Quant Assignment 2

Kristy Henrich

9/13/2020

Assignment setup

Load libraries for assignment

```
library(tidyverse)
library(ggplot2)
library(ggthemes)
```

Read csv file from assignment 1

```
tract_data <- read_csv("bayareaCA2018.csv")
```

```
## Warning: Missing column names filled in: 'X1' [1]
```

```
## Parsed with column specification:
## cols(
##   X1 = col_double(),
##   county = col_character(),
##   maj_re = col_character(),
##   maj_units_rented = col_logical(),
##   pct_no_internet = col_double(),
##   med_incomeE = col_double(),
##   avg_hh_sizeE = col_double()
## )
```

Assignment instructions & variables

Instructions: 1. Continuous variables: calculate sample mean, sample standard deviation, 95-percent confidence interval for population mean, and interquartile range. Create a histogram to illustrate each variable's distribution. 2. Categorical variables: calculate 95-percent confidence interval for proportion of population in each category.

Variables at census tract level: 1. Categorical 1. Bay Area county : county 2. Majority race / ethnicity of the population : maj_re 3. Whether the majority of the units are occupied by renters vs. owners : unit_type 2. Continuous 1. Percent of population that has no internet access : no_internet 2. Median income : med_income 3. Average household size : avg_hh_size

Distribution of continuous variables

Descriptive statistics

Calculate summary statistics for each continuous variable

```
no_internet_summary <-  summary(tract_data$pct_no_internet)
no_internet_sd <- sd(tract_data$pct_no_internet, na.rm = TRUE)
no_internet_conf_int <- t.test(tract_data$pct_no_internet)

paste("The sample mean for percent of households without internet access is",
      prettyNum(no_internet_summary["Mean"]*100, digits=3),"%")
```

```
## [1] "The sample mean for percent of households without internet access is 8.87 %"
```

```
paste("The sample median for percent of households without internet access is",
      prettyNum(no_internet_summary["Median"]*100, digits=3),"%")
```

```
## [1] "The sample median for percent of households without internet access is 7.23 %"
```

```
paste("The sample standard deviation for percent of households without internet access is",
      prettyNum(no_internet_sd*100, digits=3),"%")
```

```
## [1] "The sample standard deviation for percent of households without internet access is 6.65
%"
```

```
paste("The sample interquartile range for percent of houesholds without internet access is",
      prettyNum(no_internet_summary["1st Qu."]*100, digits=3), "% to",
      prettyNum(no_internet_summary["3rd Qu."]*100, digits=3), "%")
```

```
## [1] "The sample interquartile range for percent of houesholds without internet access is 4.24
% to 11.8 %"
```

```
income_summary <- summary(tract_data$med_incomeE)
income_sd <- sd(tract_data$med_incomeE, na.rm = TRUE)
income_conf_int <- t.test(tract_data$med_incomeE)

paste("The sample mean for median income is $",
      prettyNum(income_summary["Mean"], big.mark = ",", digits=3))
```

```
## [1] "The sample mean for median income is $ 47,926"
```

```
paste("The sample median for median income is $",
      prettyNum(income_summary["Median"], big.mark = ",", digits=3))
```

```
## [1] "The sample median for median income is $ 43,721"
```

```
paste("The sample standard deviation for median income is $",
      prettyNum(income_sd, big.mark = ",", digits = 3))
```

```
## [1] "The sample standard deviation for median income is $ 20,757"
```

```
paste("The sample interquartile range for median income is $",
      prettyNum(income_summary["1st Qu."], big.mark = ",", digits=5),"to $",
      prettyNum(income_summary["3rd Qu."], big.mark = ",", digits=5))
```

```
## [1] "The sample interquartile range for median income is $ 31,824 to $ 60,338"
```

```
hh_size_summary <- summary(tract_data$avg_hh_sizeE)
hh_size_sd <- sd(tract_data$avg_hh_sizeE, na.rm = TRUE)
hh_size_conf_int <- t.test(tract_data$avg_hh_sizeE)

paste("The sample mean for average household size is",
      prettyNum(hh_size_summary["Mean"], digits=3), "people")
```

```
## [1] "The sample mean for average household size is 2.83 people"
```

```
paste("The sample median for average household size is",
      prettyNum(hh_size_summary["Median"], digits=3),"people")
```

```
## [1] "The sample median for average household size is 2.78 people"
```

```
paste("The sample standard deviation for average household size is",
      prettyNum(hh_size_sd, digits = 3),"people")
```

```
## [1] "The sample standard deviation for average household size is 0.604 people"
```

```
paste("The sample interquartile range for average household size is",
      prettyNum(hh_size_summary["1st Qu."], digits=3), "people to",
      prettyNum(hh_size_summary["3rd Qu."], digits=3), "people")
```

```
## [1] "The sample interquartile range for average household size is 2.45 people to 3.17 people"
```

Display in formatted table

```
summary_table <- tibble(
  Statistic = c("Sample mean",
                "Median",
                "Standard deviation",
                "Interquartile range",
                "Population mean (95% confidence)"),
  `Households without internet (%)` = c(paste(prettyNum(no_internet_summary["Mean"]*100, digits=
3),"%"),
  paste(prettyNum(no_internet_summary["Median"]*100, digits=3),"%"),
  paste(prettyNum(no_internet_sd*100, digits=3),"%"),
  paste(prettyNum(no_internet_summary["1st Qu."]*100, digits=3), "% to",
  prettyNum(no_internet_summary["3rd Qu."]*100, digits=3), "%"),
  paste(prettyNum(no_internet_conf_int$conf.int[1]*100, digits=3), "% to",
        prettyNum(no_internet_conf_int$conf.int[2]*100, digits=3), "%")),
  `Median income ($)` =
c(prettyNum(income_summary["Mean"], big.mark = ",", digits=3),
  prettyNum(income_summary["Median"], big.mark = ",", digits=3),
  prettyNum(income_sd, big.mark = ",", digits = 3),
  paste(prettyNum(income_summary["1st Qu."], big.mark = ",", digits=5),"to",
        prettyNum(income_summary["3rd Qu."], big.mark = ",", digits=5)),
  paste(prettyNum(income_conf_int$conf.int[1], big.mark = ",", digits=3), "to",
        prettyNum(income_conf_int$conf.int[2], big.mark = ",", digits=3))),
  `Average household size (people)` =
c(prettyNum(hh_size_summary["Mean"], digits=3),
  prettyNum(hh_size_summary["Median"], digits=3),
  prettyNum(hh_size_sd, digits = 3),
  paste(hh_size_summary["1st Qu."],"to", hh_size_summary["3rd Qu."]),
  paste(prettyNum(hh_size_conf_int$conf.int[1], digits=3), "to",
        prettyNum(hh_size_conf_int$conf.int[2], digits=3))))

knitr::kable(summary_table, caption = "Characteristics of census tracts in nine-county Bay Area"
)
```

Characteristics of census tracts in nine-county Bay Area

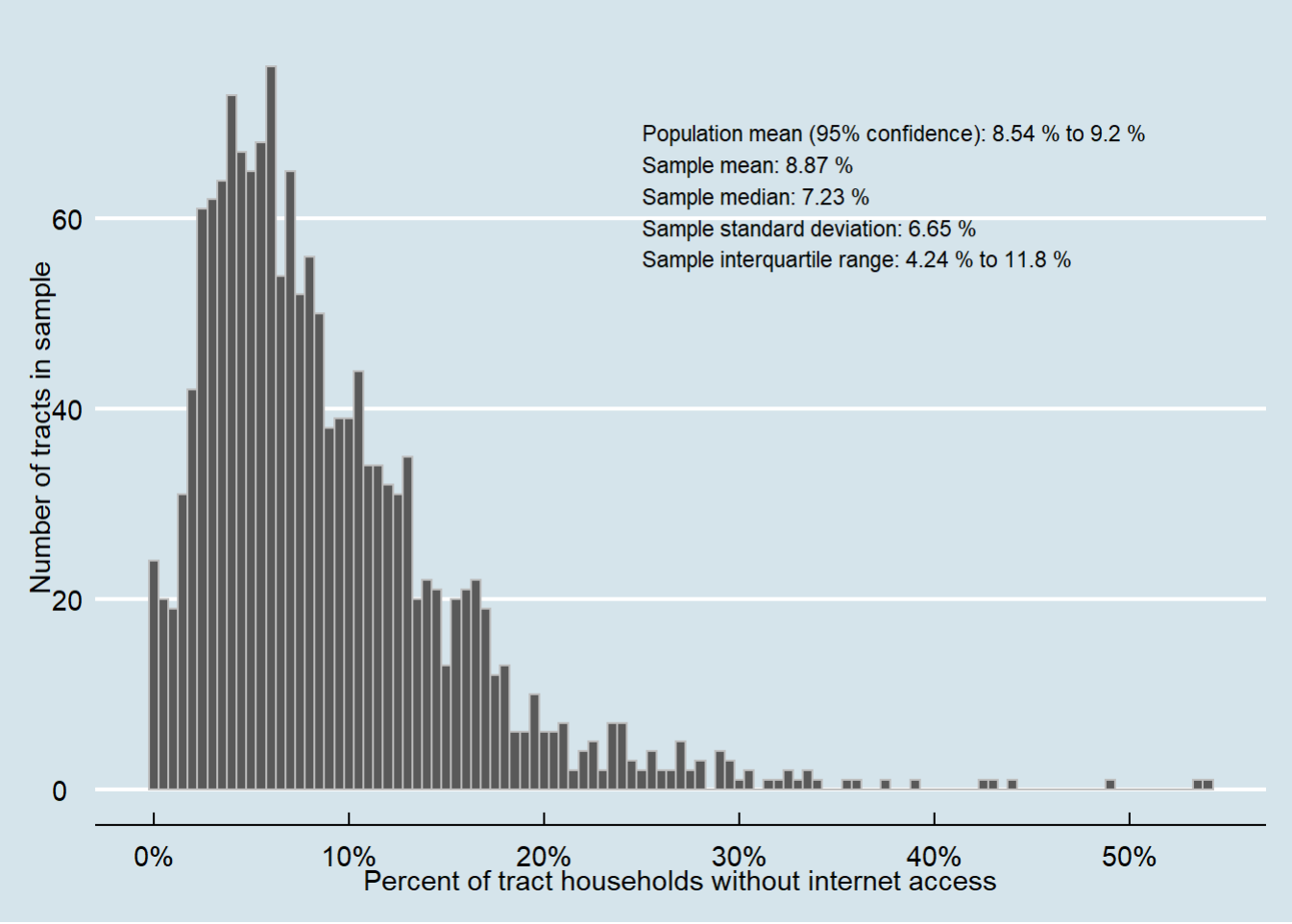
Statistic	Households without internet (%)	Median income ($)	Average household size (people)
Sample mean	8.87 %	47,926	2.83
Median	7.23 %	43,721	2.78
Standard deviation	6.65 %	20,757	0.604
Interquartile range	4.24 % to 11.8 %	31,824 to 60,338	2.45 to 3.17
Population mean (95% confidence)	8.54 % to 9.2 %	46,901 to 48,951	2.8 to 2.86

Histograms

Create histograms for each of the three distributions - 1. percent of households without internet access 2. median income 3. average household size

Percent of households without internet access

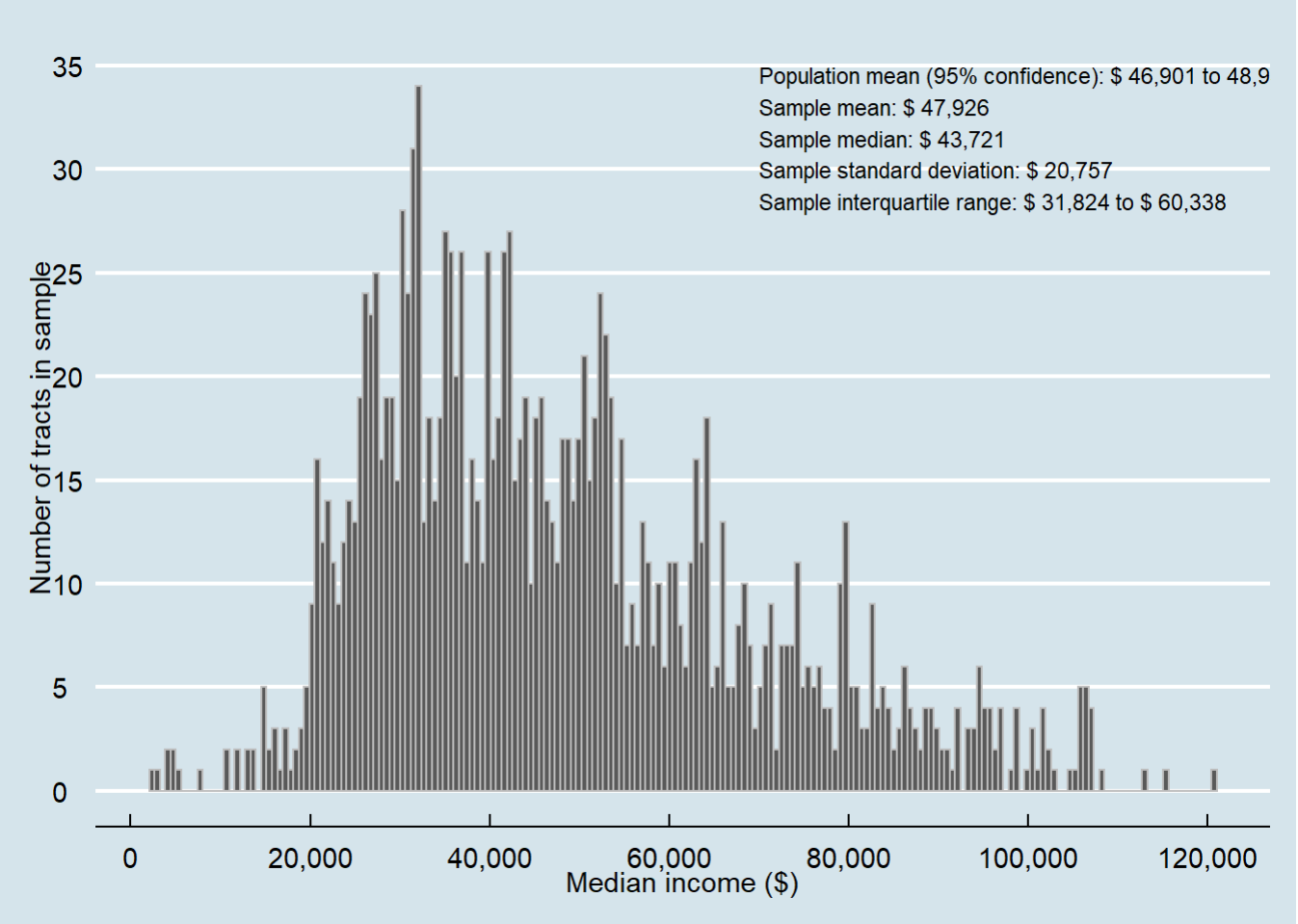
```
ggplot(tract_data, aes(x = pct_no_internet)) +
  geom_histogram(color = "gray", binwidth = 0.005) +
  theme_economist() +
  scale_x_continuous(name = "Percent of tract households without internet access",
                     breaks = breaks <- seq(0, 1, by = .1),
                     labels = paste(breaks*100, "%", sep = "")) +
  scale_y_continuous(name = "Number of tracts in sample") +
  annotate("text", x = .25, y = 70, size = 3,
           label = paste("Population mean (95% confidence):",
                         prettyNum(no_internet_conf_int$conf.int[1]*100, digits=3), "% to",
                         prettyNum(no_internet_conf_int$conf.int[2]*100, digits=3), "%",
                         "\nSample mean:",
                         prettyNum(no_internet_summary["Mean"]*100, digits=3),"%",
                         "\nSample median:",
                         prettyNum(no_internet_summary["Median"]*100, digits=3),"%",
                         "\nSample standard deviation:",
                         prettyNum(no_internet_sd*100, digits=3),"%",
                         "\nSample interquartile range:",
                         prettyNum(no_internet_summary["1st Qu."]*100, digits=3), "% to",
                         prettyNum(no_internet_summary["3rd Qu."]*100, digits=3), "%"),
           hjust = 0, vjust = 1)
```



The distribution of percent of tract households without internet access is pretty normal with a skew to the right. The median value is lower than the mean as there are some outliers with a higher percent of households without internet access - thus skewing the data to the right.

Median income

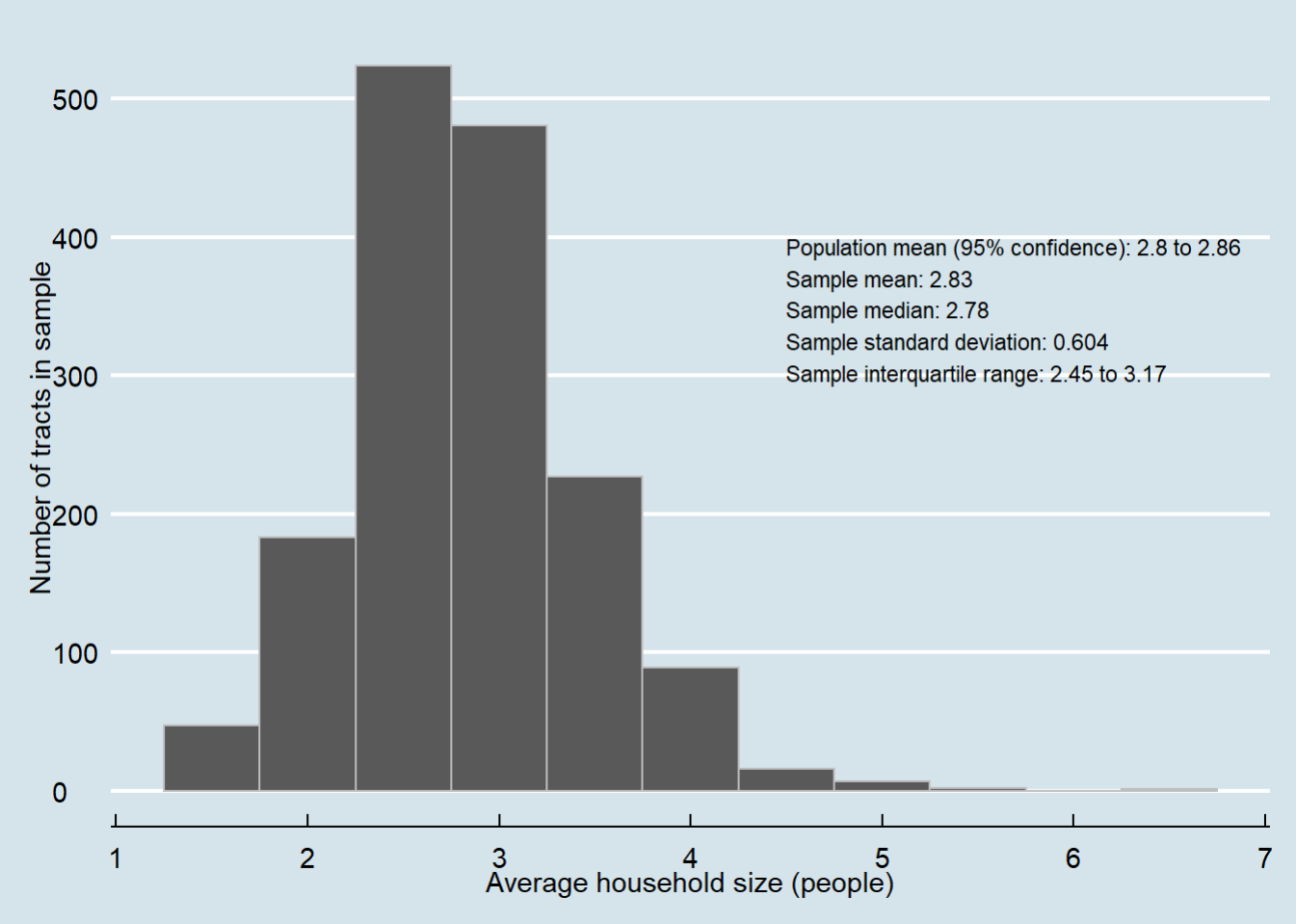
```
options(scipen = 999)
ggplot(tract_data, aes(x = med_incomeE)) +
  geom_histogram(color = "gray", bins = 200) +
  theme_economist() +
  scale_x_continuous(name = "Median income ($)",
                     breaks = breaks <- seq(0, 120000, by = 20000),
                     labels = scales::comma) +
  scale_y_continuous(name = "Number of tracts in sample",
                     breaks = breaks <- seq(0, 40, by = 5),
                     labels = paste(breaks)) +
  annotate("text", x = 70000, y = 35, size = 3,
           label = paste("Population mean (95% confidence): $",
                         prettyNum(income_conf_int$conf.int[1], big.mark = ",", digits=3), "to",
                         prettyNum(income_conf_int$conf.int[2], big.mark = ",", digits=3),
                         "\nSample mean: $",
                         prettyNum(income_summary["Mean"], big.mark = ",", digits=3),
                         "\nSample median: $",
                         prettyNum(income_summary["Median"], big.mark = ",", digits=3),
                         "\nSample standard deviation: $",
                         prettyNum(income_sd, big.mark = ",", digits = 3),
                         "\nSample interquartile range: $",
                         prettyNum(income_summary["1st Qu."], big.mark = ",", digits=5),"to $",
                         prettyNum(income_summary["3rd Qu."], big.mark = ",", digits=5)),
           hjust = 0, vjust = 1)
```



This distribution also feels fairly normal with a skew to the right. There are many more values to the right in this distribution than the one above for percent of households without internet access, driving the mean up from the median by over $4,000.

Average household size

```
ggplot(tract_data, aes(x = avg_hh_sizeE)) +
  geom_histogram(color = "gray", binwidth = .5) +
  theme_economist() +
  scale_x_continuous(name = "Average household size (people)",
                     breaks = breaks <- seq(0, 8, by = 1),
                     labels = paste(breaks)) +
  scale_y_continuous(name = "Number of tracts in sample") +
  annotate("text", x = 4.5, y = 400, size = 3,
           label = paste("Population mean (95% confidence):",
                         prettyNum(hh_size_conf_int$conf.int[1], digits=3), "to",
                         prettyNum(hh_size_conf_int$conf.int[2], digits=3),
                         "\nSample mean:",
                         prettyNum(hh_size_summary["Mean"], digits=3),
                         "\nSample median:",
                         prettyNum(hh_size_summary["Median"], digits=3),
                         "\nSample standard deviation:",
                         prettyNum(hh_size_sd, digits = 3),
                         "\nSample interquartile range:",
                         prettyNum(hh_size_summary["1st Qu."], digits=5),"to",
                         prettyNum(hh_size_summary["3rd Qu."], digits=5)),
           hjust = 0, vjust = 1)
```



The histogram of average household size looks like a negative binomial distribution with the median just below the mean. There are a few outliers past 4 that slightly bring the average up but not substantially.

Distribution of categorial variables

Show three tables, one for the proportions of each category of each categorical variable 1. Bay Area county 2. Majority race / ethnicity of the population 3. Whether the majority of the units are occupied by renters vs. owners

Bay Area county

T-test for each individual Bay Area county.

```
alameda <- t.test(tract_data$county == "Alameda")
contra_costa <- t.test(tract_data$county == "ContraCosta")
marin <- t.test(tract_data$county == "Marin")
napa <- t.test(tract_data$county == "Napa")
san_francisco <- t.test(tract_data$county == "San Francisco")
san_mateo <- t.test(tract_data$county == "San Mateo")
santa_clara <- t.test(tract_data$county == "Santa Clara")
solano <- t.test(tract_data$county == "Solano")
sonoma <- t.test(tract_data$county == "Sonoma")
```

Display the results in a formatted table

```
county <- tibble(`County` = c("Alameda", "ContraCosta", "Marin", "Napa",
                              "San Francisco", "San Mateo",
                              "Santa Clara", "Solano", "Sonoma"),
                 `Sample proportion` =
                  c(paste(prettyNum(alameda$estimate*100,
                                    digits = 3), "%"),
                    paste(prettyNum(contra_costa$estimate*100,
                                    digits = 3), "%"),
                    paste(prettyNum(marin$estimate*100,
                                    digits = 3), "%"),
                    paste(prettyNum(napa$estimate*100,
                                    digits = 3), "%"),
                    paste(prettyNum(san_francisco$estimate*100,
                                    digits = 3), "%"),
                    paste(prettyNum(san_mateo$estimate*100,
                                    digits = 3), "%"),
                    paste(prettyNum(santa_clara$estimate*100,
                                    digits = 3), "%"),
                    paste(prettyNum(solano$estimate*100,
                                    digits = 3), "%"),
                    paste(prettyNum(sonoma$estimate*100,
                                    digits = 3), "%")),
                 `95-percent confidence interval` =
                  c(paste(prettyNum(alameda$conf.int[1]*100,
                                    digits = 3), "% to ",
                          prettyNum(alameda$conf.int[2]*100,
                                    digits = 3), "%"),
                    paste(prettyNum(contra_costa$conf.int[1]*100,
                                    digits = 3), "% to ",
                          prettyNum(contra_costa$conf.int[2]*100,
                                    digits = 3), "%"),
                    paste(prettyNum(marin$conf.int[1]*100,
                                    digits = 3), "% to ",
                          prettyNum(marin$conf.int[2]*100,
                                    digits = 3), "%"),
                    paste(prettyNum(napa$conf.int[1]*100,
                                    digits = 3), "% to ",
                          prettyNum(napa$conf.int[2]*100,
                                    digits = 3), "%"),
                    paste(prettyNum(san_francisco$conf.int[1]*100,
                                    digits = 3), "% to ",
                          prettyNum(san_francisco$conf.int[2]*100,
                                    digits = 3), "%"),
                    paste(prettyNum(san_mateo$conf.int[1]*100,
                                    digits = 3), "% to ",
                          prettyNum(san_mateo$conf.int[2]*100,
                                    digits = 3), "%"),
                    paste(prettyNum(santa_clara$conf.int[1]*100,
                                    digits = 3), "% to ",
                          prettyNum(santa_clara$conf.int[2]*100,
                                    digits = 3), "%"),
                    paste(prettyNum(solano$conf.int[1]*100,
                                    digits = 3), "% to ",
```

```
                                  prettyNum(solano$conf.int[2]*100,
                                            digits = 3), "%"),
                            paste(prettyNum(sonoma$conf.int[1]*100,
                                            digits = 3), "% to ",
                                  prettyNum(sonoma$conf.int[2]*100,
                                            digits = 3), "%")))
knitr::kable(county, caption = "Census tracts in the nine-county Bay Area")
```

Census tracts in the nine-county Bay Area

County	Sample proportion	95-percent confidence interval
Alameda	22.7 %	20.7 % to 24.8 %
ContraCosta	13.1 %	11.4 % to 14.8 %
Marin	3.53 %	2.62 % to 4.43 %
Napa	2.52 %	1.75 % to 3.29 %
San Francisco	12.4 %	10.8 % to 14 %
San Mateo	9.95 %	8.48 % to 11.4 %
Santa Clara	23.4 %	21.3 % to 25.5 %
Solano	6.05 %	4.87 % to 7.22 %
Sonoma	6.3 %	5.1 % to 7.49 %

Race/Ethnicity

T-test for each individual race/ethnicity category

```
nh_black <- t.test (tract_data$maj_re == "nh_black")
hs_latino <- t.test (tract_data$maj_re == "hisp_latino")
no_majority <- t.test (tract_data$maj_re == "no_maj_re")
nh_white <- t.test (tract_data$maj_re == "nh_white")
nh_asian <- t.test (tract_data$maj_re == "nh_asian")
native_am <- t.test (tract_data$maj_re == "native_am")
hawaiian_PI <- t.test (tract_data$maj_re == "hawaiian_PI")
multi <- t.test (tract_data$maj_re == "multi")
other_nh <- t.test (tract_data$maj_re == "other_nh")
```

Create a formatted table with proportions and confidence intervals

```
race_shares <- tibble(`Majority Race` = c("Hispanic/Latino",
                                  "White (not Hispanic/Latino)",
                                  "No Majority",
                                  "Asian (not Hispanic/Latino)",
                                  "Black (not Hispanic/Latino)",
                                  "Native American (not Hispanic/Latino)",
                                  "Hawaiian/Pacific Islander (not Hispanic/Latino)",
                                  "Multiple races",
                                  "Other (not Hispanic/Latino)"),
                `Sample proportion` = c(paste(prettyNum(hs_latino$estimate*100, digits = 3),
"%"),
  paste(prettyNum(nh_white$estimate*100, digits = 3), "%"),
  paste(prettyNum(no_majority$estimate*100, digits = 3), "%"),
  paste(prettyNum(nh_asian$estimate*100, digits = 3), "%"),
  paste(prettyNum(nh_black$estimate*100, digits = 3), "%"),
  paste(prettyNum(native_am$estimate*100, digits = 3), "%"),
  paste(prettyNum(hawaiian_PI$estimate*100, digits = 3), "%"),
  paste(prettyNum(multi$estimate*100, digits = 3), "%"),
  paste(prettyNum(other_nh$estimate*100, digits = 3), "%")),
                `95-percent confidence interval` =
c(paste(prettyNum(hs_latino$conf.int[1]*100, digits=3), "% to ",
        prettyNum(hs_latino$conf.int[2]*100, digits=3), "%", sep=""),
  paste(prettyNum(nh_white$conf.int[1]*100, digits=3), "% to ",
        prettyNum(nh_white$conf.int[2]*100, digits=3), "%", sep=""),
  paste(prettyNum(no_majority$conf.int[1]*100, digits=3), "% to ",
        prettyNum(no_majority$conf.int[2]*100, digits=3), "%", sep=""),
  paste(prettyNum(nh_asian$conf.int[1]*100, digits=3), "% to ",
        prettyNum(nh_asian$conf.int[2]*100, digits=3), "%", sep=""),
  paste(prettyNum(nh_black$conf.int[1]*100, digits=3), "% to ",
        prettyNum(nh_black$conf.int[2]*100, digits=3), "%", sep=""),
  paste("no data"),
  paste("no data"),
  paste("no data"),
  paste("no data")))

knitr::kable(race_shares, caption = "Census tracts in nine-county Bay Area")
```

Census tracts in nine-county Bay Area

Majority Race	Sample proportion	95-percent confidence interval
Hispanic/Latino	8.88 %	7.48% to 10.3%
White (not Hispanic/Latino)	38.8 %	36.4% to 41.2%
No Majority	39.4 %	37% to 41.8%
Asian (not Hispanic/Latino)	12.5 %	10.9% to 14.2%
Black (not Hispanic/Latino)	0.441 %	0.115% to 0.767%
Native American (not Hispanic/Latino)	0 %	no data
Hawaiian/Pacific Islander (not Hispanic/Latino)	0 %	no data

Majority Race	Sample proportion	95-percent confidence interval
Multiple races	0 %	no data
Other (not Hispanic/Latino)	0 %	no data

Units rented vs. owned

T-test for each individual rent vs. own category

```
maj_rented <- t.test(tract_data$maj_units_rented)
maj_owned <- t.test(!tract_data$maj_units_rented)
```

Display the results in a formatted table

```
for_shares <- tibble(`Majority rented vs. owned` = c("Rented","Owned"),
                     `Sample proportion` =
                       c(paste(prettyNum(maj_rented$estimate*100,
                                         digits = 3), "%"),
                         paste(prettyNum(maj_owned$estimate*100,
                                         digits = 3), "%")),
                      `95-percent confidence interval` =
                        c(paste(prettyNum(maj_rented$conf.int[1]*100,
                                          digits = 3), "% to ",
                                prettyNum(maj_rented$conf.int[2]*100,
                                          digits = 3), "%"),
                          paste(prettyNum(maj_owned$conf.int[1]*100,
                                          digits = 3), "% to ",
                                prettyNum(maj_owned$conf.int[2]*100,
                                          digits = 3), "%")))
knitr::kable(for_shares, caption = "Census tracts in nine-county Bay Area")
```

Census tracts in nine-county Bay Area

Majority rented vs. owned	Sample proportion	95-percent confidence interval
Rented	36.7 %	34.3 % to 39.1 %
Owned	63.3 %	60.9 % to 65.7 %