Lecture 16 - Tests of Two Means

Sta 111

Colin Rundel

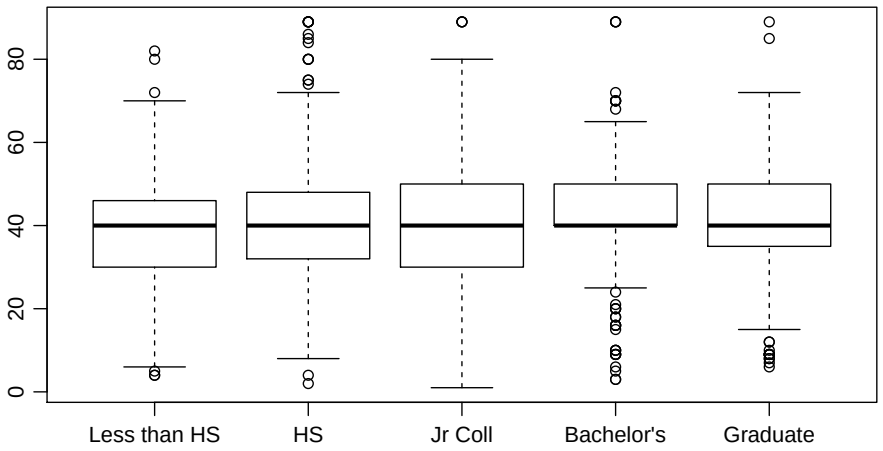
June 6, 2014

Example - GSS

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 we say about the relationship between educational attainment and hours worked per week?

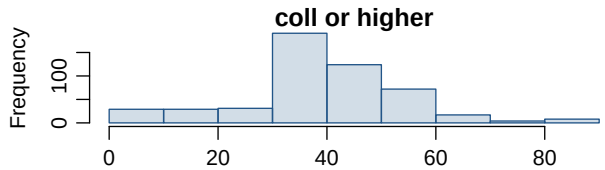
Collapsing levels

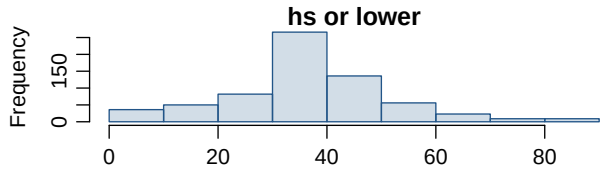
- Say we are only interested the difference between the number of hours worked per week by college and non-college graduates.
- We can combine the levels of education into:
  - `hs or lower` ← less than high school or high school
  - `coll or higher` ← junior college, bachelor's, and graduate
- Here is how you can do this in R:

```
# create a new empty variable
gss$edu = NA

# if statements to determine levels of new variable
gss$edu[gss$degree == "LESS THAN HIGH SCHOOL" |
        gss$degree == "HIGH SCHOOL"] = "hs or lower"
gss$edu[gss$degree == "JUNIOR COLLEGE" |
        gss$degree == "BACHELOR" |
        gss$degree == "GRADUATE"] = "coll or higher"

# make sure new variable is categorical
gss$edu = as.factor(gss$edu)
```

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:*
   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:*
      Since the sample is random, the college graduates in the sample are independent of those with a HS degree or lower.
2. *Sample size / skew:*
   Both distributions look reasonably symmetric, and the sample sizes are at least 30, therefore we can assume that the sampling distribution of number of hours worked per week by college graduates and those with HS degree or lower are nearly normal. Hence the sampling distribution of the average difference will be nearly normal as well.

Confidence interval for difference between two means

- All confidence intervals have the same form:

$$point\ estimate \pm ME$$

- Always, $ME = critical\ value \times SE\ of\ point\ estimate$
- In this case the point estimate is $\bar{x}_1 - \bar{x}_2$
- The CLT tells us that both $\bar{x}_1$ and $\bar{x}_2$ should have a normal distributions, and we are assuming there is independence between groups.
- Therefore, the difference between two means will be Normal with

$$\bar{x}_1 - \bar{x}_2 \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

$$SE_{(\bar{x}_1 - \bar{x}_2)} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \approx \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

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

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.

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

Since n is very large here 1172 we can use $Z^\star$ instead of $t^\star$.

What about smaller samples?

For smaller values of n_1 and n_2 when σ_1 and σ_2 are unknown we should be using a t distribution instead of the Z distribution for our critical values.

The t distribution for the difference of two means gets quite complicated (due to the need to scale for differing sample sizes and variances of group 1 and 2). The book gives the following

$$T = \frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{\sqrt{\left(\frac{(n_1-1)}{n_1 n_2} s_1^2 + \frac{(n_2-1)}{n_1 n_2} s_2^2\right)\left(\frac{n_1+n_2}{n_1+n_2-2}\right)}} \approx \frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{\sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}}$$

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1-1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{m-1}\left(\frac{s_2^2}{n_2}\right)^2} \approx \min(n_1 - 1, n_2 - 1)$$

Redoing the Confidence interval for the difference

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.

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

$$df = min(505 - 1, 667 - 1) = 504 \qquad t^\star_{0.95,\, df=504} = 1.96$$

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

We are 95% confident that college grads work on average between 0.66 and 4.14 more hours 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: $\mu_{coll} = \mu_{hs} \rightarrow \mu_{coll} - \mu_{hs} = 0$
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: $\mu_{coll} \neq \mu_{hs} \rightarrow \mu_{coll} - \mu_{hs} \neq 0$
There is a difference in the average number of hours worked per week by college graduates and those with a HS degree or lower.

Parameter and point estimate

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

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

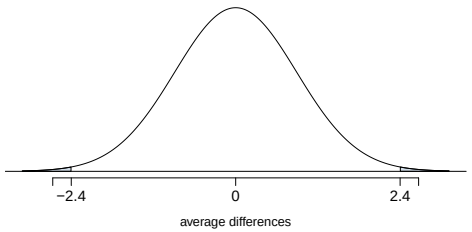
- *Point estimate:* Average difference between the number of hours worked per week by *sampled* college graduates and those with a HS degree or lower

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

Calculating the test-statistic and the p-value

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

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



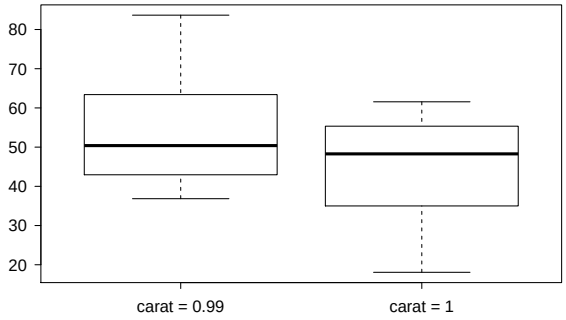
$$\begin{aligned} T_{df=504} &= \frac{(\bar{x}_{coll} - \bar{x}_{hs}) - 0}{SE_{(\bar{x}_{coll} - \bar{x}_{hs})}} \\ &= \frac{2.4}{0.89} = 2.70 \\ p - value &= P(T > 2.70 \cup T < -2.70) < 0.0 \end{aligned}$$

Since the p-value is small, 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.

Example - 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 the `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 and Conditions

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 ($_{pt99}$)?

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

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

Test statistic

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

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

In Context

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

What is the correct *df* for this hypothesis test?

p-value

What is the correct p-value for the hypothesis test?

$$T = -2.508$$

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

Synthesis

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

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 that would be equivalent to our hypothesis test?

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.

Inference using difference of two small sample means

- If σ_1 and/or σ_2 are unknown, the 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
  - reasonable sample size relative to the skew for both groups
- Hypothesis testing:

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

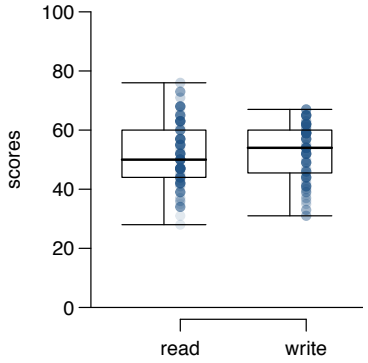
- Confidence interval:

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

Example - Reading and Writing

200 randomly selected high school 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?

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



Do you think reading and writing scores are independent?

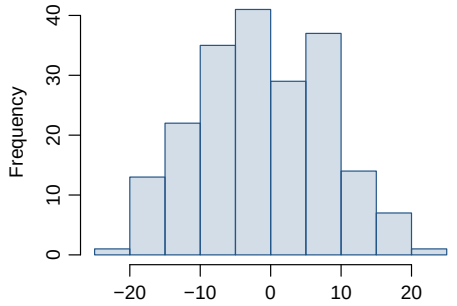
Analyzing paired data

- 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.

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

- 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?

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.

	diff
$\bar{x}$	-0.545
s	8.89
n	200

$$T_{df=199} = \frac{\bar{X} - \mu}{SE} = \frac{-0.545 - 0}{8.89/\sqrt{200}} = -0.877$$

p-value > 0.2