Hypothesis Testing for Numerical Variable

Portland State University
USP 634 Data Analysis I
Spring 2018

Hypothesis Testing

Remember when...

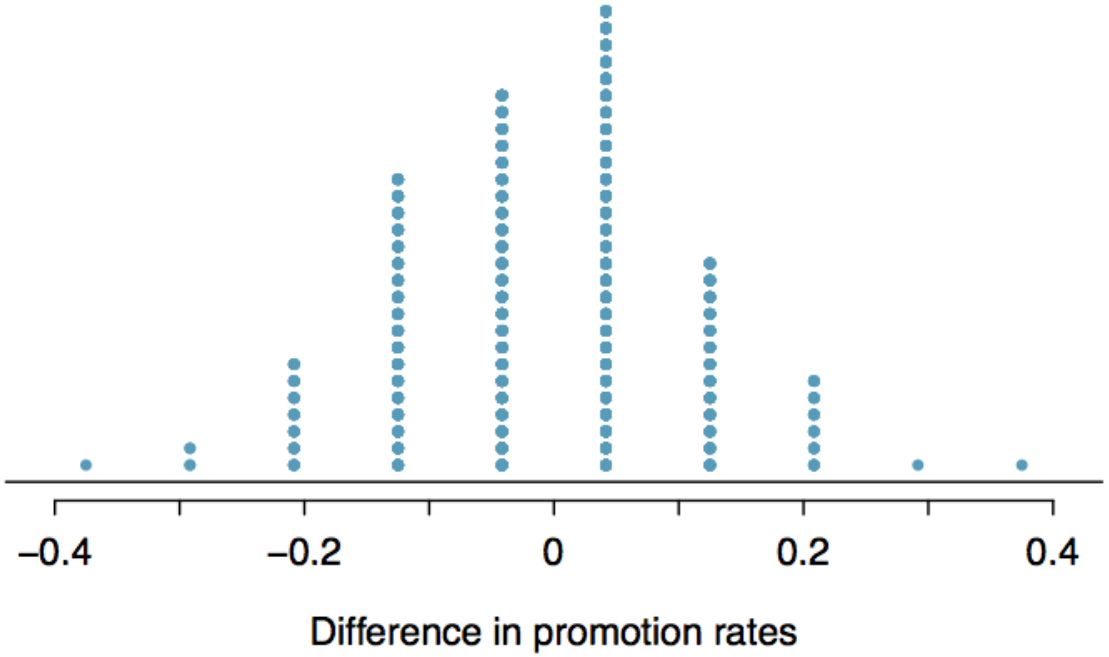
		Promotion		
		Promoted	Not Promoted	Total
Gender	Male	21	3	24
	Female	14	10	24
	Total	35	13	48

P(promotion | males) = 21 / 24 = 0.88
P(promotion | females) = 14 / 24 = 0.58

Possible explanations:

- Promotion and gender are independent, no gender discrimination, observed difference in proportions is simply due to chance.
  → null (nothing is going on)
- Promotion and gender are dependent, there is gender discrimination, observed difference in proportions is not due to chance.
  → alternative (something is going on)

Result



Since it was quite unlikely to obtain results like the actual data or something more extreme in the simulations (male promotions being 30% or more higher than female promotions), we decided to reject the null hypothesis in favor of the alternative.

Recap: hypothesis testing framework

We start with a null hypothesis (H_0) that represents the status quo.

We also have an alternative hypothesis (H_A) that represents our research question, i.e. what we're testing for.

We conduct a hypothesis test under the assumption that the null hypothesis is true, either via simulation or traditional methods based on the central limit theorem.

If the test results suggest that the data do not provide convincing evidence for the alternative hypothesis, we stick with the null hypothesis. If they do, then we reject the null hypothesis in favor of the alternative.

We'll formally introduce the hypothesis testing framework using an example on testing a claim about a population mean.

Average number of exclusive relationships

A random sample of 50 college students were asked how many exclusive relationships they have been in so far. This sample yielded a mean of 3.2 and a standard deviation of 1.74. Estimate the true average number of exclusive relationships using this sample.

$\bar{x} = 3.2$ $s = 1.74$

The approximate 95% confidence interval is defined as

point estimate ± 2 x SE

$SE = s / \sqrt{n} = 1.74 / \sqrt{50} \approx 0.25$

$\bar{x} \pm 2 \times SE \rightarrow 3.2 \pm 2 \times 0.25$

$\rightarrow (3.2 - 0.5, 3.2 + 0.5)$

$\rightarrow (2.7, 3.7)$

Testing hypotheses using confidence intervals

Earlier we calculated a 95% confidence interval for the average number of exclusive relationships college students have been in to be (2.7, 3.7). Based on this confidence interval, do these data support the hypothesis that college students on average have been in more than 3 exclusive relationships.

The associated hypotheses are:
H_0: $\mu = 3$: College students have been in 3 exclusive relationships, on average
H_A: $\mu > 3$: College students have been in more than 3 exclusive relationships, on average

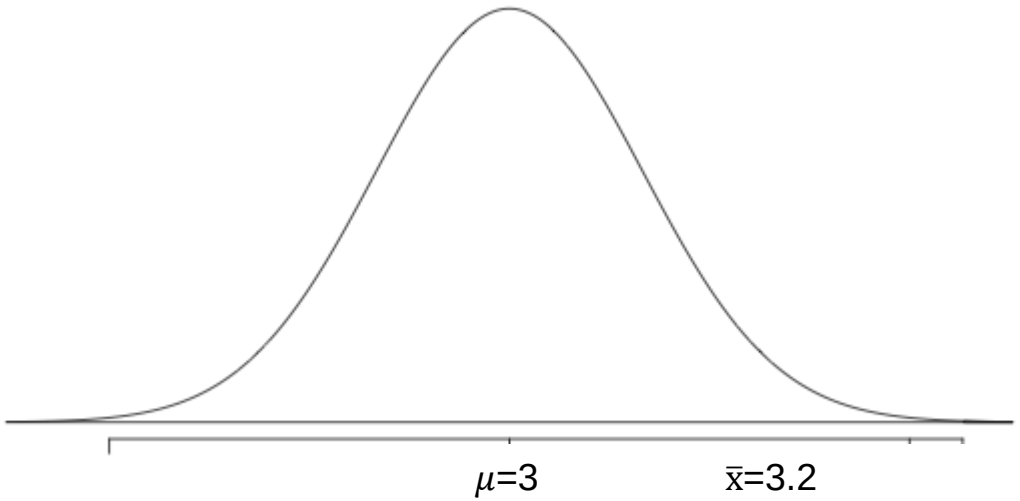
- Since the null value is included in the interval, we do not reject the null hypothesis in favor of the alternative.
- This is a quick-and-dirty approach for hypothesis testing. However it doesn't tell us the likelihood of certain outcomes under the null hypothesis, i.e. the p-value, based on which we can make a decision on the hypotheses.

p-values

- We then use this test statistic to calculate the p-value, the probability of observing data at least as favorable to the alternative hypothesis as our current data set, if the null hypothesis were true.
- If the p-value is low (lower than the significance level, α, which is usually 5%) we say that it would be very unlikely to observe the data if the null hypothesis were true, and hence reject H_0.
- If the p-value is high (higher than α) we say that it is likely to observe the data even if the null hypothesis were true, and hence do not reject H_0.

p-values

p-value: probability of observing data at least as favorable to H_A as our current data set (a sample mean greater than 3.2), if in fact H_0 were true (the true population mean was 3).



$$P(\bar{x} > 3.2 \mid \mu = 3) = P(Z > .8) = 0.22$$

Number of exclusive relationship - Making a decision

p-value = 0.22

- If the true average of the number of exclusive relationship is 3, there is only 22% chance of observing a random sample of 50 college students who have on average 3.2 or more exclusive relationship.
- This is a high probability for us to think that a sample mean of 3.2 or more exclusive relationship is likely to happen simply by chance.

Since p-value is ***high***(higher than 5%) we fail to reject H_0.

The data do not provide sufficient evidence to reject the claim that that college students have 3 exclusive relationship on average.

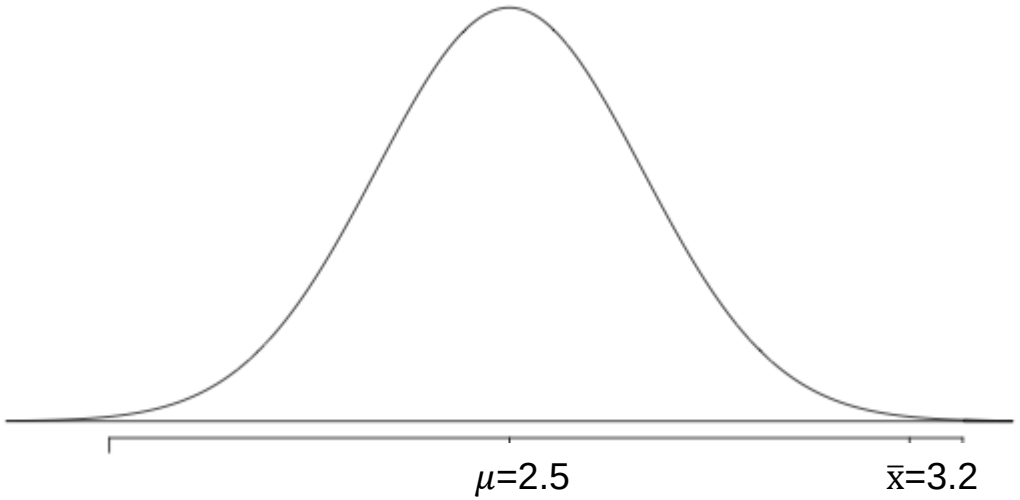
The difference between 3 exclusive relationship and observed sample mean of 3.2 may be due to chance or sampling variability.

Testing hypotheses

If instead the hypotheses are:

H_0: $\mu = 2.5$: College students have been in 2.5 exclusive relationships, on average

H_A: $\mu > 2.5$: College students have been in more than 2.5 exclusive relationships, on average



$$P(\bar{x} > 3.2 \mid \mu = 2.5) = P(Z > 2.8) = 0.0026$$

Number of exclusive relationship - Making a decision

p-value = 0.0026

If the true average of the number of exclusive relationship is 2.5, there is only 0.26% chance of observing a random sample of 50 college students who have on average 3.2 or more exclusive relationship.

- This is a very low probability for us to think that a sample mean of 3.2 or more exclusive relationship is likely to happen simply by chance.

Since p-value is *low*(lower than 5%) we reject H_0.

The data provide convincing evidence that college students have more than 2.5 exclusive relationship on average.

The difference between the 2.5 exclusive relationship and observed sample mean of 3.2 is not due to chance or sampling variability.

Practice

A poll by the National Sleep Foundation found that college students average about 7 hours of sleep per night. A sample of 169 college students taking an introductory statistics class yielded an average of 6.88 hours, with a standard deviation of 0.94 hours. Assuming that this is a random sample representative of all college students *(bit of a leap of faith?)*, a hypothesis test was conducted to evaluate if college students on average sleep *less than* 7 hours per night. The p-value for this hypothesis test is 0.0485. Which of the following is correct?

a) Fail to reject H_0, the data provide convincing evidence that college students sleep less than 7 hours on average.

b) Reject H_0, the data provide convincing evidence that college students sleep less than 7 hours on average.

c) Reject H_0, the data prove that college students sleep more than 7 hours on average.

d) Fail to reject H_0, the data do not provide convincing evidence that college students sleep less than 7 hours on average.

e) Reject H_0, the data provide convincing evidence that college students in this sample sleep less than 7 hours on average.

Practice

A poll by the National Sleep Foundation found that college students average about 7 hours of sleep per night. A sample of 169 college students taking an introductory statistics class yielded an average of 6.88 hours, with a standard deviation of 0.94 hours. Assuming that this is a random sample representative of all college students *(bit of a leap of faith?)*, a hypothesis test was conducted to evaluate if college students on average sleep *less than* 7 hours per night. The p-value for this hypothesis test is 0.0485. Which of the following is correct?

a) Fail to reject H_0, the data provide convincing evidence that college students sleep less than 7 hours on average.

b) Reject H_0, the data provide convincing evidence that college students sleep less than 7 hours on average.

c) Reject H_0, the data prove that college students sleep more than 7 hours on average.

d) Fail to reject H_0, the data do not provide convincing evidence that college students sleep less than 7 hours on average.

e) Reject H_0, the data provide convincing evidence that college students in this sample sleep less than 7 hours on average.

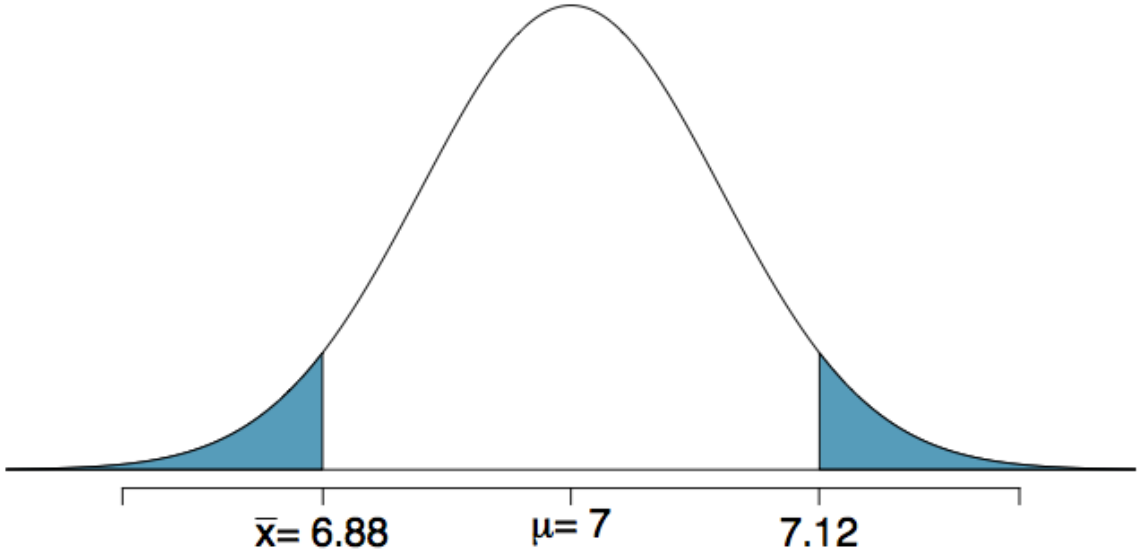
Two-sided hypothesis testing with p-values

If the research question was “Do the data provide convincing evidence that the average amount of sleep college students get per night is *different* than the national average?”, the alternative hypothesis would be different.

H_0: $\mu = 7$

H_A: $\mu \neq 7$

Hence the p-value would change as well:



p-value
$= 0.0485 \times 2$
$= 0.097$

Decision errors

Hypothesis tests are not flawless.

- In the court system innocent people are sometimes wrongly convicted, and the guilty sometimes walk free.
- Similarly, we can make a wrong decision in statistical hypothesis tests as well.
- The difference is that we have the tools necessary to quantify how often we make errors in statistics (*when the conditions for hypothesis testing are met*).

Decision errors (cont.)

There are two competing hypotheses: the null and the alternative. In a hypothesis test, we make a decision about which might be true, but our choice might be incorrect.

		Decision	
		fail to reject H_0	reject H_0
Truth	H_0 true	✓	
	H_A true		✓

Decision errors (cont.)

There are two competing hypotheses: the null and the alternative. In a hypothesis test, we make a decision about which might be true, but our choice might be incorrect.

		Decision	
		fail to reject H_0	reject H_0
Truth	H_0 true	✓	*Type 1 Error*
	H_A true		✓

A Type 1 Error is rejecting the null hypothesis when H_0 is true.

Decision errors (cont.)

There are two competing hypotheses: the null and the alternative. In a hypothesis test, we make a decision about which might be true, but our choice might be incorrect.

		Decision	
		fail to reject H_0	reject H_0
Truth	H_0 true	✓	*Type 1 Error*
	H_A true	*Type 2 Error*	✓

A Type 1 Error is rejecting the null hypothesis when H_0 is true.

A Type 2 Error is failing to reject the null hypothesis when H_A is true.

We (almost) never know if H_0 or H_A is true, but we need to consider all possibilities.

Hypothesis Test as a trial

If we again think of a hypothesis test as a criminal trial then it makes sense to frame the verdict in terms of the null and alternative hypotheses:

H_0: Defendant is innocent

H_A: Defendant is guilty

Which type of error is being committed in the following circumstances?

Declaring the defendant innocent when they are actually guilty

Declaring the defendant guilty when they are actually innocent

Hypothesis Test as a trial

If we again think of a hypothesis test as a criminal trial then it makes sense to frame the verdict in terms of the null and alternative hypotheses:

H_0: Defendant is innocent

H_A: Defendant is guilty

Which type of error is being committed in the following circumstances?

Declaring the defendant innocent when they are actually guilty

Type 2 error

Declaring the defendant guilty when they are actually innocent

Hypothesis Test as a trial

If we again think of a hypothesis test as a criminal trial then it makes sense to frame the verdict in terms of the null and alternative hypotheses:

H_0: Defendant is innocent

H_A: Defendant is guilty

Which type of error is being committed in the following circumstances?

Declaring the defendant innocent when they are actually guilty

Type 2 error

Declaring the defendant guilty when they are actually innocent

Type 1 error

Which error do you think is the worse error to make?

"better that ten guilty persons escape than that one innocent suffer"
- William Blackstone

Type 1 error rate

As a general rule we reject H_0 when the p-value is less than 0.05, i.e. we use a significance level of 0.05, $\alpha = 0.05$.

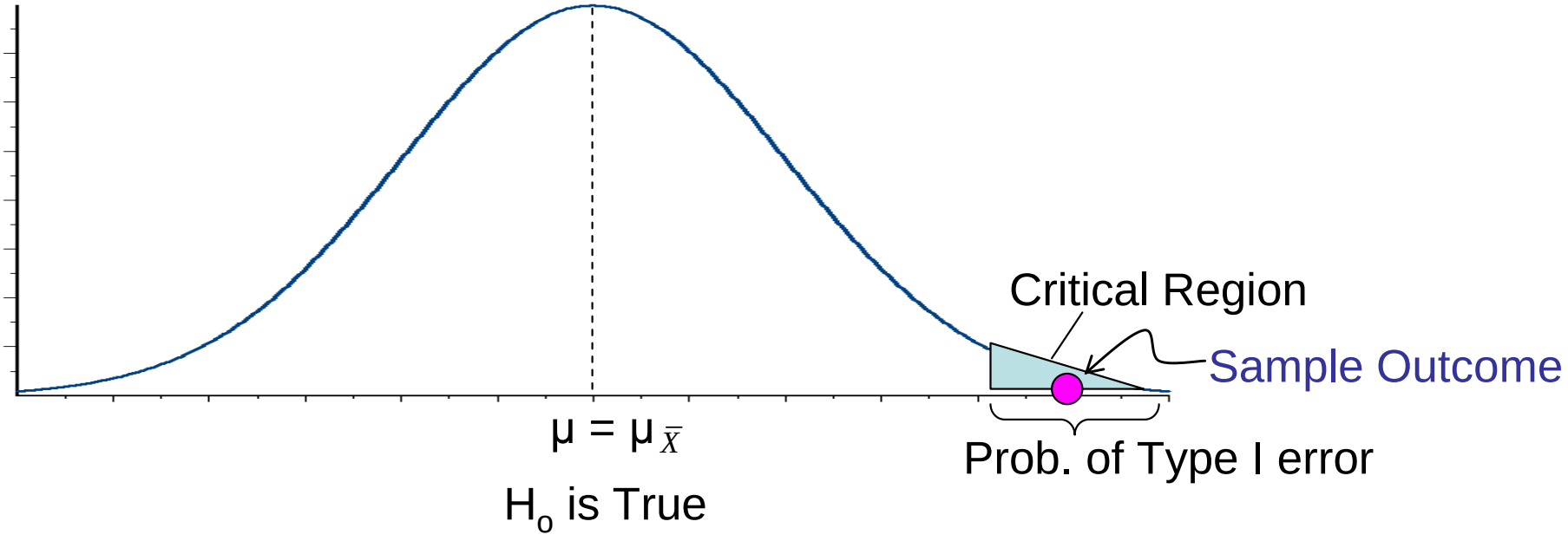
This means that, for those cases where H_0 is actually true, we do not want to incorrectly reject it more than 5% of those times.

In other words, when using a 5% significance level there is about 5% chance of making a Type 1 error if the null hypothesis is true.

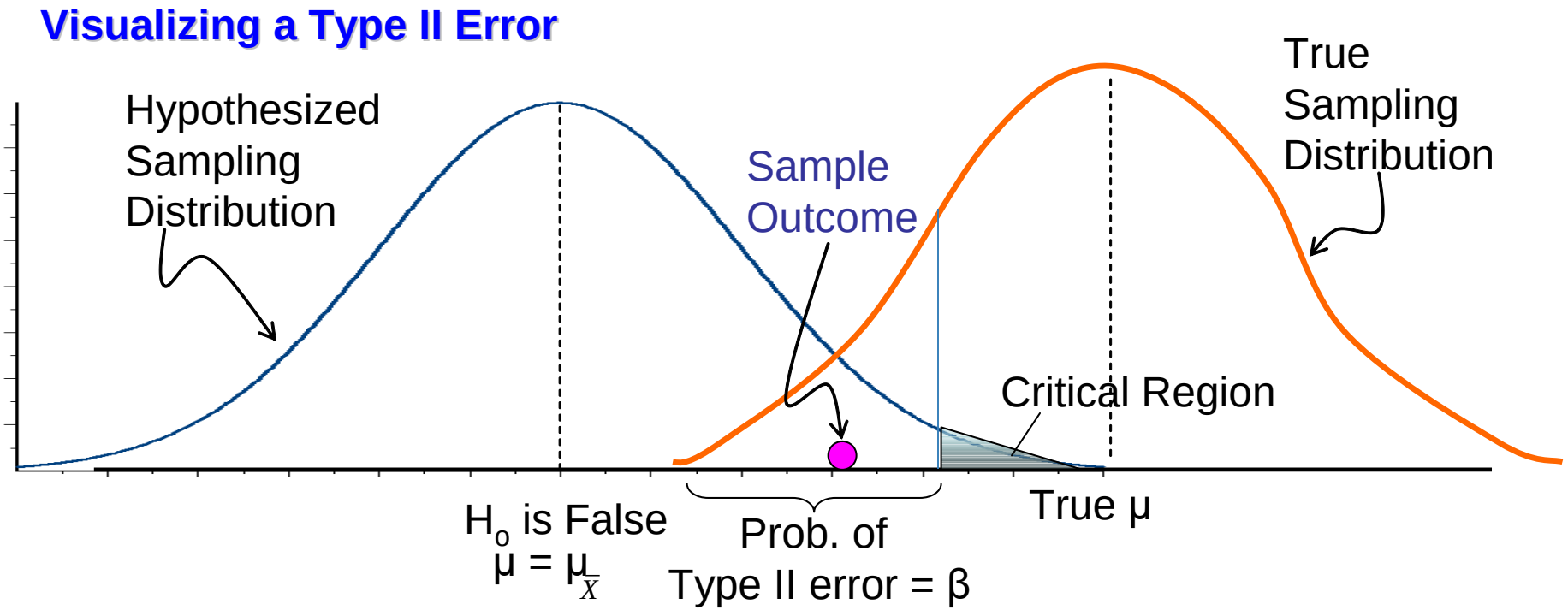
$$P(\text{Type 1 error} \mid H_0 \text{ true}) = \alpha$$

This is why we prefer small values of α -- increasing α increases the Type 1 error rate.

Visualizing a Type I error



Visualizing a Type II error



Choosing a significance level

Choosing a significance level for a test is important in many contexts, and the traditional level is 0.05. However, it is often helpful to adjust the significance level based on the application.

We may select a level that is smaller or larger than 0.05 depending on the consequences of any conclusions reached from the test.

If making a Type 1 Error is dangerous or especially costly, we should choose a small significance level (e.g. 0.01). Under this scenario we want to be very cautious about rejecting the null hypothesis, so we demand very strong evidence favoring H_A before we would reject H_0.

If a Type 2 Error is relatively more dangerous or much more costly than a Type 1 Error, then we should choose a higher significance level (e.g. 0.10). Here we want to be cautious about failing to reject H_0 when the null is actually false.

Recap: Hypothesis testing framework

1. Set the hypotheses.
2. Calculate the point estimate
3. Check assumptions and conditions.
4. Calculate a test statistic and a p-value.
5. Make a decision, and interpret it in context of the research question.

Recap: Hypothesis testing for a population mean

1. Set the hypotheses
   H_0: μ = null value
   H_A: μ < or > or ≠ null value

2. Calculate the point estimate

3. Check assumptions and conditions

- Independence: random sample/assignment, 10\% condition when sampling without replacement
- Normality: nearly normal population or n ≥ 30, no extreme skew -- or use the t distribution (Ch 5)

4. Calculate a test statistic and a p-value (draw a picture!)

$$Z = \frac{\bar{x} - \mu}{SE}, \text{ where } SE = \frac{s}{\sqrt{n}}$$

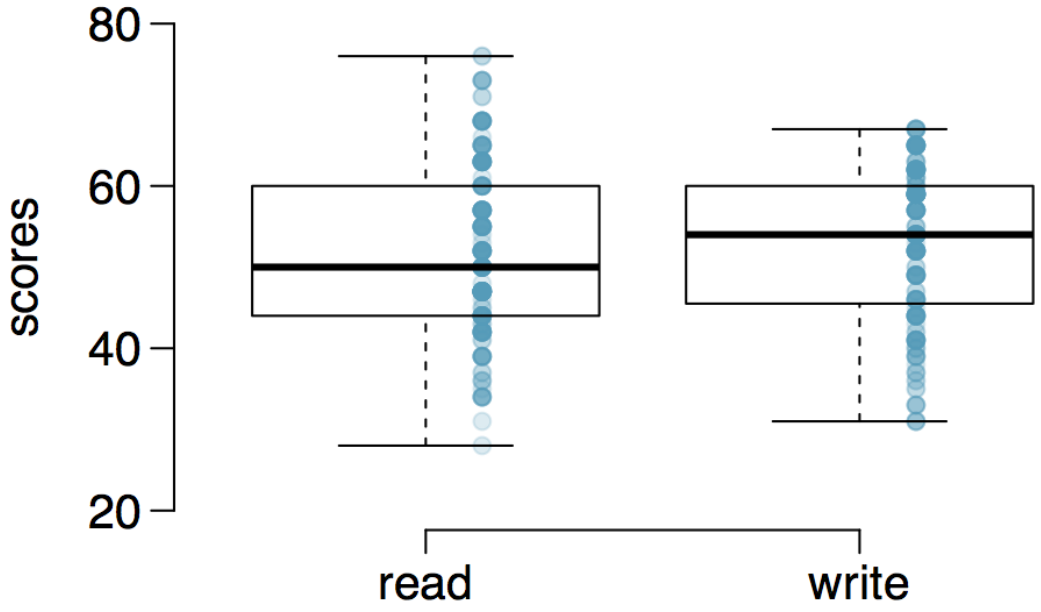
5. Make a decision, and interpret it in context

- If p-value < α, reject H_0, data provide evidence for H_A
- If p-value > α, do not reject H_0, data do not provide evidence for H_A

Paired Data

Paired observations

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?



Paired observations

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

Paired observations

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

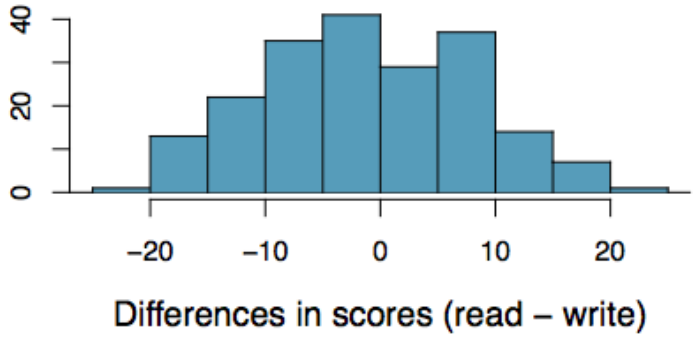
When two sets of observations have this special correspondence (not independent), they are said to be paired.

To analyze paired data, it is often useful to look at the difference in outcomes of each pair of observations.

diff = read - 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.

$$\mu_{diff}$$

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?

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.

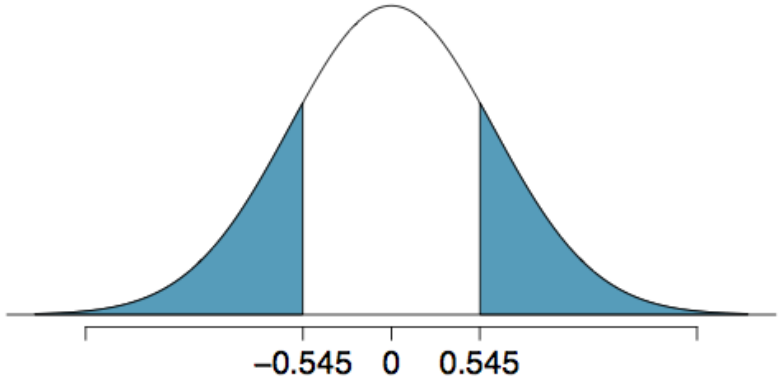
Checking assumptions & conditions

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 $\alpha = 0.05$.



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

$$= \frac{-0.545}{0.628} = -0.87$$

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

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 ↔ 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 ↔ 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

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

Krizek, K., 2007. Pretest-Posttest Strategy for Researching Neighborhood-Scale Urban Form and Travel Behavior. Transportation Research Record 1722, pp 48-55.

TABLE 3 Transitions for Mean Differences (Standard Deviations) and *p*-Values for Paired-Sample *t*-Tests by Treatment Case

TREATMENT CASES	Δ TRIP DIST (vkt)	Δ TRIP MIN	Δ TOUR DIST (vkt)	Δ TOUR MIN	ΔTRIPS per TOUR	Δ % TRIPS Alt Mode
High to medium	-0.13 (4.83)	-1.15 (6.59)	2.19 (20.20)	3.73 (31.44)	0.364 (1.49)	**-0.099 (0.25)**
p-value	.899	.421	.616	.584	.264	**.035**
High to	**3.48 (7.21)**	2.52 (8.23)	8.24 (28.03)	3.00 (41.00)	-0.034 (1.07)	-0.009 (0.33)
Low	**.006**	.071	.082	.659	.848	.432
Medium to	-1.36 (4.26)	-0.81 (6.36)	-6.21 (16.08)	-8.82 (27.52)	-0.35 (1.83)	0.079 (0.26)
High	.149	.557	.085	.148	.369	0.082
Medium to	**4.35 (11.54)**	**4.59 (9.64)**	**13.05 (38.83)**	**12.47 (41.52)**	0.12 (1.31)	-0.028 (0.22)
Low	**.001**	**.000**	**.003**	**.007**	.402	0.121
Low to	-1.25 (6.41)	-0.24 (9.81)	-6.85 (17.11)	-5.45 (23.91)	-0.021 (0.61)	0.072 (0.19)
High	0.404	.913	.098	.333	.880	.060
Low to	**-2.86 (7.83)**	**-3.65 (11.21)**	**-15.02 (38.71)**	**-16.96 (45.88)**	-0.28 (1.24)	0.021 (0.24)
Medium	**.008**	**.017**	**.005**	**.007**	0.095	.275
NO TREATMENT						
High to	9.71 (4.77)	1.77 (8.43)	0.076 (15.07)	6.87 (25.59)	0.27 (1.13)	-0.054 (0.37)
high	.999	.203	.975	.106	.151	.191
Medium to	-1.7 (8.20)	-1.18 (11.02)	-1.22 (31.33)	0.22 (40.91)	0.11 (1.08)	0.022 (0.28)
medium	.168	.476	.794	.971	.517	.303
Low to	**2.03 (11.91)**	**2.36 (13.09)**	**5.10 (35.95)**	**5.67 (43.07)**	0.0007 (1.41)	-0.003 (0.15)
low	**.011**	**.007**	**.034**	**.049**	.994	.373

Mean differences are calculated by post-test travel measure minus pre-test measure.
Bold type indicates significant at $p < 0.05$ level.
All tests are two tail, except "Δ % TRIPS alt mode" which is one-tail because of the undisputed directional hypothesis

Difference of Two Means

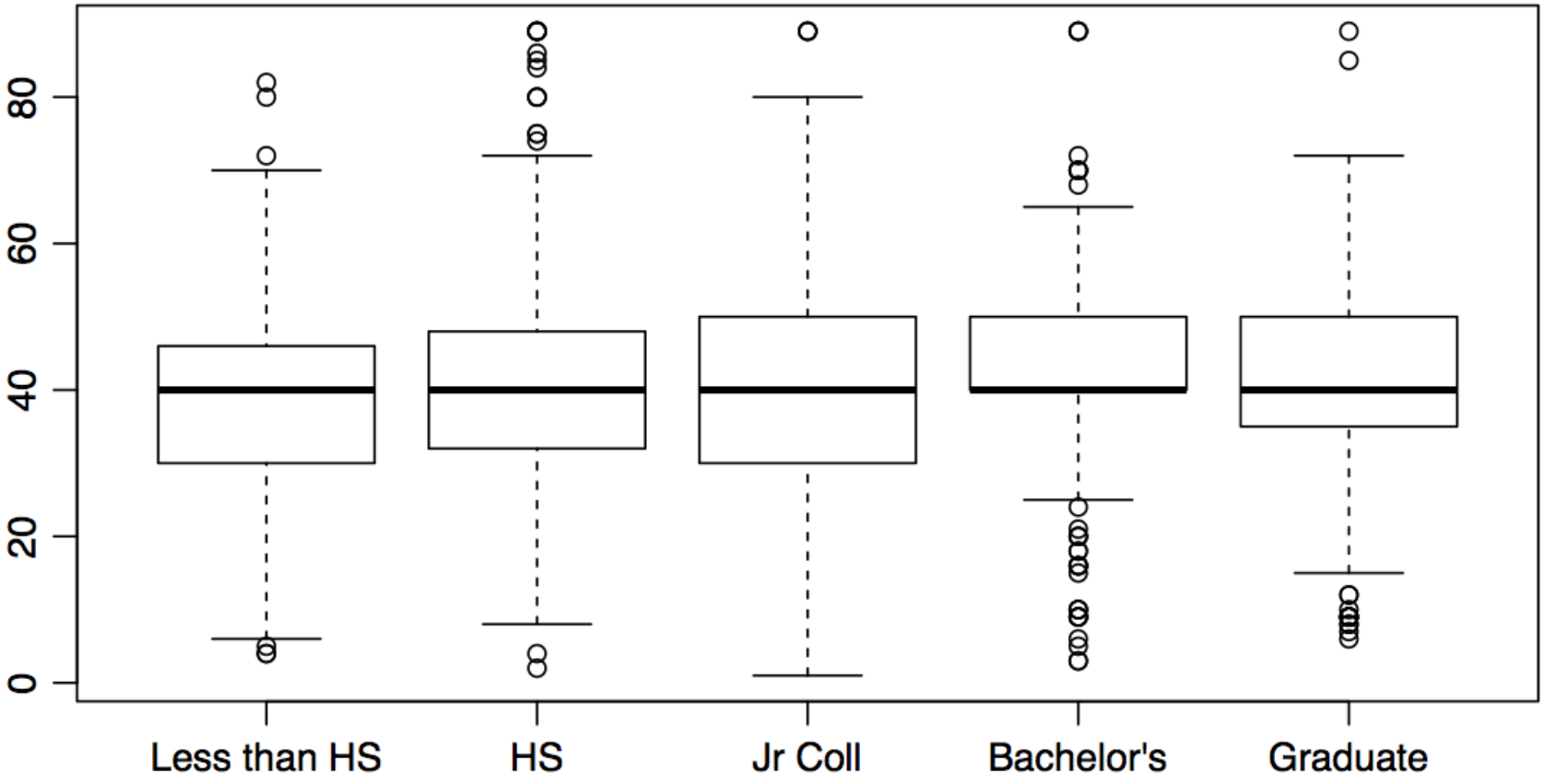
Degree and hours worked

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

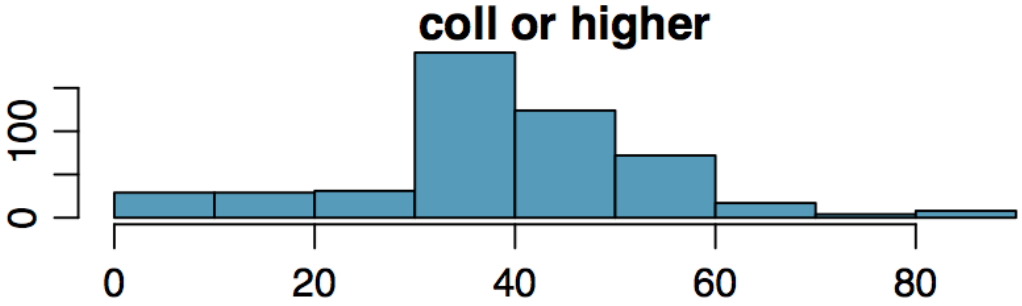
Exploratory analysis

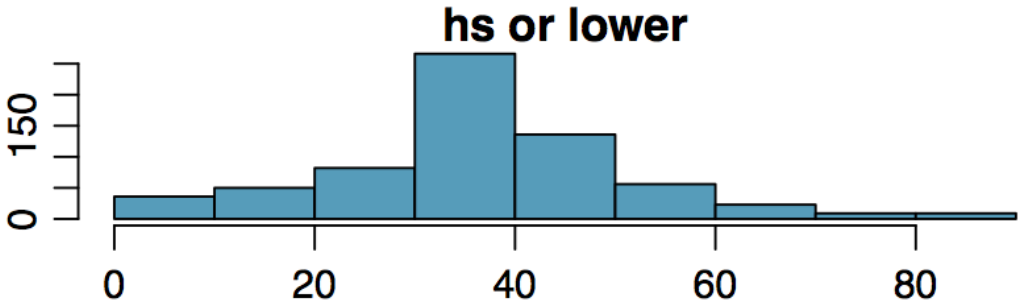
What can you say about the relationship between educational attainment and hours worked per week?



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 \red{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 ← 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

3. Sample size / skew

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

Confidence interval for difference between two means

All confidence intervals have the same form:

point estimate ± ME

And all ME = critical value x 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*

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

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

$$= \sqrt{\frac{15.14^2}{505} + \frac{15.12^2}{667}}$$

$$= 0.89$$

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?

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

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

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

$$upper\ tail = 1 - 0.9965 = 0.0035$$

$$p - value = 2 \times 0.0035 = 0.007$$

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.

t distribution and t-test

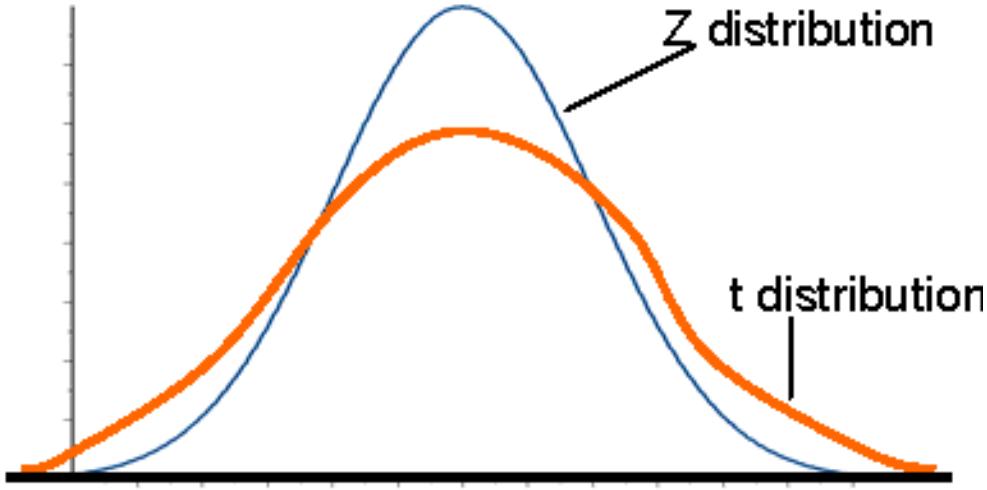
t Statistic: Small sample, σ unknown

- So far we've assumed either σ is known or sample is large.
- Sometimes sample is small, and usually *σ* is unknown.
- When using *s* to estimate *σ* and when the sample is small (<100), the test statistic is distributed not as a unit normal (i.e., not a Z distribution) but as a student's "*t*" distribution.

Different shaped distributions

● The *t* distribution is a probability distribution of sample means for small samples with unknown standard error. Like the Z, it is used for confidence interval estimate and hypothesis testing.

t distribution is a little more conservative – more difficult to reject Ho (i.e., the critical region is farther out) – because we only have sample information



t statistic vs. *Z* statistic

- The test statistic for *t* is calculated almost the same way as the Z statistic.

Z statistic: large sample

$$Z = \frac{\bar{X} - \mu}{s/\sqrt{N}}$$

t statistic: small sample

$$t = \frac{\bar{X} - \mu}{s/\sqrt{N}}$$

where $s = \sqrt{\frac{\sum(\bar{X}-\mu)^2}{N-1}}$

t **distribution**

- As N gets larger, the t distribution looks more and more normal (like Z).
- The **degrees of freedom** are the number of cases (N) minus the number of parameters used to estimate the test statistic (usually one or two, sometimes more)
- Usually use it only for *small samples*. Even when **σ** is unknown, can usually use standardized normal (**Z**) distribution when N>100 (though ***t*** distribution is always more conservative thus automatically used by R and other Stat packages). As N gets larger, **s** is a more accurate estimate of **σ**.
- So use the *t* for small samples; more importantly, understand that significantly greater uncertainty (lower precision) arises when relying on small samples

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

	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

↓

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

$s_{diff} = 1176$

$n = 10$

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.

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

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.

Using R:

```
> 2 * pt(4.94, df = 9, lower.tail = FALSE)
[1] 0.000802239
```

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.

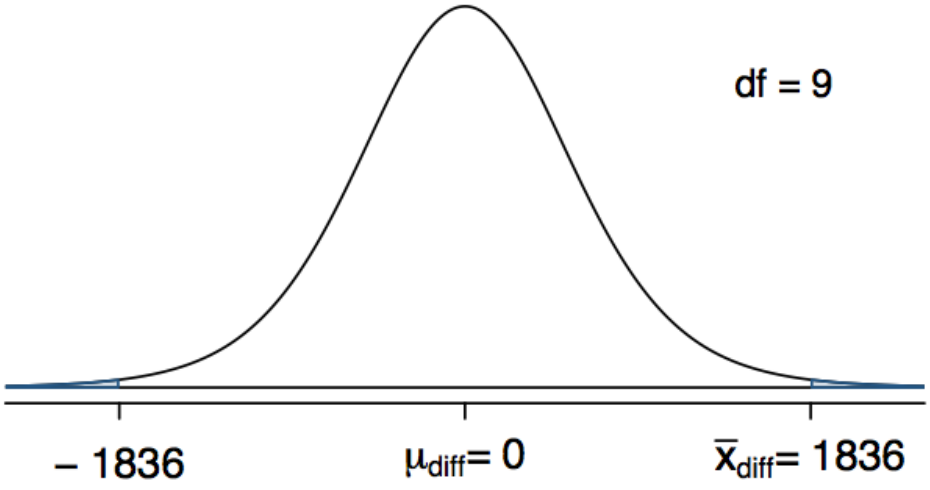
Finding the p-value

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 stat = 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 6th and 13th.

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

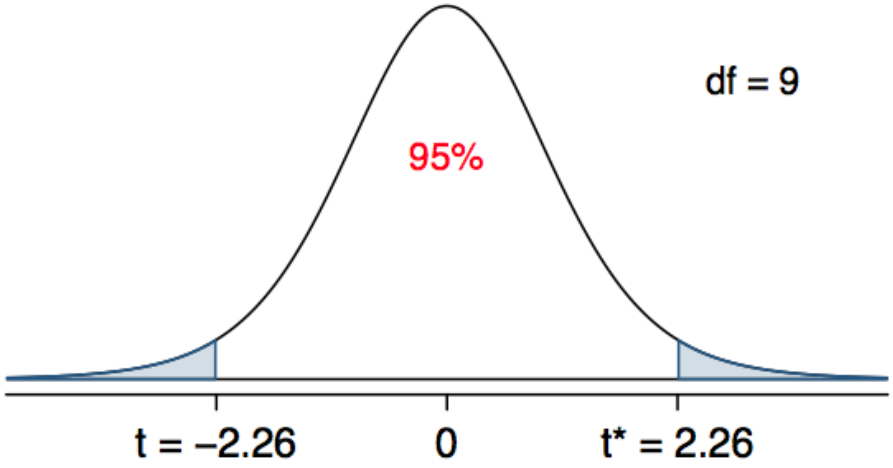
point estimate ± ME

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* (as opposed to a z*).

point estimate ± t* x SE

Finding the critical t (t*)



n = 10, df = 10 - 1 = 9,
t* 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 confidence interval 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?

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

a) $1836 \pm 1.96 \times 372$
b) $1836 \pm 2.26 \times 372$
c) $1836 \pm 2.26 \times 1176$

Constructing a confidence interval 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?

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

a) $1836 \pm 1.96 \times 372$

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

c) $1836 \pm 2.26 \times 1176$

Interpreting the CI

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

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

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?

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

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.

Recap: Inference using a small sample mean

If n < 30, sample means follow a t distribution with SE = s/√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.

Statistical vs Practical Significance

Test the hypothesis H_0: $\mu = 10$ vs. H_A: $\mu > 10$ for the following 6 samples. Assume $\sigma = 2$.

$\bar{x}$	10.05	10.1	10.2
$n = 30$	$p - value = 0.45$	$p - value = 0.39$	$p - value = 0.29$
$n = 5000$	$p - value = 0.04$	$p - value = 0.0002$	$p - value \approx 0$

When n is large, even small deviations from the null (small effect sizes), which may be considered practically insignificant, can yield statistically significant results.

Statistical vs Practical Significance

Real differences between the point estimate and null value are easier to detect with larger samples.

However, very large samples will result in statistical significance even for tiny differences between the sample mean and the null value (effect size), even when the difference is not practically significant.

This is especially important to research: if we conduct a study, we want to focus on finding meaningful results (we want observed differences to be real, but also large enough to matter).

The role of a statistician is not just in the analysis of data, but also in planning and design of a study.