Chapter 5: Inference for numerical data

Yingbo Li

Southern Methodist University

STAT 2331

Outline

Grade inflation in SMU?

In 2007 the average GPA of students at SMU was 3.13. From our class survey, a sample of 19 respondents yielded an average GPA of 3.39 with a standard deviation of 0.49. Assuming that this sample is random and representative of all students, do these data provide convincing evidence that the average GPA has changed over the last decade?

Question

1. What are the hypotheses?

(a) $H_0 : \bar{x} = 3.39$
$H_A : \bar{x} \neq 3.39$

(b) $H_0 : \mu = 3.13$
$H_A : \mu \neq 3.13$

(c) $H_0 : \mu = 3.13$
$H_A : \mu > 3.13$

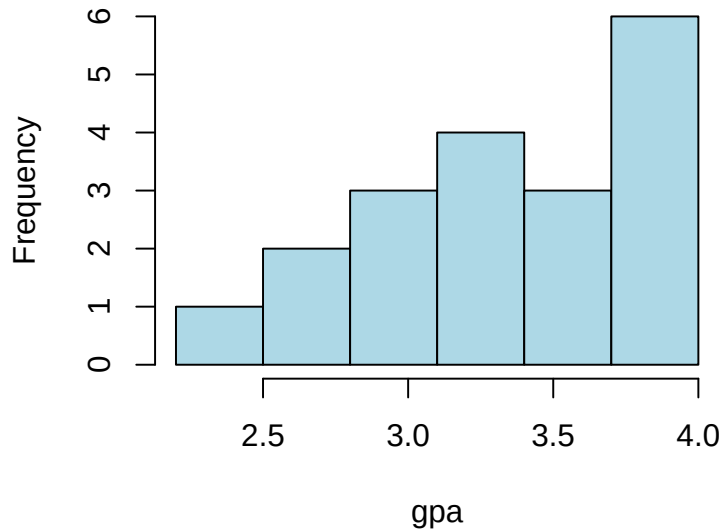
(d) $H_0 : \bar{x} = 3.13$
$H_A : \bar{x} > 3.13$

www.gradeinflation.com

Assumptions and conditions

- *Independence:*
  - We are told to assume that sample is random and representative of all students.
  - Sample size $n = 19 < 10\%$ of all SMU students.
- *Normality:*

- The sample distribution does not appear to be extremely skewed, so we can assume that the it doesn't come from an extremely skewed population.

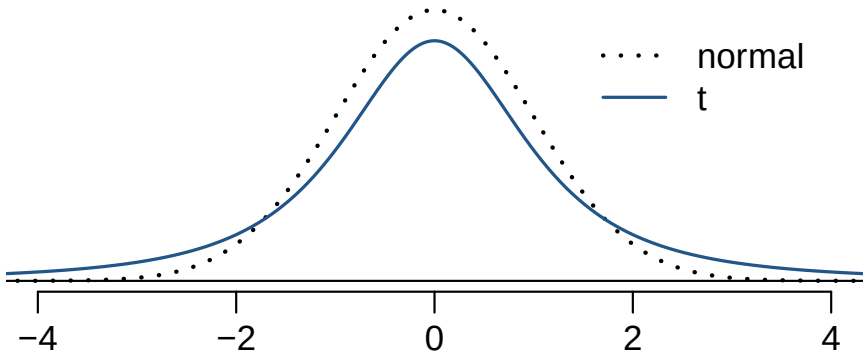


- But $n = 19 < 30!$

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

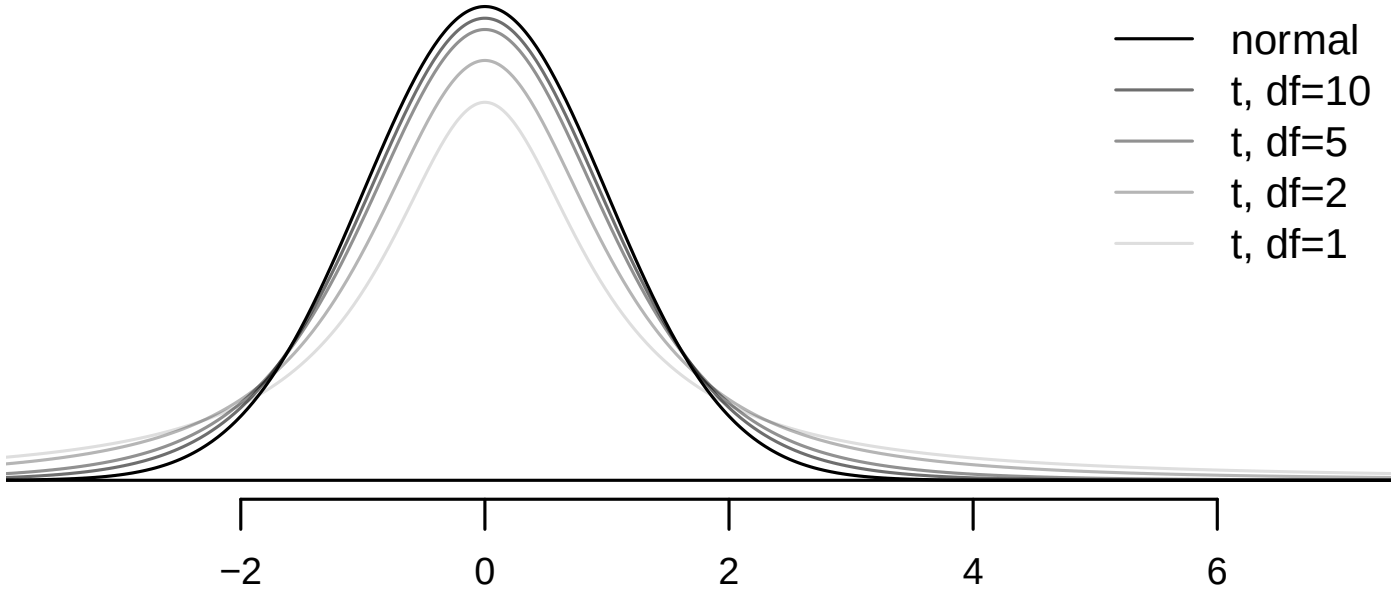
The t distribution

- When working with small sample means, and the population standard deviation σ 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 estimating the standard error.



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.

History: Student's t-distribution and t-test

William Sealy Gosset (1876-1937)

- worked for the Guinness brewery in Dublin, Ireland
- devised the t-test as an economical way to monitor the quality of stout
- Guinness forbade its chemists from publishing their findings
- published the t-test under the pseudonym of "Student", in 1908.

https://en.wikipedia.org/wiki/Student%27s_t-test

Finding the test statistic

Test statistic for inference on a small sample mean

The test statistic for inference on a small sample ($n < 30$) mean is the T statistic with $df = n - 1$.

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

in context...

$$\begin{aligned} point\ estimate &= \bar{x} = 3.39 \\ SE &= \frac{s}{\sqrt{n}} = \frac{0.49}{\sqrt{19}} = 0.112 \\ T &= \frac{3.39 - 3.13}{0.112} = 2.32 \end{aligned}$$

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

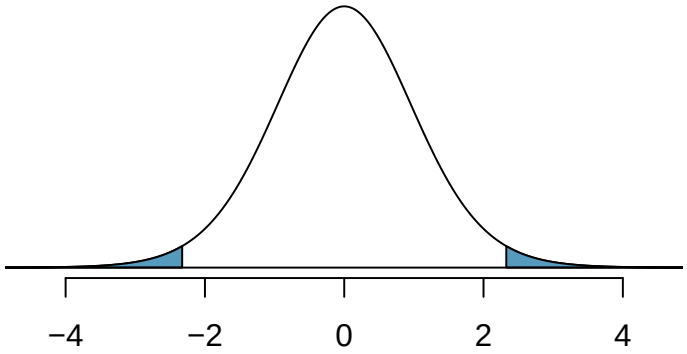
Finding the p-value using the t table

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 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



$T = 2.32$,
$0.02 <$ p-value < 0.05.

What is the conclusion of the hypothesis test?

The data provide convincing evidence to suggest the average gpa has changed over the last decade.

What if the T statistic is 3.00?
p-value < 0.01

Confidence interval for a small sample mean

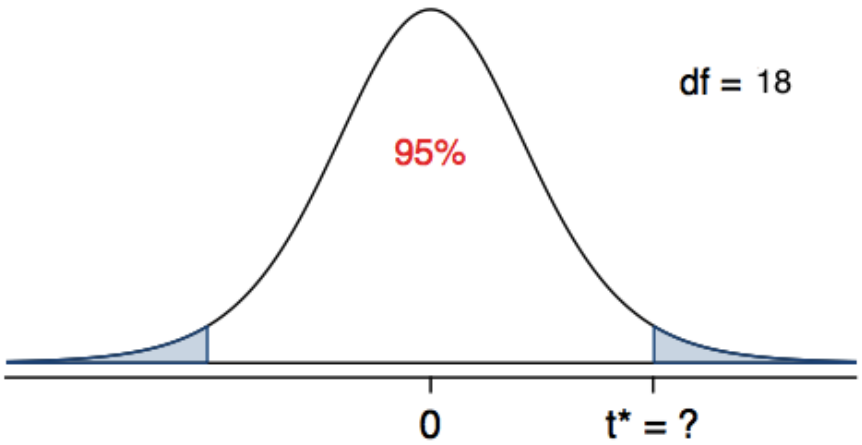
- Confidence intervals are always of the form

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

- 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 ($t^\star$)



$n = 19$, $df = 19 - 1 = 18$, $t^\star$ is at the intersection of row $df = 18$ 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	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

Constructing a CI for a small sample mean

Question

2. Which of the following is the correct calculation of a 95% confidence interval for the average gpa of SMU stutdents

$$\bar{x} = 3.39 \qquad s = 0.49 \qquad n = 19 \qquad SE = 0.112$$

(a) $3.39 \pm 1.96 \times 0.112$

(b) $3.39 \pm 2.10 \times 0.112$ $\rightarrow$ *(3.155, 3.625)*

(c) $3.39 \pm 1.96 \times 0.49$

(d) $3.39 \pm 2.10 \times 0.49$

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 3.13.

Recap: Inference using a small sample mean

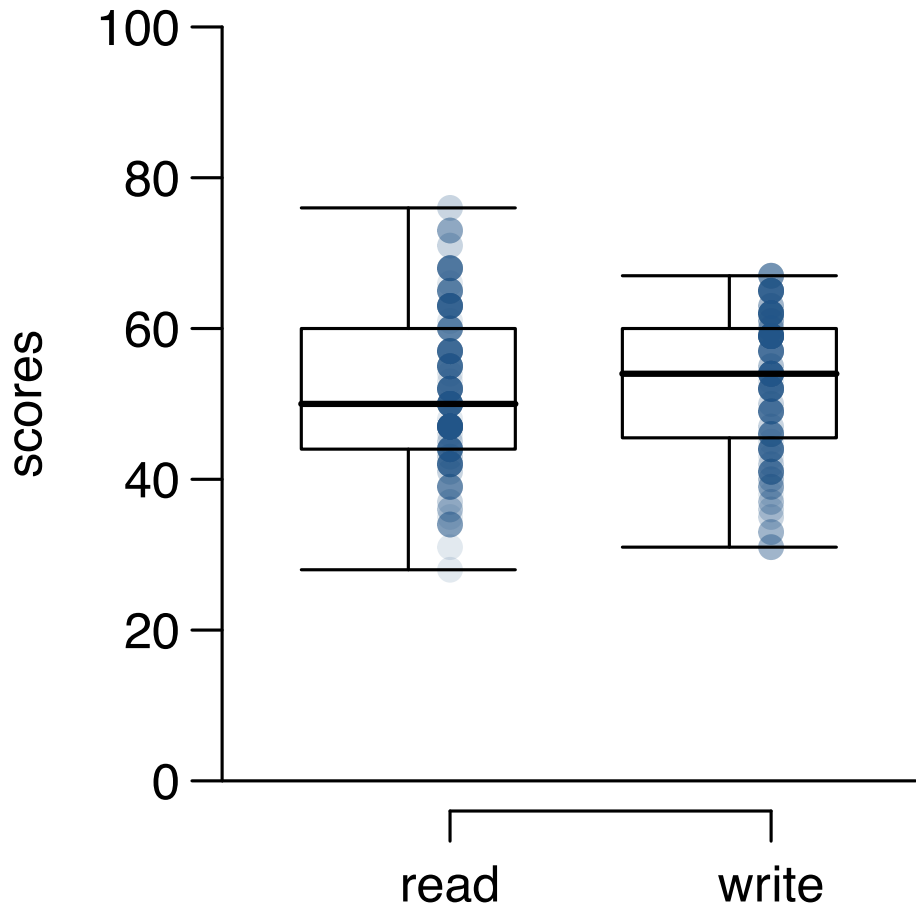
- If $n < 30$ and σ is unknown, sample means follow a t distribution with $SE = \frac{s}{\sqrt{n}}$.
- Assumptions and conditions:
  - independence of observations (often verified by a random sample, and if sampling without replacement, $n < 10\%$ of population)
  - no extreme skew
- Hypothesis testing:

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

- Confidence interval:

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

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?



Question

3. 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*

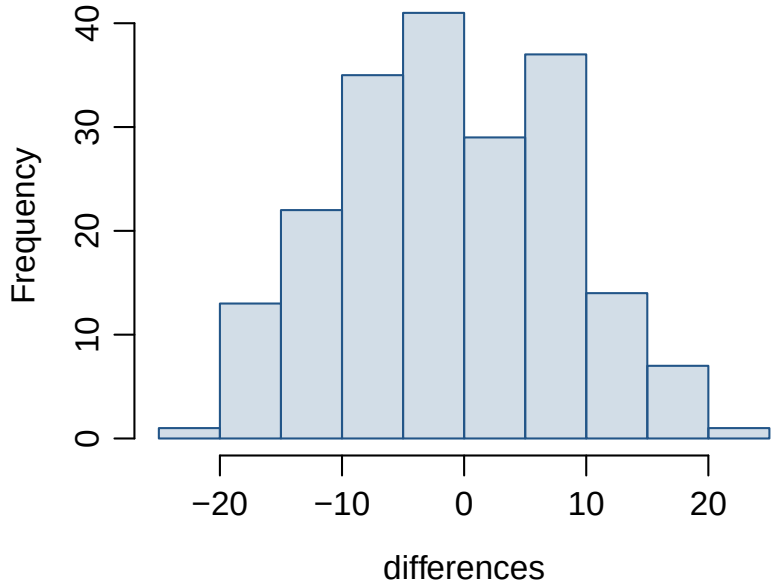
Analyzing paired data

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

$$\mu_{\text{diff}}$$

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

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

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?

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.

$$\mu_{\text{diff}} = 0$$

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

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

T-test for a population mean

- 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.
- Here we have a large sample $n = 200$. But since we use $SE \approx \frac{s}{\sqrt{n}}$, we can still use the t-distribution in the hypothesis test.

Note: To make statistical inference (hypothesis testing, confidence intervals) on one or two population means, we can always use the t-distribution regardless of the sample size.

Checking assumptions & conditions

Question

4. 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.*

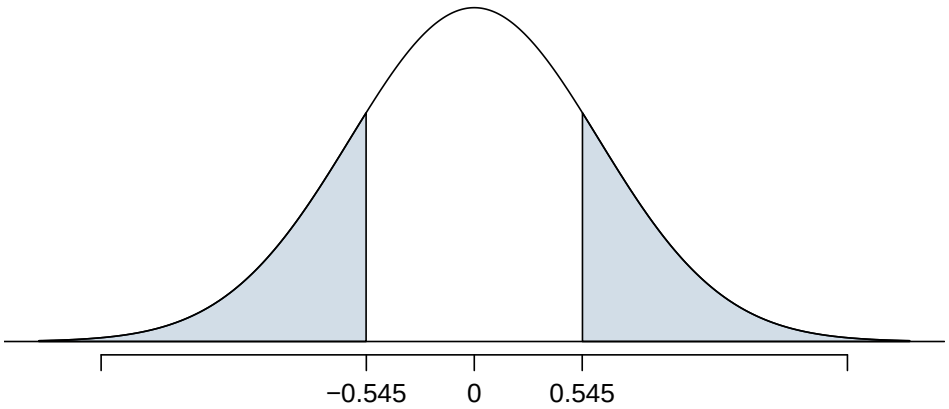
(b) The distribution of differences is bimodal, therefore we cannot continue with the hypothesis test.

(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% of 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?



$$df = n - 1 = 199$$

$$T_{df} = \frac{-0.545 - 0}{\frac{8.887}{\sqrt{200}}} = -0.87$$

p-value > 0.2

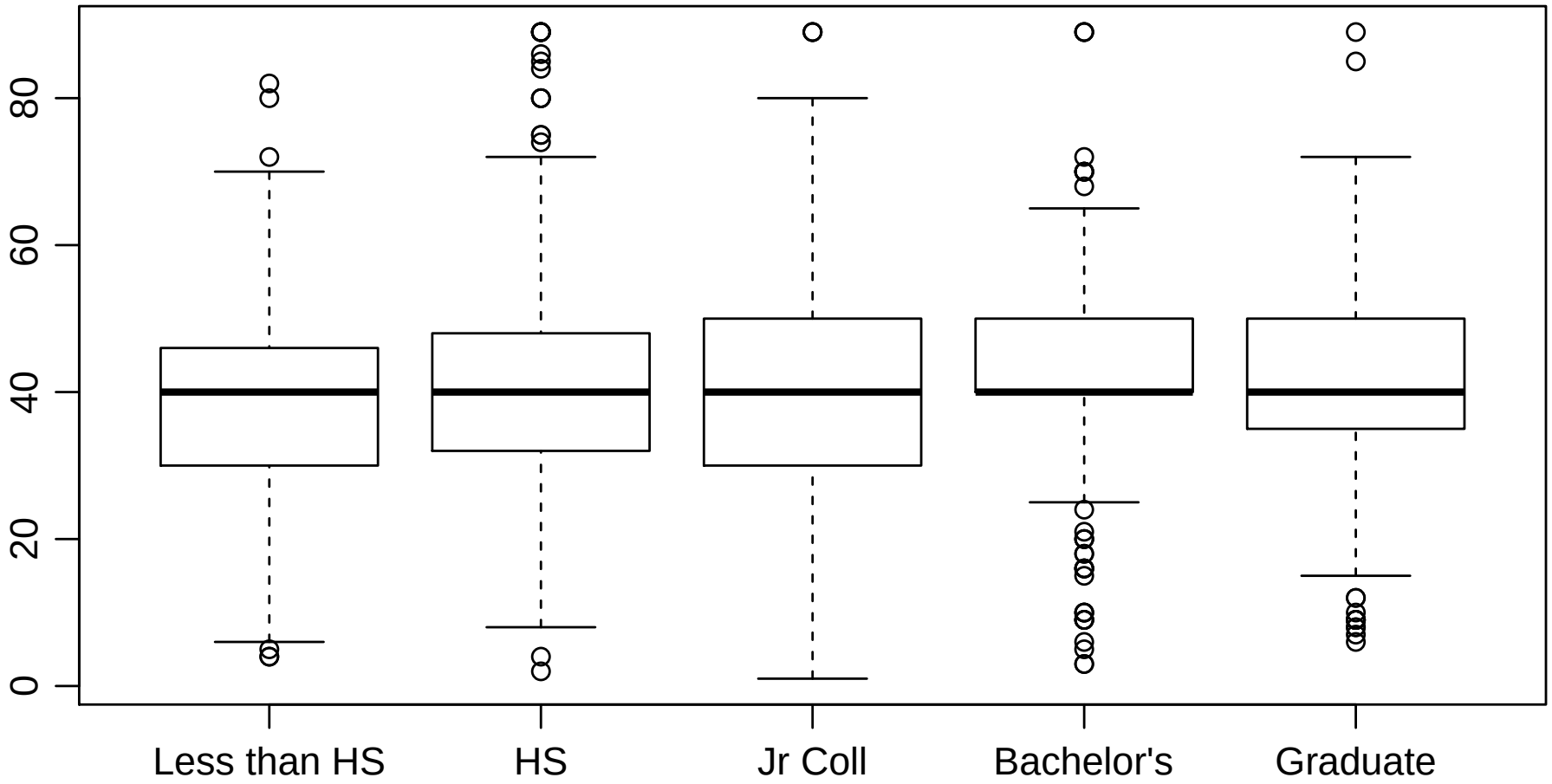
We do not reject H_0. The data do not provide convincing evidence that the reading and writing scores are on average different.

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?

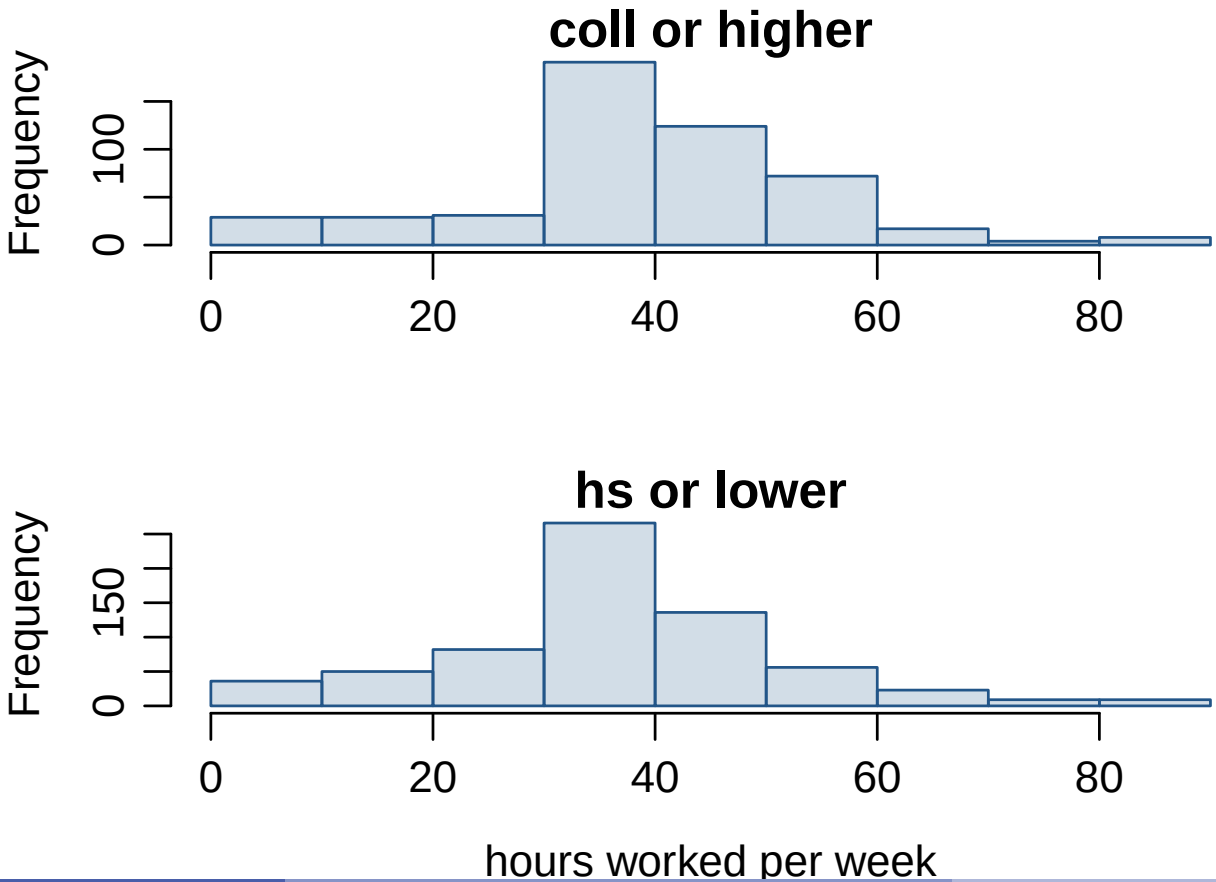


Collapsing levels into two

- Say we are only interested the difference between the number of hours worked per week by college and non-college graduates.
- Then we can collapse the levels of the categorical variable into two as:
  - `hs or lower` $\leftarrow$ less than high school or high school
  - `coll or higher` $\leftarrow$ 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 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.
   - $505 < 10\%$ 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. *Normality:*
   Both distributions look only mildly skewed, and the sample sizes are large, therefore we can assume that the sampling distribution of average number of hours worked per week by college graduates and that with HS degree or lower are nearly normal. Hence the sampling distribution of the average difference will be nearly normal as well.

Confidence interval for difference between two means

- All confidence intervals have the same form:

$$point\ estimate \pm ME$$

- And all $ME = critical\ value \times SE\ of\ point\ estimate$
- In this case the point estimate is $\bar{x}_1 - \bar{x}_2$
- Since we will use t-distribution for population mean(s) regardless of the sample size, the critical value is $t^\star$, with degrees of freedom

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

- 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

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

Confidence interval for the difference (cont.)

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

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

$$df = \min(505 - 1, 667 - 1) = 504$$

$$t^\star_{504} \approx t^\star_{500} = 1.96$$

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

Interpretation of a confidence interval for the difference

Question

5. Which of the following is the best interpretation of the confidence interval we just calculated? We are 95% confident that ...

(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.

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.

Test statistic

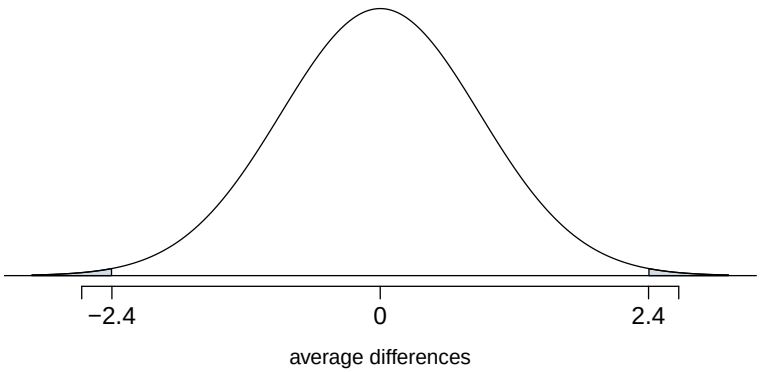
Test statistic for inference on the difference of two sample means

The test statistic for inference on the difference of two sample means is the T statistic.

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

where

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



$$\begin{aligned} T_{504} &= \frac{(\bar{x}_{coll} - \bar{x}_{hs}) - 0}{SE_{(\bar{x}_{coll} - \bar{x}_{hs})}} \\ &= \frac{2.4}{0.89} = 2.70 \\ \text{p-value} &< 0.01 \end{aligned}$$

Conclusion of the test

Question

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

(a) There is a less that 1% 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.

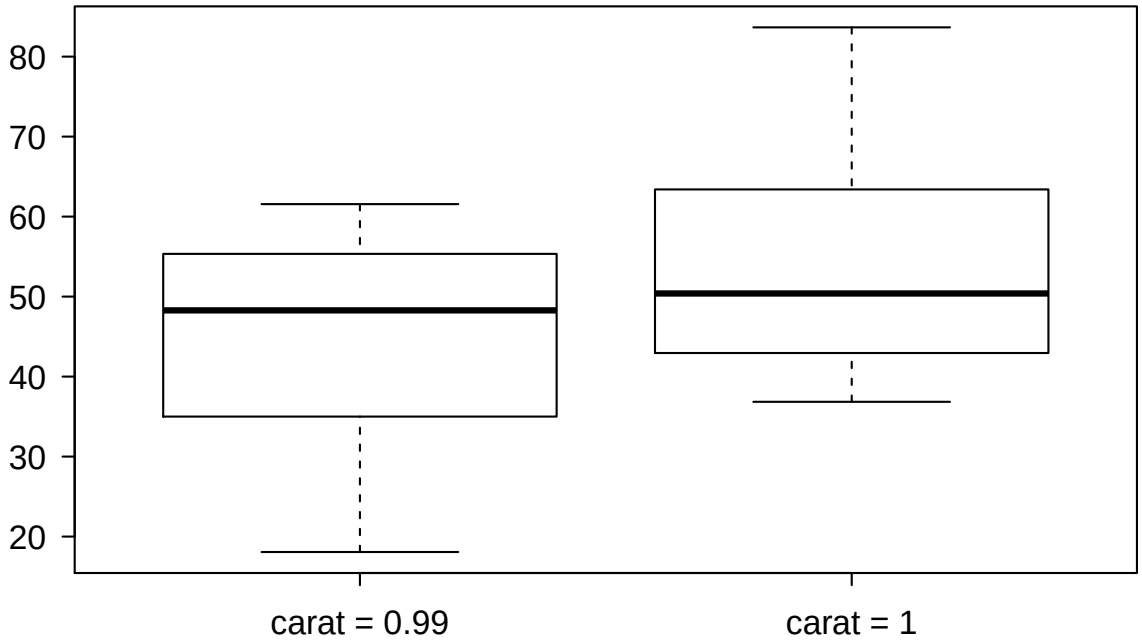
(b) *Since the p-value is low, we reject H_0. The data provide convincing evidence of a difference between the average number of hours worked per week by college graduates and those with a HS degree or lower.*

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

Diamonds

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

Data



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

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.

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

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

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

Hypotheses

Question

7. Which of the following is the correct set of hypotheses for testing if the average point price of 1 carat diamonds is higher than the average point price of 0.99 carat diamonds?

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

(b) $H_0 : \mu_{\text{diff}} = 0$
$H_A : \mu_{\text{diff}} < 0$

(c) $H_0 : \mu_{pt99} = \mu_{pt100}$
$H_A : \mu_{pt99} < \mu_{pt100}$

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

Assumptions & conditions

Question

8. 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 cars 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 (cont.)

	0.99 carat	*1 carat*
	pt99	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.)

Question

9. 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

Question

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

$$T = -2.508$$

(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

Conclusion

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

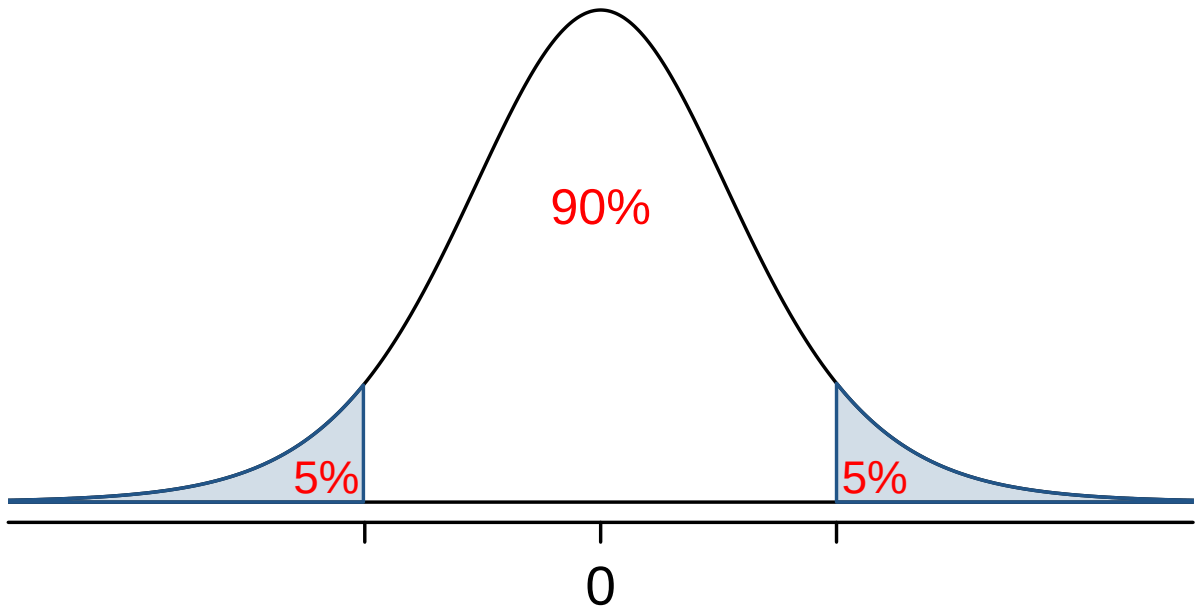
- *p-value is small so reject H_0. The data provide convincing evidence to suggest that the point price of 0.99 carat diamonds is lower than the point price of 1 carat diamonds.*
- *Maybe buy a 0.99 carat diamond? It looks like a 1 carat, but is much cheaper.*

Equivalent confidence level

Question

11. What is the equivalent confidence level (CL) 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?

$$CL = 90\%$$

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

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

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

Interpret this interval in context.

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 sample means

- Difference between the sample means follow a t distribution with $SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$.
- Assumptions and conditions:
  - independence within groups (often verified by a random sample, and if sampling without replacement, $n < 10\%$ of population)
  - independence between groups
  - 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:

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