A1: Brain Size and Intelligence

Last name: Khambaita
First name: Sunil
Student ID: 1000285924
Course section: STA302H1F-L0101

Oct. 4, 2016

Q1: t-test for MRIcount between high and low intellegince groups

Type your concise and clear answer here.

- Decribe the null hypothesis, which seems okay to be tested using t-test

Solution:

Let's call μ_{high} : The mean MRI count for the high intelligence group.

Also μ_{low} : The mean MRI count for the low intelligence group.

The null hypothesis is:

$$\begin{cases} H_0 : \mu_{high} = \mu_{low} \\ H_a : \mu_{high} \neq \mu_{low} \end{cases}$$

- what t-test should I use? and the assumptions for the t-test I use.

Solution:

We will use the Welch Two Sample t-test, we will assume that the variance of the MRI accounts in the two groups are different, will also assume that normality is maintained which means that the data set will be well modelled by a normal distribution.

- give the test statistics, degrees of freedom, and the significance

Solution:

p-value = 0.1344
t-value = 1.53
degrees of freedom = 37.324
significance = 0.05

- make a conclusion about either to reject the hypothesis or unable to reject

Solution:

We fail to reject the null hypothesis since the p-value (0.1344) is greater than α (0.05)

- What is the conclusion in the context of this study

Solution:

From our results of the t-test, we failed to reject the null hypothesis which states that the mean of the highIQ MRI scans and the lowIQ MRI scans are equal.
We can also see that our 95% confidence interval includes the null value (i.e. zero), this implies there is no statistically significant difference between the two groups of highIQ MRI and lowIQ MRI.
In simple terms, we can conclude the size of your brain is **not** an indicator of mental capacity.

Q2: correlation analysis among the MRI count and IQ variables

Correlations of the IQ measurements with MRI count (p-value for test of $\rho = 0$ is in brackets):

-	Full data	High-IQ group	low-IQ group
FSIQ	0.3576(0.0235)	0.5482853(0.01231)	0.5273002(0.01689)
VIQ	0.3374777(0.0332)	0.4066862(0.07516)	0.1463655(0.5381)
PIQ	0.3868173(0.01367)	0.2012682(0.3948)	0.5861888(0.006602)

From the correlation analysis, there are slight things to note:

Our given null hypothesis in the correlation test is: true correlation is equal to 0. This basically states that there is no relationship between the MRI count and the three IQ scores of the different data sets.

- Three values on the correlation analysis table help validate with the null hypothesis conclusion from the t test.

In the technical aspect, when we observe the High-IQ group with VIQ and PIQ. Low-IQ group with VIQ. The p-value we received is higher than 0.05 which goes to show that we fail to reject the null hypothesis in these given scenarios and therefore agreeing that our true correlation is equal to 0, this means that there is no relationship with the MRI count. This would coincide with the null hypothesis from the t test because if there is no relationship then there is a chance the means of the High-IQ and Low-IQ would have equal means.

- The High-IQ group (VIQ) and Low-IQ group (VIQ) shows no relationship

From the point given in the last paragraph, we can also notice that if we are to observe between the high-IQ (VIQ) and low-IQ (VIQ) sections, both failed to reject the null hypothesis. This implies that the higher your VIQ doesn't necessarily mean the higher your MRI as there is no relationship from the correlation test.

- Majority of the remaining values on the correlation analysis table guide us in rejecting the null hypothesis.

All the other remaining values used in the correction test have a p-value lower than 0.05 which leads us to reject the null hypothesis. Which means that the true correlation is not equal to 0 for majority of the values on the table. This implies that there is some sort of relationship with the null hypothesis

- The full data (uncategorized) overall rejects the null hypothesis

Another observation which is also good to note, is that despite some categorized values (high-IQ and low-IQ) are failing to reject the null hypothesis, if we were to only observe the first column which takes into account the full data itself without any categorization, there is clearly a relationship between the MRI count as the whole first column rejects the null hypothesis and states that there is some sort of relationship.

- Due to the size of the correlation, confidence is not significant enough for it to be conclusive

In the technical aspect, the confidence in a relationship is not only just determined from the number given by the correlation coefficient but also from the number of pairs in our data, since we do not have a lot of pairs on our correlation analysis then the coefficient needs to be as close as 1 or -1 as possible for it to be deemed "statistically significant". However, if we had many pairs, our results we had could have a chance of being deemed conclusive. This also practically implies that we can not be confident with our data as being conclusive with the null hypothesis. Hence rejecting the null hypothesis seems ideal.

Q3

Type your answer here. Please make it as concise as possible.

- Does the result of the t-test in question 1 agree with the relevant correlation in question 2? Why or why not?

Solution:

In question 1 we assumed that that the mean in our MRI count of the low-IQ group and high-IQ group were equal. We proceeded and performed a t-test and confirmed that we failed to reject the null hypothesis since the p-value (0.1344) was greater than α (0.05). We finally concluded that, the size of your brain is **not** an indicator of mental capacity.

In question 2 we performed a correlation test between the MRI count and the three IQ scores, we also assumed that the true correlation is equal to zero, if true correlation is equal to zero this means that there is no relationship between the size of your brain and your IQ i.e. Mental capacity. If we failed to reject the null hypothesis this would have implied the question 1 results and question 2 results agree with each other. However, majority of the results received in question 2 rejected the null hypothesis and went on to show that there is in fact a relationship between your size of brain and mental capacity. If we look specifically at the first row of our table of results in the correlation test we can see that FSIQ in the full data, high-IQ, low-IQ all have p-values of lower than 0.05 which gave us no other option but to reject the null hypothesis and accept the alternative hypothesis: true correlation is not equal to 0. Which infact means that there is a relationship, and that the means of the MRI count in the high-IQ and low-IQ cannot equal each other.

With this information, we can deduce that question 1 and question 2 results do **not** agree with each other.

Solution:

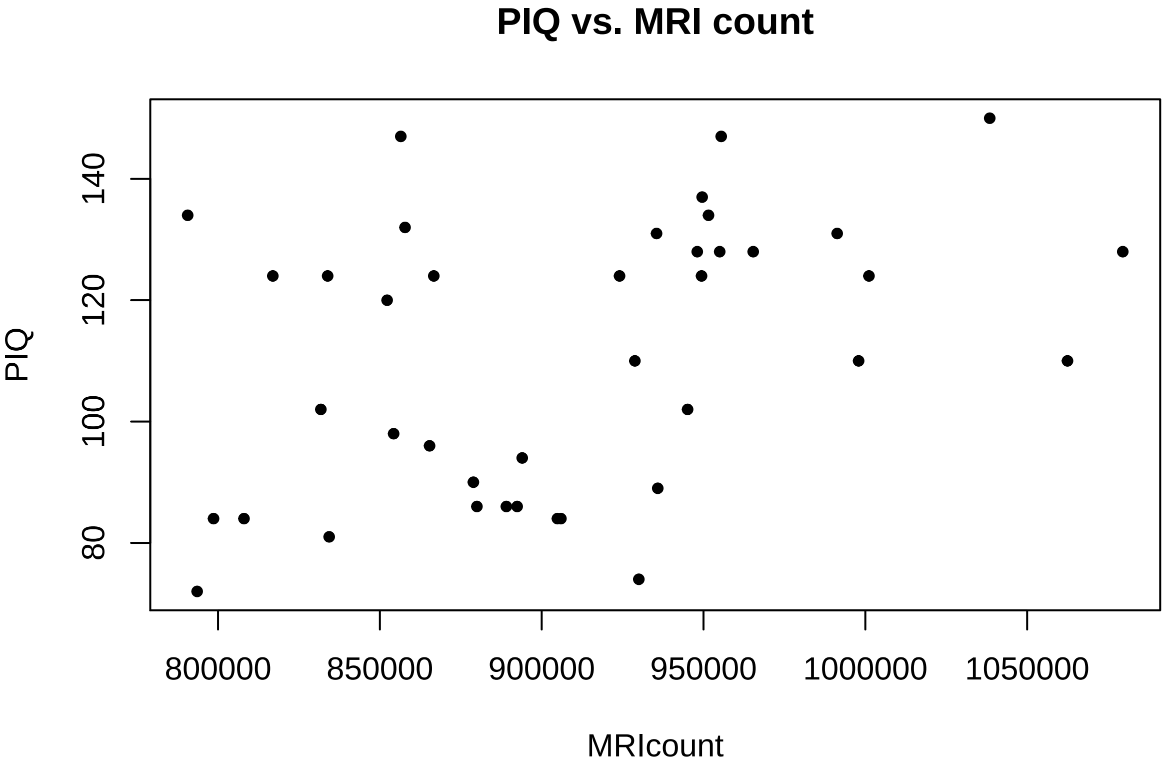
- Under the assumption that the data have a bivariate normal distribution, why would the correlation be the preferred analysis ? (The answer to the last part requires you to use your common sense. People in medical science often like to dichotomized variables, such as high IQ group and low IQ groups, while statisticians tend to utilize all the information given).

Solution:

Correlation is simply the most preferred analysis because we can see that with dichotomizing variables we are asserting that there is a straight line effect between one variable and another, in our case we are dichotomizing high-IQ and low-IQ with the t-test, in the t-test we are making an assumption that these are the only values that matter. But we aren't considering ones IQ who is between 104 and 129 so all this extra information is lost.

Using correlation we are using all the data given to us to provide our conclusion this information is clearly more trustable as it's making conclusions based on the complete data set as opposed to categories from the t-test.

Q4(a) Scatter plot of PIQ versus MRI count



Q4(b) Regression analysis for two groups

Regression	R^2	Intercept (b_0)	Slope(b_1)	MSE	p-value for $H_0 : \beta_1 = 0$
High-IQ groups	0.04051	1.100e+02 (110)	2.265e-05	73.513476	0.3948
Low-IQ groups	0.3436	1.636e+00 (1.636)	1.003e-04	88.755241	0.006602

i.) We use the slope of a regression line to represent the rate of change of PIQ as the MRI changes since the value of b_1 (slope) is so low this is clear to show that as the MRI count increases there is a very very minor change in the PIQ which is an indication of no relationship between PIQ and MRI count.

ii.) Comparing the two R^2 values, the regression model for the high-IQ group is 4.051%, the regression model for the low-IQ group is 34.36%. By default, the higher the R^2, the better the model fits our data since the more variance that would be accounted for by the regression model, the closer the data points will fall to the fitted regression line. This makes the **low-IQ group the better fit if we use R^2 as our criteria**.

iii.) For MSE the values closer to zero, the better, as lower values indicate a better fit. In our two models we see that the high-IQ group has a MSE value of 73.513476. Whilst the low-IQ group has a MSE value of 88.755241. This makes the **high-IQ group the better fit if we use MSE as our criteria.**

iv.) The R^2 is a measure to analyze how close data are to the fitted regression line. In our case, it's basically the percentage of the PIQ variation that is explained by our PIQ vs MRI count model. The MSE measures the average of the squares of the errors/deviations, which is basically the difference between the estimator and what is estimated.

MSE is used throughout in analysis of data, it is useful for hypothesis testing, inferences and confidence intervals. This is something that we use all the time compared to R^2

While R^2 is a useful tool, it is hard to determine if the coefficient estimates and predictions are biased, which is a reason to why we look at the residual plots (this is what the MSE takes into account).

Another issue with R^2 is that it increases everytime you add a predictor to a model, even if this due to chance alone. This will result in a model having a higher R^2 which will seem it has a better fit whilst its simply because of more terms.

R^2 is also misleading since correlation does not imply causation. Also by default, people tend to believe a high R^2 is always ideal, however, after some research it is also good to note that low R^2 values are entirely anticipated in some fields. "For example, any field that attempts to predict human behavior, such as psychology, typically has R-squared values lower than 50%. Humans are simply harder to predict than, say, physical processes."

The coefficient of determination R^2 can greatly be affected by just one data point (or a few data points) which will cause a wrong conclusion while with MSE, squaring always gives a positive value and emhasizes the large differences.

I would say R^2 is mainly useful as a validation technique in order to validate your analysis given you have already used MSE together with other techniques and came up with a conclusion, we can use R^2 to validate these claims but would not solely use it by itself.

Apendix: Source R code

```r
# ---------> complete and run the following code for this assignment  <-------
#
# R code for STA302 or STA1001H1F assignment 1
# copyright by Sunil Khambaita
# date: Oct 4, 2016
#

## Load in the data set
brain = read.table("BrainData.csv",header=T);


## create an indicator for high-IQ (value =1) and low-IQ (value=0)
brain$highIQ = ifelse(brain$FSIQ>=130,1, 0)

## or
highIQ <- ifelse(brain$FSIQ>=130,1, 0)


sub_high <- subset(brain, brain$highIQ == 1)
## or
sub_high <- subset(brain, highIQ == 1)

sub_low <- subset(brain, brain$highIQ == 0)

## Q1: t-test on MRI count between high- and low IQ groups

# subsets by female and male
male_dat <- subset(brain, brain$Gender=="Male")
female_dat <- subset(brain, brain$Gender=="Female")

# testing equality of mean of weight in female and male.
t.test(male_dat$Weight, female_dat$Weight);


## Q2: correlation analysis
# cor.test() : missing value is suppressed, default setting:
# - find correlation between MRI count and 3 IQ variables

## get a correlation matrix
cor(male_dat[,2:3])

# get a single correlation value
cor(male_dat[,2], male_dat[,3])
cor(male_dat[,2], male_dat[,3], use = "pairwise.complete")


cor.test(male_dat[,5], male_dat[,3])
cor(male_dat[,5], male_dat[,3], use = "pairwise.complete")


# - find correlation between MRI count and 3 IQ variables in high-IQ group

cor(sub_high[,2:5], use="pairwise.complete")
cor(sub_high[,2], sub_high[,3])

# - find correlation between MRI count and 3 IQ variables in low-IQ group
```

```r
## Q4:
# - Scatterplot of PIG vs MRI count

# scatter plot of PIQ versus MRI count
#complete the following plot() command to get the scatter plot
#plot()

## I am plotting some other variables, you should change them to the variables you are required to plot.
# Also, you should modify your title name and you are not required to add regression line unless the question asks

plot(sub_high$FSIQ, sub_high$MRIcount, main ="My graph", xlab="FSIQ", ylab = "MRI");
abline(lm(sub_high$MRIcount~sub_high$PIQ));
```