Unit 4: Inference for numerical variables
Lecture 1: Bootstrap, paired, and two sample

Statistics 101

Mine Çetinkaya-Rundel

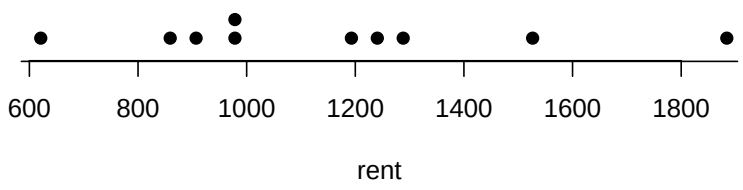
February 26, 2013

Announcements

- Extra credit due Thursday at the beginning of class
- RA4 on Thursday, covers all of Unit 4, including what we're doing today
- Grades on Sakai will not reflect dropped grades etc., those will be taken into consideration at the end of the semester – were not taken into consideration for (unofficial) midterm course grades
- Peer eval feedback:
  - All team members should contribute to the final lab report – not including a plot etc. should not be one person's responsibility.
  - In instances where there is a confusion even if you get a clicker question right, please stop me so that we can review it, chances are others are confused too so it'll be for everyone's benefit.
  - Getting to lab on time and sitting together in class are crucial.

Rent in Durham

A random sample of 10 housing units were chosen on *http://raleigh.craigslist.org* after subsetting posts with the keyword "durham". The dot plot below shows the distribution of the rents of these apartments. Can we apply the methods we have learned so far to construct a confidence interval using these data. Why or why not?



Bootstrapping

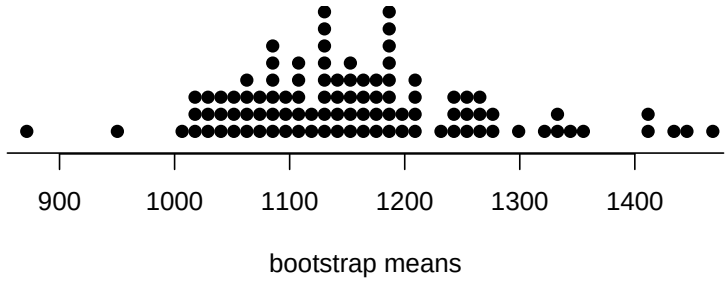
- An alternative approach to constructing confidence intervals is *bootstrapping*.
- This term comes from the phrase "pulling oneself up by one's bootstraps", which is a metaphor for accomplishing an impossible task without any outside help.
- In this case the ~~im~~possible task is estimating a population parameter, and we'll accomplish it using data from only the given sample.

Bootstrapping

- Bootstrapping works as follows:
  (1) take a bootstrap sample - a random sample taken with replacement from the original sample, of the same size as the original sample
  (2) calculate the bootstrap statistic - a statistic such as mean, median, proportion, etc. computed on the bootstrap samples
  (3) repeat steps (1) and (2) many times to create a bootstrap distribution - a distribution of bootstrap statistics
- The 95% bootstrap confidence interval is estimated by the cutoff values for the middle 95% of the bootstrap distribution.

Rent in Durham - bootstrap interval

The dot plot below shows the distribution of means of 100 bootstrap samples from the original sample. Estimate the 90% bootstrap confidence interval based on this bootstrap distribution.



Randomization testing for a mean

- We can also use a simulation method to conduct the same test.
- This is very similar to bootstrapping, i.e. we randomly sample with replacement from the sample, but this time we shift the bootstrap distribution to be centered at the null value.
- The p-value is then defined as the proportion of simulations that yield a sample mean at least as favorable to the alternative hypothesis as the observed sample mean.

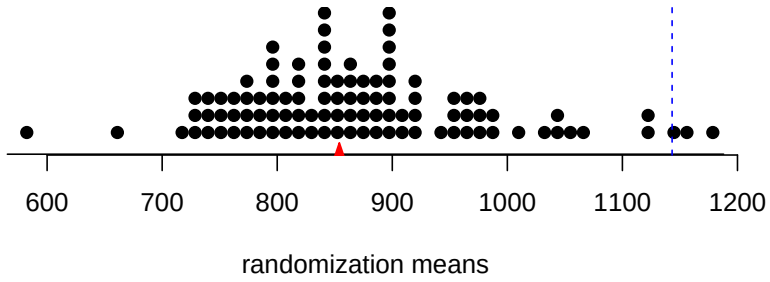
Rent in Durham - randomization testing

According to *rentjungle.com* the average rent for an apartment in Durham is \$854. Your random sample had a mean of \$1143.2. Does this sample provide convincing evidence that the article's estimate is an underestimate?

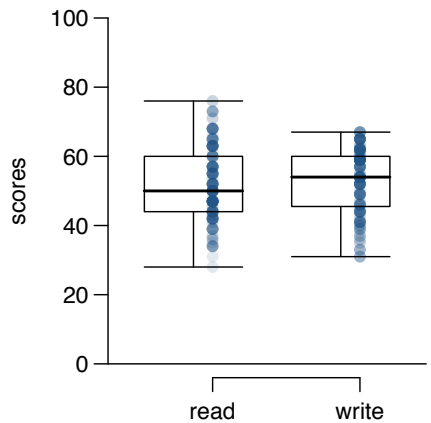
$H_0: \mu = \$854$

$H_A: \mu > \$854$

p-value: proportion of simulations where the simulated sample mean is at least as extreme as the one observed. → 3 / 100 = 0.03



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?



Clicker question

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

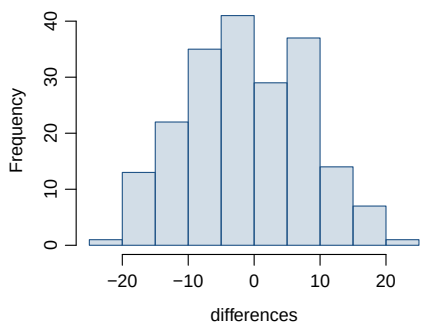
Analyzing paired data

- 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.

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

- 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.

Checking assumptions & conditions

Clicker question

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.

Application exercise:
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. Which of the below is the closest p-value for evaluating a difference between the average scores on the two exams?

(a) 20%
(b) 40%
(c) 5%
(d) 48%
(e) 95%

Interpretation of p-value

Clicker question

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

Clicker question

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

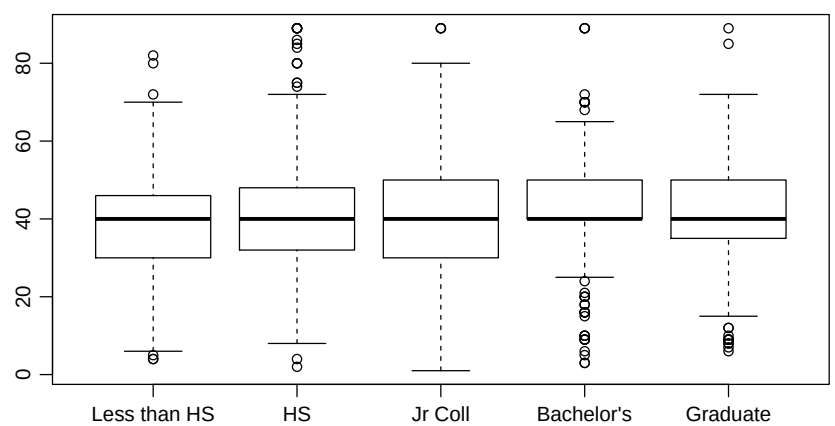
(c) cannot tell from the information given

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?



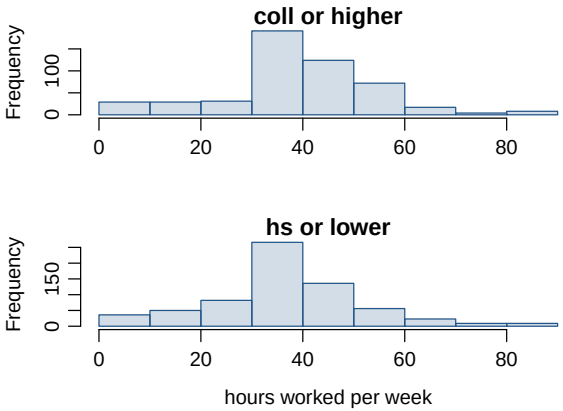
Collapsing levels into two

- Say we are only interested the difference between the number of hours worked per week by college and non-college graduates.
- Then we 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
- 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:*
   - *Within groups:*
     - both samples are random
     - 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.
   - *Between groups:* ← *new!*

     Since the sample is random, we have no reason to believe that the college graduates in the sample would not be 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:

$$point\ estimate \pm ME$$

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

Standard error of the difference between two sample means

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

Let's put things in context

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

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

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

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.

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

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

Interpretation of a confidence interval for the difference

Clicker question

Which of the following is the <u>best</u> 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.

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?

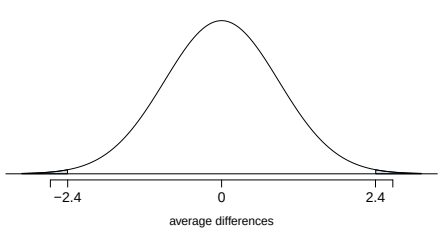
H_0: $\mu_{coll} = \mu_{hs}$
There is no difference in the average number of hours worked per week by college graduates and those with a HS degree or lower. Any observed difference between the sample means is due to natural sampling variation (chance).

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

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



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

Conclusion of the test

Clicker question

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, 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.