|  |
| --- |
|  |
| Exploratory Data Analysis |
| Swiss Data |
| Saukaurya DuttA |
| **Roll no. -400420** |
| **12/4/2016** |

|  |
| --- |
|  |

**Description of the data**

We are given a dataset “**swiss**” in R and we have to perform exploratory data analysis for the given dataset. The data is about “**Swiss Fertility and Socioeconomic Indicators (1888)”** and talks about standardized fertility measure and socioeconomic indicators for each of 47 french speaking provinces of Switzerland at about 1888. The dataset contains 6 variables, namely Fertility, Agriculture, Examination, Education, Catholic and Infant. Mortality.

The description of the six variables is given as follows:-

1. **Fertility** – It gives the common standardized fertility measure.
2. **Agriculture** – It gives the % of males involved in agriculture as an occupation.
3. **Examination** – It gives the % draftees receiving highest mark on army examination.
4. **Education** – It gives the % draftees receiving highest mark on army examination.
5. **Catholic** – It gives the % of Christians that are catholic.
6. **Infant.Mortality** – It gives the live births who live less than 1 year.

All variables, but ‘Fertility’ give the proportions of the population.

Switzerland, in 1888, was entering a period known as the demographic transition; i.e., its fertility was beginning to fall from the high level typical of underdeveloped countries. The data collected are for 47 French-speaking “provinces” at about 1888. Here, all variables are scaled to *[0, 100]*, where in the original, all but "Catholic" were scaled to*[0, 1]*.

Project “16P5”, pages 549–551 in Mosteller, F. And Tukey, J. W. (1977) Data Analysis and Regression: A Second Course in Statistics. Addison-Wesley, Reading Mass indicating their source as “Data used by permission of Franice van de Walle. Office of Population Research, Princeton University, 1976. Unpublished data, assembled under NICHD contract number No 1-HD-O-2077.”

The data looks as follows: - (We present the first six provinces of the data. It can be obtained by using the head() function.)

> head(swiss)

Fertility Agriculture Examination Education Catholic Infant.Mortality

Courtelary 80.2 17.0 15 12 9.96 22.2

Delemont 83.1 45.1 6 9 84.84 22.2

Franches-Mnt 92.5 39.7 5 5 93.40 20.2

Moutier 85.8 36.5 12 7 33.77 20.3

Neuveville 76.9 43.5 17 15 5.16 20.6

Porrentruy 76.1 35.3 9 7 90.57 26.6

**Note**

Files for all 182 districts in 1888 and other years have been available at<http://opr.princeton.edu/archive/eufert/switz.html> or <http://opr.princeton.edu/archive/pefp/switz.asp>.

They state that variables Examination and Education are averages for 1887, 1888 and 1889.

**Exploratory data analysis of single variables**

**Theory**

**Data Summarization**

**Measures of central Tendency**

The **arithmetic mean** of a variable is the sum of all values of the variable divided by the total no. of the observations of that variable. It is obtained using the mean() function in R.

The **median** of an ordered set of values is that value in that ordered set which divides the set into two equal parts. It is obtained using the median() function.

**Measures of Dispersion**

The **standard deviation(s.d)** is a measure of the dispersion of a set of data from its mean. The more spread apart the data, the higher the deviation. Standard deviation is calculated as the square root of variance. However, in R we have a function sd() to calculate the standard deviation of a variable.

The **range** of a variable is defined as the difference of the maximum and the minimum value. We calculate range in R by using the range() function which returns the maximum and the minimum value of the variable.

The **Inter-Quartile range (IQR)** is defined as the difference between the upper and lower quartiles,

IQR = *Q*3 −  *Q*1. In other words, the IQR is the 1st quartile subtracted from the 3rd quartile. In R, we can calculate the inter-quartile range by using the IQR() function.

**Data Visualization**

**Box-Plot**

In descriptive statistics, a **box plot** or **boxplot** is a convenient way of graphically depicting groups of numerical data through their quartiles. Box plots may also have lines extending vertically from the boxes (*whiskers*) indicating variability outside the upper and lower quartiles, hence the terms **box-and-whisker plot** and **box-and-whisker diagram**. From the box plots one can identify the skewness, dispersion,median, the IQR, outliers in the data. In R we can draw boxplots using the boxplot() function and the outliers can be obtained using the statement boxplot()$out.

**Histogram**

A **histogram** is a graphical representation of the distribution of numerical data. It is an estimate of the probability distribution of a continuous variable (quantitative variable) and was first introduced by Karl Pearson. In R we can draw a histogram by using hist() function and the lines() function will draw a frequency polygon over the histogram.

**Stem and Leaf Display**

A **stem and leaf display** is often called a stem plot but the latter term often refers to another chart type. A simple stem plot may refer to plotting a matrix of y values onto a common x axis, and identifying the common x value with a vertical line, and the individual y values with symbols on the line. Unlike histograms, stem and leaf displays retain the original data to at least two significant digits, and put the data in order, thereby easing the move to order-based inference and non-parametric statistics.

A basic stem and leaf display contains two columns separated by a vertical line. The left column contains the stems and the right column contains the leaves. In R we draw a stem and leaf display using the stem() function.

**Testing for Normality**

**Q-Q Plot**

In statistics, a **Q–Q plot** ("Q" stands for *quantile*) is a probability plot, which is a graphical method for comparing two probability distributions by plotting their quantiles against each other. First, the set of intervals for the quantiles is chosen. A point (*x*, *y*) on the plot corresponds to one of the quantiles of the second distribution (*y*-coordinate) plotted against the same quantile of the first distribution (*x*-coordinate). Thus the line is a parametric curve with the parameter which is the (number of the) interval for the quantile.

If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line *y* = *x*. If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line *y* = *x*. Q–Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions.

In R Q-Q plot can be obtained by using the function qqnorm() and qqline().

**Kolmogrov-Smirnov Test**

H0: The data follow a particular distribution(Normal in this case).

H1:The data donot follow the specified distribution.

The Kolmogrov-Smirnov test statistic is defined as,

Dn=sup(|Fn(x)-F(x)|)

Where F is the theoretical cumulative distribution of the distribution being tested, which must be a continuous distribution (i.e, no discrete distributions such as the Binomial or Poisson), and it must be fully specified (i.e, location, scale and shape parameters cannot be estimated from the data). We reject the null hypothesis if the observed value of the test statistic is greater than the critical value at α level of significance. In R we can perform the K-S test by using the function ks.test ().

**Shapiro-Wilk Test**

H0 : The samples come from a parent population with Normal distribution.

H1 : The samples do not come from a parent population with Normal distribution.

The Shapiro-Wilk test is a test of normality. Null hypothesis checks whether the sample came from a normally distributed population. The test statistic is :



W=

where x(i)(with parentheses enclosing the subscript index i) is the i-th order statistic, i.e., the i-th smallest number in the sample. x̅ is the sample mean. The constants a i are given by  where m=(m1,m2,…,mn)T and m1,…mn are the expected values of the order statistics of independent and identically distributed random variables sampled from the standard normal distribution, and V is the covariance matrix of those order statistics. The user may reject the null hypothesis if W is below a predetermined threshold.

**For the Shapiro-Wilk statistic**:

* If p is more than 0.05 {α}, we can be 95% {(1−α)100%} certain that the data are normally distributed.(In other words, we fail to reject the null hypothesis.)
* If p is less than 0.05 {α}, we can be 95% {(1 − α)100%} certain that the data are not normally distributed. (In other words, we reject the null hypothesis.)

We can perform this test in R using the shapiro.test() function.

**Fertility**

The variable “Fertility” gives the common standardized fertility measure of the 47 french speaking provinces in Switzerland in the year 1888.

The data type (discrete/continuous) of the variable can be obtained by using the str() function. If the function returns “num” then the variable is continuous or if it returns “int” then it is of discrete type.

> str(swiss$Fertility)

num [1:47] 80.2 83.1 92.5 85.8 76.9 76.1 83.8 92.4 82.4 82.9 ...

Thus, the data type of the variable “Fertility” is **continuous**.

**Measures of Central Tendency**

The **arithmetic mean** is,

> mean(swiss$Fertility)

[1] 70.14255

The **median** is,

> median(swiss$Fertility)

[1] 70.4

From the above results we can say that the mean of “Fertility “ is 70.14255 and median is 70.4.

**Measures of Dispersion**

The **standard deviation(s.d)** is,

> sd(swiss$Fertility)

[1] 12.4917

The **range** is,

> range(swiss$Fertility)

[1] 35.0 92.5

Thus, range is 57.5

The **Inter-Quartile range (IQR)** is,

> IQR(swiss$Fertility)

[1] 13.75

Thus for the variable Fertility we can say that the variable is well dispersed around its central value and the **s.d** is **12.4917**, **range** is **57.5** and **IQR** is **13.75**.

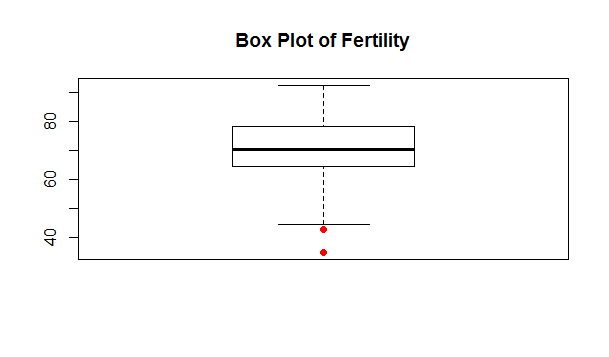
**Data Visualization**

**Box-Plot**

> boxplot(swiss$Fertility,main="Box Plot of Fertility",outcol="red",pch=19)

> boxplot(swiss$Fertility)$out

[1] 35.0 42.8

From the boxplot we can say about the symmetricity of the data by checking the no. of dots from the upper hinge to the 3rd quartile and the no. of dots from the lower hinge to the 1st quartile. If the no. of dots are equal, then we can say the data is symmetric. Also by looking at the position of the median we can conclude about the symmetricity, i.e if the median is about the centre of the box then the data is symmetric. Thus, from the above box plot we conclude that the data is not symmetric and contains outliers which are marked in red dots and the values are 35.0 and 42.8.

**Histogram**

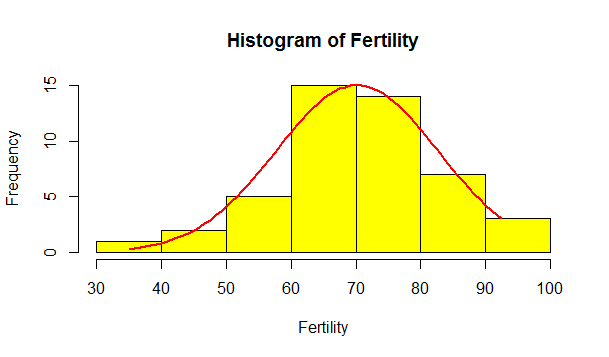
> h1<-hist(swiss$Fertility,main="Histogram of Fertility",xlab="Fertility",col="yellow")

> x<-seq(min(swiss$Fertility),max(swiss$Fertility),length=70)

> y<-dnorm(x,mean(swiss$Fertility),sd(swiss$Fertility))

> y<-y\*diff(h1$mids[1:2])\*length(swiss$Fertility)

> lines(x,y,col="red",lwd=2)



From the histogram also we find that the data of the Fertility variable is not symmetric and is negatively skewed.

**Stem and Leaf Display**

> stem(swiss$Fertility)

The decimal point is 1 digit(s) to the right of the |

3 | 5

4 | 35

5 | 46778

6 | 124455556678899

7 | 01223346677899

8 | 0233467

9 | 223

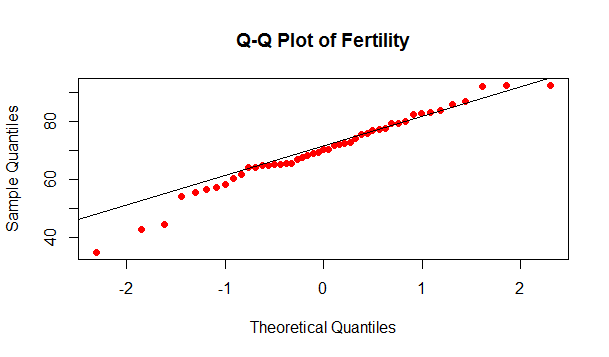
From the above diagram, we find that the modal class is 60-70 i.e, the mode of the variable lies in this class.

**Testing for Normality**

**Q-Q plot**

> qqnorm(swiss$Fertility,main="Q-Q Plot of Fertility",col="red",pch=19)

> qqline(swiss$Fertility)

****From the above plot, we conclude that the values of the variable lies more or less close to the normal line. Hence, we can say that the data is from a normal distribution.

**Kolmogrov-Smirnov Test**

> ks.test(swiss$Fertility,"pnorm",mean(swiss$Fertility),sd(swiss$Fertility))

One-sample Kolmogorov-Smirnov test

data: swiss$Fertility

D = 0.1015, p-value = 0.7179

alternative hypothesis: two-sided

From the results of the test, we find that the p-value>α=0.05. Thus, we accept the null hypothesis at 95% level of significance i.e, the data follows a normal distribution.

**Shapiro-Wilk Test**

> shapiro.test(swiss$Fertility)

Shapiro-Wilk normality test

data: swiss$Fertility

W = 0.9731, p-value = 0.3449

From the results of the test, we find that the p-value>α=0.05. Thus, we accept the null hypothesis at 95% level of significance i.e, the samples come from a parent population with normal distribution.

**Agriculture**

The variable “Agriculture” gives the percentage of males involved in agriculture as an occupation of the 47 french speaking provinces in Switzerland in the year 1888.

The data type (discrete/continuous) of the variable can be obtained by using the str() function. If the function returns “num” then the variable is continuous or if it returns “int” then it is of discrete type.

|  |
| --- |
| > str(swiss$Agriculture)  num [1:47] 17 45.1 39.7 36.5 43.5 35.3 70.2 67.8 53.3 45.2 ... |
|  |
| |  | | --- | |  | |

Thus, the data type of the variable “Agriculture” is **continuous**.

**Measures of Central Tendency**

The **arithmetic mean** is,

> mean(swiss$Agriculture)

[1] 50.65957

The **median** is,

> median(swiss$Agriculture)

[1] 54.1

From the above results we can say that the variable Agriculture “ have mean 50.65957 and median 54.1.

**Measures of Dispersion**

The **standard deviation(s.d)** is,

> sd(swiss$Agriculture)

[1] 22.71122

The **range** is,

> range(swiss$Agriculture)

[1] 1.2 89.7

Thus, range is 88.5

The **Inter-Quartile range (IQR)** is,

> IQR(swiss$Agriculture)

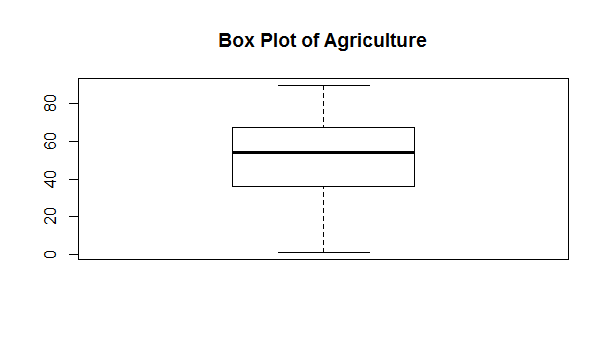
[1] 31.75

Thus for the variable Agriculture we can say that the variable is well dispersed around its central value and the **s.d** is **22.71122**, **range** is **88**.**5** and **IQR** is **31.75**.

**Data Visualization**

**Box-Plot**

> boxplot(swiss$Agriculture,main="Box Plot of Agriculture",outcol="red",pch=19)

> boxplot(swiss$Agriculture)$out

numeric(0)

The no. of dots above the box and below the box are not same, and also the median is not exactly in the middle of the box, so the data is slightly left skewed but from the above diagram we find that there are no outliers in the data.

**Histogram**

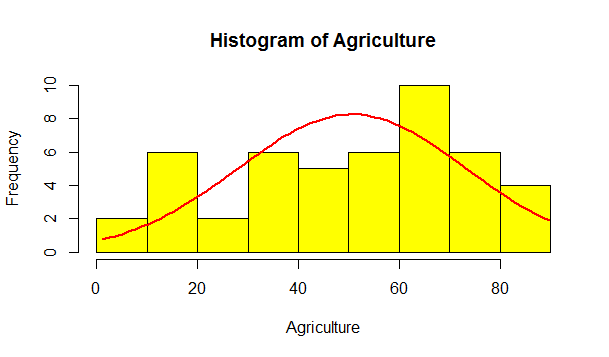
> h1<-hist(swiss$Agriculture,main="Histogram of Agriculture",xlab="Agriculture",col="yellow")

> x<-seq(min(swiss$Agriculture),max(swiss$Agriculture),length=70)

> y<-dnorm(x,mean(swiss$Agriculture),sd(swiss$Agriculture))

> y<-y\*diff(h1$mids[1:2])\*length(swiss$Agriculture)

> lines(x,y,col="red",lwd=2)

****From the histogram, we can conclude that the data is not purely symmetric but it is slightly skewed.

**Stem and Leaf Display**

> stem(swiss$Agriculture)

The decimal point is 1 digit(s) to the right of the |

0 | 18

1 | 577899

2 | 78

3 | 45788

4 | 04557

5 | 013458

6 | 01123455889

7 | 013368

8 | 556

9 | 0

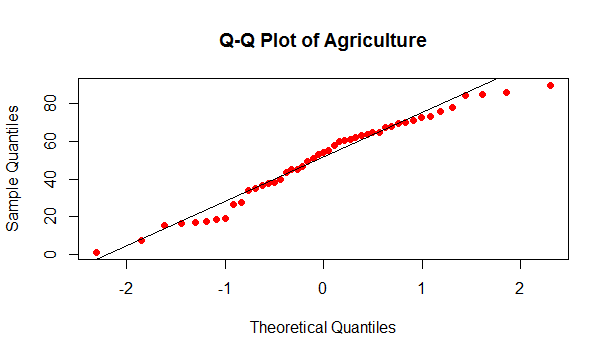
From the above display, we find that the modal class is 60-70, and hence the rest of the values are distributed around this class.

**Testing for Normality**

**Q-Q Plot**

> qqnorm(swiss$Agriculture,main="Q-Q Plot of Agriculture",col="red",pch=19)

> qqline(swiss$Agriculture)

****

From the above plot, we find that values of the variable lie more or less close to the normal line. Hence we conclude that the data comes from a normal distribution.

**Kolmogrov-Smirnov Test**

>ks.test(swiss$Agriculture,"pnorm",mean(swiss$Agriculture),sd(swiss$Agriculture))

One-sample Kolmogorov-Smirnov test

data: swiss$Agriculture

D = 0.1031, p-value = 0.6613

alternative hypothesis: two-sided

From the above test, we find that the p-value>α=0.05, thus we accept the null hypothesis at 95% level of significance and conclude that the data follows a normal distribution.

**Shapiro-Wilk Test**

> shapiro.test(swiss$Agriculture)

Shapiro-Wilk normality test

data: swiss$Agriculture

W = 0.9664, p-value = 0.193

From the above results, we find that the p-value>α=0.05, thus we accept the null hypothesis at 95% level of significance and conclude that the samples coming from a parent population has a normal distribution.

**Examination**

The variable “Examination” gives the percentage of draftees receiving highest mark in an army examination of the 47 french speaking provinces in Switzerland in the year 1888.

The data type (discrete/continuous) of the variable can be obtained by using the str() function. If the function returns “num” then the variable is continuous or if it returns “int” then it is of discrete type.

|  |
| --- |
| > str(swiss$Examination)  int [1:47] 15 6 5 12 17 9 16 14 12 16 ... |
|  |
| |  | | --- | |  |   Thus the data type of the variable is **discrete**. |

**Measures of Central Tendency**

The **arithmetic mean** is,

> mean(swiss$Examination)

[1] 16.48936

The **median** is,

> median(swiss$Examination)

[1] 16

From the above results we can say that the mean is 16.48936 and median is 16.

**Measures of Dispersion**

The **standard deviation(s.d)** is,

> sd(swiss$Examination)

[1] 7.977883

The **range** is,

> range(swiss$Examination)

[1] 3 37

Thus, range is 34.

The **Inter-Quartile range (IQR)** is,

> IQR(swiss$Examination)

[1] 10

Thus for the variable Examination we can say that the variable is well dispersed around its central value and the **s.d** is **7.977883**, **range** is **34** and **IQR** is **10**.

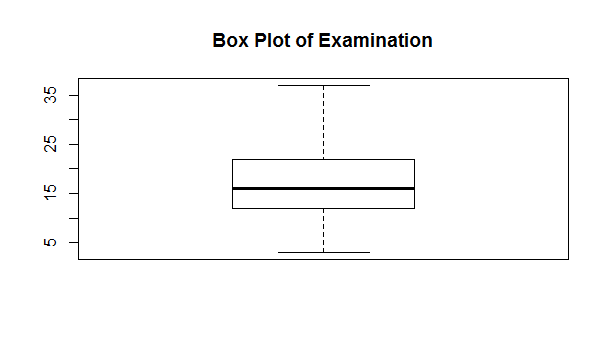
**Data Visualization**

**Box-Plot**

> boxplot(swiss$Examination,main="Box Plot of Examination",outcol="red",pch=19)

> boxplot(swiss$Examination)$out

numeric(0)

****

From the above diagram, we conclude that the no of dots below and above the box are not equal and also the median is not at the centre of the box, thus showing lack of symmetry in the data but in this data there are no outliers present.

**Histogram**

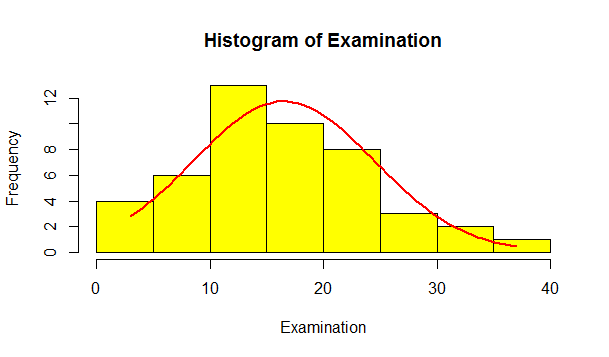
> h1<-hist(swiss$Examination,main="Histogram of Examination",xlab="Examination",col="yellow")

> x<-seq(min(swiss$Examination),max(swiss$Examination),length=70)

> y<-dnorm(x,mean(swiss$Examination),sd(swiss$Examination))

> y<-y\*diff(h1$mids[1:2])\*length(swiss$Examination)

> lines(x,y,col="red",lwd=2)

****From the above diagram, we conclude that the data is slightly positiively skewed and has a tail in the right.

**Stem and Leaf Display**

> stem(swiss$Examination)

The decimal point is 1 digit(s) to the right of the |

0 | 33

0 | 55667799

1 | 2222344444

1 | 555666677899

2 | 0122222

2 | 55669

3 | 1

3 | 57

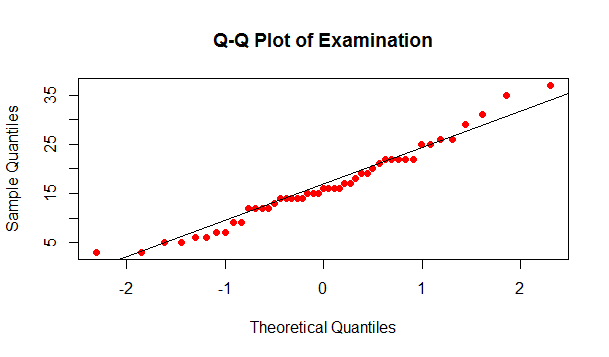
From the above diagram, we conclude that the modal class is 10-15 and the values of the variable are distributed around this class.

**Testing for Normality**

**Q-Q Plot**

> qqnorm(swiss$Examination,main="Q-Q Plot of Examination",col="red",pch=19)

> qqline(swiss$Examination)

****

From the above plot, we find that the values of the variable are more or less close to the normal line. Hence the data come from a normal distribution.

**Kolmogrov-Smirnov Test**

>ks.test(swiss$Examination,"pnorm",mean(swiss$Examination),sd(swiss$Examination))

One-sample Kolmogorov-Smirnov test

data: swiss$Examination

D = 0.0989, p-value = 0.7472

alternative hypothesis: two-sided

We can easily see that p-value>α=0.05, thus we accept the null hypothesis at 95% level of significance and conclude that the data follows a normal distribution.

**Shapiro-Wilk Test**

> shapiro.test(swiss$Examination)

Shapiro-Wilk normality test

data: swiss$Examination

W = 0.9696, p-value = 0.2563

Since, p-value>α=0.05, thus we accept the null hypothesis at 95% level of significance and conclude that the data comes from a parent population which follows a normal distribution.

**Education**

The variable “Education” gives the percentage of education beyond primary school for draftees of the 47 french speaking provinces in Switzerland in the year 1888.

The data type (discrete/continuous) of the variable can be obtained by using the str() function. If the function returns “num” then the variable is continuous or if it returns “int” then it is of discrete type.

|  |
| --- |
| > str(swiss$Education)  int [1:47] 12 9 5 7 15 7 7 8 7 13 ...  The data type of the variable is discrete. |
|  |
|  |

**Measures of Central Tendency**

The **arithmetic mean** is,

> mean(swiss$Education)

[1] 10.97872

The **median** is,

> median(swiss$Education)

[1] 8

From the above results we can say that the mean is 10.97872 and median is 8 .

**Measures of Dispersion**

The **standard deviation(s.d)** is,

> sd(swiss$Education)

[1] 9.615407

The **range** is,

> range(swiss$Education)

[1] 1 53

Thus, range is 52.

The **Inter-Quartile range (IQR)** is,

> IQR(swiss$Education)

[1] 6

Thus for the variable Education we can say that the variable is well dispersed around its central value and the **s.d** is **9.615407**, **range** is **52** and **IQR** is **6**.

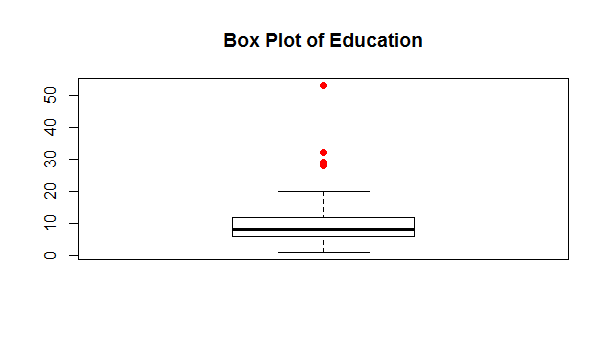
**Data Visualization**

**Box-Plot**

> boxplot(swiss$Education,main="Box Plot of Education",outcol="red",pch=19)

> boxplot(swiss$Education)$out

[1] 28 32 53 29 29

****From the above diagram, we conclude that the no. of dots above and below the box are not equal and also the median is not at all present at the centre of the box, hence the data is not at all symmetric and there is presence of some outliers also marked by the red dots.

**Histogram**

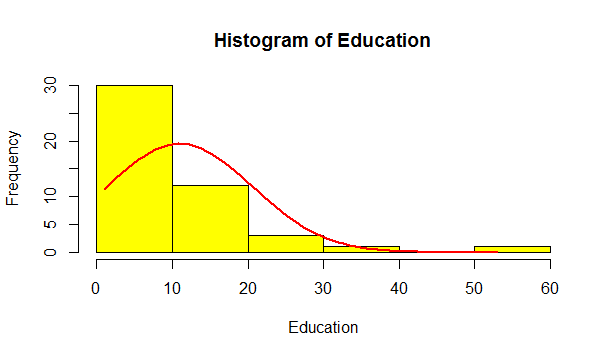
> h1<-hist(swiss$Education,main="Histogram of Education",xlab="Education",col="yellow")

> x<-seq(min(swiss$Education),max(swiss$Education),length=70)

> y<-dnorm(x,mean(swiss$Education),sd(swiss$Education))

> y<-y\*diff(h1$mids[1:2])\*length(swiss$Education)

> lines(x,y,col="red",lwd=2)

****Clearly, from the histogram we can say that the data is positively skewed and have a long right tail.

**Stem and Leaf Display**

> stem(swiss$Education)

The decimal point is 1 digit(s) to the right of the |

0 | 1222333355666677777778888999

1 | 0012222233359

2 | 0899

3 | 2

4 |

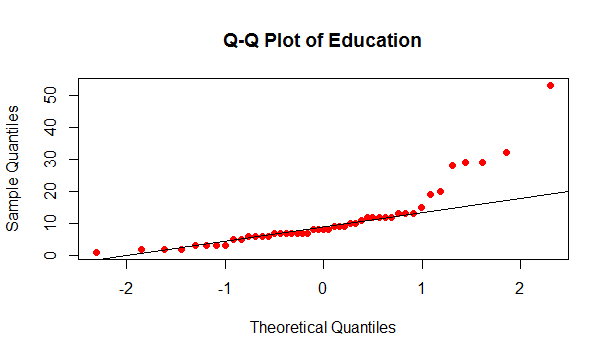
5 | 3

The modal class is from 0-10 as most of the values are present in this class, and this figure also tells about the skewness in the data.

**Testing for Normality**

**Q-Q Plot**

> qqnorm(swiss$Education,main="Q-Q Plot of Education",col="red",pch=19)

> qqline(swiss$Education)

Clearly from the plot, we observe that the values of the variable are scattered from the normal line, thus telling us that the data donot follow a normal distribution.

**Kolmogrov-Smirnov Test**

> ks.test(swiss$Education,"pnorm",mean(swiss$Education),sd(swiss$Education))

One-sample Kolmogorov-Smirnov test

data: swiss$Education

D = 0.2465, p-value = 0.006603

alternative hypothesis: two-sided

Since p-value<α=0.05, we reject the null hypothesis at 95% level of significance, and conclude that the data donot follow a normal distribution.

**Shapiro-Wilk Test**

> shapiro.test(swiss$Education)

Shapiro-Wilk normality test

data: swiss$Education

W = 0.7482, p-value = 1.312e-07

Since, p-value<α=0.05, we reject the null hypothesis at 95% level of significance and conclude that the data comes from a parent population which donot follow a normal distribution.

**Catholic**

The variable “Catholic” gives the percentage of catholics in the 47 french speaking provinces in Switzerland in the year 1888.

The data type (discrete/continuous) of the variable can be obtained by using the str() function. If the function returns “num” then the variable is continuous or if it returns “int” then it is of discrete type.

|  |
| --- |
| > str(swiss$Catholic)  num [1:47] 9.96 84.84 93.4 33.77 5.16 ... |
|  |
| |  | | --- | |  |   Thus the data type of the variable is continuous. |

**Measures of Central Tendency**

The **arithmetic mean** is,

> mean(swiss$Catholic)

[1] 41.14383

The **median** is,

|  |
| --- |
| > median(swiss$Catholic)  [1] 15.14 |
|  |
| |  | | --- | |  | |

From the above results we can say that the mean is 41.14383 and median is 15.14 .

**Measures of Dispersion**

The **standard deviation(s.d)** is,

> sd(swiss$Catholic)

[1] 41.70485

The **range** is,

> range(swiss$Catholic)

[1] 2.15 100.00

Thus, range is 97.85.

The **Inter-Quartile range (IQR)** is,

> IQR(swiss$Catholic)

[1] 87.93

Thus, for the variable Fertility we can say that the variable is well dispersed around its central value and the **s.d** is **41.70485**, **range** is **97.85** and **IQR** is **87.93**.

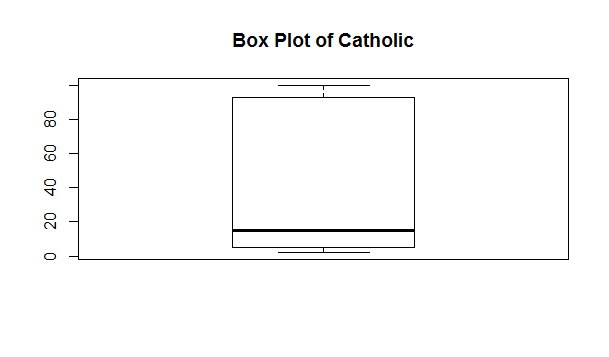
**Data Visualization**

**Box-Plot**

> boxplot(swiss$Catholic,main="Box Plot of Catholic",outcol="red",pch=19)

> boxplot(swiss$Catholic)$out

numeric(0)

****From the above diagram, we conclude that as the median is not at the centre of the box, we cannot say that the data is symmetric, however there is no presence of some outlier in the data.

**Histogram**

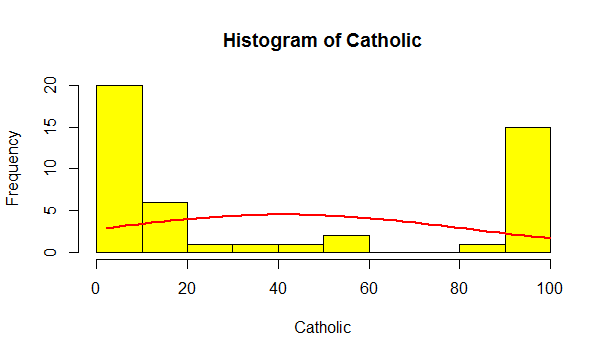
h1<-hist(swiss$Catholic,main="Histogram of Catholic",xlab="Catholic",col="yellow")

> x<-seq(min(swiss$Catholic),max(swiss$Catholic),length=70)

> y<-dnorm(x,mean(swiss$Catholic),sd(swiss$Catholic))

> y<-y\*diff(h1$mids[1:2])\*length(swiss$Catholic)

> lines(x,y,col="red",lwd=2)

****

From the histogram, we can conclude that the data is not at all symmetric.

**Stem and leaf display**

> stem(swiss$Catholic)

The decimal point is 1 digit(s) to the right of the |

0 | 2223333445555566899

1 | 0124578

2 | 4

3 | 4

4 | 2

5 | 08

6 |

7 |

8 | 5

9 | 113377889999

10 | 000

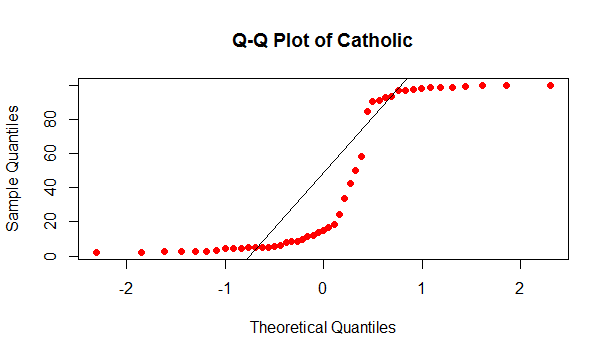
From the above display, we say that 0-10 is the modal class, and thus the values of the variable are distributed around this class.

**Testing for Normality**

**Q-Q Plot**

> qqnorm(swiss$Catholic,main="Q-Q Plot of Catholic",col="red",pch=19)

> qqline(swiss$Catholic)

****Clearly, we see that the values of the variable are scattered from the normal line, thus we conclude that the data doesnot follow a normal distribution.

**Kolmogrov-Smirnov Test**

> ks.test(swiss$Catholic,"pnorm",mean(swiss$Catholic),sd(swiss$Catholic))

One-sample Kolmogorov-Smirnov test

data: swiss$Catholic

D = 0.2599, p-value = 0.003488

alternative hypothesis: two-sided

Since, p-value<α=0.05, we reject the null hypothesis at 95% level of significance, and conclude that the data donot follow a normal distribution.

**Shapiro-Wilk Test**

> shapiro.test(swiss$Catholic)

Shapiro-Wilk normality test

data: swiss$Catholic

W = 0.7463, p-value = 1.205e-07

Since, p-value<α=0.05, we reject the null hypothesis at 95% level of significance and conclude that the data comes from a parent population which donot follow a normal distribution.

**Infant Mortality**

The variable “Infant.Mortality” gives the live births who live less than a year of the 47 french speaking provinces in Switzerland in the year 1888.

The data type (discrete/continuous) of the variable can be obtained by using the str() function. If the function returns “num” then the variable is continuous or if it returns “int” then it is of discrete type.

|  |
| --- |
|  |
| > str(swiss$Infant.Mortality)  num [1:47] 22.2 22.2 20.2 20.3 20.6 26.6 23.6 24.9 21 24.4 ... |
| |  | | --- | |  |   Thus the data type of the variable is continuous. |

**Measures of Central Tendency**

The **arithmetic mean** is,

> mean(swiss$Infant.Mortality)

[1] 19.94255

The **median** is,

> median(swiss$Infant.Mortality)

[1] 20

From the above results we can say that the mean is 19.94255 and median is 20.

**Measures of Dispersion**

The **standard deviation(s.d)** is,

> sd(swiss$Infant.Mortality)

[1] 2.912697

The **range** is,

> range(swiss$Infant.Mortality)

[1] 10.8 26.6

Thus, range is 15.8.

The **Inter-Quartile range (IQR)** is,

> IQR(swiss$Infant.Mortality)

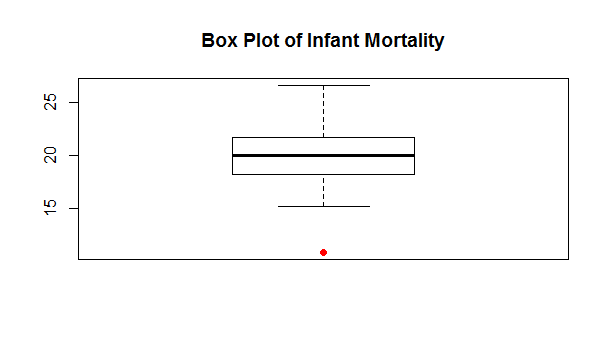
[1] 3.55

Thus, for the variable Infant Mortality we can say that the variable is not so well dispersed around its central value in comparison to the other variables and the **SD** is **2.912697**, **range** is 1**5.8** and **IQR** is **3.55**.

**Data Visualization**

**Box-Plot**

> boxplot(swiss$Infant.Mortality,main="Box Plot of Infant Mortality",outcol="red",pch=19)

> boxplot(swiss$Infant.Mortality)$out

[1] 10.8

From the above plot, we can conclude that the data is more or less symmetric however there is a presence of an outlier marked by a red dot in the graph.

**Histogram**

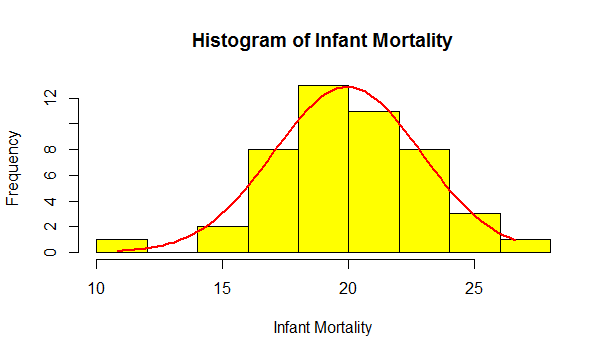
> h1<-hist(swiss$Infant.Mortality,main="Histogram of Infant Mortality",xlab="Infant Mortality",col="yellow")

> x<-seq(min(swiss$Infant.Mortality),max(swiss$Infant.Mortality),length=70)

> y<-dnorm(x,mean(swiss$Infant.Mortality),sd(swiss$Infant.Mortality))

> y<-y\*diff(h1$mids[1:2])\*length(swiss$Infant.Mortality)

> lines(x,y,col="red",lwd=2)

****

It is observable from the histogram that the data is more or less symmetric but has a slight long tail in the left.

**Stem and Leaf Display**

> stem(swiss$Infant.Mortality)

The decimal point is at the |

10 | 8

12 |

14 | 13

16 | 33578

18 | 0001237913458

20 | 00022233569002

22 | 22457068

24 | 459

26 | 6

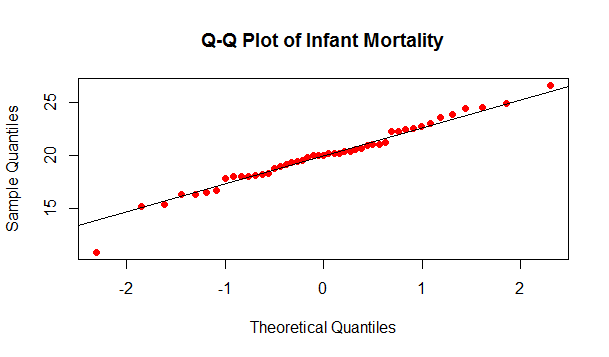
It is observable that the modal class is 20-22 and the values of the variable are distributed around this class.

**Testing for Normality**

**Q-Q Plot**

> qqnorm(swiss$Infant.Mortality,main="Q-Q Plot of Infant Mortality",col="red",pch=19)

> qqline(swiss$Infant.Mortality)

****Clearly, the values of the variable are more or less close to the normal line, thus the data follows a normal distribution.

**Kolmogrov-Smirnov Test**

>ks.test(swiss$Infant.Mortality,"pnorm",mean(swiss$Infant.Mortality),sd(swiss$Infant.Mortality))

One-sample Kolmogorov-Smirnov test

data: swiss$Infant.Mortality

D = 0.0822, p-value = 0.9086

alternative hypothesis: two-sided

Since, p-value>α=0.05, we accept the null hypothesis at 95% level of significance and conclude that the data follows a normal distribution.

**Shapiro-Wilk Test**

> shapiro.test(swiss$Infant.Mortality)

Shapiro-Wilk normality test

data: swiss$Infant.Mortality

W = 0.9776, p-value = 0.4978

Since, p-value>α=0.05, we accept the null hypothesis at 95% level of significance, and conclude that the data comes from a parent population which follows a normal distribution.

**Exploratory data analysis of multiple variables**

**Theory**

**Correlation**

**Correlation** is a measure of extent of linear association between two variables. The Pearsons’s product-moment correlation coefficient between two random variables X and Y is given as,



In R we can calculate correlation using the cor() function.

**Scatter Plot**

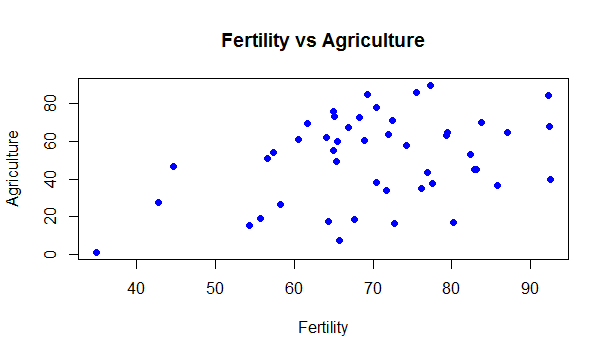
A **scatter plot** is just a graphical representation of the linear relation between two random variables, i.e. the correlation between the two variables. We use the plot() function for drawing a scatter plot.

**Fertility vs Agriculture**

> plot(swiss$Fertility,swiss$Agriculture,main="Fertility vs Agriculture",xlab="Fertility",ylab="Agriculture",col="blue",pch=19)

> cor(swiss$Fertility,swiss$Agriculture)

[1] 0.3530792

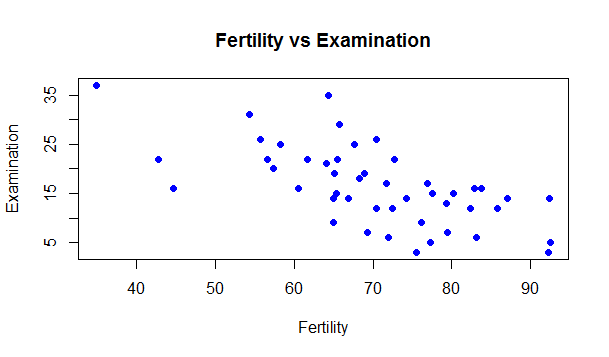
****

By looking at the scatter plot and the correlation we can conclude that the two variables are positively correlated.

**Fertility vs Examination**

> plot(swiss$Fertility,swiss$Examination,main="Fertility vs Examination",xlab="Fertility",ylab="Examination",col="blue",pch=19)

> cor(swiss$Fertility,swiss$Examination)

[1] -0.6458827

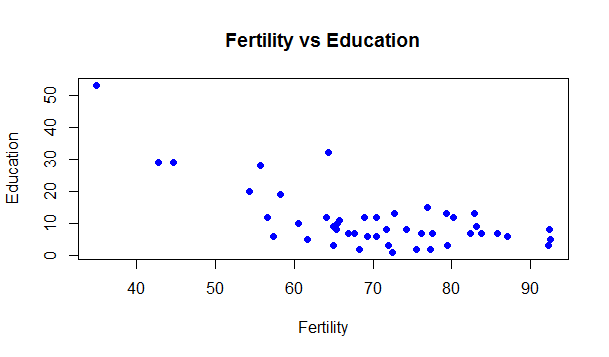
By looking at the scatter plot and the correlation we can conclude that the two variables are highly negatively correlated.

**Fertility vs Education**

> plot(swiss$Fertility,swiss$Education,main="Fertility vs Education",xlab="Fertility",ylab="Education",col="blue",pch=19)

> cor(swiss$Fertility,swiss$Education)

[1] -0.6637889

****

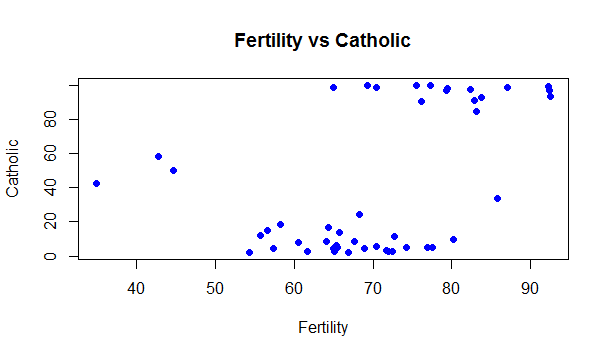
By looking at the scatter plot and the correlation we can conclude that the two variables are highly negatively correlated.

**Fertility vs Catholic**

> plot(swiss$Fertility,swiss$Catholic,main="Fertility vs Catholic",xlab="Fertility",ylab="Catholic",col="blue",pch=19)

> cor(swiss$Fertility,swiss$Catholic)

[1] 0.4636847

****

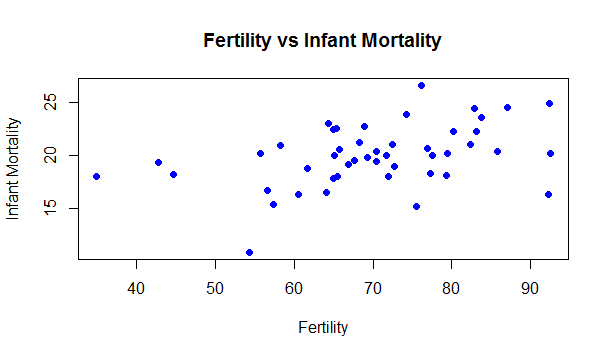
By looking at the scatter plot and the correlation we can conclude that the two variables are positively correlated.

**Fertility vs Infant Mortality**

> plot(swiss$Fertility,swiss$Infant.Mortality,main="Fertility vs Infant Mortality",xlab="Fertility",ylab="Infant Mortality",col="blue",pch=19)

> cor(swiss$Fertility,swiss$Infant.Mortality)

[1] 0.416556

****

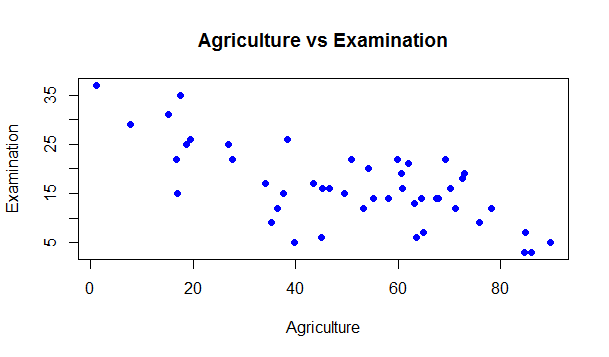
By looking at the scatter plot and the correlation we can conclude that the two variables are positively correlated.

**Agriculture vs Examination**

> plot(swiss$Agriculture,swiss$Examination,main="Agriculture vs Examination",xlab="Agriculture",ylab="Examination",col="blue",pch=19)

> cor(swiss$Agriculture,swiss$Examination)

[1] -0.6865422

****

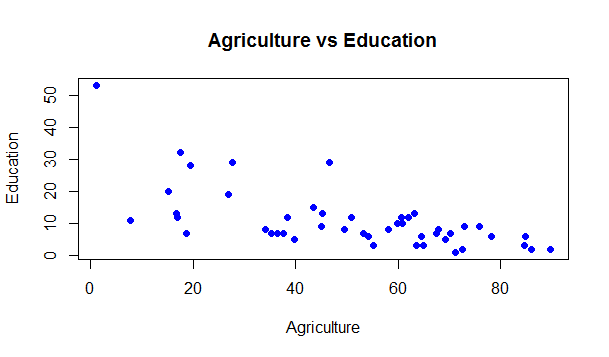
By looking at the scatter plot and the correlation we can conclude that the two variables are highly negatively correlated.

**Agriculture vs Education**

> plot(swiss$Agriculture,swiss$Education,main="Agriculture vs Education",xlab="Agriculture",ylab="Education",col="blue",pch=19)

> cor(swiss$Agriculture,swiss$Education)

[1] -0.6395225

****

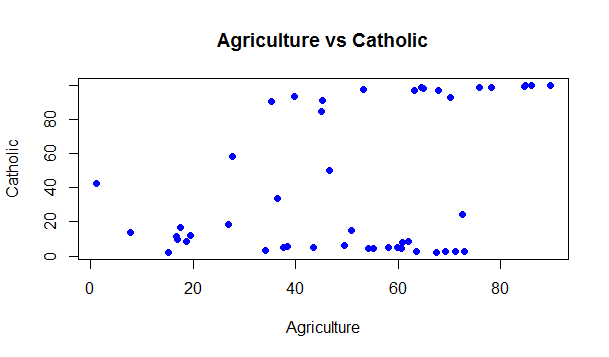
By looking at the scatter plot and the correlation we can conclude that the two variables are highly negatively correlated.

**Agriculture vs Catholic**

> plot(swiss$Agriculture,swiss$Catholic,main="Agriculture vs Catholic",xlab="Agriculture",ylab="Catholic",col="blue",pch=19)

> cor(swiss$Agriculture,swiss$Catholic)

[1] 0.4010951

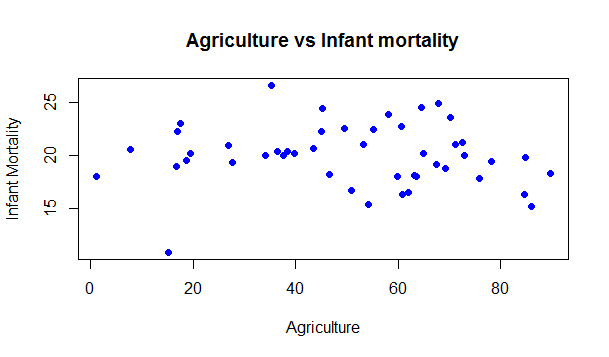
****

By looking at the scatter plot and the correlation we can conclude that the two variables are positively correlated.

**Agriculture vs Infant Mortality**

> plot(swiss$Agriculture,swiss$Infant.Mortality,main="Agriculture vs Infant mortality",xlab="Agriculture",ylab="Infant Mortality",col="blue",pch=19)

> cor(swiss$Agriculture,swiss$Infant.Mortality)

[1] -0.06085861

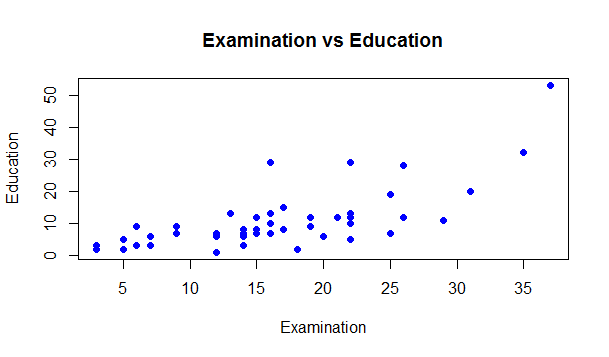
By looking at the scatter plot and the correlation we can conclude that the two variables are weakly negatively correlated.

**Examination vs Education**

> plot(swiss$Examination,swiss$Education,main="Examination vs Education",xlab="Examination",ylab="Education",col="blue",pch=19)

> cor(swiss$Examination,swiss$Education)

[1] 0.6984153

****

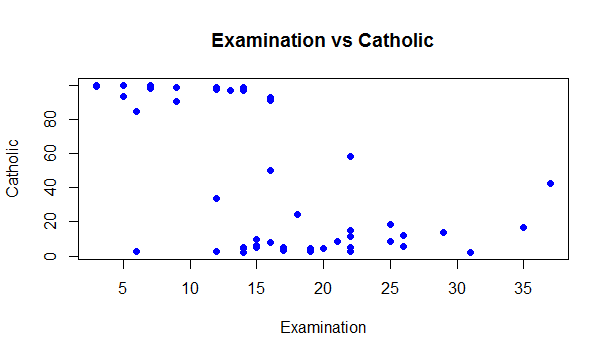
By looking at the scatter plot and the correlation we can conclude that the two variables are highly positively correlated.

**Examination vs Catholic**

> plot(swiss$Examination,swiss$Catholic,main="Examination vs Catholic",xlab="Examination",ylab="Catholic",col="blue",pch=19)

> cor(swiss$Examination,swiss$Catholic)

[1] -0.5727418

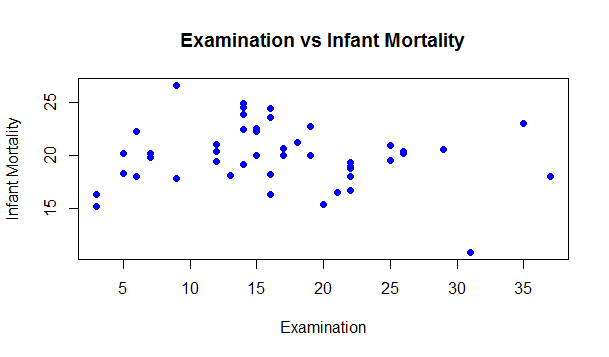
****

By looking at the scatter plot and the correlation we can conclude that the two variables are negatively correlated.

**Examination vs Infant Mortality**

> plot(swiss$Examination,swiss$Infant.Mortality,main="Examination vs Infant Mortality",xlab="Examination",ylab="Infant Mortality",col="blue",pch=19)

> cor(swiss$Examination,swiss$Infant.Mortality)

[1] -0.1140216

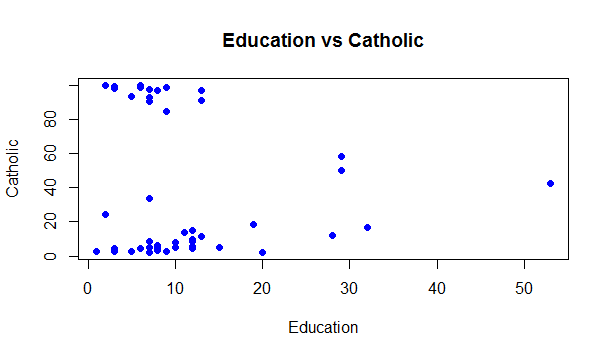
By looking at the scatter plot and the correlation we can conclude that the two variables are weakly negatively correlated.

**Education vs Catholic**

> plot(swiss$Education,swiss$Catholic,main="Education vs Catholic",xlab="Education",ylab="Catholic",col="blue",pch=19)

> cor(swiss$Education,swiss$Catholic)

[1] -0.1538589

****

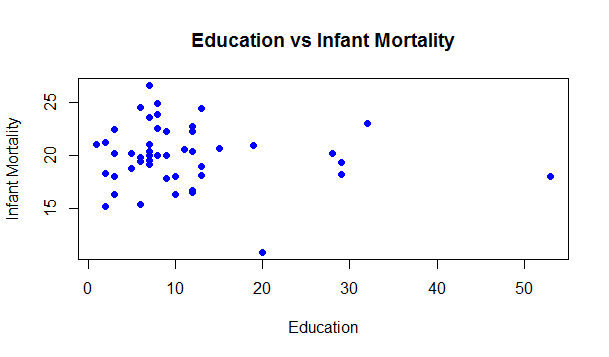
By looking at the scatter plot and the correlation we can conclude that the two variables are weakly negatively correlated.

**Education vs Infant Mortality**

> plot(swiss$Education,swiss$Infant.Mortality,main="Education vs Infant Mortality",xlab="Education",ylab="Infant Mortality",col="blue",pch=19)

> cor(swiss$Education,swiss$Infant.Mortality)

[1] -0.09932185

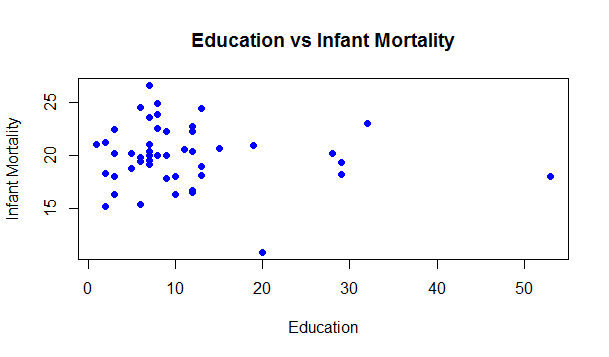
****

By looking at the scatter plot and the correlation we can conclude that the two variables are weakly negatively correlated or we can consider it to have zero correlation.

**Catholic vs Infant Mortality**

> plot(swiss$Catholic,swiss$Infant.Mortality,main="Catholic vs Infant Mortality",xlab="Catholic",ylab="Infant Mortality",col="blue",pch=19)

> cor(swiss$Catholic,swiss$Infant.Mortality)

[1] 0.1754959

By looking at the scatter plot and the correlation we can conclude that the two variables are weakly positively correlated or we can consider it to have zero correlation.

**Regression Analysis**

Let y be a dependent variable and x be a predictor, then the least square linear regression equation of y on x is given by,

Y=a+bx+e, where e is the error in predicting y.

and a=y̅-bx̅ , b= cov(x,y)/var(x).

We are to perform regression of agriculture on fertility. From the above scatter plot we observe that the two variables are positively correlated. Here the dependent variable is agriculture and the predictor is the fertility variable.

In R we can perform linear regression by using the lm() function.

> model<-lm(swiss$Agriculture~swiss$Fertility)

> summary(model)

Call:

lm(formula = swiss$Agriculture ~ swiss$Fertility)

Residuals:

Min 1Q Median 3Q Max

-40.116 -17.753 4.836 15.775 34.781

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 5.6326 18.0601 0.312 0.7566

swiss$Fertility 0.6419 0.2536 2.532 0.0149 \*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 21.48 on 45 degrees of freedom

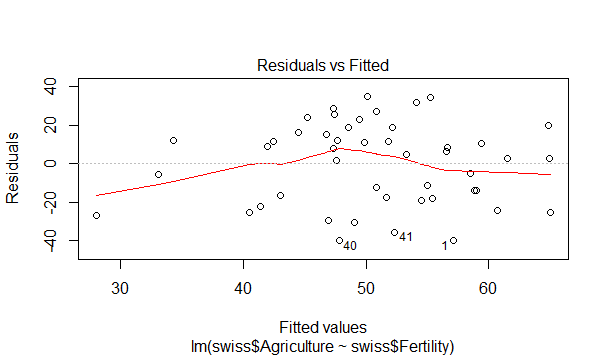
Multiple R-squared: 0.1247, Adjusted R-squared: 0.1052

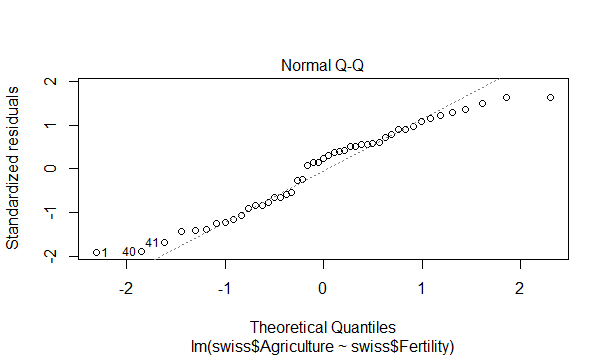
F-statistic: 6.409 on 1 and 45 DF, p-value: 0.01492

Hence the regression equation is , y=5.6326+0.6419\*x

This summary statistics tells us a number of things. One of them is the model p-Value (bottom last line) and the p-Value of individual predictor variables (extreme right column under ‘Coefficients’). The p-Values are very important because, We can consider a linear model to be statistically significant only when both these p-Values are less that the pre-determined statistical significance level, which is ideally 0.05. This is visually interpreted by the significance stars at the end of the row. The more the stars beside the variable’s p-Value, the more significant the variable is. Here we can see that the coefficient of the Fertility is more significant as compared to the intercept.

> plot(model)



****

From the residuals vs fitted graph if the no. of points above and below the fitted line(red line) is more or less equal, then the linearity condition holds. Clearly, here the same thing happens and we conclude that the linearity condition is satisfied. Also, from the Q-Q plot we see that the values are more or less close to the normal line, hence the data comes from a normal distribution, thus satisfying the normality assumption.

**Code**

#Fertility

str(swiss$Fertility)

#Measures of Central Tendency

mean(swiss$Fertility)

median(swiss$Fertility)

#measures of Dispersion

sd(swiss$Fertility)

IQR(swiss$Fertility)

range(swiss$Fertility)

#Data Visualization

boxplot(swiss$Fertility,main="Box Plot of Fertility",outcol="red",pch=19)

boxplot(swiss$Fertility)$out

stem(swiss$Fertility)

h1<-hist(swiss$Fertility,main="Histogram of Fertility",xlab="Fertility",col="yellow")

x<-seq(min(swiss$Fertility),max(swiss$Fertility),length=70)

y<-dnorm(x,mean(swiss$Fertility),sd(swiss$Fertility))

y<-y\*diff(h1$mids[1:2])\*length(swiss$Fertility)

lines(x,y,col="red",lwd=2)

#Testing for normality

qqnorm(swiss$Fertility,main="Q-Q Plot of Fertility",col="red",pch=19)

qqline(swiss$Fertility)

ks.test(swiss$Fertility,"pnorm",mean(swiss$Fertility),sd(swiss$Fertility))

shapiro.test(swiss$Fertility)

#Agriculture

str(swiss$Agriculture)

#Measures of Central Tendency

mean(swiss$Agriculture)

median(swiss$Agriculture)

#measures of Dispersion

sd(swiss$Agriculture)

IQR(swiss$Agriculture)

range(swiss$Agriculture)

#Data Visualization

boxplot(swiss$Agriculture,main="Box Plot of Agriculture",outcol="red",pch=19)

boxplot(swiss$Agriculture)$out

stem(swiss$Agriculture)

h1<-hist(swiss$Agriculture,main="Histogram of Agriculture",xlab="Agriculture",col="yellow")

x<-seq(min(swiss$Agriculture),max(swiss$Agriculture),length=70)

y<-dnorm(x,mean(swiss$Agriculture),sd(swiss$Agriculture))

y<-y\*diff(h1$mids[1:2])\*length(swiss$Agriculture)

lines(x,y,col="red",lwd=2)

#Testing for normality

qqnorm(swiss$Agriculture,main="Q-Q Plot of Agriculture",col="red",pch=19)

qqline(swiss$Agriculture)

ks.test(swiss$Agriculture,"pnorm",mean(swiss$Agriculture),sd(swiss$Agriculture))

shapiro.test(swiss$Agriculture)

#Examination

str(swiss$Examination)

#Measures of Central Tendency

mean(swiss$Examination)

median(swiss$Examination)

#measures of Dispersion

sd(swiss$Examination)

IQR(swiss$Examination)

range(swiss$Examination)

#Data Visualization

boxplot(swiss$Examination,main="Box Plot of Examination",outcol="red",pch=19)

boxplot(swiss$Examination)$out

stem(swiss$Examination)

h1<-hist(swiss$Examination,main="Histogram of Examination",xlab="Examination",col="yellow")

x<-seq(min(swiss$Examination),max(swiss$Examination),length=70)

y<-dnorm(x,mean(swiss$Examination),sd(swiss$Examination))

y<-y\*diff(h1$mids[1:2])\*length(swiss$Examination)

lines(x,y,col="red",lwd=2)

#Testing for normality

qqnorm(swiss$Examination,main="Q-Q Plot of Examination",col="red",pch=19)

qqline(swiss$Examination)

ks.test(swiss$Examination,"pnorm",mean(swiss$Examination),sd(swiss$Examination))

shapiro.test(swiss$Examination)

#Education

str(swiss$Education)

#Measures of Central Tendency

mean(swiss$Education)

median(swiss$Education)

#measures of Dispersion

sd(swiss$Education)

IQR(swiss$Education)

range(swiss$Education)

#Data Visualization

boxplot(swiss$Education,main="Box Plot of Education",outcol="red",pch=19)

boxplot(swiss$Education)$out

stem(swiss$Education)

h1<-hist(swiss$Education,main="Histogram of Education",xlab="Education",col="yellow")

x<-seq(min(swiss$Education),max(swiss$Education),length=70)

y<-dnorm(x,mean(swiss$Education),sd(swiss$Education))

y<-y\*diff(h1$mids[1:2])\*length(swiss$Education)

lines(x,y,col="red",lwd=2)

#Testing for normality

qqnorm(swiss$Education,main="Q-Q Plot of Education",col="red",pch=19)

qqline(swiss$Education)

ks.test(swiss$Education,"pnorm",mean(swiss$Education),sd(swiss$Education))

shapiro.test(swiss$Education)

#Catholic

str(swiss$Catholic)

#Measures of Central Tendency

mean(swiss$Catholic)

median(swiss$Catholic)

#measures of Dispersion

sd(swiss$Catholic)

IQR(swiss$Catholic)

range(swiss$Catholic)

#Data Visualization

boxplot(swiss$Catholic,main="Box Plot of Catholic",outcol="red",pch=19)

boxplot(swiss$Catholic)$out

stem(swiss$Catholic)

h1<-hist(swiss$Catholic,main="Histogram of Catholic",xlab="Catholic",col="yellow")

x<-seq(min(swiss$Catholic),max(swiss$Catholic),length=70)

y<-dnorm(x,mean(swiss$Catholic),sd(swiss$Catholic))

y<-y\*diff(h1$mids[1:2])\*length(swiss$Catholic)

lines(x,y,col="red",lwd=2)

#Testing for normality

qqnorm(swiss$Catholic,main="Q-Q Plot of Catholic",col="red",pch=19)

qqline(swiss$Catholic)

ks.test(swiss$Catholic,"pnorm",mean(swiss$Catholic),sd(swiss$Catholic))

shapiro.test(swiss$Catholic)

#Infant Mortality

str(swiss$Infant.Mortality)

##Measures of central Tendency

mean(swiss$Infant.Mortality)

median(swiss$Infant.Mortality)

##Measures of Dispersion

sd(swiss$Infant.Mortality)

IQR(swiss$Infant.Mortality)

range(swiss$Infant.Mortality)

#Data Visualization

boxplot(swiss$Infant.Mortality,main="Box Plot of Infant Mortality",outcol="red",pch=19)

boxplot(swiss$Infant.Mortality)$out

stem(swiss$Infant.Mortality)

h1<-hist(swiss$Infant.Mortality,main="Histogram of Infant Mortality",xlab="Infant Mortality",col="yellow")

x<-seq(min(swiss$Infant.Mortality),max(swiss$Infant.Mortality),length=70)

y<-dnorm(x,mean(swiss$Infant.Mortality),sd(swiss$Infant.Mortality))

y<-y\*diff(h1$mids[1:2])\*length(swiss$Infant.Mortality)

lines(x,y,col="red",lwd=2)

#Testing for Normality

qqnorm(swiss$Infant.Mortality,main="Q-Q Plot of Infant Mortality",col="red",pch=19)

qqline(swiss$Infant.Mortality)

ks.test(swiss$Infant.Mortality,"pnorm",mean(swiss$Infant.Mortality),sd(swiss$Infant.Mortality))

shapiro.test(swiss$Infant.Mortality)

#EDA for multiple variables

plot(swiss$Fertility,swiss$Agriculture,main="Fertility vs Agriculture",xlab="Fertility",ylab="Agriculture",col="blue",pch=19)

cor(swiss$Fertility,swiss$Agriculture)

plot(swiss$Fertility,swiss$Examination,main="Fertility vs Examination",xlab="Fertility",ylab="Examination",col="blue",pch=19)

cor(swiss$Fertility,swiss$Examination)

plot(swiss$Fertility,swiss$Education,main="Fertility vs Education",xlab="Fertility",ylab="Education",col="blue",pch=19)

cor(swiss$Fertility,swiss$Education)

plot(swiss$Fertility,swiss$Catholic,main="Fertility vs Catholic",xlab="Fertility",ylab="Catholic",col="blue",pch=19)

cor(swiss$Fertility,swiss$Catholic)

plot(swiss$Fertility,swiss$Infant.Mortality,main="Fertility vs Infant Mortality",xlab="Fertility",ylab="Infant Mortality",col="blue",pch=19)

cor(swiss$Fertility,swiss$Infant.Mortality)

plot(swiss$Agriculture,swiss$Examination,main="Agriculture vs Examination",xlab="Agriculture",ylab="Examination",col="blue",pch=19)

cor(swiss$Agriculture,swiss$Examination)

plot(swiss$Agriculture,swiss$Education,main="Agriculture vs Education",xlab="Agriculture",ylab="Education",col="blue",pch=19)

cor(swiss$Agriculture,swiss$Education)

plot(swiss$Agriculture,swiss$Catholic,main="Agriculture vs Catholic",xlab="Agriculture",ylab="Catholic",col="blue",pch=19)

cor(swiss$Agriculture,swiss$Catholic)

plot(swiss$Agriculture,swiss$Infant.Mortality,main="Agriculture vs Infant mortality",xlab="Agriculture",ylab="Infant Mortality",col="blue",pch=19)

cor(swiss$Agriculture,swiss$Infant.Mortality)

plot(swiss$Examination,swiss$Education,main="Examination vs Education",xlab="Examination",ylab="Education",col="blue",pch=19)

cor(swiss$Examination,swiss$Education)

plot(swiss$Examination,swiss$Catholic,main="Examination vs Catholic",xlab="Examination",ylab="Catholic",col="blue",pch=19)

cor(swiss$Examination,swiss$Catholic)

plot(swiss$Examination,swiss$Infant.Mortality,main="Examination vs Infant Mortality",xlab="Examination",ylab="Infant Mortality",col="blue",pch=19)

cor(swiss$Examination,swiss$Infant.Mortality)

plot(swiss$Education,swiss$Infant.Mortality,main="Education vs Infant Mortality",xlab="Education",ylab="Infant Mortality",col="blue",pch=19)

cor(swiss$Education,swiss$Infant.Mortality)

plot(swiss$Education,swiss$Catholic,main="Education vs Catholic",xlab="Education",ylab="Catholic",col="blue",pch=19)

cor(swiss$Education,swiss$Catholic)

plot(swiss$Catholic,swiss$Infant.Mortality,main="Catholic vs Infant Mortality",xlab="Catholic",ylab="Infant Mortality",col="blue",pch=19)

cor(swiss$Catholic,swiss$Infant.Mortality)

#Regression Analysis

model<-lm(swiss$Agriculture~swiss$Fertility)

summary(model)

plot(model)