STA130H1 – Winter 2018 – Test Review

Prof. Gibbs

February 26, 2018

Purpose of today's class

- Review some material and ideas for the test
- Not 100% comprehensive. What else should you study? all lecture slides, weekly practice problems, online midterm review quiz

Clarification about content

- You are responsible for knowing the material about histograms and densities that was covered in Week 2 (which was a subset of the Week 1 material on this topic). You are not responsible for knowing the material in the Week 1 slides on this topic that was not covered in Week 2 and not covered in lecture (for example, kernel estimator).
- You are not responsible for knowing the mathematical note at the end of the Week 4 slides and you are not responsible for knowing the material on counting principles at the end of the Week 5 slides.

Structure of test

The test is a combination of:

- multiple choice
- true / false
- fill in the blanks
- short answer (explain why / apply)
- answers that require you to write some sentences

Lightning Round

Lightning Round Question 1

A clinical oncologist is investigating the efficacy of a new treatment on reduction in tumour size. She randomly assigns patients to the new treatment or old treatment and compares the mean of the reduction in tumour size between the two groups. She carries out a statistical test and the P-value is 0.001. How many of the following are valid interpretations of the P-value?

I. The probability of observing a difference between the treatment groups as large or larger than she observed if the new treatment has the same efficacy as the old treatment.
II. The probability that the new treatment works the same as the old treatment.
III. The probability that the new treatment, on average, reduces tumour size more than the old treatment. about alternative hypothesis (not ~~VAL~~ VALID)

~~A. None~~
[✓ B. One]
~~C. Two~~
~~D. Three~~

II is FALSE, II is a statement about ~~the~~ H_0 & its probability. We don't make probability statements about the null hypothesis. It's true or not. We make probability statements about the data

I is TRUE — probability of data (test statistic) given the null hypothesis is true.

Want CI for mean ~~for all~~ fish

Lightning Round Question 2

Environmental scientists want to estimate the mean mercury content in ppm of fish in a lake. They collect a random sample of 50 fish from the lake, measure the mercury content of each, calculate the average mercury content for these 50 fish and use the bootstrap to find a 99% confidence interval for the mean. The confidence interval is (0.82, 1.13). How many of the following are valid interpretations of the confidence interval?

I. We are 99% certain that each fish has approximately 0.82 to 1.13 ppm of mercury.

talking about individual fish

II. We expect 99% of the fish to have between 0.82 and 1.13 ppm of mercury.
III. We would expect about 99% of all possible sample means from this population to be in between 0.82 and 1.13 ppm of mercury.
IV. We are 99% certain that the confidence interval of (0.82, 1.13) includes the true mean of the mercury content of fish in the lake.

VALID

~~A. None~~; [✓ B. One]; ~~C. Two; D. Three; E. All four~~

What does "99% certain" mean?
~~99% of the time~~ If we took repeated samples of the same size, 99% ~~will~~ would give a CI that includes the true mean.

Lightning Round Question 3

Fill in the respective blanks:
Suppose we wish to test the null hypothesis that a Yoga method does not have an effect on blood pressure versus the alternative that it does have an effect. A _____________ error would be make by concluding that the Yoga method _____________ on blood pressure if in fact the Yoga method ____________ on blood pressure.

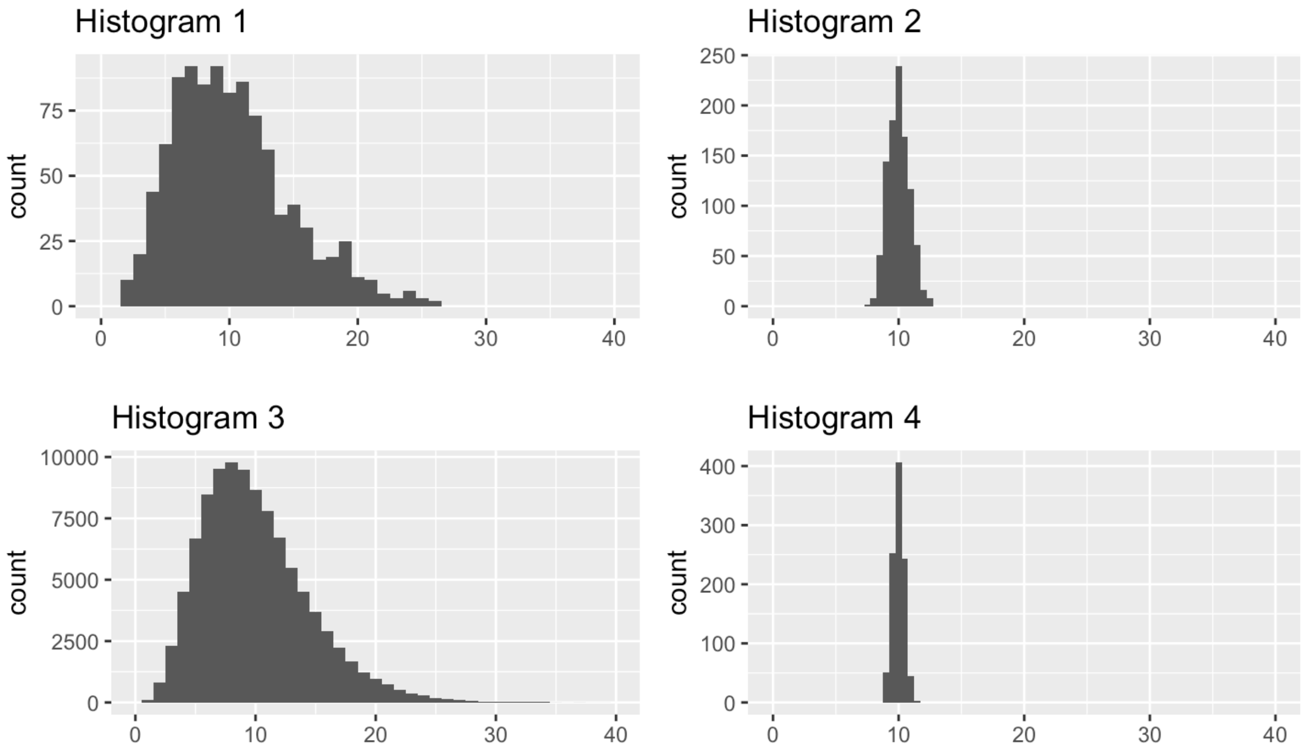
A. Type 2; does not have an effect; does have an effect
B. Type 2; does not have an effect; does not have an effect
C. Type 2; does have an effect; does not have an effect
D. Type 1; does not have an effect; does have an effect
E. Type 1; does not have an effect; does not have an effect

Lightning Round Question 4

On the next slide are 4 histograms:

- One is the histogram of the variable `x` in the population (which consists of 1,000,000 individuals).
- One is the histogram of the variable `x` for a sample of size 1000 from the population.
- One is the histogram for 1000 means of `x`, each from a sample of size 25.
- One is the histogram for 1000 means of `x`, each from a sample of size 100.

Which is which?


Histogram 1
count
0
25
50
75
0
10
20
30
40
Histogram 2
count
0
50
100
150
200
250
0
10
20
30
40
Histogram 3
count
0
2500
5000
7500
10000
0
10
20
30
40
Histogram 4
count
0
100
200
300
400
0
10
20
30
40

From the above plots, consider these two plots:

- The histogram for 1000 means of `x`, each from a sample of size 25
- The histogram for 1000 means of `x`, each from a sample of size 100

1. What is a name for what is plotted in these two plots?
2. What do these plots tell you about the effect of sample size on a confidence interval for the mean?

Lightning Round Question 5

In statistical inference, we want to make conclusions about what we think about the **theoretical world** (a scientific model or population) based on what we've observed in the **real world** (data, typically observed on a random sample). Do the following items exist in the theoretical world or the real world?

- Statistic REAL
- Parameter THEORETICAL
- Null hypothesis (and alternative hypothesis) — competing views on what theoretical world is
- ** Test statistic REAL
- * Simulated values of the test statistic under the null hypothesis THEORETICAL
- P-value IN BETWEEN need * and **
- Sampling distribution THEORETICAL (based on population)
- Bootstrap sampling distribution REAL (based on data)

Default for replace is FALSE (sampling without replacement)

Size not specified → get same size as x

Lightning Round Question 6

Suppose we have a vector of data values, **`x`**

```
x <- c(1, 3, 4, 4, 7)
```

We've used the **`sample()`** function various ways. What output is possible with each of the following commands?

```
sample(x)
```

gives all elements of x, in a random order

eg. 7, 4, 3, 4, 1

```
sample(x, replace=TRUE)
```

eg, 1, 1, 1, 1, 1

```
sample(x, size=2, replace=TRUE)
```

eg. 1, 1 ; eg. 3, 3, eg. 7, 4

```
sample(x, size=2, prob=c(0.5, 0.5, 0, 0, 0))
```

only possibilities!
1, 3
or 3, 1

Which one of these is a bootstrap sample?

sample(x, replace = TRUE)

Case Study: American Community Survey 2012

Case Study: American Community Survey 2012

The American Community Survey is conducted by the US Census Bureau each year on a random sample of 3.5 million households. Findings from the survey influence the allocation of more than $400 billion in federal and state funds. The dataset `acs12` is a random sample from the people who completed the American Community Survey in 2012.

Here is a look at the data and some of the variables we will consider later:

```
glimpse(acs12)
```

```
## Observations: 2,000
## Variables: 13
## $ income       <int> 60000, 0, NA, 0, 0, 1700, NA, NA, NA, 45000, NA, ...
## $ employment   <fctr> not in labor force, not in labor force, NA, not ...
## $ hrs_work     <int> 40, NA, NA, NA, NA, 40, NA, NA, NA, 84, NA, 23, N...
## $ race         <fctr> white, white, white, white, white, other, white,...
## $ age          <int> 68, 88, 12, 17, 77, 35, 11, 7, 6, 27, 8, 69, 69, ...
## $ gender       <fctr> female, male, female, male, female, female, male...
## $ citizen      <fctr> yes, yes, yes, yes, yes, yes, yes, yes, yes, yes...
## $ time_to_work <int> NA, NA, NA, NA, NA, 15, NA, NA, NA, 40, NA, 5, NA...
## $ lang         <fctr> english, english, english, other, other, other, ...
## $ married      <fctr> no, no, no, no, no, yes, no, no, no, yes, no, no...
## $ edu          <fctr> college, hs or lower, hs or lower, hs or lower, ...
## $ disability   <fctr> no, yes, no, no, yes, yes, no, yes, no, no, no, ...
## $ birth_qrtr   <fctr> jul thru sep, jan thru mar, oct thru dec, oct th...
```

```
table(acs12$employment)
```

```
##
## not in labor force          unemployed            employed
##                656                 106                 843
```

```
table(acs12$edu)
```

```
##
## hs or lower     college        grad
##        1439         359         144
```

Subset of observations

Case study question 1

remove observations with employment missing

Describe the data frames that are created by each of the following commands:

```
labor_force <- acs12 %>% filter(!is.na(employment)) %>%
  filter(employment == "employed" | employment == "unemployed")
```

or

fewer observations; removed people not in labor force

```
employed <- labor_force %>% filter(employment == "employed")
```

```
employed <- employed %>%
  mutate(edu2 = recode(edu, "hs or lower" = "hs_or_lower",
                       "college"="more_than_hs", "grad"="more_than_hs"))
```

add a column (new variable)

```
cat_vars <- acs12 %>% select(employment, race, gender, citizen, lang, married, edu,
                             disability, birth_qrtr)
```

select particular columns

Case study question 2

We've used these plot geometries:
`geom_bar, geom_boxplot, geom_dotplot, geom_histogram, geom_line, geom_point, geom_vline`

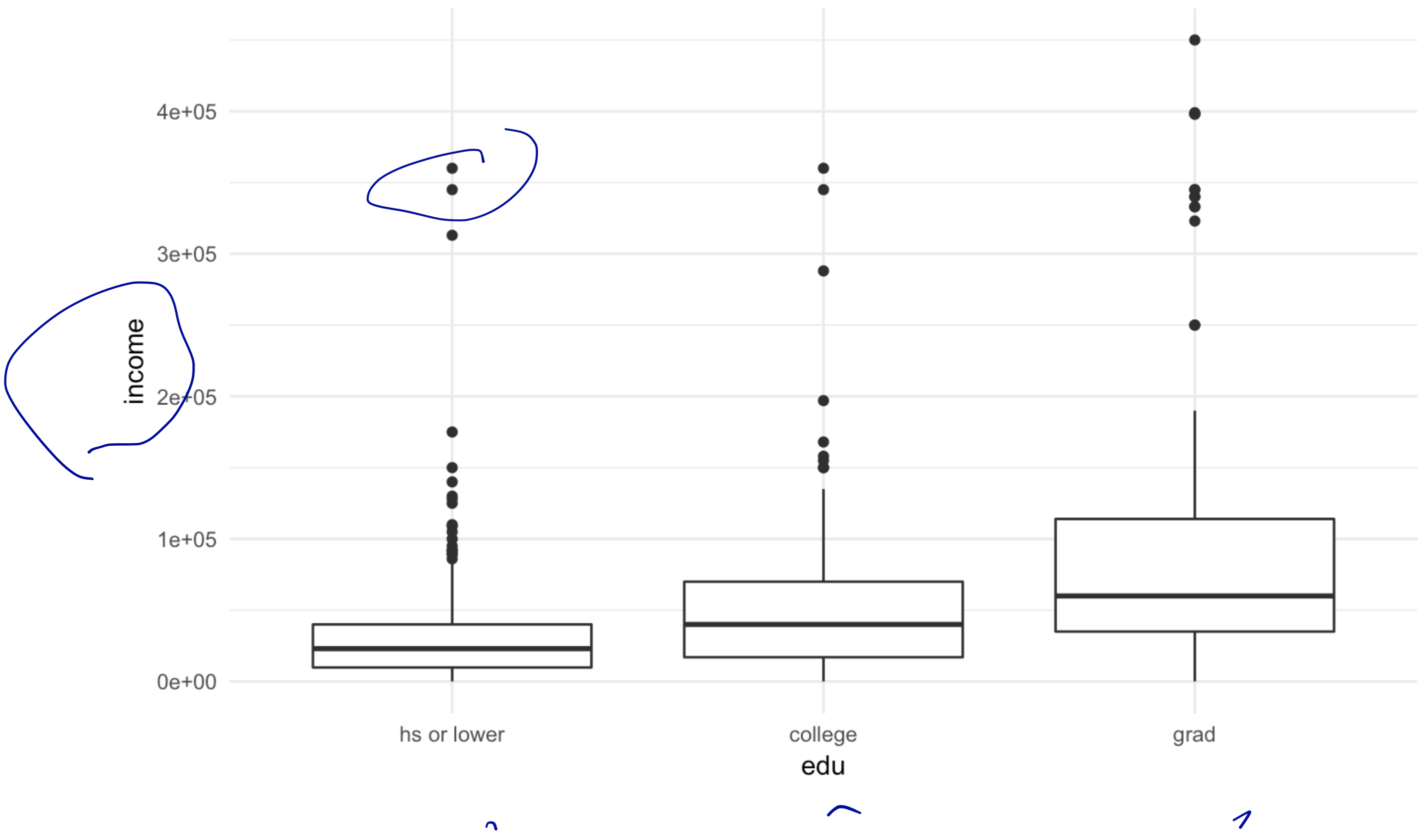
Recall this plot vocabulary:

- Bar plots: modes, frequency
- Histograms / boxplots: centre, spread, modes (unimodal, bimodal, multimodal, no mode), frequency, symmetric / left-skewed / right-skewed, outliers
- Scatterplots: strong / weak / no relationship, linear (positive or negative) / nonlinear relationshiop, outliers, clusters

On the next several slides are a number of plots, each constructed from the dataset `employed`. For each:

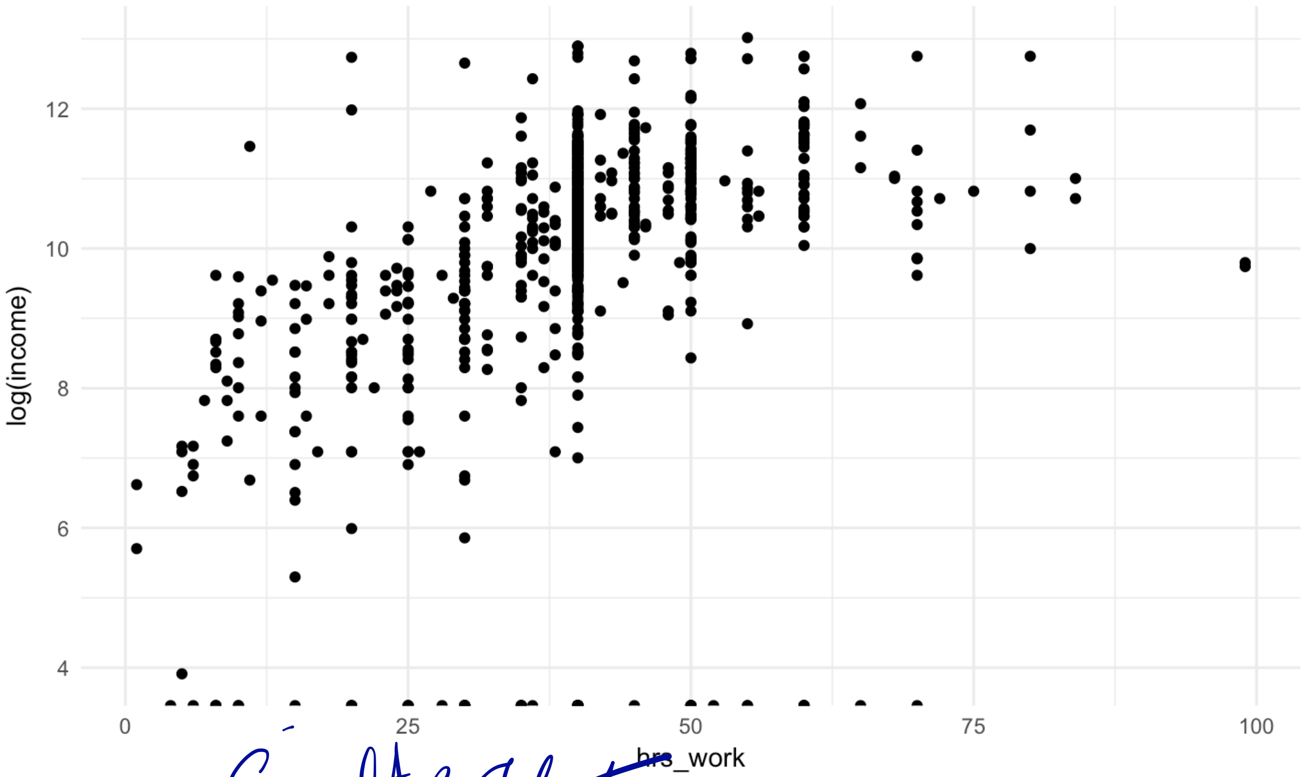
- What type of plot is it?
- What `ggplot` geometry is used?
- What is the purpose of the plot?
- Describe the distribution(s) of the variable(s).

Plot 1



Plot 2

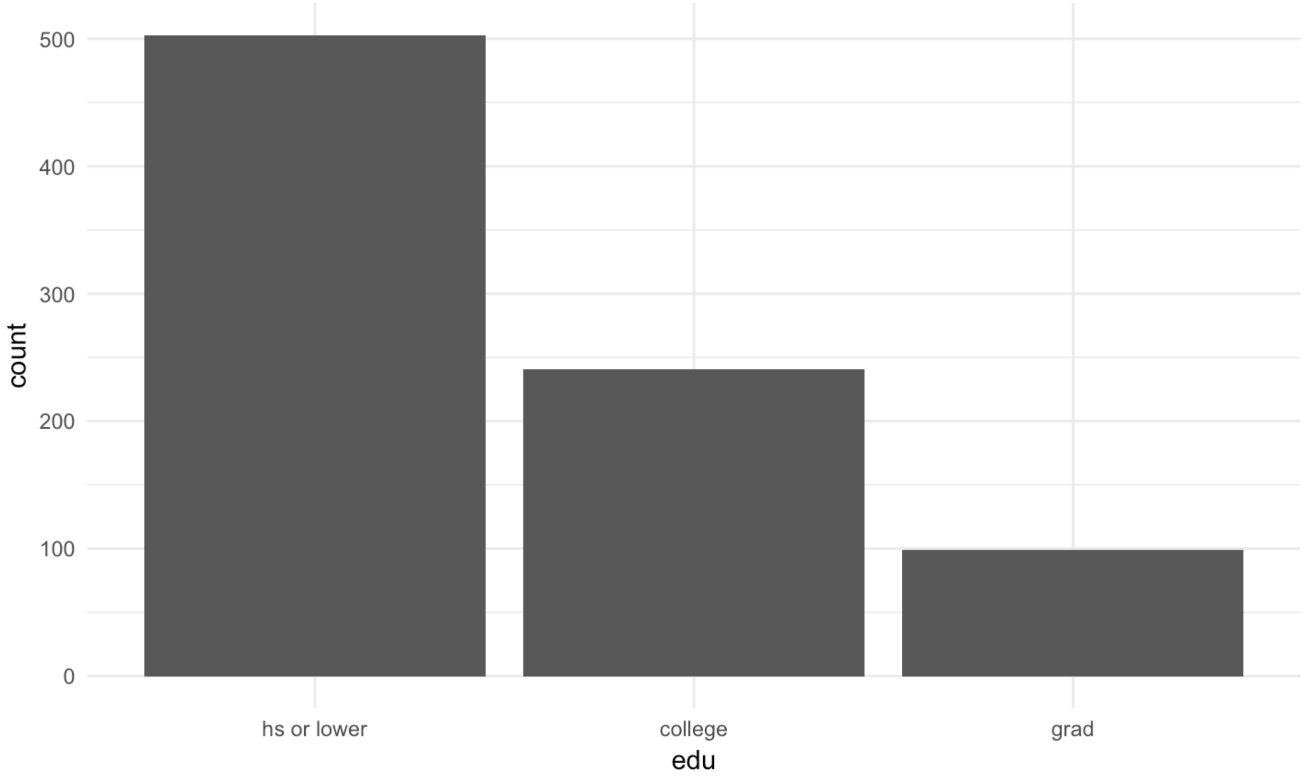
Why use a log transformation?



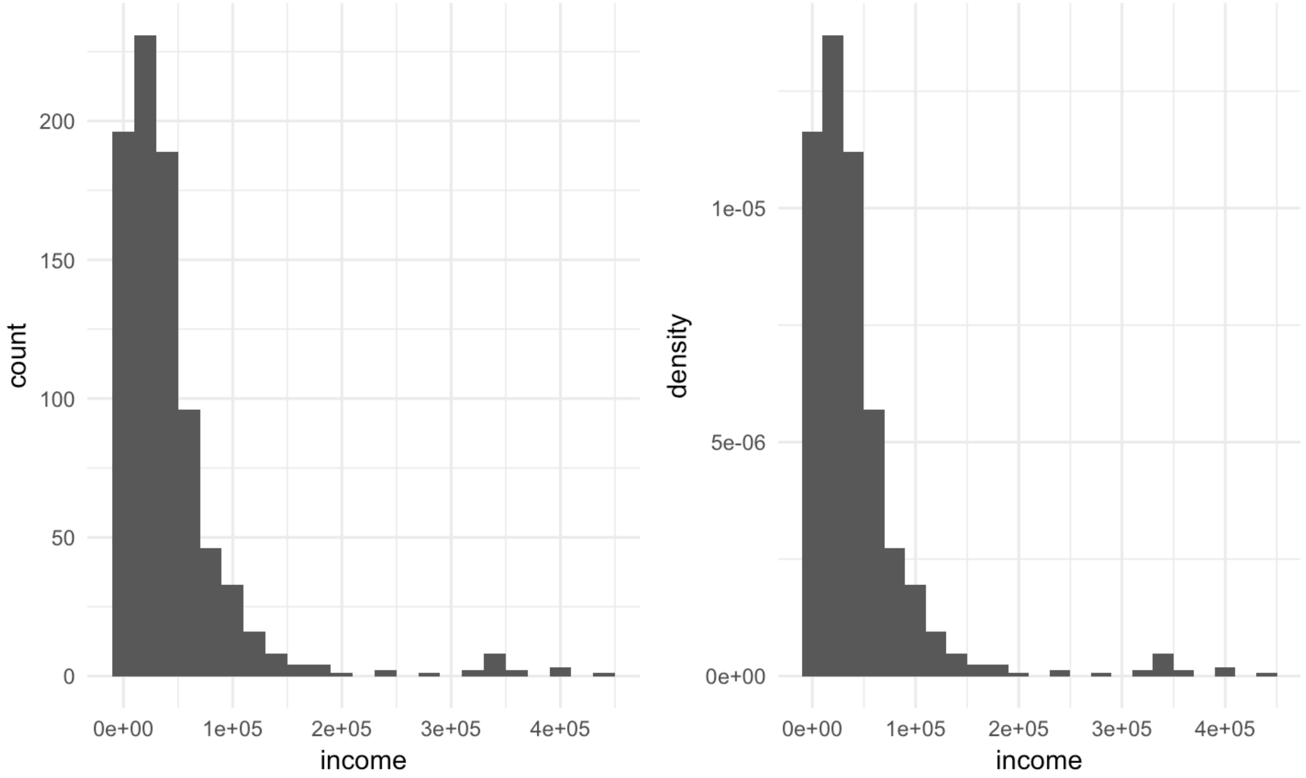
Scatterplot

geometry: geom_point

Plot 3



Plot 4



What is the difference in these histograms? How do you get the second from the first?

Case study question 3

We've looked at simulations to:

1. See how a statistic calculated from a population might vary from sample to sample (what is this called?)

   sampling distribution

2. Estimate 1. when we only have one set of data (what is this called?)

   bootstrap sampling distribution

3. See the distribution of possible values of a statistic under an assumption (what is this assumption called?). Two cases of this we considered:

   null hypothesis

   - a. simulate outcomes for a proportion (how did we do this under our assumption?)

     flip a coin with probabilities specified by H_0

   - b. simulate the difference in a statistic between groups (how did we do this under our assumption?)

     randomly shuffle group label

Below is code for three simulations. For each:

- What is the purpose of the simulation (from the 5 choices above)?
- Is the simulation used for a test or for a confidence interval or neither?
- For the dotplot of simulated values, where is its centre?

If the purpose of the simulation is to …

- … find a confidence interval for a parameter, describe how you would estimate a 90% confidence interval from the dotplot of simulated values.
- … carry out an hypothesis test, what is the null hypothesis? Estimate the P-value from the values plotted in the dotplot. What is your conclusion?

Some statistics that might be useful

```
labor_force  %>% group_by(employment) %>% summarise(n_group = n()) %>%
  mutate(percent = n_group / sum(n_group))
```

```
## # A tibble: 2 x 3
##   employment n_group   percent
##       <fctr>   <int>     <dbl>
## 1 unemployed     106 0.1116965
## 2   employed     843 0.8883035
```

```
employed %>% group_by(edu2) %>% summarise(mean(income))
```

```
## # A tibble: 2 x 2
##           edu2 `mean(income)`
##         <fctr>          <dbl>
## 1  hs_or_lower       29963.08
## 2 more_than_hs       65010.21
```

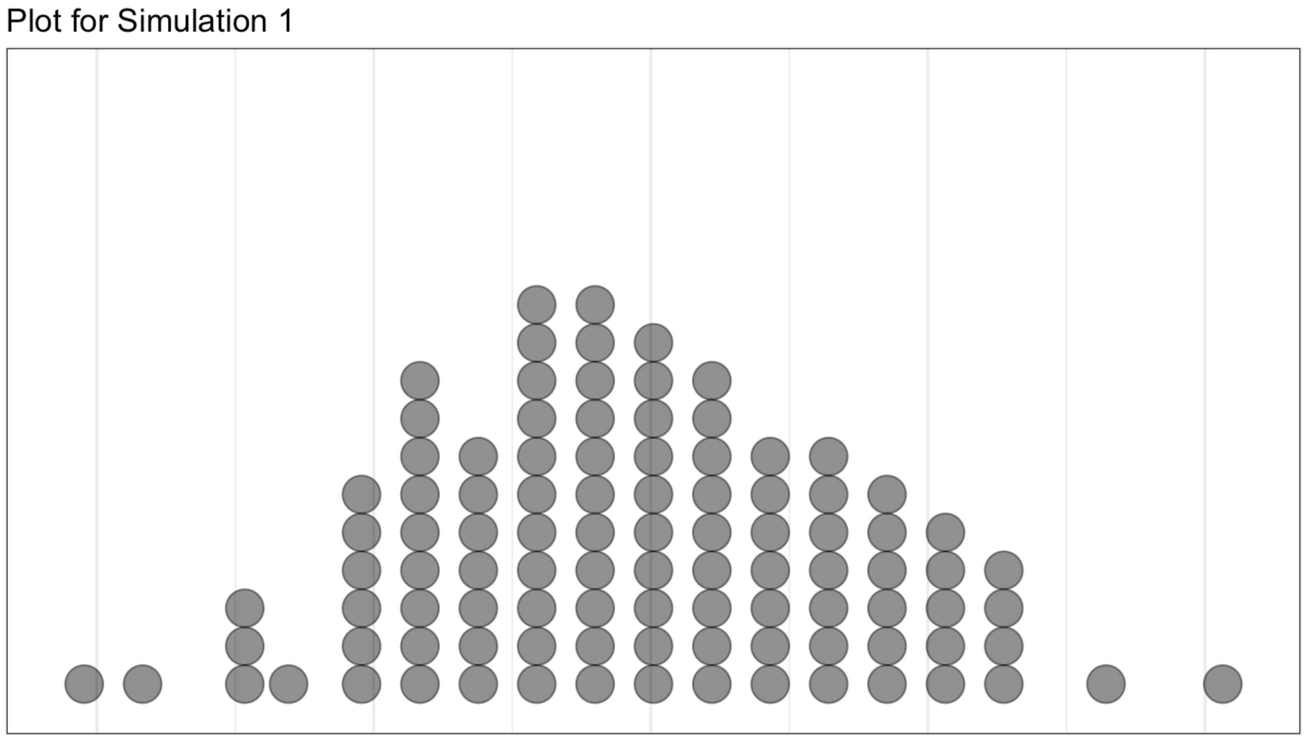
Simulation 1

```r
repetitions <- 100
x <- rep(NA, repetitions)

n <- as.numeric(labor_force %>% summarize(n()))

for (i in 1:repetitions)
{
  sim <- sample(c("unemployed", "employed"), size=n, prob=c(0.089, 1-0.089), replace=TRUE)
  sim_stat <- sum(sim == "unemployed") / n
  x[i] <- as.numeric(sim_stat)
}
```

```
## `stat_bindot()` using `bins = 30`. Pick better value with `binwidth`.
```



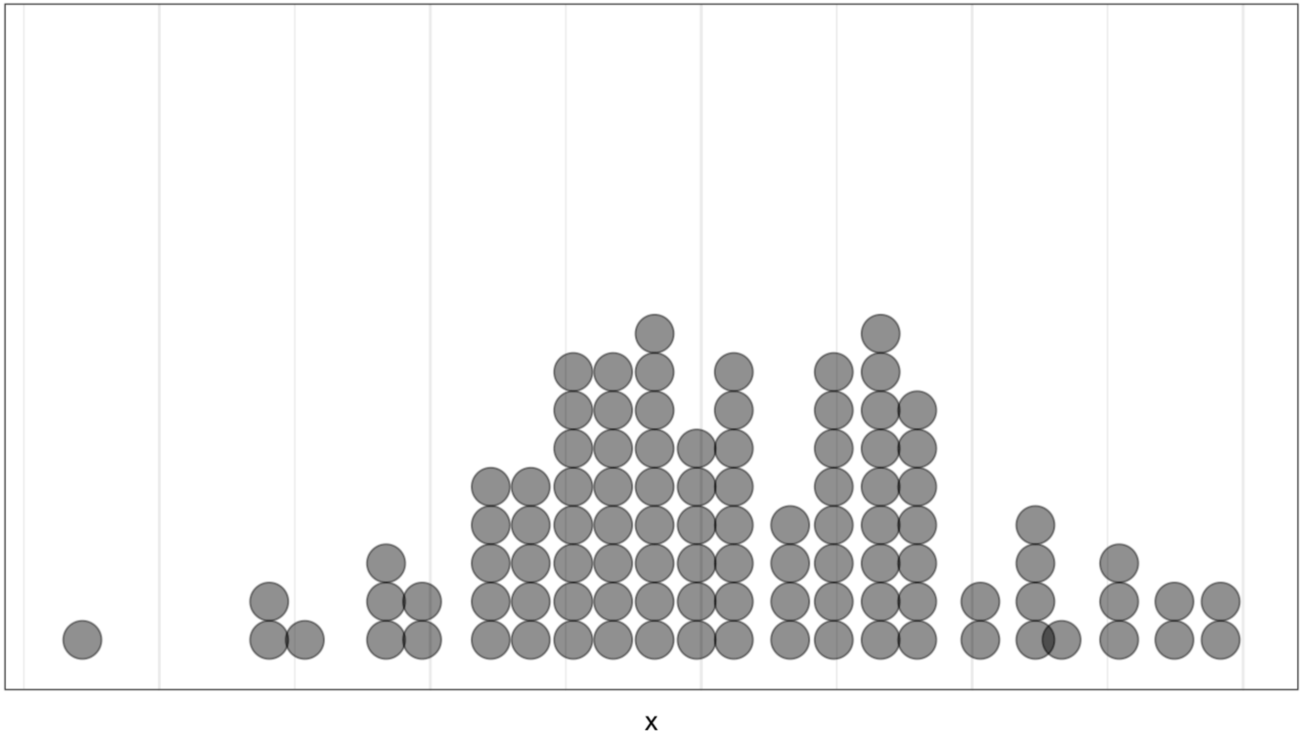
Note: The grid marks are 0.005 apart on the horizontal axis.

Simulation 2

```r
repetitions <- 100
x <- rep(NA, repetitions)

for (i in 1:repetitions)
{
  sim <- employed %>% mutate(edu2 = sample(edu2))
  sim_stat <- sim %>% group_by(edu2) %>%
                    summarise(means = mean(income)) %>%
                    summarise(diff(means))
  x[i] <- as.numeric(sim_stat)
}
```

Plot for Simulation 2



Note: The grid marks are 2500 apart on the horizontal axis.

Simulation 3

```
repetitions <- 100
x <- rep(NA, repetitions)

n <- as.numeric(labor_force %>% summarize(n()))

for (i in 1:repetitions)
{
  sim <- labor_force %>% sample_n(size = n, replace=TRUE)
  x[i] <- as.numeric(sim %>% filter(employment == "unemployed") %>%
                                   summarize(n())) / n
}
```

bootstrap samples

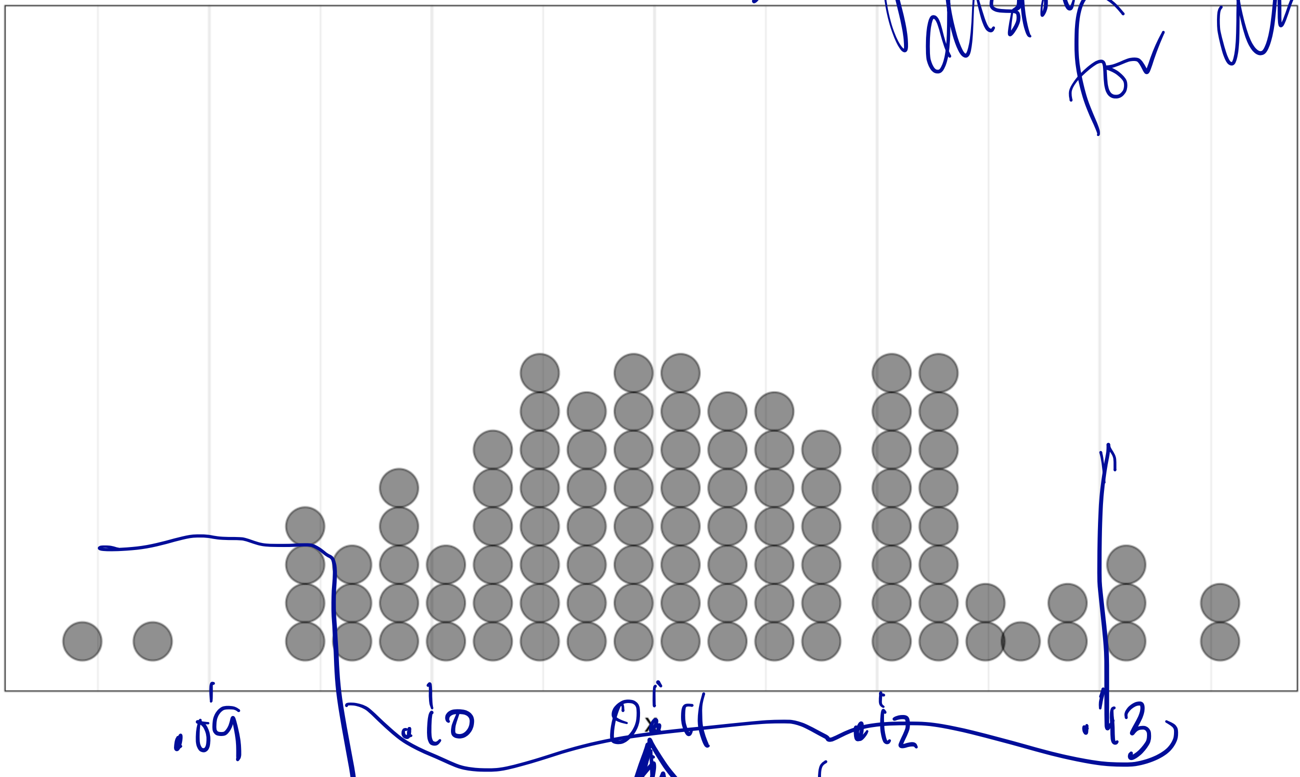
calculate

unemployment rate

CI for this

Bootstrap Sampling distribution for unemployment rate

Plot for Simulation 3



Note: The grid marks are 0.005 apart on the horizontal axis.

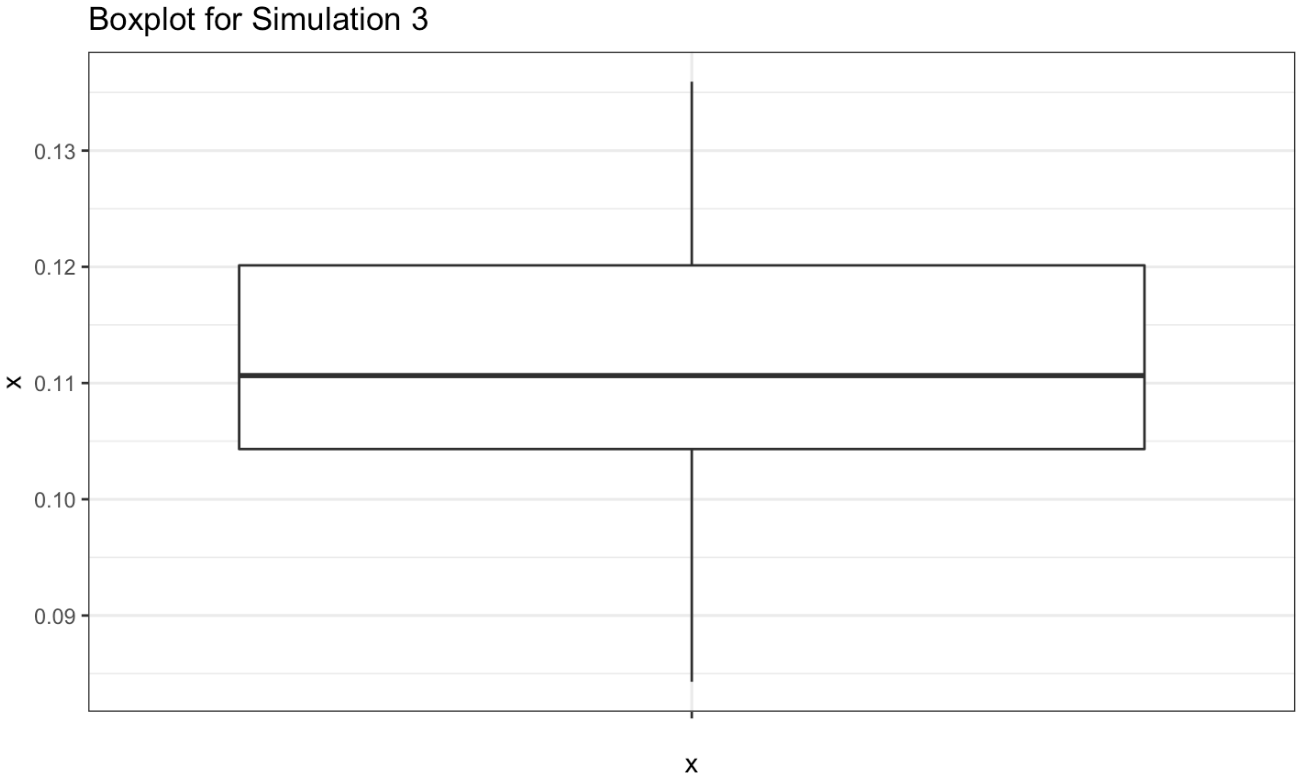
middle 90% of data

unemployment from our data

90% CI

~(0.095, 0.13)

Below is a boxplot for Simulation 3, plotting the same data as the histogram. Can you use it to estimate a confidence interval? For any confidence level?


Boxplot for Simulation 3
x
x