Sq	F value	Pr(>F)
(Group)	depth	2	16.96	8.48	6.13	0.0063
(Error)	Residuals	27	37.33	1.38		
	Total	29	54.29			

Sum of squares total, SST

Measures the variability between groups

$$SST = \sum_{i=1}^{n} (x_i - \bar{x})$$

where x_i represent each observation in the dataset.

$$\begin{aligned} SST &= (3.8 - 5.1)^2 + (4.8 - 5.1)^2 + (4.9 - 5.1)^2 + \cdots + (5.2 - 5.1)^2 \\ &= (-1.3)^2 + (-0.3)^2 + (-0.2)^2 + \cdots + (0.1)^2 \\ &= 1.69 + 0.09 + 0.04 + \cdots + 0.01 \\ &= 54.29 \end{aligned}$$

		Df	Sum Sq	Mean Sq	F value	Pr(>F)
(Group)	depth	2	16.96	8.48	6.13	0.0063
(Error)	Residuals	27	*37.33*	1.38		
	Total	29	54.29			

Sum of squares error, SSE

Measures the variability within groups:

$$SSE = SST - SSG$$

		Df	Sum Sq	Mean Sq	F value	Pr(>F)
(Group)	depth	2	16.96	8.48	6.13	0.0063
(Error)	Residuals	27	*37.33*	1.38		
	Total	29	54.29			

Sum of squares error, SSE

Measures the variability within groups:

$$SSE = SST - SSG$$

$$SSE = 54.29 - 16.96 = 37.33$$

		Df	Sum Sq	Mean Sq	F value	Pr(>F)
(*G*roup)	depth	2	16.96	*8.48*	6.13	0.0063
(*E*rror)	Residuals	27	37.33	*1.38*		
	*T*otal	29	54.29			

Mean square error

Mean square error is calculated as sum of squares divided by the degrees of freedom.

		Df	Sum Sq	Mean Sq	F value	Pr(>F)
(Group)	depth	2	16.96	*8.48*	6.13	0.0063
(Error)	Residuals	27	37.33	*1.38*		
	Total	29	54.29			

Mean square error

Mean square error is calculated as sum of squares divided by the degrees of freedom.

$$MSG = 16.96/2 = 8.48$$

		Df	Sum Sq	Mean Sq	F value	Pr(>F)
(Group)	depth	2	16.96	*8.48*	6.13	0.0063
(Error)	Residuals	27	37.33	*1.38*		
	Total	29	54.29			

Mean square error

Mean square error is calculated as sum of squares divided by the degrees of freedom.

$$\begin{aligned} MSG &= 16.96/2 = 8.48 \\ MSE &= 37.33/27 = 1.38 \end{aligned}$$

		Df	Sum Sq	Mean Sq	F value	Pr(>F)
(Group)	depth	2	16.96	8.48	*6.14*	0.0063
(Error)	Residuals	27	37.33	1.38		
	Total	29	54.29			

Test statistic, F value

As we discussed before, the F statistic is the ratio of the between group and within group variability.

$$F = \frac{MSG}{MSE}$$

		Df	Sum Sq	Mean Sq	F value	Pr(>F)
(Group)	depth	2	16.96	8.48	*6.14*	0.0063
(Error)	Residuals	27	37.33	1.38		
	Total	29	54.29			

Test statistic, F value

As we discussed before, the F statistic is the ratio of the between group and within group variability.

$$F = \frac{MSG}{MSE}$$

$$F = \frac{8.48}{1.38} = 6.14$$

		Df	Sum Sq	Mean Sq	F value	Pr(>F)
(Group)	depth	2	16.96	8.48	*6.14*	0.0063
(Error)	Residuals	27	37.33	1.38		
	Total	29	54.29			

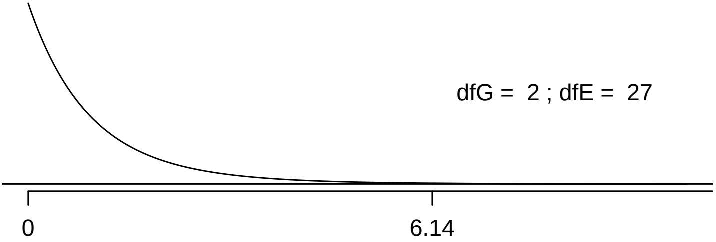
p-value

p-value is the probability of at least as large a ratio between the "between group" and "within group" variability, if in fact the means of all groups are equal. It's calculated as the area under the F curve, with degrees of freedom df_G and df_E, above the observed F statistic.

		Df	Sum Sq	Mean Sq	F value	Pr(>F)
(Group)	depth	2	16.96	8.48	*6.14*	0.0063
(Error)	Residuals	27	37.33	1.38		
	Total	29	54.29			

p-value

p-value is the probability of at least as large a ratio between the "between group" and "within group" variability, if in fact the means of all groups are equal. It's calculated as the area under the F curve, with degrees of freedom df_G and df_E, above the observed F statistic.



Conclusion - in context

What is the conclusion of the hypothesis test?

The data provide convincing evidence that the average aldrin concentration

(a) is different for all groups.

(b) on the surface is lower than the other levels.

(c) is different for at least one group.

(d) is the same for all groups.

Conclusion - in context

What is the conclusion of the hypothesis test?

The data provide convincing evidence that the average aldrin concentration

(a) is different for all groups.

(b) on the surface is lower than the other levels.

(c) *is different for at least one group.*

(d) is the same for all groups.

Conclusion

- If p-value is small (less than α), reject H_0. The data provide convincing evidence that at least one mean is different from (but we can't tell which one).

Conclusion

- If p-value is small (less than α), reject H_0. The data provide convincing evidence that at least one mean is different from (but we can't tell which one).
- If p-value is large, fail to reject H_0. The data do not provide convincing evidence that at least one pair of means are different from each other, the observed differences in sample means are attributable to sampling variability (or chance).

(1) independence

Does this condition appear to be satisfied?

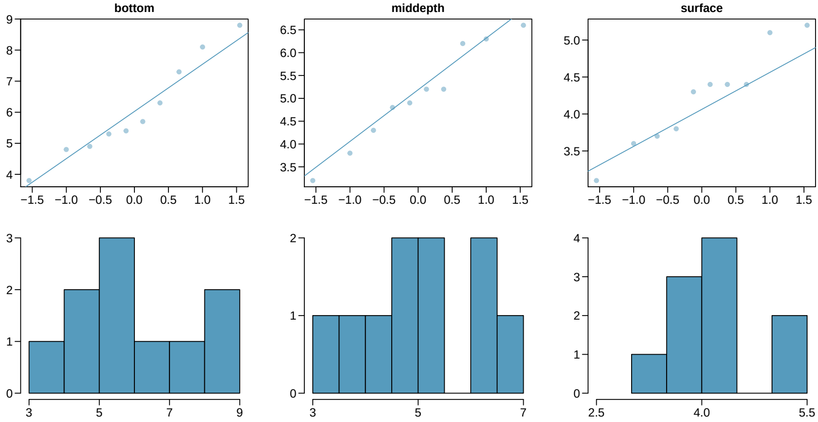
(1) independence

Does this condition appear to be satisfied?

In this study the we have no reason to believe that the aldrin concentration won't be independent of each other.

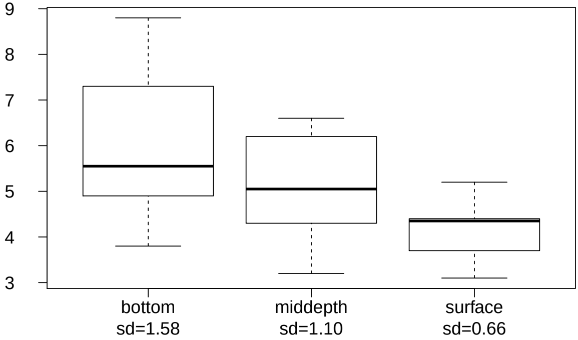
(2) approximately normal

Does this condition appear to be satisfied?



(3) constant variance

Does this condition appear to be satisfied?



Which means differ?

- Earlier we concluded that at least one pair of means differ. The natural question that follows is “which ones?”

Which means differ?

- Earlier we concluded that at least one pair of means differ. The natural question that follows is “which ones?”
- We can do two sample t tests for differences in each possible pair of groups.

Which means differ?

- Earlier we concluded that at least one pair of means differ. The natural question that follows is "which ones?"
- We can do two sample t tests for differences in each possible pair of groups.

Can you see any pitfalls with this approach?

Which means differ?

- Earlier we concluded that at least one pair of means differ. The natural question that follows is "which ones?"
- We can do two sample t tests for differences in each possible pair of groups.

Can you see any pitfalls with this approach?

- When we run too many tests, the Type 1 Error rate increases.
- This issue is resolved by using a modified significance level.

Multiple comparisons

- The scenario of testing many pairs of groups is called *multiple comparisons*.

Multiple comparisons

- The scenario of testing many pairs of groups is called *multiple comparisons*.
- The *Bonferroni correction* suggests that a more *stringent* significance level is more appropriate for these tests:

$$\alpha^\star = \alpha / K$$

  where K is the number of comparisons being considered.

Multiple comparisons

- The scenario of testing many pairs of groups is called *multiple comparisons*.
- The *Bonferroni correction* suggests that a more *stringent* significance level is more appropriate for these tests:

$$\alpha^\star = \alpha / K$$

  where K is the number of comparisons being considered.
- If there are k groups, then usually all possible pairs are compared and $K = \frac{k(k-1)}{2}$.

Determining the modified α

In the aldrin data set depth has 3 levels: bottom, mid-depth, and surface. If $\alpha = 0.05$, what should be the modified significance level for two sample t tests for determining which pairs of groups have significantly different means?

(a) $\alpha^* = 0.05$

(b) $\alpha^* = 0.05/2 = 0.025$

(c) $\alpha^* = 0.05/3 = 0.0167$

(d) $\alpha^* = 0.05/6 = 0.0083$

Determining the modified α

In the aldrin data set depth has 3 levels: bottom, mid-depth, and surface. If $\alpha = 0.05$, what should be the modified significance level for two sample t tests for determining which pairs of groups have significantly different means?

(a) $\alpha^* = 0.05$

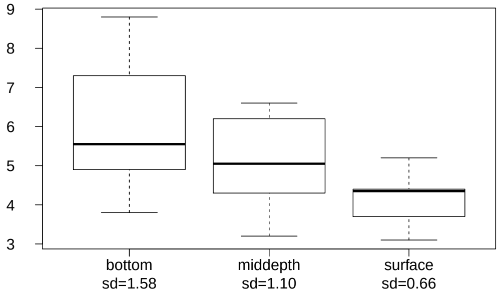
(b) $\alpha^* = 0.05/2 = 0.025$

(c) $\alpha^* = 0.05/3 = 0.0167$

(d) $\alpha^* = 0.05/6 = 0.0083$

Which means differ?

Based on the box plots below, which means would you expect to be significantly different?



(a) bottom & surface

(b) bottom & mid-depth

(c) mid-depth & surface

(d) bottom & mid-depth; mid-depth & surface

(e) bottom & mid-depth; bottom & surface; mid-depth & surface

Which means differ? (cont.)

If the ANOVA assumption of equal variability across groups is satisfied, we can use the data from all groups to estimate variability:

- Estimate any within-group standard deviation with $\sqrt{MSE}$, which is s_{pooled}
- Use the error degrees of freedom, $n - k$, for t-distributions

Difference in two means: after ANOVA

$$SE = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \approx \sqrt{\frac{MSE}{n_1} + \frac{MSE}{n_2}}$$

Is there a difference between the average aldrin concentration at the bottom and at mid depth?

	n	mean	sd
bottom	10	*6.04*	1.58
middepth	10	*5.05*	1.10
surface	10	4.2	0.66
overall	30	5.1	1.37

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
depth	2	16.96	8.48	6.13	0.0063
Residuals	*27*	37.33	*1.38*		
Total	29	54.29			

$$T_{df_E} = \frac{(\bar{x}_{bottom} - \bar{x}_{middepth})}{\sqrt{\frac{MSE}{n_{bottom}} + \frac{MSE}{n_{middepth}}}}$$

Is there a difference between the average aldrin concentration at the bottom and at mid depth?

	n	mean	sd
bottom	10	*6.04*	1.58
middepth	10	*5.05*	1.10
surface	10	4.2	0.66
overall	30	5.1	1.37

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
depth	2	16.96	8.48	6.13	0.0063
Residuals	*27*	37.33	*1.38*		
Total	29	54.29			

$$T_{df_E} = \frac{(\bar{x}_{bottom} - \bar{x}_{middepth})}{\sqrt{\frac{MSE}{n_{bottom}} + \frac{MSE}{n_{middepth}}}}$$

$$T_{27} = \frac{(6.04 - 5.05)}{\sqrt{\frac{1.38}{10} + \frac{1.38}{10}}} = \frac{0.99}{0.53} = 1.87$$

Is there a difference between the average aldrin concentration at the bottom and at mid depth?

	n	mean	sd
bottom	10	*6.04*	1.58
middepth	10	*5.05*	1.10
surface	10	4.2	0.66
overall	30	5.1	1.37

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
depth	2	16.96	8.48	6.13	0.0063
Residuals	*27*	37.33	*1.38*		
Total	29	54.29			

$$T_{df_E} = \frac{(\bar{x}_{bottom} - \bar{x}_{middepth})}{\sqrt{\frac{MSE}{n_{bottom}} + \frac{MSE}{n_{middepth}}}}$$

$$T_{27} = \frac{(6.04 - 5.05)}{\sqrt{\frac{1.38}{10} + \frac{1.38}{10}}} = \frac{0.99}{0.53} = 1.87$$

$$0.05 < p - value < 0.10$$ (two-sided)

Is there a difference between the average aldrin concentration at the bottom and at mid depth?

	n	mean	sd
bottom	10	*6.04*	1.58
middepth	10	*5.05*	1.10
surface	10	4.2	0.66
overall	30	5.1	1.37

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
depth	2	16.96	8.48	6.13	0.0063
Residuals	*27*	37.33	*1.38*		
Total	29	54.29			

$$T_{df_E} = \frac{(\bar{x}_{bottom} - \bar{x}_{middepth})}{\sqrt{\frac{MSE}{n_{bottom}} + \frac{MSE}{n_{middepth}}}}$$

$$T_{27} = \frac{(6.04 - 5.05)}{\sqrt{\frac{1.38}{10} + \frac{1.38}{10}}} = \frac{0.99}{0.53} = 1.87$$

$$0.05 < p - value < 0.10 \quad \text{(two-sided)}$$

$$\alpha^\star = 0.05/3 = 0.0167$$

Is there a difference between the average aldrin concentration at the bottom and at mid depth?

	n	mean	sd
bottom	10	*6.04*	1.58
middepth	10	*5.05*	1.10
surface	10	4.2	0.66
overall	30	5.1	1.37

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
depth	2	16.96	8.48	6.13	0.0063
Residuals	*27*	37.33	*1.38*		
Total	29	54.29			

$$T_{df_E} = \frac{(\bar{x}_{bottom} - \bar{x}_{middepth})}{\sqrt{\frac{MSE}{n_{bottom}} + \frac{MSE}{n_{middepth}}}}$$

$$T_{27} = \frac{(6.04 - 5.05)}{\sqrt{\frac{1.38}{10} + \frac{1.38}{10}}} = \frac{0.99}{0.53} = 1.87$$

$$0.05 < p - value < 0.10 \quad \text{(two-sided)}$$

$$\alpha^{\star} = 0.05/3 = 0.0167$$

Fail to reject H_0, the data do not provide convincing evidence of a difference between the average aldrin concentrations at bottom and mid depth.

Pairwise comparisons

Is there a difference between the average aldrin concentration at the bottom and at surface?

Pairwise comparisons

Is there a difference between the average aldrin concentration at the bottom and at surface?

$$T_{df_E} = \frac{(\bar{x}_{bottom} - \bar{x}_{surface})}{\sqrt{\frac{MSE}{n_{bottom}} + \frac{MSE}{n_{surface}}}}$$

Pairwise comparisons

Is there a difference between the average aldrin concentration at the bottom and at surface?

$$T_{df_E} = \frac{(\bar{x}_{bottom} - \bar{x}_{surface})}{\sqrt{\frac{MSE}{n_{bottom}} + \frac{MSE}{n_{surface}}}}$$

$$T_{27} = \frac{(6.04 - 4.02)}{\sqrt{\frac{1.38}{10} + \frac{1.38}{10}}} = \frac{2.02}{0.53} = 3.81$$

Pairwise comparisons

Is there a difference between the average aldrin concentration at the bottom and at surface?

$$
\begin{aligned}
T_{df_E} &= \frac{(\bar{x}_{bottom} - \bar{x}_{surface})}{\sqrt{\frac{MSE}{n_{bottom}} + \frac{MSE}{n_{surface}}}} \\
T_{27} &= \frac{(6.04 - 4.02)}{\sqrt{\frac{1.38}{10} + \frac{1.38}{10}}} = \frac{2.02}{0.53} = 3.81 \\
p - value &< 0.01 \qquad \textit{(two-sided)}
\end{aligned}
$$

Pairwise comparisons

Is there a difference between the average aldrin concentration at the bottom and at surface?

$$
\begin{aligned}
T_{df_E} &= \frac{(\bar{x}_{bottom} - \bar{x}_{surface})}{\sqrt{\frac{MSE}{n_{bottom}} + \frac{MSE}{n_{surface}}}} \\
T_{27} &= \frac{(6.04 - 4.02)}{\sqrt{\frac{1.38}{10} + \frac{1.38}{10}}} = \frac{2.02}{0.53} = 3.81 \\
p - value &< 0.01 \quad \textit{(two-sided)} \\
\alpha^\star &= 0.05/3 = 0.0167
\end{aligned}
$$

Pairwise comparisons

Is there a difference between the average aldrin concentration at the bottom and at surface?

$$
\begin{aligned}
T_{df_E} &= \frac{(\bar{x}_{bottom} - \bar{x}_{surface})}{\sqrt{\frac{MSE}{n_{bottom}} + \frac{MSE}{n_{surface}}}} \\
T_{27} &= \frac{(6.04 - 4.02)}{\sqrt{\frac{1.38}{10} + \frac{1.38}{10}}} = \frac{2.02}{0.53} = 3.81 \\
p - value &< 0.01 \qquad \textit{(two-sided)} \\
\alpha^{\star} &= 0.05/3 = 0.0167
\end{aligned}
$$

Reject H_0, the data provide convincing evidence of a difference between the average aldrin concentrations at bottom and surface.