Correlation

Put your name here

Put the date here

Introduction

In this assignment we will learn how to run a correlation analysis. Correlation measures the strength of the linear relationship between two numerical variables.

Instructions

Presumably, you have already created a new project and downloaded this file into it. From the `Run` menu above, select `Run All` to run all existing code chunks.

When prompted to complete an exercise or demonstrate skills, you will see the following lines in the document:

ANSWER

These lines demarcate the region of the R Markdown document in which you are to show your work.

Sometimes you will be asked to add your own R code. That will appear in this document as a code chunk with a request for you to add your own code, like so:

```
# Add code here
```

Be sure to remove the line `# Add code here` when you have added your own code. You should run each new code chunk you create by clicking on the dark green arrow in the upper-right corner of the code chunk.

Sometimes you will be asked to type up your thoughts. That will appear in the document with the words, "Please write up your answer here." Be sure to remove the line "Please write up your answer here" when you have written up your answer. In these areas of the assignment, please use contextually meaningful full sentences/paragraphs (unless otherwise indicated) and proper spelling, grammar, punctuation, etc. This is not R code, but rather a free response section where you talk about your analysis and conclusions. You may need to use inline R code in these sections.

When you are finished with the assignment, knit to PDF and proofread the PDF file **carefully**. Do not download the PDF file from the PDF viewer; rather, you should export the PDF file to your computer by selecting the check box next to the PDF file in the Files pane, clicking the `More` menu, and then clicking `Export`. Submit your assignment according to your professor's instructions.

Load Packages

We load the standard `mosaic` package as well as the `reshape2` package for the `tips` data and the `OIdata` package for the `state` data. The `broom` package gives us tidy output.

```
library(reshape2)
library(OIdata)
data(state)
library(broom)
library(mosaic)
```

We set the seed to make our results reproducible.

```
set.seed(11)
```

Research question

Is there is a correlation between the size of a restaurant bill and the size of the tip?

Correlation

The word correlation describes a linear relationship between two numerical variables. As long as certain conditions are met, we can calculate a statistic called the Pearson correlation coefficient, denoted R.[1] This value will be some number between -1 and 1. Coefficients close to zero indicate little or no correlation, coefficients close to 1 indicate strong positive correlation, and coefficients close to -1 indicate strong negative correlation. In between, we often use words like weak, moderately weak, moderate, and moderately strong. There are no exact cutoffs for when such words apply. You must learn from experience how to judge scatterplots and R values to make such determinations.

Let's examine a data set called `tips` from the `reshape2` package. Since there is also a `tips` data set in the `openintro` package, we'll use a trick to make sure we get the right one. The double colon is placed between the name of the package and the name of the data frame:

```
tips <- reshape2::tips
str(tips)
```

```
## 'data.frame':    244 obs. of  7 variables:
##  $ total_bill: num  17 10.3 21 23.7 24.6 ...
##  $ tip       : num  1.01 1.66 3.5 3.31 3.61 4.71 2 3.12 1.96 3.23 ...
##  $ sex       : Factor w/ 2 levels "Female","Male": 1 2 2 2 1 2 2 2 2 2 ...
##  $ smoker    : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
##  $ day       : Factor w/ 4 levels "Fri","Sat","Sun",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ time      : Factor w/ 2 levels "Dinner","Lunch": 1 1 1 1 1 1 1 1 1 1 ...
##  $ size      : int  2 3 3 2 4 4 2 4 2 2 ...
```

These 244 observations were collected by one waiter over a period of a few months working in a restaurant. Our research question asks us to consider the variables `tip` and `total_bill`.

If all we wanted was the value of R, we could find it by using the `cor` command.

```
R <- cor(tip ~ total_bill, data = tips)
R
```

```
## [1] 0.6757341
```

Although the `cor` command accepts the "tilde" notation, the order doesn't matter; correlation is symmetric, so the R value is the same independent of the choice of response and explanatory variables.

This sample correlation R is an estimate of the true population correlation, called ρ, the Greek letter "rho". A typical null hypothesis is that there is no correlation between the two variables—in other words, $\rho = 0$. Under that assumption, the sampling distribution is somewhat complicated. Although the sample correlations don't follow a simple distribution, if we calculate

$$t = \frac{R - \rho}{\sqrt{\frac{1-R^2}{n-2}}} = \frac{R}{\sqrt{\frac{1-R^2}{n-2}}},$$

[1] In most places, the Pearson correlation coefficient is denoted by a lowercase r, but the OpenIntro book uses an uppercase R.

then the values of t follow a Student t distribution with $n-2$ degrees of freedom. (The last step above takes into account the fact that the null value for ρ is zero.)

We can verify this with a basic simulation. First, we shuffle the values of `total_bill` to remove any association with tips to simulate the assumption of the null hypothesis. Here are a few examples:

```
cor(tip ~ shuffle(total_bill), data = tips)
```

```
## [1] 0.005853122
```

```
cor(tip ~ shuffle(total_bill), data = tips)
```

```
## [1] 0.007911685
```

```
cor(tip ~ shuffle(total_bill), data = tips)
```

```
## [1] 0.1027245
```

We use the `do` command to do this a bunch of times.

```
sims <- do(2000) * cor(tip ~ shuffle(total_bill), data = tips)
tail(sims)
```

```
##              cor
## 1995 -0.02872252
## 1996 -0.03308273
## 1997 -0.05312519
## 1998 -0.07350161
## 1999  0.05731321
## 2000  0.07000983
```

The t scores follow a Student t distribution, not the correlations themselves, so we have to calculate the t scores. We use the `mutate` command[2] to compute the t score for each row of `sims`. The number 242 is $n-2$.

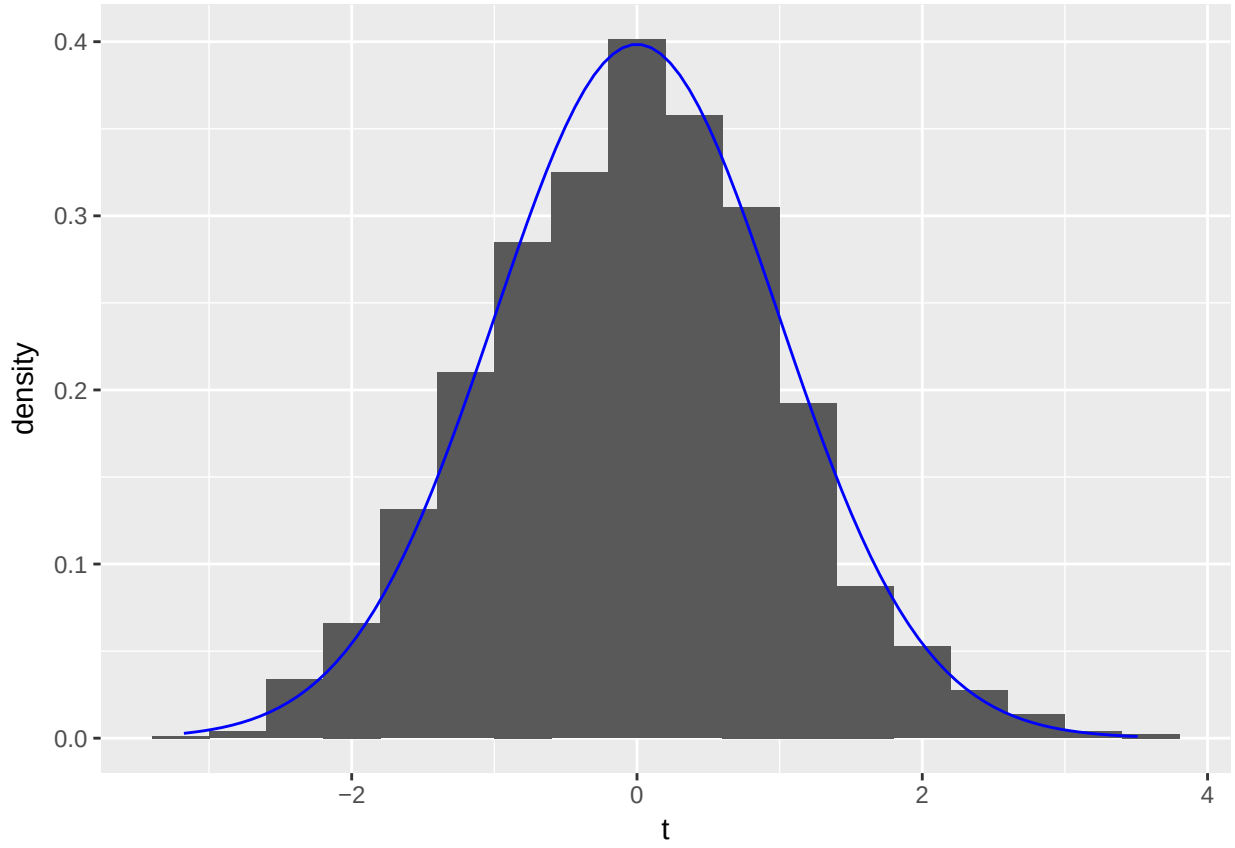
```
sims <- mutate(sims, t = cor/(sqrt((1 - cor^2)/242)))
tail(sims)
```

```
##              cor          t
## 1995 -0.02872252 -0.4470020
## 1996 -0.03308273 -0.5149284
## 1997 -0.05312519 -0.8276027
## 1998 -0.07350161 -1.1465179
## 1999  0.05731321  0.8930522
## 2000  0.07000983  1.0917763
```

Now we can graph the simulated values. We superimpose the t distribution with $df = 242$ to show that it's a pretty good fit.

```
# Don't worry about the syntax here.
# You won't need to know how to do this on your own.
ggplot(sims, aes(x = t)) +
    geom_histogram(aes(y = ..density..), binwidth = 0.4) +
    stat_function(fun = "dt", args = list(df = 242), color = "blue")
```

[2]The `mutate` command comes from a package called `dplyr` which is helpfully loaded for you when you load the `mosaic` package.



Inference for correlation

Calculating a correlation coefficient blindly can be dangerous. Without meeting certain conditions, the value of R could be incredibly misleading. R (the software) will gladly compute R (the correlation coefficient) for any data, whether appropriate or not. Therefore, we will follow our inferential rubric to decide if there is a statistically significant relationship between the tip and the corresponding bill. In truth, the entire inferential rubric is probably overkill for such a simple question. Nevertheless, the rubric does ensure that we take care to identify our hypotheses and check conditions.

In addition to the standard "Random" and "10%" conditions, we introduce two new conditions. First, we need to know that the association is linear. Nonlinear relationships can exist, but the R value makes no sense for such situations. Finally, we need to check for outliers. These two conditions should be checked by looking at a scatterplot.

Exploratory data analysis

Use data documentation (help files, code books, Google, etc.), the View command, the str command, and other summary functions to understand the data.

[You should type `?tips` at the Console to read the help file and use `View` to look at the spreadsheet view of the data.]

```
str(tips)
```

```
## 'data.frame':    244 obs. of  7 variables:
```

```
##  $ total_bill: num  17 10.3 21 23.7 24.6 ...
##  $ tip       : num  1.01 1.66 3.5 3.31 3.61 4.71 2 3.12 1.96 3.23 ...
##  $ sex       : Factor w/ 2 levels "Female","Male": 1 2 2 2 1 2 2 2 2 2 ...
##  $ smoker    : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
##  $ day       : Factor w/ 4 levels "Fri","Sat","Sun",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ time      : Factor w/ 2 levels "Dinner","Lunch": 1 1 1 1 1 1 1 1 1 1 ...
##  $ size      : int  2 3 3 2 4 4 2 4 2 2 ...
```

```
head(tips)
```

```
##   total_bill  tip    sex smoker day   time size
## 1      16.99 1.01 Female     No Sun Dinner    2
## 2      10.34 1.66   Male     No Sun Dinner    3
## 3      21.01 3.50   Male     No Sun Dinner    3
## 4      23.68 3.31   Male     No Sun Dinner    2
## 5      24.59 3.61 Female     No Sun Dinner    4
## 6      25.29 4.71   Male     No Sun Dinner    4
```

We can also look at each numerical variable individually with `favstats`:

```
favstats(tips$tip)
```

```
##  min Q1 median     Q3 max     mean       sd   n missing
##    1  2    2.9 3.5625  10 2.998279 1.383638 244       0
```

```
favstats(tips$total_bill)
```

```
##   min      Q1 median      Q3   max     mean       sd   n missing
##  3.07 13.3475 17.795 24.1275 50.81 19.78594 8.902412 244       0
```

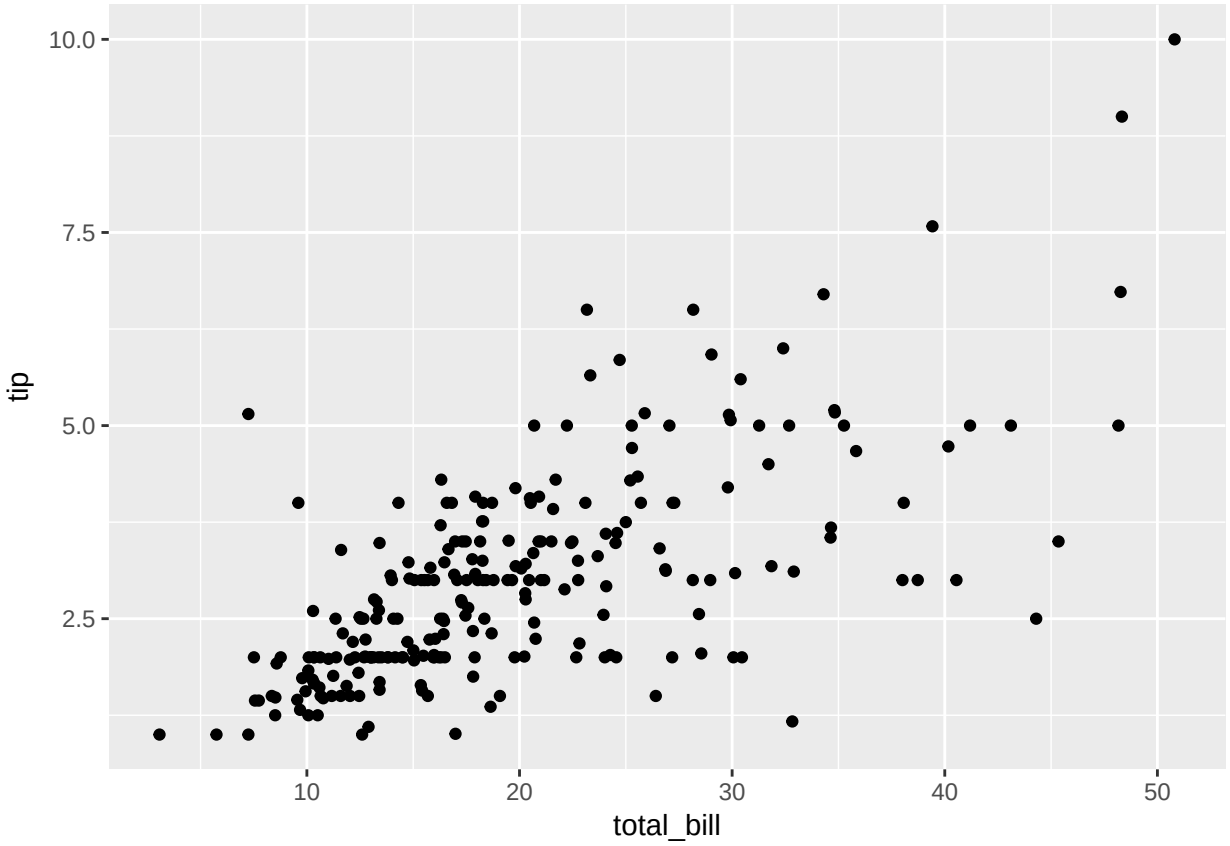
Prepare the data for analysis. [Not always necessary.]

The two variables of interest are coded correctly as numerical variables in the `tips` data frame, so we don't need to do anything for this step.

Make tables or plots to explore the data visually.

The appropriate plot for two numerical variables is a scatterplot. We are thinking of `tip` as the response variable and `total_bill` as the explanatory variable.

```
ggplot(tips, aes(y = tip, x = total_bill)) +
    geom_point()
```



There does appear to be a moderate, positive association between these variables.

Exercise

There appears to be a pattern of unusual bunching in the lower left part of the graph. Can you explain this pattern? (Hint: in the spreadsheet view produced by `View`, sort the `tip` variable.)

ANSWER

Please write up your answer here.

Hypotheses

Identify the sample (or samples) and a reasonable population (or populations) of interest.

The sample consists of 244 meals a waiter served over the course of several months working at a restaurant. The population is presumably all meals this waiter might ever serve at this restaurant. (It would not make sense to include other servers or other restaurants in this population as bills and tips vary widely from person to person and restaurant to restaurant.)

Express the null and alternative hypotheses as contextually meaningful full sentences.

H_0 : There is no correlation between the tip and the total bill.

H_A : There is a correlation between the tip and the total bill.

Express the null and alternative hypotheses in symbols (when possible).

$H_0 : \rho = 0$

$H_A : \rho \neq 0$.

Commentary: We are performing a two-sided test here. One could perform a one-sided test if the question of interest was about a positive or a negative correlation specifically. Unless otherwise specified, though, the default is to run a two-sided test.

Model

Identify the sampling distribution model.

We use a t model with 242 degrees of freedom.

Check the relevant conditions to ensure that model assumptions are met.

- Random
  - This is not a random sample, but over several months, it seems reasonable that this is representative of this waiter's experiences at this restaurant.
- 10%
  - Assuming the waiter works at this restaurant for several years, 244 meals is probably less than 10% of all meals he will serve.
- Linear association
  - The scatterplot shows a reasonably linear pattern.
- Outliers
  - We don't see any significant outliers in the scatterplot. There are a few dots here and there that are a little far from the main cloud, but nothing that worries us too much, especially given the large sample size.

Commentary: No data will ever line up in a perfect straight line. The "linear association" condition is meant to suggest that the "cloud of dots" should be more or less in a straight pattern moving across the plot. We are most concerned here with checking that the pattern does not curve substantially, and this does not appear to. As with any outliers, judge them based on how far away from the data cloud they are, and keep in mind that outliers tend to be more influential when sample sizes are small.

Mechanics

Compute and report the test statistic.

```
tips_test <- cor.test(tip ~ total_bill, data = tips)
tips_test_tidy <- tidy(tips_test)
tips_test_tidy
```

```
##    estimate statistic                                        p.value parameter
## 1 0.6757341  14.26035 0.000000000000000000000000000000000006692471       242
##    conf.low conf.high                                method alternative
## 1 0.6011647 0.7386372 Pearson's product-moment correlation   two.sided
```

```
t <- tips_test_tidy$statistic
t
```

```
## [1] 14.26035
```

The t score is 14.260355.

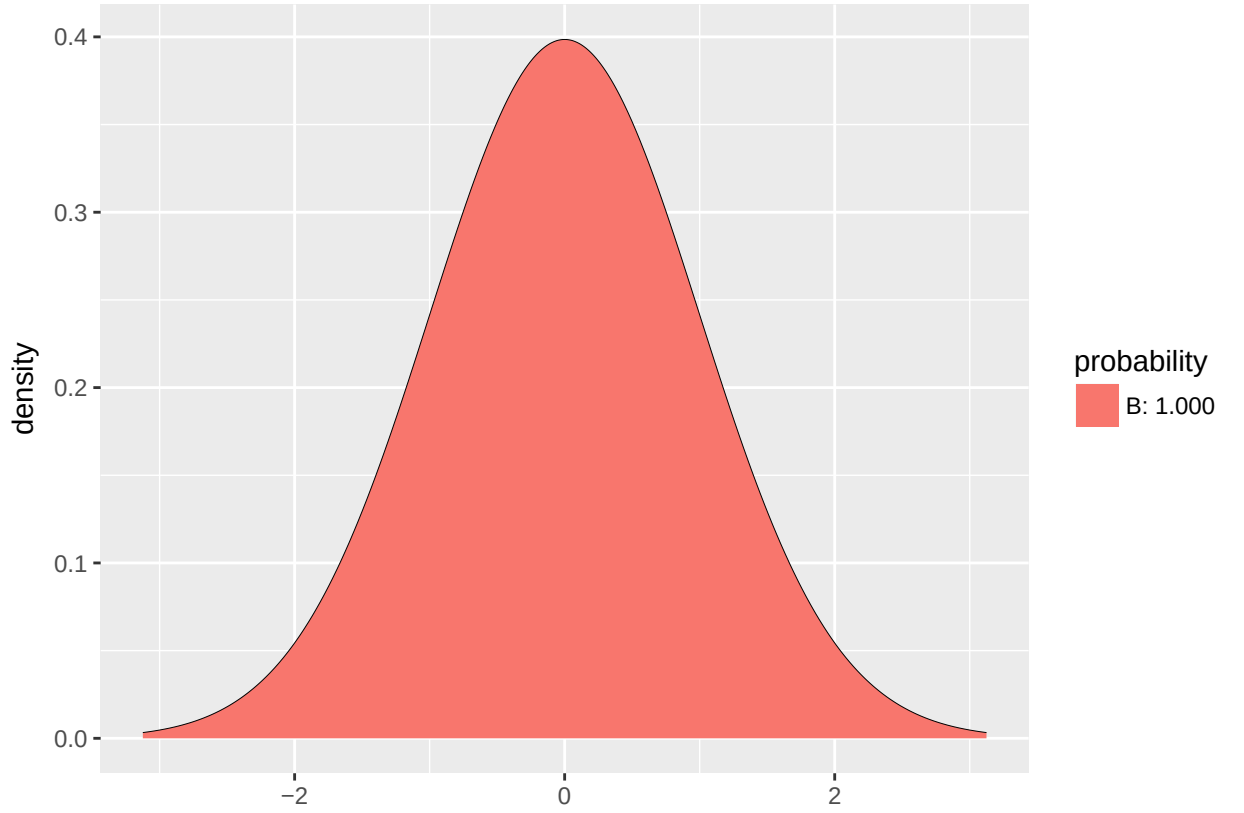
Commentary: Although `cor.test` accepts the "tilde" notation, the order doesn't matter; correlation is symmetric, so the R value is the same independent of the choice of response and explanatory variables.

The output of the `cor.test` function is very similar to other hypothesis tests in R. The correlation coefficient is the `estimate`. Everything else is straightforward: the t score (`statistic`), P-value (`p.value`), degrees of freedom (`parameter`), and confidence interval (`conf.low` and `conf.high`).

You may have noticed that this is an insanely large t score. This is typical of correlation tests. If there is enough visual evidence of a correlation in the scatterplot, the R value will be pretty far from 0. That's why the full rubric for inference is somewhat overkill for questions about correlation.

Plot the null distribution.

```
pdist("t", df = tips_test_tidy$parameter,
      q = c(-t, t),
      invisible = TRUE)
```



Calculate and report the P-value.

```
P <- tips_test_tidy$p.value
P
```

```
## [1] 0.00000000000000000000000000000000006692471
```

$P < 0.001$

Commentary: $P < 0.001$ is quite the understatement. The P-value has 33 zeros after the decimal point!

Conclusion

State the statistical conclusion.

We reject the null hypothesis.

State (but do not overstate) a contextually meaningful conclusion.

There is sufficient evidence that there is a correlation between the tip and the total bill (for this waiter at this restaurant).

Identify the possibility of either a Type I or Type II error and state what making such an error means in the context of the hypotheses.

If we've made a Type I error, then there is actually no correlation between tips and the total bill, but this data shows one.

Confidence interval

Check the relevant conditions to ensure that model assumptions are met.

All the conditions have already been checked.

Calculate the confidence interval.

```
tips_test_tidy$conf.low
```

```
## [1] 0.6011647
```

```
tips_test_tidy$conf.high
```

```
## [1] 0.7386372
```

State (but do not overstate) a contextually meaningful interpretation.

We are 95% confident that the true correlation between tips and total bill (for this waiter at this restaurant) is captured in the interval (0.6011647, 0.7386372).

Your turn

The `state` data from the `OIdata` package has a number of variables collected from various sources. There are 51 rows, representing the 50 states and the District of Columbia. Run a correlation test to determine if the percentage of the state's population that smokes is correlated with the median household income in each state.

There is something unusual about this example that you will need to consider in your answer. The sample is 50 states and the District of Columbia. The population is tricky though because these states don't represent some larger groups of states; we already have all the states in our data. One can think of this data, though, as a snapshot of what was true in each state at one point in time. Therefore, the population can be thought of as similar measurements taken at other times.

This also makes it difficult to check conditions. We do not have a random sample of states (as we have all of them), but remember that we're thinking of this as a random sample across a number of years in which we might have gathered this data. Having said that, I imagine that median income goes up every year with inflation, so it may or may not be representative of other years. The 10% condition also requires some thought.

The rubric outline is reproduced below. You may refer to the worked example above and modify it accordingly. Remember to strip out all the commentary. That is just exposition for your benefit in understanding the steps, but is not meant to form part of the formal inference process.

Another word of warning: the copy/paste process is not a substitute for your brain. You will often need to modify more than just the names of the data frames and variables to adapt the worked examples to your own work. Do not blindly copy and paste code without understanding what it does. And you should **never** copy and paste text. All the sentences and paragraphs you write are expressions of your own analysis. They must reflect your own understanding of the inferential process.

Exploratory data analysis

Use data documentation (help files, code books, Google, etc.), the View command, the str command, and other summary functions to understand the data.

ANSWER

```
# Add code here to understand the data.
```

Prepare the data for analysis. [Not always necessary.]

ANSWER

```
# Add code here to prepare the data for analysis.
```

Make tables or plots to explore the data visually.

ANSWER

```
# Add code here to make tables or plots.
```

Hypotheses

Identify the sample (or samples) and a reasonable population (or populations) of interest.

ANSWER

Please write up your answer here.

Express the null and alternative hypotheses as contextually meaningful full sentences.

ANSWER

H_0 : Null hypothesis goes here.

H_A : Alternative hypothesis goes here.

Express the null and alternative hypotheses in symbols (when possible).

ANSWER

$H_0 : math$

$H_A : math$

Model

Identify the sampling distribution model.

ANSWER

Please write up your answer here.

Check the relevant conditions to ensure that model assumptions are met.

ANSWER

Please write up your answer here. (Some conditions may require R code as well.)

Mechanics

Compute and report the test statistic.

ANSWER

```
# Add code here to compute the test statistic.
```

Please write up your answer here.

Plot the null distribution.

ANSWER

```
# Add code here to plot the null distribution.
```

Calculate and report the P-value.

ANSWER

```
# Add code here to calculate the P-value.
```

Please write up your answer here.

Conclusion

State the statistical conclusion.

ANSWER

Please write up your answer here.

State (but do not overstate) a contextually meaningful conclusion.

ANSWER

Please write up your answer here.

Identify the possibility of either a Type I or Type II error and state what making such an error means in the context of the hypotheses.

ANSWER

Please write up your answer here.

Confidence interval

Check the relevant conditions to ensure that model assumptions are met.

ANSWER

Please write up your answer here. (Some conditions may require R code as well.)

Calculate the confidence interval.

ANSWER

```
# Add code here to calculate the confidence interval.
```

State (but do not overstate) a contextually meaningful interpretation.

ANSWER

Please write up your answer here.