STATProject

Contents

Abstract

This report is the demonstration of the solutions to some statistical questions regarding a sample population of male and female and their physical attributes like height, weight and physical activities. Arising questions should be if there is any relationship between height and weight, do male and female differ in heights significantly and if male participate in physical activity than the opposite gender, etc.

The approach to yield any valid solutions should be drawn from statistical tests. Throughout the report, these questions are answered based on statistical results.

Introduction

Based on a given data set that outlines some physical attributes of male and female sample population, solutions to the following research questions will be discussed:

- Is there a linear relationship between height and weight?
- Is the mean height of male and female the same? You can assume equal variances between male and female heights.
- Is there any association between gender and the amount of physical activity?

The three statistical tests that would be discussed are linear regression, t test and the χ^2 test.

Based on a first person statistical experiment on a particular data set 'project.csv,' there is a linear relationship between height and weight. The mean height of male and female are different. Lastly, the gender does not seem to affect each individuals physical activities. The last remark is that the report is compatible with a different data set, if the structure is identical. This means these results may change.

Research question 1

Methods

The first research question is whether there is a linear relationship between heights and weights of the sample population. The test should be able to detect the linear relationship β between the two variables. In this way, the linear regression test should be used.

The method to secure the required data is simple. Out of 5 of the variables ID, gender, height, weight, phys, just taking height and weight and inserting them into the lm() function (R Core Team (2021)) would yield the required results. This is the actual code: lm(height~weight, dat). The use of the data set is given from a simple designation: dat <- read_csv("project.csv").

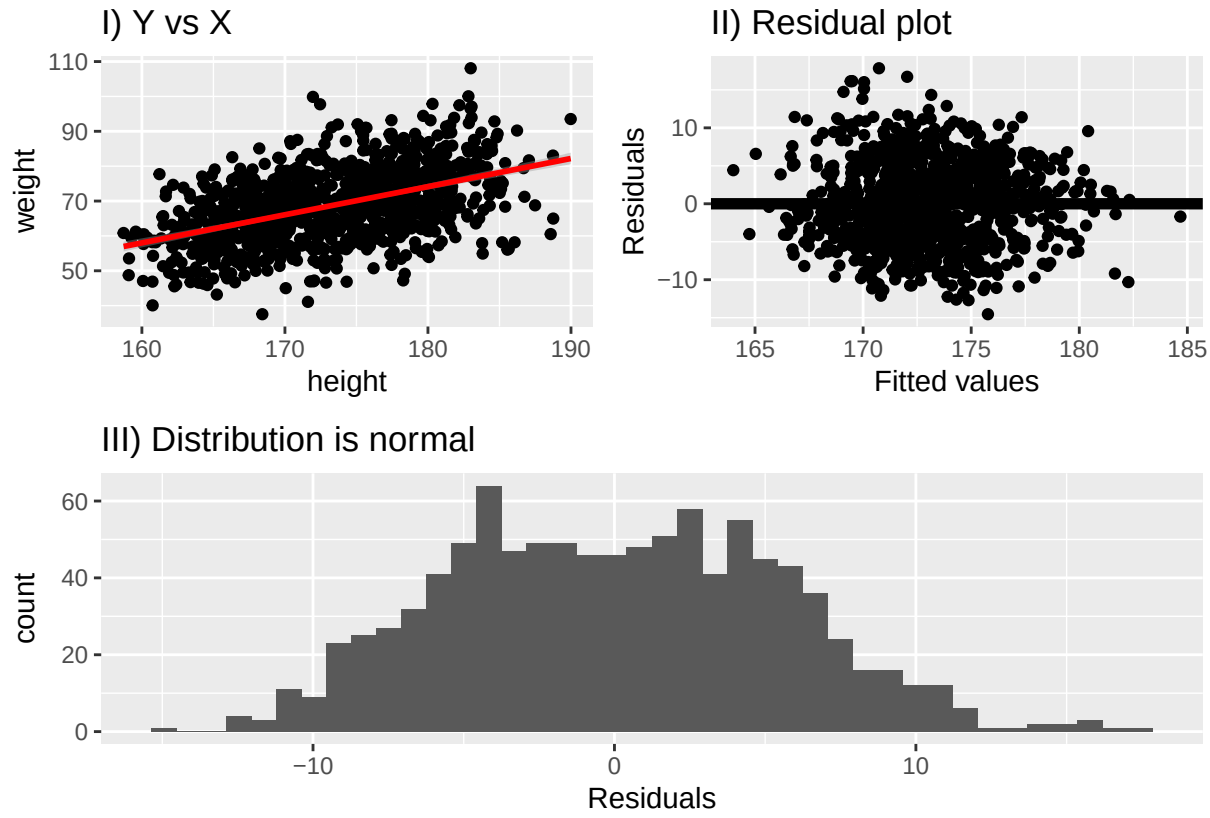
```
## # A tibble: 1,000 x 5
##    ID    gender height weight phys
##    <chr> <chr>   <dbl>  <dbl> <chr>
##  1 ID1   Male     184.   84.0 None
##  2 ID2   Male     178.   61.3 Moderate
##  3 ID3   Female   167.   68.0 None
##  4 ID4   Male     174.   67.9 Moderate
##  5 ID5   Male     175.   63.6 Moderate
##  6 ID6   Female   167.   68.6 Moderate
##  7 ID7   Male     183.   79.1 Moderate
##  8 ID8   Male     178.   73.9 None
##  9 ID9   Male     184.   80.2 Moderate
## 10 ID10  Male     172.   97.8 Moderate
## 11 ID11  Male     177.   71.2 Intense
## 12 ID12  Female   164.   65.0 Intense
## 13 ID13  Male     176.   63.1 Moderate
## 14 ID14  Female   165.   55.8 Intense
## 15 ID15  Female   175.   62.4 Intense
## 16 ID16  Male     175    65.9 None
## 17 ID17  Male     182.   71.3 Intense
## 18 ID18  Female   172.   83.5 Moderate
## # ... with 982 more rows
```

The above is the 'dat' data set.

Results

Exploratory Data Analysis

Making a proper linear regression decision requires the linearity between the two variables, constant variance and normality of the residuals. In fact, these are all assumed to hold already. The following graphs must show the linearity, evenly dispersed residuals and a bell shaped histogram of the residuals vs fitted values:



Once again, the I Y vs X plot must show the dots lining up forming a linear pattern. The II residual graph must show the dots dispersed evenly around the 0 horizontal line, which confirms that the residuals vary around the red linear line in I). Otherwise, the residuals' equal variance assumption is broken. The III graph must show the bell-shape to confirm the normality of the residuals.

Once it is clear that these assumptions are met, the P value can then be computed directly using the r function lm().

Outcome

The β is being tested, so the sampling distribution is: $\hat{\beta} \sim N(\beta, \frac{\sigma^2}{S_{xx}})$.

$H_0 : \beta = 0$ against $H_1 : \beta \neq 0$

We assume the linear regression model is appropriate: $Y_i = \alpha + \beta X_i + \epsilon_i$, where ϵ_i are independent and identically distributed $N(0, \sigma^2)$.

Test statistic: $\tau = \frac{\hat{\beta}}{s_{Y|X}/\sqrt{S_{xx}}} \sim t_{n-2}$ under H_0.

$\tau_{obs} = 132.3145054$

P value $= 1.6279622 \times 10^{-60}$

Conclusion

The computed P value is $1.6279622 \times 10^{-60}$. With this P value, it is possible to make the conclusion about the test:

```
## REJECT H0:  1.627962e-60  < 0.05
##
## There is a relationship between height and weight: As the P-value is very small,
## we have very strong evidence to reject H0. I.E. very strong evidence that the
## slope parameter is significant and there is a relationship between the height
## and weight of the sample population.
```

Research question 2

Methods

The next test is about the mean heights of male and female. It is to test if they are same or not. Since the issue is about two different means from two different samples, and the conclusion must be made about them being equal or not equal, the two sample t test should be used. The data set contains information about male and female heights:

```
## # A tibble: 1,000 x 5
##    ID    gender height weight phys
##    <chr> <chr>   <dbl>  <dbl> <chr>
##  1 ID1   Male     184.   84.0 None
##  2 ID2   Male     178.   61.3 Moderate
##  3 ID3   Female   167.   68.0 None
##  4 ID4   Male     174.   67.9 Moderate
##  5 ID5   Male     175.   63.6 Moderate
##  6 ID6   Female   167.   68.6 Moderate
##  7 ID7   Male     183.   79.1 Moderate
##  8 ID8   Male     178.   73.9 None
##  9 ID9   Male     184.   80.2 Moderate
## 10 ID10  Male     172.   97.8 Moderate
## 11 ID11  Male     177.   71.2 Intense
## 12 ID12  Female   164.   65.0 Intense
## 13 ID13  Male     176.   63.1 Moderate
## 14 ID14  Female   165.   55.8 Intense
## 15 ID15  Female   175.   62.4 Intense
## # ... with 985 more rows
```

Using filter and select functions (Wickham et al. (2021)), two new variables can be assigned to contain only male heights and female heights. Then, it is possible to run a t test function using the two variables: t.test(datm, datf, var.equal = TRUE). It is given that the equal variance assumption holds. This means var.equal is TRUE.

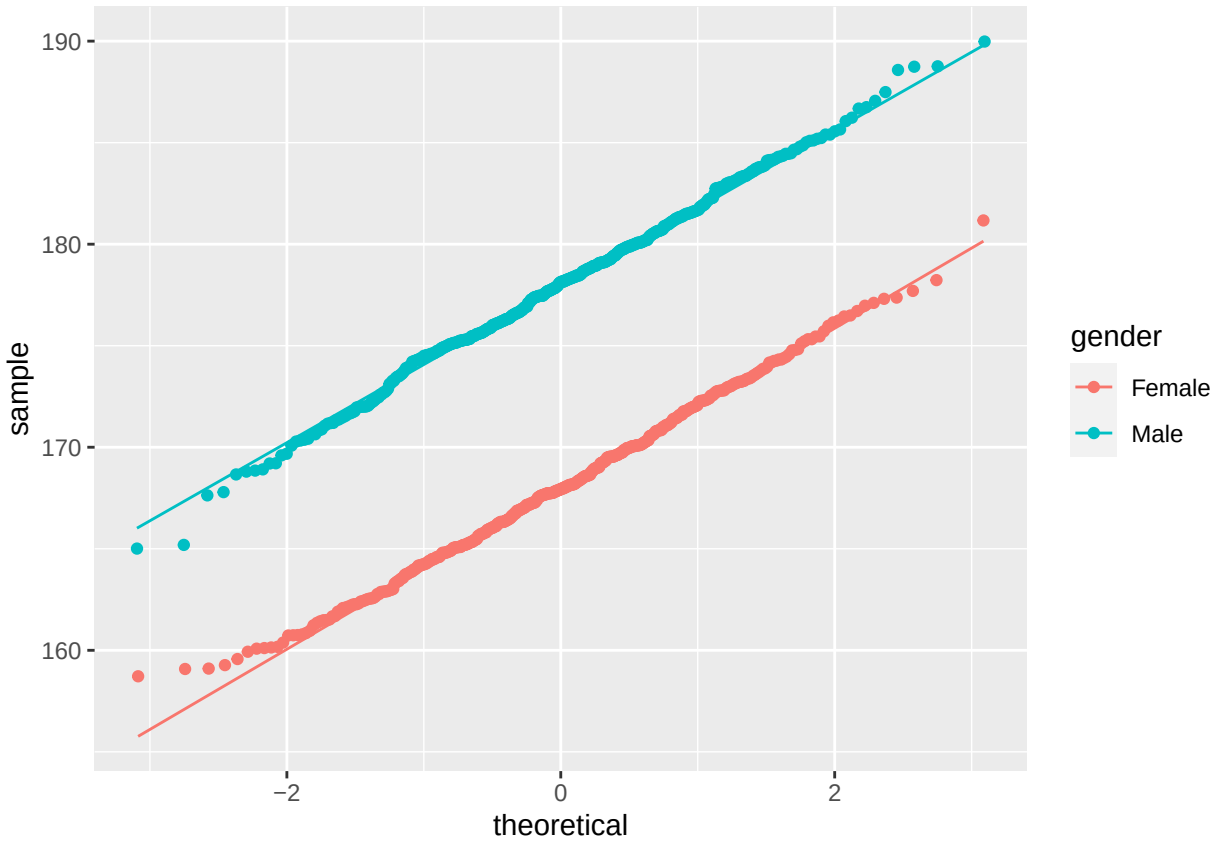
```
## # A tibble: 507 x 1
##    height
##     <dbl>
##  1   184.
##  2   178.
##  3   174.
##  4   175.
##  5   183.
##  6   178.
##  7   184.
##  8   172.
##  9   177.
## 10   176.
## 11   175
## 12   182.
## # ... with 495 more rows
```

The above is an example of the male height data frame. It only contains one column which is required to work with the t.test(). The data frame is assigned as 'datm' and the other female data frame is named 'datf.'

Results

Exploratory Data Analysis

2 conditions must meet to carry out this test. The normal distribution and equal variance must hold. Firstly, the QQ plot can be examined to confirm the normality assumption:



The above graph must show that the dots are lined around the sample qq line, which confirms that they are normally distributed. Once this is checked, the assumption is confirmed.

On top of this, the equal variance is assumed. The larger standard deviation divided by the smaller must not exceed 2. The larger is 3.9596539 and the smaller is 3.8744566. The division gives 1.0219895.

Outcome

The t test is based on equal variances assumption. Assuming the null hypothesis $H_0 : \mu_1 = \mu_2$, the resulting sampling distribution is $\frac{\overline{X_1}-\overline{X_2}}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim N(0,1)$.

$H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_1 \neq \mu_2$

Test statistic: $\tau = \frac{\overline{X_1}-\overline{X_2}}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim t_{n_1+n_2-2}$, if H_0 is true.

$\tau_{obs} = 39.9471643$.

P value $= 3.3213865 \times 10^{-209}$

Conclusion

The t.test(datm, datf, var.equal = TRUE) function outputs the P value ($3.3213865 \times 10^{-209}$) with which the conclusion can be made:

```
## REJECT H0:  3.321387e-209  < 0.05
##
## The mean height of male and female are NOT the same: As the P-value is very
## small, we have very strong evidence to reject H0. I.E. very strong evidence that
## the mean height of male is not the same as the mean height of female.
```

Research question 3

Methods

The last test is to see if male and female have different amounts of physical activity. In other words, if gender affects the amount of physical activity. In terms of statistics, this is equivalent to saying there is association between the two variables gender and physical activity. Intuitively, the test of independence should be the suitable test in this context, the χ^2 test.

Preparing the right form of data to make the chi.test() function work is not straight forward in this case. This is because the data frame only contains information in a way that is not yet appropriate for a χ^2 test:

```
## # A tibble: 1,000 x 5
##    ID    gender height weight phys
##    <chr> <chr>   <dbl>  <dbl> <chr>
##  1 ID1   Male     184.   84.0 None
##  2 ID2   Male     178.   61.3 Moderate
##  3 ID3   Female   167.   68.0 None
##  4 ID4   Male     174.   67.9 Moderate
##  5 ID5   Male     175.   63.6 Moderate
##  6 ID6   Female   167.   68.6 Moderate
##  7 ID7   Male     183.   79.1 Moderate
##  8 ID8   Male     178.   73.9 None
##  9 ID9   Male     184.   80.2 Moderate
## 10 ID10  Male     172.   97.8 Moderate
## 11 ID11  Male     177.   71.2 Intense
## 12 ID12  Female   164.   65.0 Intense
## 13 ID13  Male     176.   63.1 Moderate
## 14 ID14  Female   165.   55.8 Intense
## 15 ID15  Female   175.   62.4 Intense
## 16 ID16  Male     175    65.9 None
## 17 ID17  Male     182.   71.3 Intense
## 18 ID18  Female   172.   83.5 Moderate
## # ... with 982 more rows
```

The ideal form of the data frame is to have Male and Female as the horizontal marginal variables, and the three degrees of physical activity as the vertical marginal variables like the following:

```
## # A tibble: 3 x 3
##   `Physical activity` Male  Female
##   <chr>               <lgl> <lgl>
## 1 None                NA    NA
## 2 Moderate            NA    NA
## 3 Intense             NA    NA
```

Using filter, select and count (Wickham et al. (2021)), all 6 types of data such as Male with no physical activity and Male with a moderate physical activity, etc were retrieved. The following is the example of the Male with no physical activity:

```
## # A tibble: 127 x 1
##    phys
##    <chr>
##  1 None
```

```
##  2 None
##  3 None
##  4 None
##  5 None
##  6 None
##  7 None
##  8 None
##  9 None
## 10 None
## 11 None
## 12 None
## 13 None
## 14 None
## 15 None
## # ... with 112 more rows
```

All of these 6 data frames are converted into pure integers using the count() function, which counts how many rows there are in each data frames. For instance, the above data frame converts into 127.

The consolidation gives:

```
## # A tibble: 3 x 3
##   `Physical activity`  Male Female
##   <chr>               <int>  <int>
## 1 None                  127    116
## 2 Moderate              255    242
## 3 Intense               125    135
```

Results

Exploratory Data Analysis

Since the test is based on the normal approximation, all entries (O_{ij} and E_{ij}) in the table below must be at least 5:

```
## # A tibble: 3 x 3
##   `Physical activity`  Male Female
##   <chr>               <int>  <int>
## 1 None                  127    116
## 2 Moderate              255    242
## 3 Intense               125    135
```

Assuming they are all greater than or equal to 5, the P value can then be generated with the chisq.test() function.

Outcome

A χ^2 distribution is constructed by squaring a single standard normal distribution: $Q \sim \chi_i^2$ where Q is an example of a χ^2 distribution. Then $Q = Z^2$ where $Z \sim N(0, 1)$.

H_0 : the two variables are independent against each other. H_1 : not H_0.

The Pearson's χ^2 test-statistic (without continuity correction) for the test of independence is: $\tau = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim \chi^2_{(r-1)(c-1)}$, under H_0.

$\tau_{obs} = 1.0267993$
P value = 0.5984576

Conclusion

The P value is 0.5984576. Based on this P value, the conclusion to the test can be made:

```
## DO NOT REJECT H0:  0.5984576  > 0.05
##
## Gender does NOT affect the amount of physical activity: As the P-value is large,
## we have no evidence to reject H0. I.E. no evidence that the two variables are
## dependent against each other. The two variables are independent against each
## other and there is no association between gender and the amount of physical
## acitivity.
```

References

R Core Team. 2021. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller. 2021. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.