HW7_Madhu

Madhu Jagdale

3/29/2021

```
library(tidyverse)
```

```
-- Attaching packages --------------------------------------- tidyverse 1.3.0
--

v ggplot2 3.3.3     v purrr   0.3.4
v tibble  3.0.5     v dplyr   1.0.5
v tidyr   1.1.2     v stringr 1.4.0
v readr   1.4.0     v forcats 0.5.1

Warning: package 'dplyr' was built under R version 4.0.4

-- Conflicts ------------------------------------------ tidyverse_conflicts()
--
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
```

Learning objectives

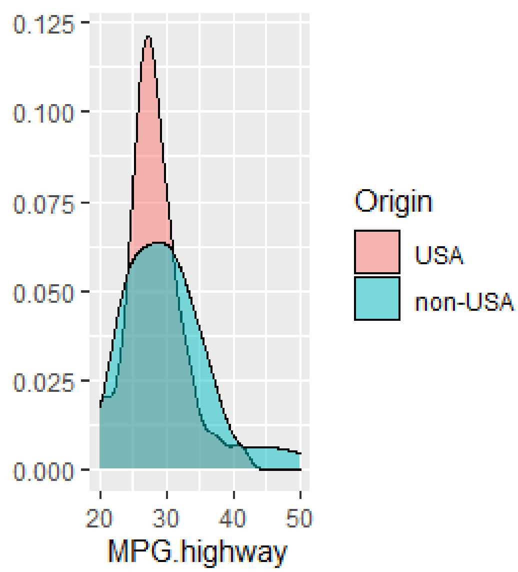
```
Cars93 <- as_tibble(MASS::Cars93)
head(Cars93)
```

```
# A tibble: 6 x 27
  Manufacturer Model Type  Min.Price Price Max.Price MPG.city MPG.highway
  <fct>        <fct> <fct>     <dbl> <dbl>     <dbl>    <int>       <int>
1 Acura        Inte~ Small      12.9  15.9      18.8       25          31
2 Acura        Lege~ Mids~      29.2  33.9      38.7       18          25
3 Audi         90    Comp~      25.9  29.1      32.3       20          26
4 Audi         100   Mids~      30.8  37.7      44.6       19          26
5 BMW          535i  Mids~      23.7  30        36.2       22          30
6 Buick        Cent~ Mids~      14.2  15.7      17.3       22          31
# ... with 19 more variables: AirBags <fct>, DriveTrain <fct>, Cylinders
<fct>,
#   EngineSize <dbl>, Horsepower <int>, RPM <int>, Rev.per.mile <int>,
#   Man.trans.avail <fct>, Fuel.tank.capacity <dbl>, Passengers <int>,
#   Length <int>, Wheelbase <int>, Width <int>, Turn.circle <int>,
#   Rear.seat.room <dbl>, Luggage.room <int>, Weight <int>, Origin <fct>,
#   Make <fct>
```

Testing means between two groups

Here is a command that generates density plots of `MPG.highway` from the Cars93 data. Separate densities are constructed for US and non-US vehicles.

```
qplot(data = Cars93, x = MPG.highway,
      fill = Origin, geom = "density", alpha = I(0.5))
```



(a) Using the Cars93 data and the `t.test()` function, run a t-test to see if average `MPG.highway` is different between US and non-US vehicles. *Interpret the results*

```
MPG.highway.t.test <- t.test(MPG.highway ~ Origin, data = Cars93)
MPG.highway.t.test


    Welch Two Sample t-test

data:  MPG.highway by Origin
t = -1.7545, df = 75.802, p-value = 0.08339
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -4.1489029  0.2627918
sample estimates:
    mean in group USA mean in group non-USA
             28.14583              30.08889
```

Try doing this both using the formula style input and the x, y style input.

```
with(Cars93, t.test(x = MPG.highway[Origin == "USA"], y = MPG.highway[Origin
== "non-USA"]))


    Welch Two Sample t-test
```

```
data:  MPG.highway[Origin == "USA"] and MPG.highway[Origin == "non-USA"]
t = -1.7545, df = 75.802, p-value = 0.08339
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -4.1489029  0.2627918
sample estimates:
mean of x mean of y
 28.14583  30.08889
```

(b) What is the confidence interval for the difference? Interpret this confidence interval.

```
MPG.highway.t.test$conf.int

[1] -4.1489029  0.2627918
attr(,"conf.level")
[1] 0.95
```

> *[Confidence interval for the difference of USA and Non USA cars is : -4.1489029 0.2627918. 95% of the time difference mean will come in the above range.]*

(c) What is the p-value of the observed data? Interpret it and draw a conclusion about the hypotheses.

```
2*pt(-abs(-1.7545),df=75.802)

[1] 0.083386

pt(-abs(-1.7545),df=75.802)

[1] 0.041693
```

> *[The p-value, or probability value, tells you how likely it is that your data could have occurred under the null hypothesis. The p-value is a proportion: p-value is 0.083, that means that 8.3% of the time and if the null hypothesis is true we can see a test statistic at least as extreme as the one we found.]*

(d) In this test, what were the two hypotheses? What was the test statistic? Are you rejecting H_0 for large or small values of this statistic, or is it a two-tailed test? *Explain*.

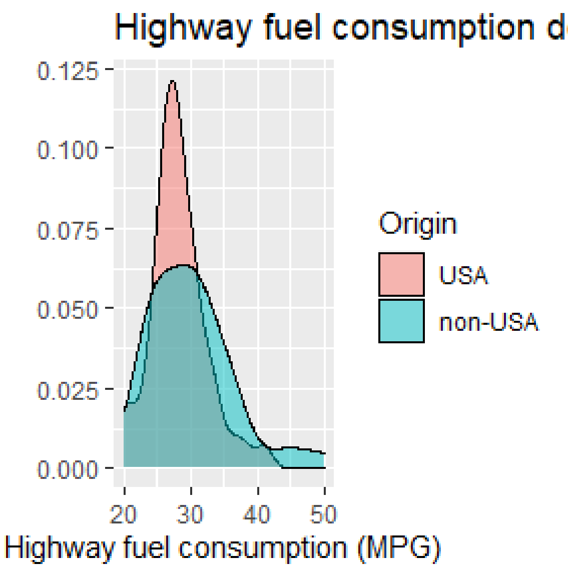
> *[Null hypothesis : true difference in means is equal to 0 and alternative hypothesis : true difference in means is not equal to 0. Z-score was the test statistic and it is a two tailed test.]*

Is the data normal?

(a) Modify the density plot code provided in problem 1 to produce a plot with better axis labels. Also add a title.

```
qplot(data = Cars93, x = MPG.highway,
      fill = Origin, geom = "density", alpha = I(0.5),
```

```
    xlab = "Highway fuel consumption (MPG)",
    main = "Highway fuel consumption density plots")
```

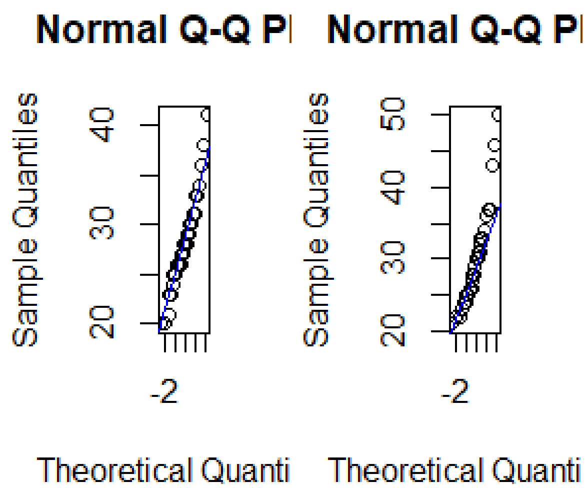


(b) Does the data look to be normally distributed? If not, describe why.

> *[The non-USA MPG.highway data looks quite far from normally distributed. This distribution appears to have a heavier upper tail.]*

(c) Construct qqplots of `MPG.highway`, one plot for each `Origin` category. Overlay a line on each plot as illustrated in lecture.

```
par(mfrow = c(1,2))
# USA cars
with(Cars93, qqnorm(MPG.highway[Origin == "USA"]))
with(Cars93, qqline(MPG.highway, col = "blue"))
# Foreign cars
with(Cars93, qqnorm(MPG.highway[Origin == "non-USA"]))
with(Cars93, qqline(MPG.highway, col = "blue"))
```



(d) Does the data look to be normally distributed? If not, describe why.

[Data for non-USA MPG highway doesn't seem to be normally distributed. Distribution for non-USA MPG highway appears to have a upper tail.]

Testing 2 x 2 tables

Doll and Hill's 1950 article studying the association between smoking and lung cancer contains one of the most important 2 x 2 tables in history.

Here's their data:

```
smoking <- as.table(rbind(c(688, 650), c(21, 59)))
dimnames(smoking) <- list(has.smoked = c("yes", "no"),
                          lung.cancer = c("yes","no"))
smoking

          lung.cancer
has.smoked yes  no
       yes 688 650
       no   21  59
```

(a) Use `fisher.test()` to test if there's an association between smoking and lung cancer.

```
smoking.fisher.test <- fisher.test(smoking)
smoking.fisher.test


    Fisher's Exact Test for Count Data

data:  smoking
p-value = 1.476e-05
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
```

```
 1.755611 5.210711
sample estimates:
odds ratio
  2.971634
```

(b) What is the odds ratio? Interpret this quantity. (Look it up, if needed!)

```
smoking.fisher.test$estimate

odds ratio
  2.971634
```

> *[Odds Ratio is a measure of the strength of association with an exposure and an outcome. odds ratio is 2.971634 it means greater odds of association with the exposure and outcome.]*

(c) Are your findings statistically significant?

```
smoking.fisher.test$p.value

[1] 1.476303e-05
```

> *[If p value is in between 0 and 1 then findings are statistically significant.Our p value is 1.476303e-05 so it is not statistically significant.]*

(d) Interpret the results of this hypothesis test. Make sure your sentences include an inline code chunk similar to the one you saw in class (do not hard-code any numerical values in the text!).

prop.test(x, n, p = NULL, alternative = c("two.sided", "less", "greater"), conf.level = 0.95, correct = TRUE)

> *[I will not reject the hypothesis of true difference in means being equal to zero at significance level $\alpha = 0.05$ because p-value is greater than 0.05 which is 0.08339]*

Writing/summarizing/theory.

- Explain the definition of a p-value. (It is a probability of an event; explain.)

  [The p-value of a distribution is here interpreted as the probability outside the smallest credibility interval or region containing a point; if no point is explicitly given, it is assumed to be zero, or the origin. p value calculated by p.value(object,point) object: The probability distribution for which the p-value should be computed. point: The point which should be included in the credibility interval or region.]

- Is reporting a p-value better than reporting a reject/not reject decision at a pre-determined significance level α for a hypothesis test? (Why/why not?)

  *[We use p-values to make conclusions in significance testing. More specifically, we compare the p-value to a significance level alpha to make conclusions about our hypotheses. H_0 is null hypothesis and H_ is an alternative hypothesis.

So reporting a p-value is better than reporting a reject/not reject decision at a pre-determined significance level α for a hypothesis test.]*

- Look back at the **Testing means between two groups** question.
  - Obtain the 95% confidence interval for the difference between `MPG.highway` in US and non-US vehicles. (Compute it again if necessary, or simply access the result of that computation using since you probably computed it above already - use an in-line code chunk!)

```
MPG.highway.t.test <- t.test(MPG.highway ~ Origin, data = Cars93)
MPG.highway.t.test


    Welch Two Sample t-test

data:  MPG.highway by Origin
t = -1.7545, df = 75.802, p-value = 0.08339
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -4.1489029  0.2627918
sample estimates:
    mean in group USA mean in group non-USA
             28.14583              30.08889
```

- Would you reject the hypothesis of true difference in means being equal to zero at significance level $\alpha = 0.05$?

  [I will not reject the hypothesis of true difference in means being equal to zero at significance level $\alpha = 0.05$ *because p-value is greater than 0.05 which is 0.08339]*

- **[bonus]** Relate the last two answers: the confidence interval and the rejection region. What do you conclude? Explain your answer carefully, and display any code you need to do these computations and write a conclusion.

  [There is a 95% chance that choosing a random sample from this population results in a confidence interval of 28.14583 and 30.08889 which contains the true value being estimated.If p-value is less than alpha value then we reject the hypothesis. In above example 0.05 is the value of alpha which is also a reject region but our p value is 0.08339 so I will not reject the hypothesis of true difference in means.]