Lab 8 - Inference for numerical data

Sanaz Saadatifar

3/23

Lab report

Load data

```
acs <- read_csv("https://dyurovsky.github.io/85309/data/lab8/acs.csv")
```

```
## Rows: 2000 Columns: 13
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (9): employment, race, gender, citizen, lang, married, edu, disability, ...
## dbl (4): income, hrs_work, age, time_to_work
```

```
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Set a seed

Exercise 1:

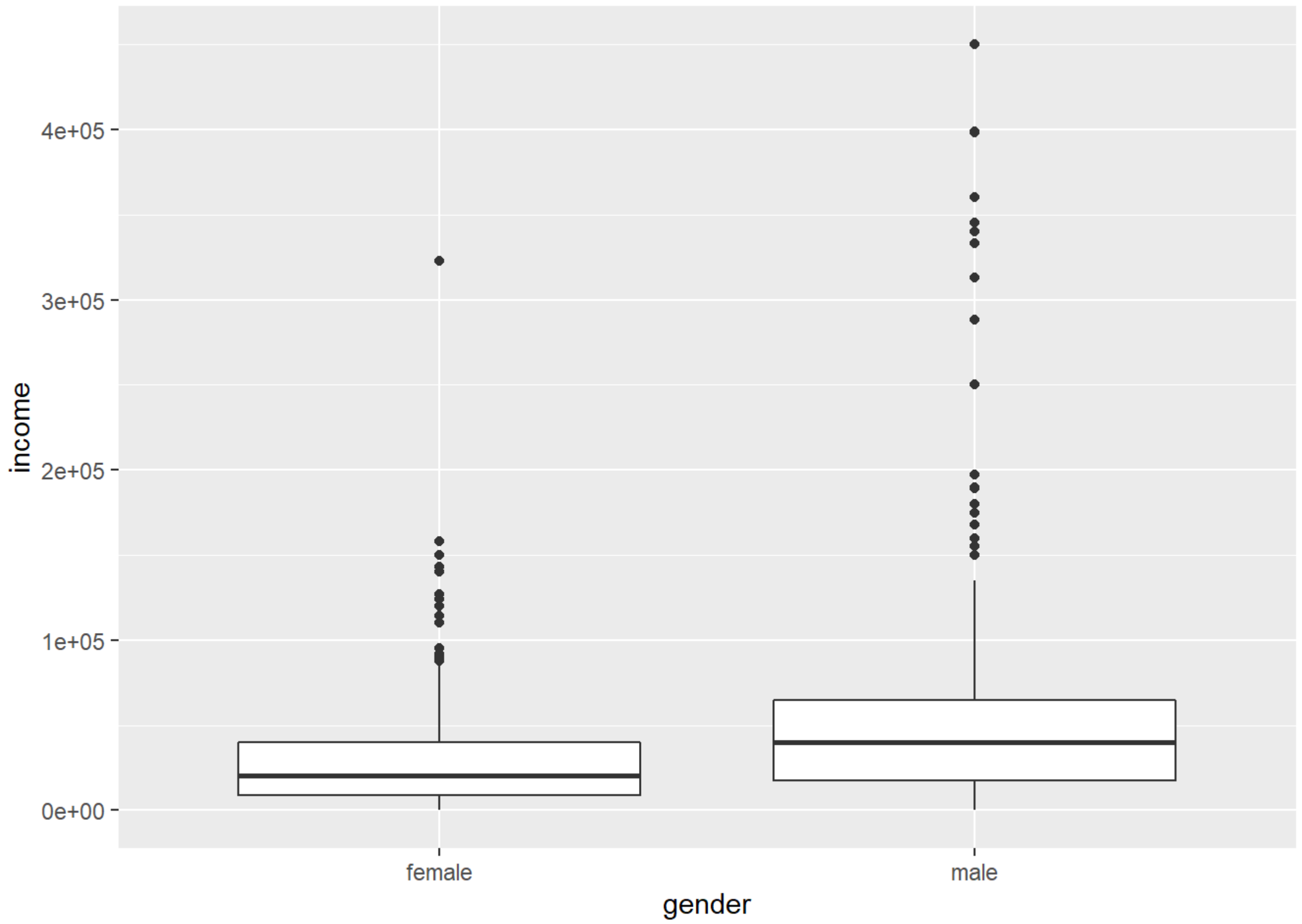
```
acs_emp <- acs %>%
  filter(employment == "employed")

employed_prop <- nrow(acs_emp)/nrow(acs)
employed_prop
```

```
## [1] 0.4215
```

42.15 percent of population are employed.

Exercise 2:

```
ggplot(acs_emp, aes(x = gender, y = income)) +
  geom_boxplot()
```



```
acs_emp %>%
  group_by(gender) %>%
  summarise(xbar = mean(income),
            s = sd(income),
            n = n())
```

```
## # A tibble: 2 x 4
##   gender   xbar      s     n
##   <chr>   <dbl>  <dbl> <int>
## 1 female 29244. 32026.   373
## 2 male   55887. 68768.   470
```

The mean salary of men appears to be higher. So is the standard deviation of their salaries. It looks like maybe there are more outlier in the group of men than women as well, which could be having a large effect on the mean.

Exercise 3:

```
acs_emp_female <- acs_emp %>%
  filter(gender == "female")

acs_emp_male <- acs_emp %>%
  filter(gender == "male")

t.test(acs_emp_male$income, acs_emp_female$income)
```

```
##
##  Welch Two Sample t-test
##
## data:  acs_emp_male$income and acs_emp_female$income
## t = 7.4437, df = 694.94, p-value = 2.903e-13
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  19615.96 33671.11
## sample estimates:
## mean of x mean of y
##  55887.23  29243.70
```

```
t.test(income~gender, data = acs_emp)
```

```
##
##  Welch Two Sample t-test
##
## data:  income by gender
## t = -7.4437, df = 694.94, p-value = 2.903e-13
## alternative hypothesis: true difference in means between group female and group male is not e
qual to 0
## 95 percent confidence interval:
##  -33671.11 -19615.96
## sample estimates:
## mean in group female   mean in group male
##             29243.70             55887.23
```

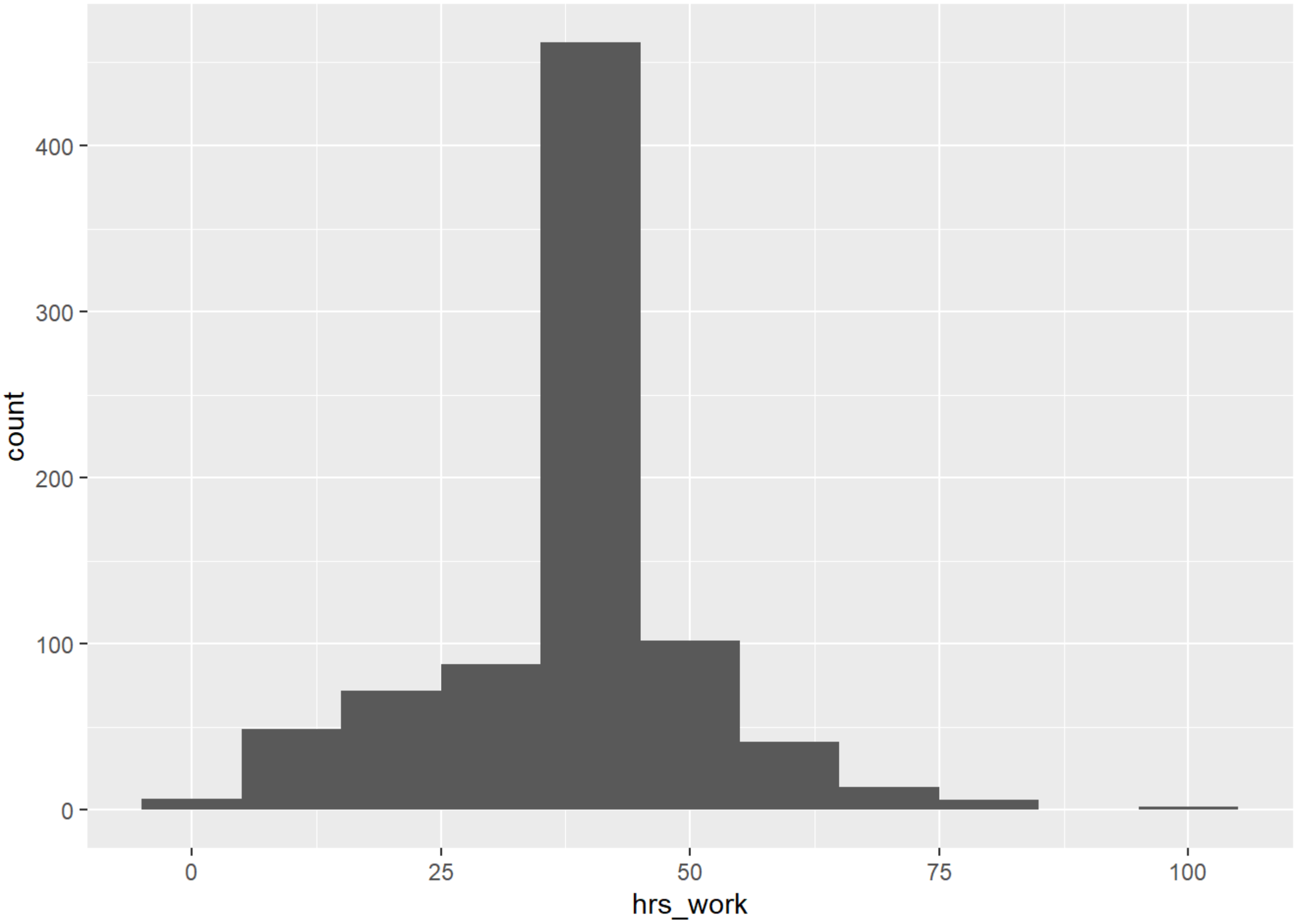
the 95% confidence interval for the difference between men’s and women’s salaries is (19615.96, 33671.11)

Exercise 4:

Because this confidence interval does not include 0, we should reject the null hypothesis that men and women make the same income on average.

Exercise 5:

```
ggplot(acs_emp, aes(x = hrs_work)) +
  geom_histogram(binwidth = 10)
```



We see a distribution centered around 40 hr/week with tails on both sides. Almost everybody seems to work around 40 hrs/week but there are people who work both more hours and people who work fewer hours who might be part-time employees.

Exercise 6:

```
t.test(acs_emp$hrs_work, mu=40)
```

```
## 
##  One Sample t-test
## 
## data:  acs_emp$hrs_work
## t = -2.4028, df = 842, p-value = 0.01649
## alternative hypothesis: true mean is not equal to 40
## 95 percent confidence interval:
##  38.05811 39.80429
## sample estimates:
## mean of x 
##   38.9312
```

H0: the sample comes from a population with a mean of 40 HA: the sample comes from a population with a mean that is not 40

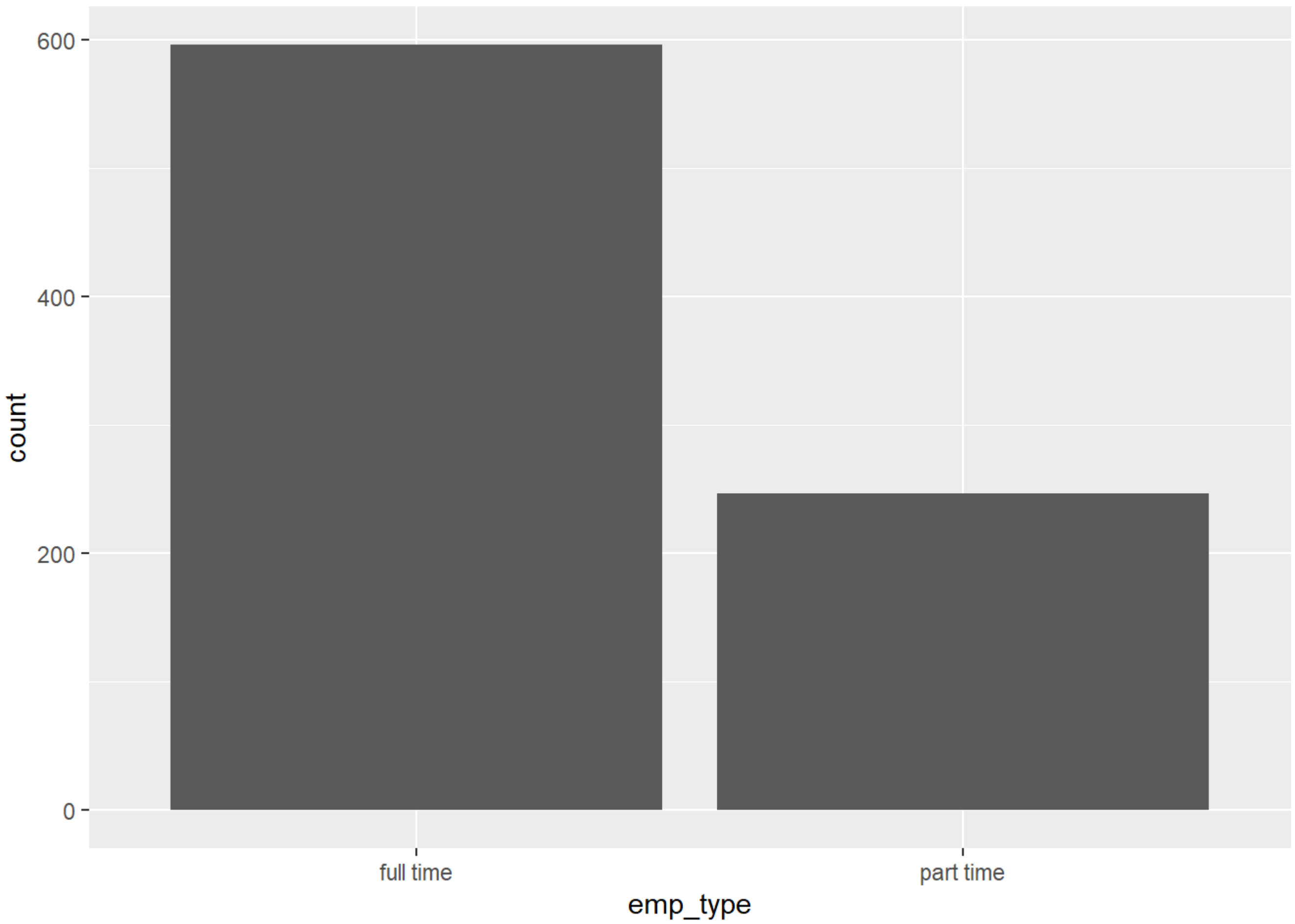
Because the P-value is less than 0.05, we reject the null hypothesis that the sample comes from a population with a mean of 40. So we might have a mixture of full-time and not full-time employees.

Exercise 7:

```
acs_type <- acs_emp %>%
  mutate(emp_type = if_else(hrs_work >= 40, "full time", "part time"))

acs_type %>%
  group_by(emp_type) %>%
  summarise(total_type = n()) %>%
  ungroup() %>%
  mutate(prop_type = total_type/sum(total_type))
```

```
## # A tibble: 2 x 3
##   emp_type  total_type prop_type
##   <chr>          <int>     <dbl>
## 1 full time        596     0.707
## 2 part time        247     0.293
```

```
ggplot(acs_type, aes(x = emp_type)) +
  geom_bar()
```



70.6% of the sample is full time and 29.3% of the sample is part-time.

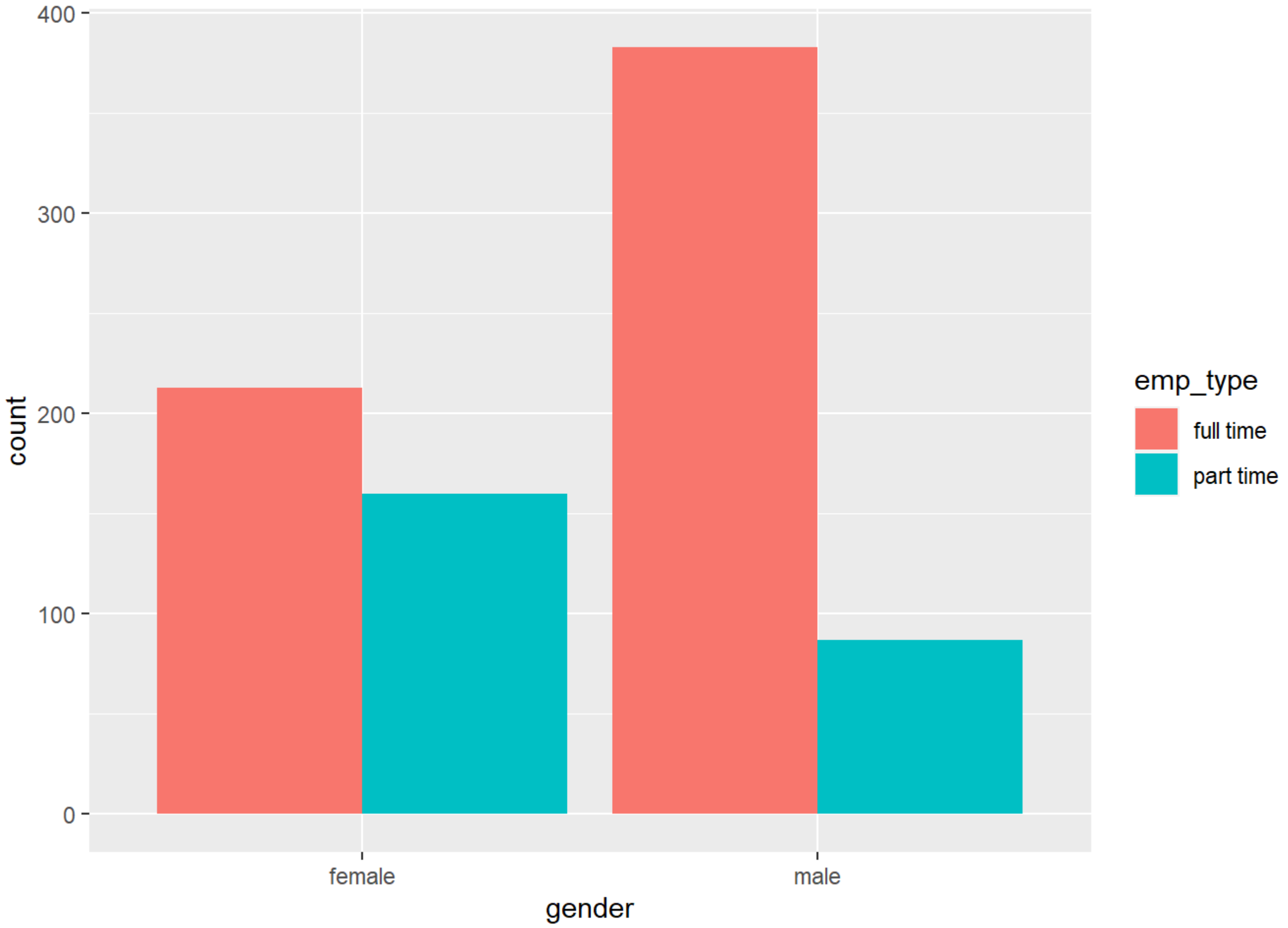
Exercise 8:

```
acs_type %>%
  group_by(emp_type, gender) %>%
  summarise(n = n()) %>%
  group_by(gender)%>%
  mutate(prop=n/sum(n))
```

```
## `summarise()` has grouped output by 'emp_type'. You can override using the `.groups` argumen
t.
```

```
## # A tibble: 4 x 4
## # Groups:   gender [2]
##   emp_type  gender     n  prop
##   <chr>     <chr>  <int> <dbl>
## 1 full time female   213 0.571
## 2 full time male     383 0.815
## 3 part time female   160 0.429
## 4 part time male      87 0.185
```

```
#ggplot(acs_type, aes(x = emp_type, fill = gender)) +
#  geom_bar()


ggplot(acs_type, aes(x = gender, fill = emp_type)) +
  geom_bar(position = "dodge")
```



women are more heavily represented among part time employees

More Practice

Exercise 9:

```
acs_emp_full_time <- acs_type %>%
  filter(emp_type == "full time")

acs_emp_part_time <- acs_type %>%
  filter(emp_type == "part time")
```

Exercise 10:

```
t.test(income~gender, data = acs_emp_full_time)
```

```
##
##  Welch Two Sample t-test
##
## data:  income by gender
## t = -5.4822, df = 590.5, p-value = 6.232e-08
## alternative hypothesis: true difference in means between group female and group male is not e
qual to 0
## 95 percent confidence interval:
##  -31825.72 -15037.15
## sample estimates:
## mean in group female   mean in group male
##             39752.11             63183.55
```

H0: there is no difference in average incomes of full time male and female employees HA: there is a difference in average incomes of full time male and female employees

We can reject the null hypothesis because the P-value is less than 0.05. and the confidence interval is (-31825.72, -15037.15) which does not include 0.

Exercise 11:

```
t.test(income~gender, data = acs_emp_part_time)
```

```
##
##  Welch Two Sample t-test
##
## data:  income by gender
## t = -1.3249, df = 97.058, p-value = 0.1883
## alternative hypothesis: true difference in means between group female and group male is not e
qual to 0
## 95 percent confidence interval:
##  -21264.16   4239.58
## sample estimates:
## mean in group female   mean in group male
##             15254.38             23766.67
```

H0: there is no difference in average incomes of part time male and female employees HA: there is a difference in average incomes of part time male and female employees

We do not have enough evidence to reject the null hypothesis because the P-value is greater than 0.05. and the confidence interval is (-21264.16 , 4239.58) which includes 0.

Exercise 12:

Working full-time/part-time can be a confounding variable in the relationship between gender and income, because men and women who are full-time employees had the difference between their income, but when they were both part-time employees, the difference between their income was not clear or at least we did not have enough evidence to prove that difference.

Exercise 13:

```
t.test(age~citizen, data = acs_emp)
```

```
##
##  Welch Two Sample t-test
##
## data:  age by citizen
## t = -4.0015, df = 58.997, p-value = 0.0001781
## alternative hypothesis: true difference in means between group no and group yes is not equal
to 0
## 95 percent confidence interval:
##  -11.006609  -3.668227
## sample estimates:
##  mean in group no mean in group yes
##          36.17647          43.51389
```

I want to know whether there is a difference between the average age of the employed citizens vs employed not-citizens H0: there is no difference between the mean of the age of the employed citizens and the mean of the age of the employed not-citizens HA: there is difference between the mean of the age of the employed citizens and the mean of the age of the employed not-citizens Numerical : age Categorial: citizenship of employed people

We can reject the null hypothesis because the P-value is less than 0.05. and the confidence interval is (-11.006609, -3.668227) which does not includes 0. Considering the means of ages in both groups we can see that no-citizen group are younger than citizen group.