Week 9 - 6103 Handout

Subhayan Mukerjee

17/3/2021

The Correlation Test

Correlation between two variables refers to the extent to which an increase in one variable is associated with an increase in the second variable. When correlation is 1, it means the two variables are perfectly correlated. An increase in X is associated with the exact amount of increase in Y. When correlation is -1, it means the two variables are perfectly negatively correlated. An increase in X is associated with the exact amount of decrease in Y.

Let's look at some trivial correlations.

```
x <- 1:10 # this is same as writing x <- c(1,2,3,4,5,6,7,8,9,10)
y <- 11:20 # this is same as writing y <- c(11,12,13,14,15,16,17,18,19,20)

cor(x,y)
```

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

```
x <- 1:10 # this is same as writing x <- c(1,2,3,4,5,6,7,8,9,10)
y <- -x # this is same as writing y <- c(-1, -2, -3, -4 ... , -10)

cor(x,y)
```

```
## [1] -1
```

It is rare, almost impossible to get perfectly 0 correlations. Even if the numbers are totally random.

```
x <- sample(1:100, 10) # sample 10 numbers randomly from between 1 and 100
y <- sample(1:100, 10) # sample another 10 numbers randomly between 1 and 100

# x and y will be different for each one of you as they are random
print(x)
```

```
##  [1] 68 39  1 34 87 43 14 82 59 51
```

```
print(y)
```

```
##  [1] 97 85 21 54 74  7 73 79 99 37
```

```r
# how about their correlation?
print(cor(x,y))
```

```
## [1] 0.4828566
```

A correlation of 0.48 is pretty high given the X and Y are random! Here's where the **significance** comes in.

Is the correlation of 0.48 significant?

```r
print(cor.test(x,y))
```

```
##
##  Pearson's product-moment correlation
##
## data:  x and y
## t = 1.5596, df = 8, p-value = 0.1575
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.2108818  0.8531185
## sample estimates:
##       cor
## 0.4828566
```

Let's try to understand the output. For each of you, the numbers in the output will be different. But the structure of the output will be the same. I will refer to the results that I got when I ran the script.

The last line tells us that the correlation is 0.48. But the p-value (3rd line) is 0.1575. In other words, assuming X and Y are independent, there's a 15.75% chance of getting a correlation of 0.48. This is obviously not enough to reject the null hypothesis that the correlation between X and Y is 0. So, we cannot say that X and Y are correlated (even when the value of the correlation is 0.48). In other words, the correlation between X and Y is not significantly different from 0.

Look at the 95% confidence interval. What does that mean? This means that if we were to sample 100 times from the distributions from which X and Y were drawn, the correlation would lie between those two values 95% of the time. This lets us look at another interpretation of the significance level and its relationship with the confidence interval:

If we are testing for a test statistic at 5% level of significance, and we find no significant difference from 0, then the 95% confidence interval would also automatically include 0. Basically, in colloqial terms this means, that if we do this test enough number of times (with different values of X and Y drawn from the same distribution), we'll find that the correlation can be negative or positive. So the null hypothesis cannot be rejected.

Repeat this exercise to see how you get different correlations and p-values every time. Unless you are very lucky, you shouldn't be getting a significant p-value.

Now let us look at a real-world example.

We will be using a dataset that comes by default in R. This dataset is called `mtcars`, and it lists various specifications of 32 popular cars. Let's look at the first few rows of the data:

```r
head(mtcars)
```

```
##                   mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
```

```
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
## Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

We shall only focus on three columns for this exercise.

1. `disp` which is the engine size (displacement)
2. `mpg` which is the mileage (miles per gallon)
3. `hp` which is the horsepower

Is there a correlation between `disp` and `hp`? Is it significant?

```
cor.test(mtcars$disp, mtcars$hp) # you can remove the print statement for commands that print their out
```

```
##
##  Pearson's product-moment correlation
##
## data:  mtcars$disp and mtcars$hp
## t = 7.0801, df = 30, p-value = 7.143e-08
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.6106794 0.8932775
## sample estimates:
##       cor
## 0.7909486
```

A high correlation of 0.709 which is also significant! (p-value is 7.143×10^{-8} which is very low (and lower than the conventional 0.05)). That `e` in the p-value is the scientific notation for "10 raised to the power".

This significant high correlation makes sense as we would expect cars that have larger engines to have higher horsepower.

How about between `hp` and 'mpg'?

```
cor.test(mtcars$hp, mtcars$mpg)
```

```
##
##  Pearson's product-moment correlation
##
## data:  mtcars$hp and mtcars$mpg
## t = -6.7424, df = 30, p-value = 1.788e-07
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.8852686 -0.5860994
## sample estimates:
##       cor
## -0.7761684
```

A correlation of -0.77 which is also very significant (p-value = 1.78×10^{-7}). This again makes sense, as we would expect cars with higher horsepower to have lower mileage as they likely consume more fuel.

Exercise

Go back to the `gapminder` dataset. Pick any year of your choice (of course, amongst the ones that exist in the `gapminder` data).

1. Which two variables (out of population, life expectancy, and GDP per capita) would you expect to be significantly positively correlated?
2. Which two variables would you expect to NOT be correlated?

Do the correlation tests and see if your findings are in line with you what you expect.

The t-test

The t-test allows us to ask a few kinds of questions:

The one-sample t-test

Is the mean of a single sample significantly different from a certain number?

The average household monthly income per household member in Singapore was $3940 in 2017/18. Here we have the monthly income of 26 migrant workers in Singapore. We want to test if their average income is significantly different from tthe above value (We all know what the answer will be, but let's see how to do the test)

```
migrant_salaries <- c(1100, 1200, 900, 685, 1320, 1055, 745, 1350, 1400,
                      1700, 1650, 1245, 1120, 1075, 980, 1025, 1550, 1945,
                      750, 1140, 1250, 1090, 1150, 940, 880, 950)

t.test(migrant_salaries,  mu = 3940) # the mu is used to tell the command to use 3940 as the value to t
```

```
##
##  One Sample t-test
##
## data:  migrant_salaries
## t = -46.509, df = 25, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 3940
## 95 percent confidence interval:
##  1038.300 1284.393
## sample estimates:
## mean of x
##  1161.346
```

Study the output. It first tells us that it is a One Sample t-test. The last line tells us the mean of x(the migrant salaries) which is $1161.346. The null hypothesis is that the mean of these salaries is NOT significantly different from 3940. The p-value (3rd line) is very low (less than 2.2×10^{-16}) implying that this mean is indeed significantly different from 3940. Thus we can reject the null hypothesis. In line 4, you will see that R also tells you which hypothesis to support: Alternative hypothesis: true mean is not equal to 3940.

Notice that in this case, we tested whether the mean is *different* from 3940. Not if the mean is *less* than 3940. The former is a two-tailed (or two-sided) test, the latter is a one-tailed test (or a one-sided). By

default, the `t.test` does a two-tailed test. In order to test the latter, we need to tell the command as shown below, using the "alternative" term.

```
t.test(migrant_salaries,  mu = 3940, alternative = "less")
```

```
##
##  One Sample t-test
##
## data:  migrant_salaries
## t = -46.509, df = 25, p-value < 2.2e-16
## alternative hypothesis: true mean is less than 3940
## 95 percent confidence interval:
##      -Inf 1263.398
## sample estimates:
## mean of x
##  1161.346
```

The `alternative = "less"` tells the command to do a one-sided test and test if the mean is less than mu. Similarly, you can test for "greater" also by using `alternative = "greater"`.

You again see that the null is rejected. What is the null in this case? That the mean migrant salaries are NOT LESS than $3940. And the alternative hypothesis is supported: that the true mean is less than 3940.

The two-sample t-test

The two-sample t-test is used to test if the means of two samples are different (two-sided) or if the mean of one sample is than/greater than (one sided) the mean of the other sample.

To illustrate, let us compare the salaries of the 26 migrant workers, with that of 15 citizens.

```
citizen_salaries <- c(3500, 4250, 2560, 1900, 6750, 7500, 3550, 2820, 6520,
                      4570, 1700, 2510, 8440, 7510, 4580)

t.test(migrant_salaries,  citizen_salaries) # there is no longer a mu
```

```
##
##  Welch Two Sample t-test
##
## data:  migrant_salaries and citizen_salaries
## t = -5.8953, df = 14.302, p-value = 3.575e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -4656.319 -2175.655
## sample estimates:
## mean of x mean of y
##  1161.346  4577.333
```

As the output shows, the two means are 1161.346 and 4577.333 respectively, and the low p-value (3.575×10^{-5}) means we can safely reject the null. Again, this was a two-sided test. We can do the corresponding one-sided test as well. What happens if use `alternative = greater`?

```r
t.test(migrant_salaries, citizen_salaries, alternative = "greater")
```

```
##
##  Welch Two Sample t-test
##
## data:  migrant_salaries and citizen_salaries
## t = -5.8953, df = 14.302, p-value = 1
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -4435.051       Inf
## sample estimates:
## mean of x mean of y
##  1161.346  4577.333
```

The p-value is 1 (very high)! This means that the null hypothesis, that the migrant salaries are NOT GREATER than citizen salaries, cannot be rejected.

And what if we use `alternative = "less"`?

```r
t.test(migrant_salaries, citizen_salaries, alternative = "less")
```

```
##
##  Welch Two Sample t-test
##
## data:  migrant_salaries and citizen_salaries
## t = -5.8953, df = 14.302, p-value = 1.787e-05
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf -2396.923
## sample estimates:
## mean of x mean of y
##  1161.346  4577.333
```

A very low p-value, yet again (1.787×10^{-5}). This means that we can reject the null that that migrant salaries are not less than the citizen salaries

The paired T-test

The paired t-test is used to test the difference in means between two sets of values of the SAME subjects.

Let's consider the salaries of the same 26 migrant workers in 2017-18 and 2018-19.

```r
migrant_salaries_201718 <- c(1100, 1200, 900, 685, 1320, 1055, 745, 1350, 1400,
                        1700, 1650, 1245, 1120, 1075, 980, 1025, 1550, 1945,
                        750, 1140, 1250, 1090, 1150, 940, 880, 950)


migrant_salaries_201819 <- c(1150, 1210, 950, 700, 1380, 1100, 800, 1380, 1450,
                        1780, 1700, 1295, 1190, 1145, 1020, 1100, 1600, 2000,
                        750, 1150, 1250, 1090, 1150, 980, 900, 980)
```

You can then ask, has their salaries increased significantly? The null in this case is: the mean change in salary is 0.

To do the paired t-test, you need to specific `paired = TRUE` when running the command.

```
t.test(migrant_salaries_201718, migrant_salaries_201819, paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  migrant_salaries_201718 and migrant_salaries_201819
## t = -7.9212, df = 25, p-value = 2.813e-08
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -48.70392 -28.60378
## sample estimates:
## mean of the differences
##               -38.65385
```

This is again a two-sided test. Try running the one-sided test, with `alternative = "less"` and see if you can interpret the answer:

```
t.test(migrant_salaries_201718, migrant_salaries_201819, paired = TRUE, alternative = "less")
```

```
##
##  Paired t-test
##
## data:  migrant_salaries_201718 and migrant_salaries_201819
## t = -7.9212, df = 25, p-value = 1.406e-08
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf -30.31852
## sample estimates:
## mean of the differences
##               -38.65385
```

Exercise

Let's turn back to the `gapminder` data.

4. Calculate the global mean life expectancy in the year 1982.
5. What test would you use to evaluate if the average life expectancy in African countries in 1982 is less than the global mean? Do the test and check.
6. What test would you use to evaluate if the average life expectancy in African countries in 1982 is different from the average life expectancy in Asian countries in 1982? Do the test and check.
7. What test would you use to evaluate if the average life expectancy in African countries has significantly increased between 1982 and 2007? Do the test and find out.

The chi-square test

So far we have been looking at ordinal variables in our statistical tests. The chi-square test lets us examine the independence of two `categorical` variables. Let's do this with an example. Download the `media_knowledge.csv` file from Luminus and read it into a dataframe.

```
df <- read.csv("data/media_knowledge.csv")
```

Look at the data by doing `head(df)`

```
head(df)
```

This is mock data from a survey where we have the news exposure levels and political knowledge levels of 82 respondents. We want to know if exposure to news and political knowledge are independent. We first construct what is called a contingency table that gives us the number of people for different combinations of news exposure and political knowledge. Store it in a variable called `tbl`.

```
tbl <- table(df$Knowledge, df$News)
```

This shows that there are 26 respondents with "frequent" exposure and high political knowledge, 3 with "none" exposure and High political knowledge and so on. Clearly you see that those with high knowledge also report to being exposed to news frequently. But how do we test this statistically?

Again, the null hypothesis is that there is news exposure and political knowledge are independent. The chi-square test will evaluate the extent to which these two categorical variables are indeed independent.

```
chisq.test(tbl)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  tbl
## X-squared = 31.502, df = 1, p-value = 1.993e-08
```

The low p-value of 1.993×10^{-8} indicates the the null hypothesis can be rejected. In other words, it does look like there is a relationship between news exposure and political knowledge.

Exercise

For this exercise we will be using a dataset in R that comes with a `package`. A `package` in R is a set of specialized commands that R doesn't give you by default, but you can access them by using the `library` command. We will be using the `MASS` package. Once you have included the package, you can access this dataset called `survey`

```
library(MASS) # include the library

head(survey) # see the dataset
```

```
##      Sex Wr.Hnd NW.Hnd W.Hnd    Fold Pulse    Clap Exer Smoke Height      M.I
## 1 Female   18.5   18.0 Right  R on L    92    Left Some Never 173.00   Metric
## 2   Male   19.5   20.5  Left  R on L   104    Left None Regul 177.80 Imperial
## 3   Male   18.0   13.3 Right  L on R    87 Neither None Occas     NA     <NA>
## 4   Male   18.8   18.9 Right  R on L    NA Neither None Never 160.00   Metric
## 5   Male   20.0   20.0 Right Neither    35   Right Some Never 165.00   Metric
## 6 Female   18.0   17.7 Right  L on R    64   Right Some Never 172.72 Imperial
```

```
##      Age
## 1 18.250
## 2 17.583
## 3 16.917
## 4 20.333
## 5 23.667
## 6 21.000
```

The `Exer` and `Smoke` columns in this dataframe tells is how frequently the respondent does exercises and smokes respectively.

8. Are those who are more likely to smoke, also likely to exercise more? In other words, are the `exer` and `smoke` variables dependent on each other?
9. What about sex differences? Are males more likely to smoke than females?
10. Are males more likely to do exercise than females?