Data Description and Descriptive Statistics with R

Dr. Emile Chimusa
Department of Integrative Biomedical Sciences
University of Cape Town

May 9, 2016

Contents

1 Features of Data Distributions

Given that the data have been appropriately displayed, the next step is to try to identify salient features represented in the graph. The acronym to remember is Center, Unusual features, Spread and Shape.

1.1 Center:

One of the most basic features of a data set is its center. Loosely speaking, the center of a data set is associated with a number that represents a middle or general tendency of the data. The sample mean is denoted $\bar{x}$ and is simply the arithmetic average of the observation:

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{1}{n}\sum_{n=1}^{n} x_i. \tag{1}$$

It is appropriate for use with data sets that are not highly skewed without extreme observations. The sample median is another popular measure of center and is denoted $\tilde{x}$. To calculate its value, first sort the data into an increasing sequence of numbers. If the data set has an odd number of observations then $\tilde{x}$ is the value of the middle observation, which lies in position $\frac{(n+1)}{2}$; otherwise, there are two middle observations and $\tilde{x}$ is the average of those middle values.

The sample median is another popular measure of center and is denoted $\tilde{x}$. To calculate its value, first sort the data into an increasing sequence of numbers. If the data set has an odd number of observations then $\tilde{x}$ is the value of the middle observation, which lies in position $\frac{(n+1)}{2}$; otherwise, there are two middle observations and $\tilde{x}$ is the average of those middle values. One desirable property of the sample median is that it is resistant to extreme observations, in the sense that the value of $\tilde{x}$ depends only the values of the middle observations, and is quite unaffected by the actual values of the outer observations in the ordered list.

Any significant changes in the magnitude of an observation x_k results in a corresponding change in the value of the mean. Hence, the sample mean is said to be sensitive to extreme observations. The trimmed mean is a measure designed to address the sensitivity of the sample mean to extreme observations. The idea is to "trim" a fraction (less than $(\frac{2}{2})$ of the observations off each end of the ordered list, and then calculate the sample mean of what remains. It denotes by $\bar{x}_{t=0.05}$.

```
> x <- c(74, 31, 95, 61, 76, 34, 23, 54, 96)
> mean(x)

[1] 60.44444

> median(x)

[1] 61

> mean(x, trim = 0.05)

[1] 60.44444
```

1.2 Order Statistics and the Sample Quantiles:

A common first step in an analysis of a data set is to sort the values. Given a data set $x_1x_2, \ldots\ldots, x_n$, we may sort the values to obtain an increasing sequence

$$x_{(1)} \leq x_{(2)} \leq x_{(3)} \leq \ldots, x_{(n)}.$$

and the resulting values are called the order statistics. The k^{th} entry in the list, $x_{(k)}$, is the k^{th} order statistic, and approximately $100(\frac{k}{n})\%$ of the observations fall below $x_{(k)}$. The order statistics give an indication of the shape of the data distribution, in the sense that a person can look at the order statistics and have an idea about where the data are concentrated, and where they are sparse. Suppose the data set has n observations. Find the sample quantile of order $p(0p < 1)$, denoted $\tilde{q}_p$, as follows:

(1) Sort the data to obtain the order statistics $x_{(1)} \leq x_{(2)} \leq x_{(3)} \leq \ldots, x_{(n)}$.

(2) calculate $(n-1)p+1$ and write it in the form $k.d$, where k is an integer and d is a decima.

(3) The sample quantile $\tilde{q}_p$ is

$$\tilde{q}_p = x_{(k)} + d(x_{(k+1)} - x(k))$$

The interpretation of $\tilde{q}_p$ is that approximately $100p\%$ of the data fall below the value $\tilde{q}_p$ and there is not a unique definition of percentiles, quartiles, etc. The most popular sample quantile is $\tilde{q}_{0.50}$, also known as the sample median, $\tilde{x}$

```
> sort(x)

[1] 23 31 34 54 61 74 76 95 96

> quantile(x, probs = c(0, 0.25, 0.37))

  0%  25%  37%
23.0 34.0 53.2
```

will return the smallest observation, the first quartile, $\tilde{q}_{0.25}$, and the 37^{th} sample quantile, $\tilde{q}_{0.37}$.

The interpretation of $\tilde{q}_p$ is that approximately $100p\%$ of the data fall below the value $\tilde{q}_p$ and there is not a unique definition of percentiles, quartiles, etc. The most popular sample quantile is $\tilde{q}_{0.50}$, also known as the sample median, $\tilde{x}$

```
> sort(x)

[1] 23 31 34 54 61 74 76 95 96
```

```
> quantile(x, probs = c(0, 0.25, 0.37))

  0%  25%  37%
23.0 34.0 53.2
```

will return the smallest observation, the first quartile, $\tilde{q}_{0.25}$, and the 37^{th} sample quantile, $\tilde{q}_{0.37}$.
Important Note: The interquartile range is defined as the difference between the third and the first quartile, that is $x_{0.75} - x_{0.25}$. It can be computed by the function

```
> IQR(x)

[1] 42
```

More specifically, the value $\frac{IQR(x)}{1.349}$ is a robust estimator of the standard deviation. The median absolute deviation (MAD) is defined as a constant times the median of the absolute deviations of the data from the median. In R it is computed by the function mad defined as the median of the sequence $|x_1 - x_{0.50}|, |x_2 - x_{0.50}, \ldots, |x_n - x_{0.50}|$ multiplied by the constant 1.4826. It equals the standard deviation in case the data come from a bell-shaped (normal) distributio. Because the interquartile range and the median absolute deviation are based on quantiles, these are robust against outliers.

1.3 Spread:

The spread of a data set is associated with its variability; data sets with a large spread tend to cover a large interval of values, while data sets with small spread tend to cluster tightly around a central value. The sample variance is denoted s^2 and is calculated

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{2}$$

The sample standard deviation is $s = \sqrt{s^2}$. The sample standard deviation is used to scale the estimate back to the measurement units of the original data and it is sensitive to extreme values.

1.4 Shape, Symmetry and Skewness:

When we speak of the shape of a data set, we are usually referring to the shape exhibited by an associated graphical display, such as a histogram. The shape can tell us a lot about any underlying structure to the data, and can help us decide which statistical procedure we should use to analyse them. The sample skewness, denoted by g_1 , is defined by

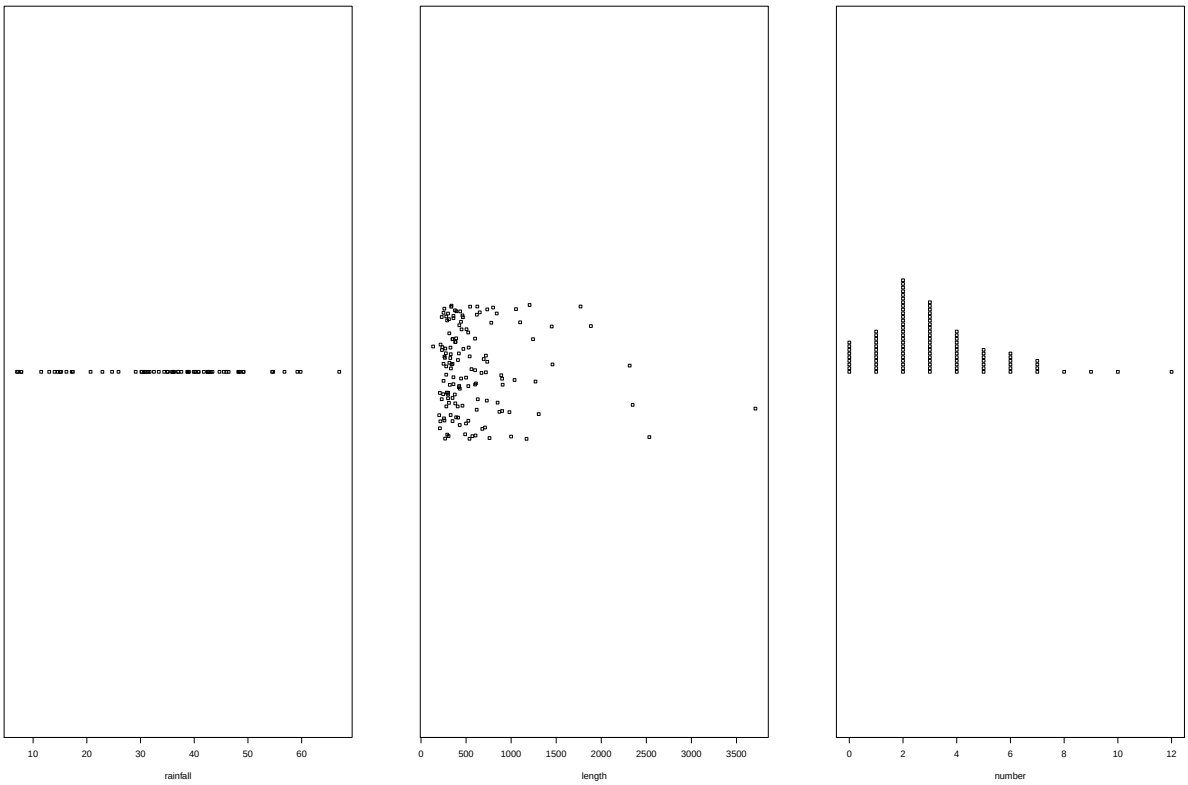
$$g_1 = \frac{1}{n}\frac{\sum_{i=1}^{n}(x_i - \bar{x})^3}{s^3} \tag{3}$$

The sample skewness can be any value $-\infty < g_1 < \infty$. A distribution is said to be right-skewed (or positively skewed) if the right tail seems to be stretched from the center. A

left-skewed (or negatively skewed) distribution is stretched to the left side. The data sets with skewness larger than $2\sqrt{\frac{6}{n}}$ in magnitude are substantially skewed, in the direction of the sign of g_1.

A symmetric distribution has a graph that is balanced about its center, in the sense that half of the graph may be reflected about a central line of symmetry to match the other half.

```
> par(mfrow=c(1,3))
> stripchart(precip, xlab = "rainfall")
> stripchart(rivers, method = "jitter", xlab = "length")
> stripchart(discoveries, method = "stack", xlab = "number")
```



1.5 Kurtosis:

Another component to the shape of a distribution is how "peaked" it is. The sample excess kurtosis, denoted by g_2, is given by

$$g_2 = \frac{1}{n}\frac{\sum_{i=1}^{n}(x_i - \bar{x})^4}{s^4} - 3$$

The sample excess kurtosis takes values $-2 \leq g_2 < \infty$. Samples with $g_2 > 0$ are called leptokurtic, and samples with $g_2 < 0$ are called platykurtic. Samples with $g_2 \simeq 0$ are called mesokurtic. An example of a platykurtic distribution is the uniform distribution. On the

other end of the spectrum are distributions with a steep peak, or spike, accompanied by heavy tails; these are called leptokurtic. Examples of leptokurtic distributions are the Laplace distribution and the logistic distribution. In between are distributions (called mesokurtic) with a rounded peak and moderately sized tails. The standard example of a mesokurtic distribution is the famous bellshaped curve, also known as the Gaussian, or normal, distribution, and the binomial distribution can be mesokurtic for specific choices of p. If $4\sqrt{\frac{6}{n}}$ then the sample excess kurtosis is substantially different from zero in the direction of the sign of g_2.

```
> install.packages('e1071',repo="http://cran.r-project.org", dep=TRUE)

The downloaded source packages are in
        '/tmp/RtmpOFUEk8/downloaded_packages'

> library(e1071)
> skewness(discoveries)

[1] 1.2076

> 2 * sqrt(6/length(discoveries))

[1] 0.4898979
```

The data are definitely skewed to the right. Let us check the sample excess kurtosis of the UKDriverDeaths data:

```
> kurtosis(UKDriverDeaths)

[1] 0.07133848

> 4 * sqrt(6/length(UKDriverDeaths))

[1] 0.7071068
```

The UKDriverDeaths data appear to be mesokurtic, or at least not substantially leptokurtic.

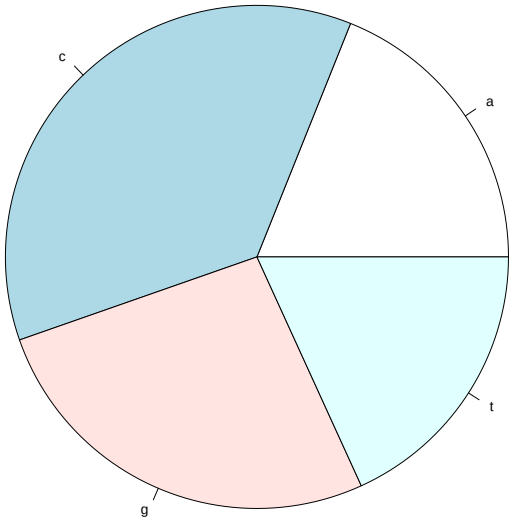
2 Basic Data Representation

2.1 Frequency table:

Discrete data occur when the values naturally fall into categories. A frequency table simply gives the number of occurrences within a category. The visualization of such data can be done using eg. pie in R.
Example: A gene consists of a sequence of nucleotides $\{A, C, G, T\}$. The number of each nucleotide can be displayed in a frequency table. This will be illustrated by the Zyxin gene which plays an important role in cell adhesion The accession number (X94991.1) of one of its variants can be found in a data base like NCBI (UniGene). The code below illustrates how

to read the sequence "X94991.1" of the species homo sapiens from GenBank, to construct a pie from a frequency table of the four nucleotides (for more details, ?pie).

```
> #setwd()
> #install.packages("ape_3.2.tar.gz")
> library(ape)
> bank <- table(read.GenBank(c("X94991.1"),as.character=TRUE))
> pie(bank)
```



2.2 stripchart:

An elementary method to visualize data is by using a so-called stripchart, by which the values of the data are represented as e.g. small boxes (for more details, ?stripchart). It is useful in combination with a factor that distinguishes members from different experimental conditions or patients groups.

Example: Many visualization methods will be illustrated by the Golub et al. (1999) data. We shall concentrate on the expression values of gene "CCND3 Cyclin D3", which are collected in row 1042 of the data matrix golub. To plot the data values one can simply use

```
> #source("http://www.bioconductor.org/biocLite.R")
> #biocLite("multtest")
> library("multtest")
> data(ALL)
> data(golub, package = "multtest")
> plot(golub[1042,], main="Plot of gene expression values of CCND3 Cyclin D3")
```

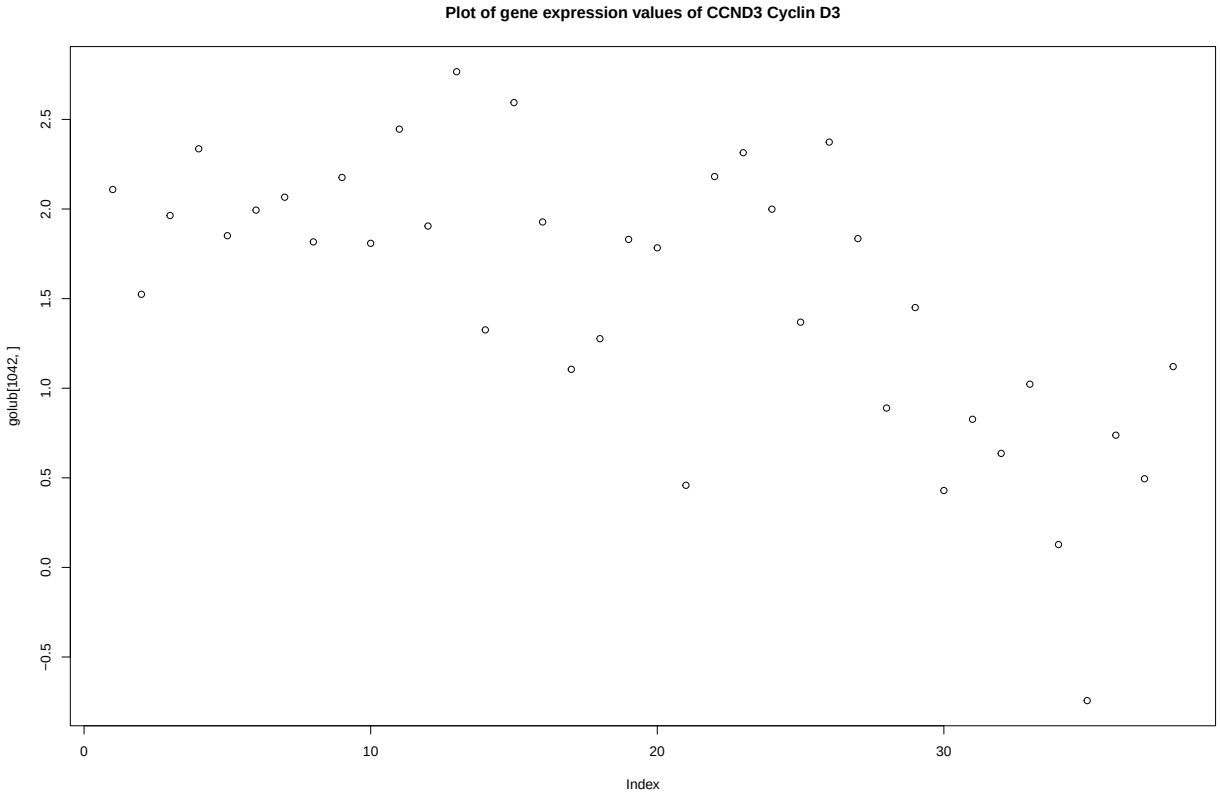


Figure 1: Plot of gene expression values of CCND3 Cyclin D3

In the resulting plot above, the vertical axis gives the size of the expression values and the horizontal axis the index of the patients (in Figure 1). It can be observed that the values for patient 28 to 38 are somewhat lower, but, indeed, the picture is not very clear because the groups are not plotted separately.

```
> data(golub, package = "multtest")
> gol.fac <- factor(golub.cl,levels=0:1, labels= c("ALL","AML"))
> stripchart(golub[1042,] ~ gol.fac, method="jitter", main="Stripchart of gene expre
```

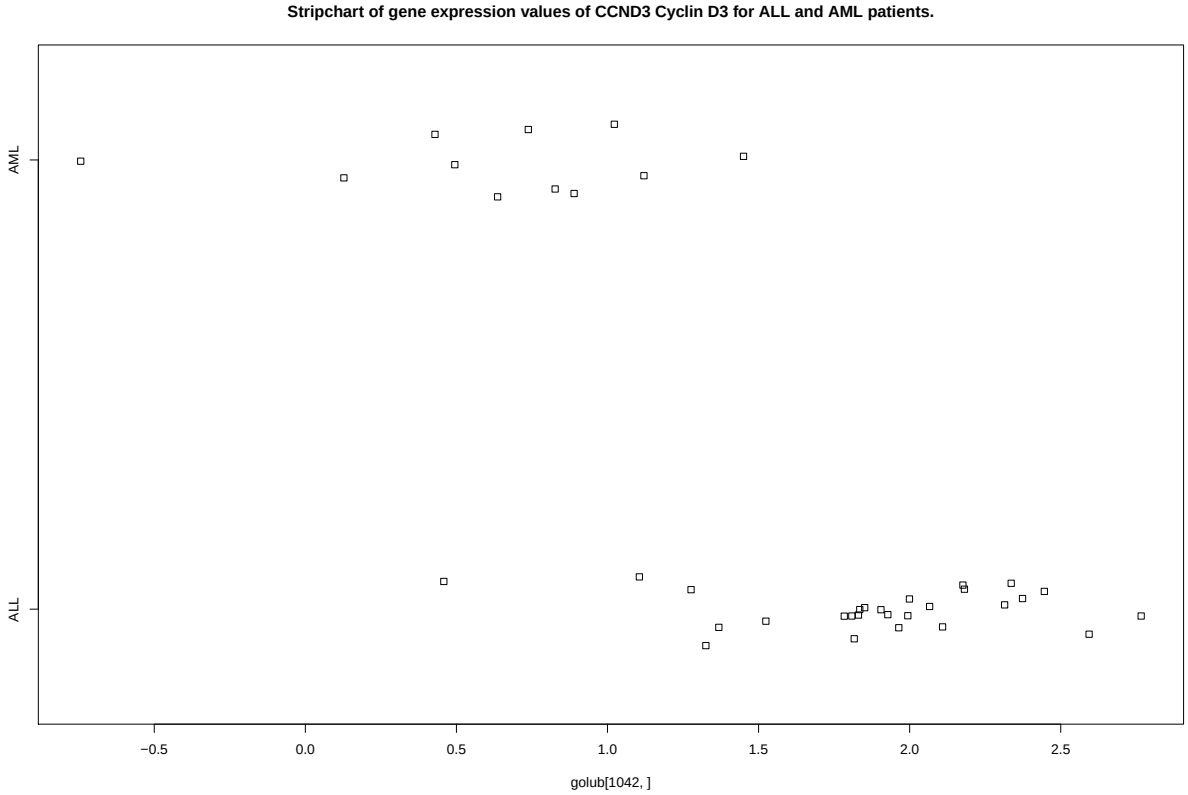


Figure 2: Stripchart of gene expression values of CCND3 Cyclin D3 for ALL and AML patients

To produce two adjacent stripcharts one for the ALL and one for the AML patients, we use the factor called gol.fa. From the above figure, it can be observed that the *CCND3* Cyclin D3 expression values of the ALL patients tend to have larger expression values than those of the AML patients (in Figure 2).

2.3 Histogram:

Another method to visualize data is by dividing the range of data values into a number of intervals and to plot the frequency per interval as a bar. Such a plot is called a histogram.
Example: A histogram of the expression values of gene "CCND3 Cyclin D3" of the acute lymphoblastic leukemia patients can be produced as follows.

```
> hist(golub[1042, gol.fac=="ALL"])
```

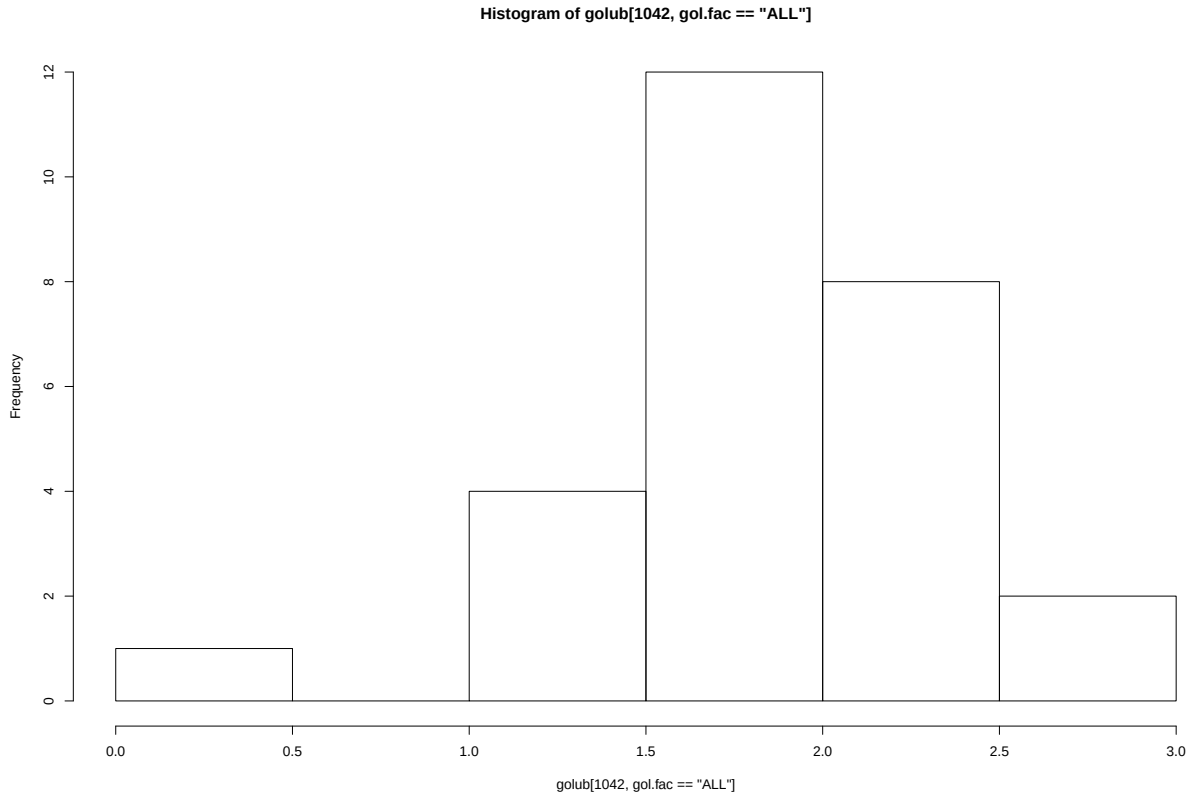


Figure 3: A histogram of the expression values of gene "CCND3 Cyclin D3" of the acute lymphoblastic leukemia patients

The function hist divides the data into 5 intervals having width equal to 0.5, see (Figure 3). Observe from the latter that one value is small and the other are more or less symmetrically distributed around the mean.

2.4 Boxplot:

It is always possible to sort n data values to have increasing order $x_{(1)} \leq x_{(2)} \leq x_{(3)} \leq \ldots, x_{(n)}$, where x_1 is the smallest, x_2 is the first-to-the smallest, etc. Let $x_{0.25}$ be a number for which it holds that 25% of the data values $x_{(1)} \leq x_{(2)} \leq x_{(3)} \leq \ldots, x_{(n)}$, where x_1 is smaller. That is, 25% of the data values lay on the left side of the number $x_{0.25}$, reason for which it is called the first quartile or the 25^{th} percentile. The second quartile is the value $x_{0.25}$ such that 50% of the data values are smaller. Similarly, the third quartile or 75^{th} percentile is the value $x_{0.75}$ such that 75% of the data is smaller. A popular method to display data is by drawing a box around the first and the third quartile (a bold line segment for the median), and the smaller line segments (whiskers) for the smallest and the largest data values. Such a data display is known as a box-and-whisker plot.

Example: A vector with gene expression values can be put into increasing order by the function sort. We shall illustrate this by the ALL expression values of gene "CCND3 Cyclin D3" in row 1042 of golub.

```
> x <- sort(golub[1042, gol.fac=="ALL"], decreasing = FALSE)
> par(mfrow=c(1,2))
> hist(x, main="ALL expression values of gene CCND3 Cyclin D3.")
> boxplot(x, main="ALL expression values of gene CCND3 Cyclin D3.")
```

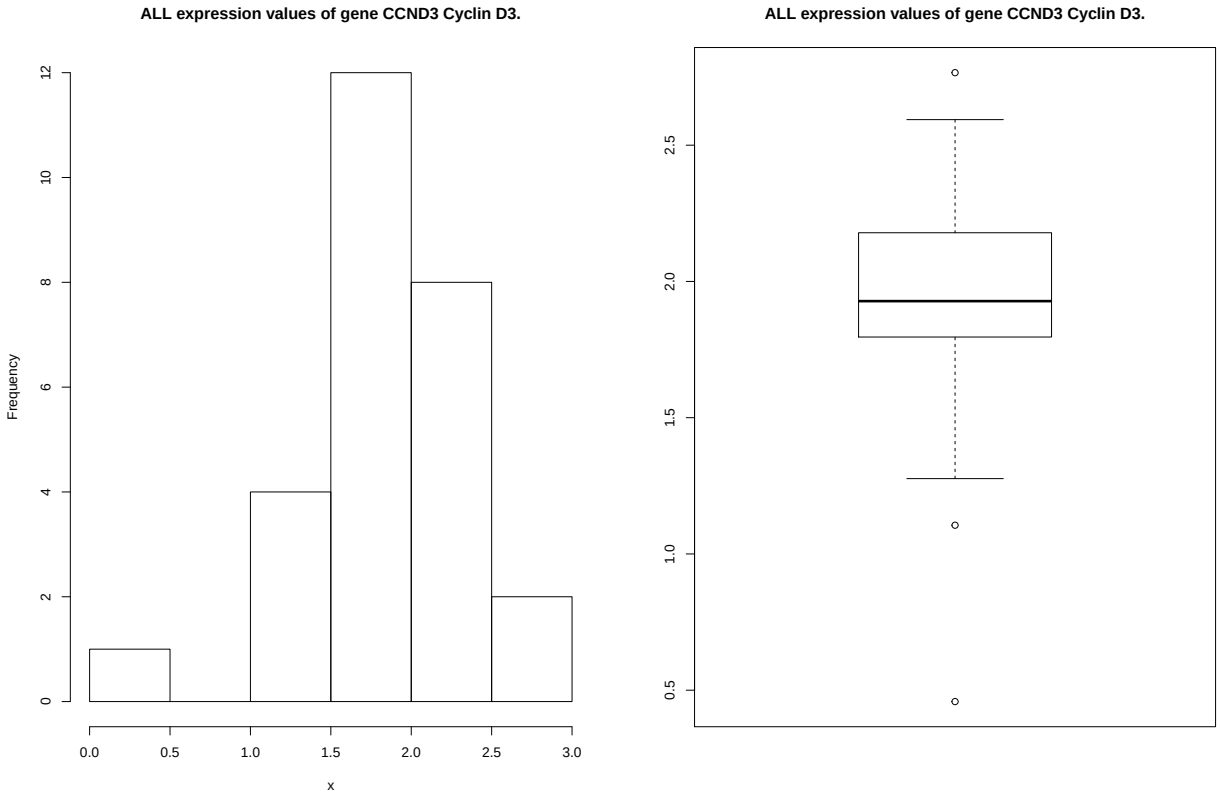


Figure 4: ALL expression values of gene CCND3 Cyclin D3

Now, a view on the distribution of the expression values of the ALL and the AML patients on gene CCND3 Cyclin D3 can be obtained by constructing two separate boxplots adjacent to one another (in Figure 5). To produce such a plot the factor gol.fac is again very useful.

```
> boxplot(golub[1042,] ~ gol.fac, main ="ALL and AML expression values of gene CCND3
```

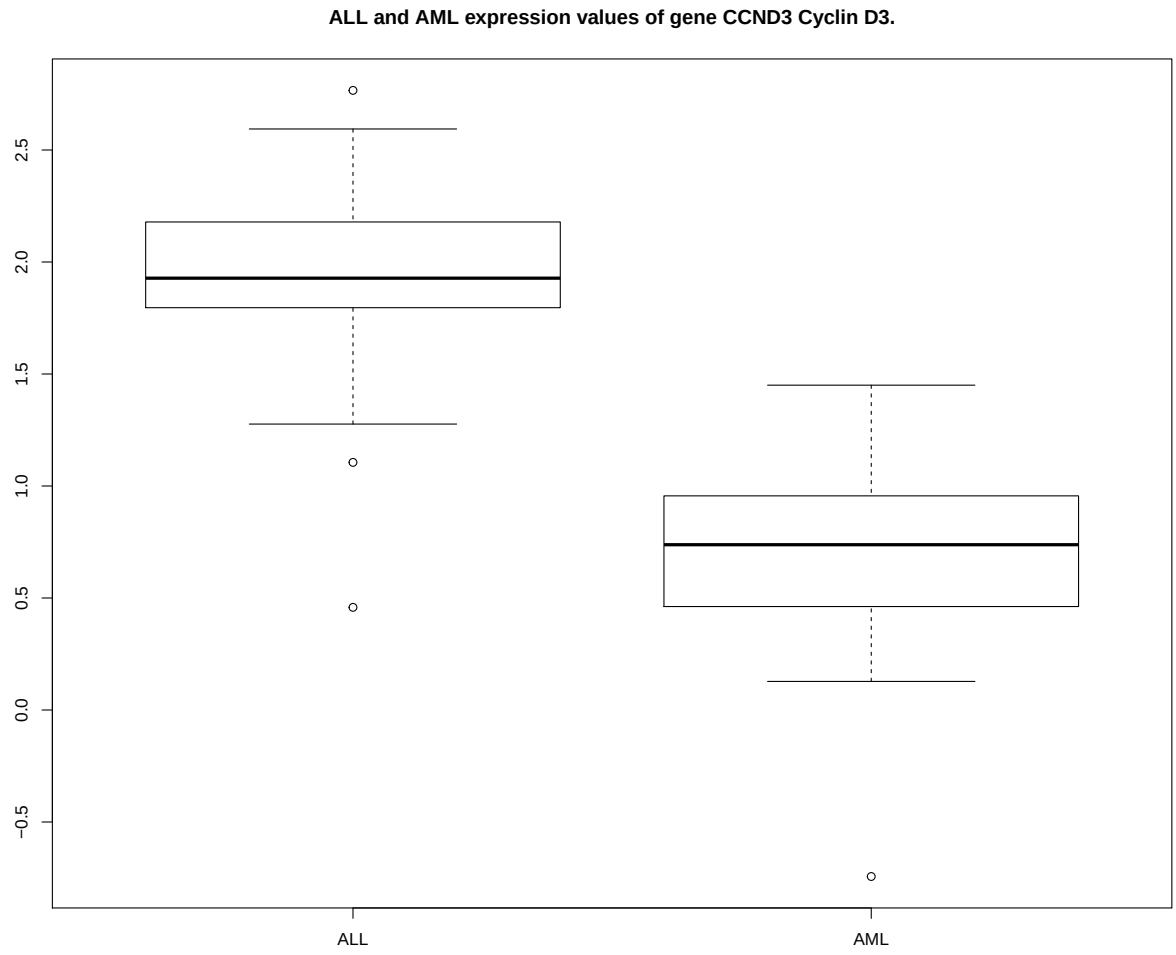


Figure 5: Boxplot:expression values of gene CCND3 Cyclin D3

From the position of the boxes in Figure 5, it can be observed that the gene expression values for ALL are larger than those for AML. Furthermore, since the two sub-boxes around the median are more or less equally wide, the data are quite symmetrically distributed around the median. To compute exact values for the quartiles we need a sequence running from 0.00 to 1.00 with steps equal to 0.25. To construct such a sequence the function seq is useful.

```
> pvec <- seq(0,1,0.25)
> quantile(golub[1042, gol.fac=="ALL"],pvec)

      0%       25%       50%       75%      100%
0.458270 1.796065 1.927760 2.178705 2.766100
```

The first quartile $x_{0.25} = 1.796$, the second $x_{0.50} = 1.928$, and the third $x_{0.75} = 2.179$. The

smallest observed expression value equals $x_{0.00} = 0.458$ and the largest $x_{1.00} = 2.77$. The latter can also be obtained by the function

```
> min(golub[1042, gol.fac=="ALL"])

[1] 0.45827
```

and

```
> max(golub[1042, gol.fac=="ALL"])

[1] 2.7661
```

or more briefly by

```
> range(golub[1042, gol.fac=="ALL"])

[1] 0.45827 2.76610
```

Outliers are data values laying far apart from the pattern set by the majority of the data values. The implementation in R of the (modified) boxplot draws such outlier points separately as small circles. A data point x is defined as an outlier point if

$$x < x_{0.25} - 1.5(x_{0.75} - x_{0.25}) \tag{4}$$

or

$$x > x_{0.75} + 1.5(x_{0.75} - x_{0.25} \tag{5}$$

From the boxplot figure above, it can be observed that there are outliers among the gene expression values of ALL patients. These are the smaller values 0.45827 and 1.10546, and the largest value 2.76610. The AML expression values have one outlier with value -0.74333. To define extreme outliers, the factor 1.5 is raised to 3.0. Note that this is a descriptive way of defining outliers instead of statistically testing for the existence of an outlier.

2.5 Quantile-Quantile (Q-Q) plot:

Generally speaking, Q-Q plot is a probability plot, which is a graphical method for comparing two probability distributions by plotting their quantiles against each other. To visualize the distribution of gene expression values is by the so-called quantile-quantile (Q-Q) plot. In such a plot the quantiles of the gene expression values are displayed against the corresponding quantiles of the normal (bell-shaped). A straight line is added representing points which correspond exactly to the quantiles of the normal distribution. By observing the extent in which the points appear on the line, it can be evaluated to what degree the data are normally distributed. That is, the closer the gene expression values appear to the line, the more likely it is that the data are normally distributed.

Example: To produce a Q-Q plot of the ALL gene expression values of CCND3 Cyclin D3 one may use the following.

```
> qqnorm(golub[1042, gol.fac=="ALL"])
> qqline(golub[1042, gol.fac=="ALL"])
```

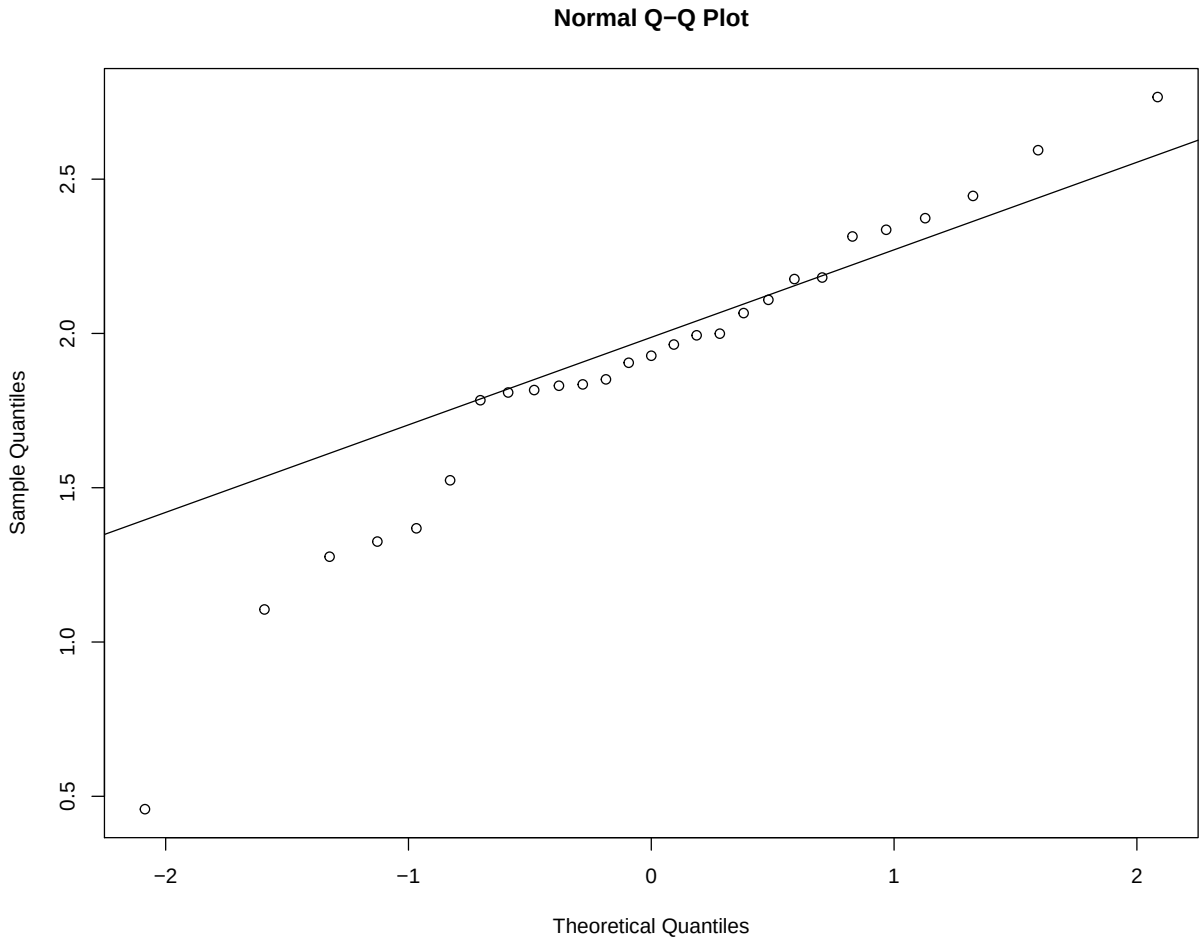


Figure 6: Boxplot:expression values of gene CCND3 Cyclin D3Q-Q plot of the ALL gene expression values of CCND3 Cyclin D3.

It can be observed from the Figure 6 above, that most of the data points are on or near the straight lin, while a few others are further away. The above example illustrates a case where the degree of non-normality is moderate so that a clear conclusion cannot be drawn.

3 Tutorial

(1) Comparing normality for two genes. Consider the gene expression values in row 790 and 66 of the Golub et al. (1999) data.

(a) Produce a boxplot for the expression values of the ALL patients and comment on the differences. Are there outliers?

(b) Produce a QQ-plot and formulate a hypothesis about the normality of the genes.

(c) Compute the mean and the median for the expression values of the ALL patients and compare these. Do this for both genes.

(2) Effect size. An important statistic to measure the effect size which is defined for a sample as x/s. It measures the mean relative to the standard deviation, so that is value is large when the mean is large and the standard deviation small.

(a) Determine the five genes with the largest effect size of the ALL patients from the Golub et al. (1999) data. Comment on their size.

(b) Invent a robust variant of the effect size and use it to answer the previous question.

(3) Plotting gene expressions "CCND3 Cyclin D3". Use the gene expressions from "CCND3 Cyclin D3" of Golub et al. (1999) collected in row 1042 of the object golub from the multtest library. After using the function plot you produce an object on which you can program.

(a) Produce a so-called stripchart for the gene expressions separately for the ALL as well as for the AML patients. Hint: Use a factor for appropriate separation.

(b) Rotate the plot to a vertical position and keep it that way for the questions to come.

(c) Color the ALL expressions red and AML blue. Hint: Use the col parameter.

(d) Add a title to the plot. Hint: Use title.

(e) Change the boxes into stars. Hint: Use the pch parameter. Hint: Store the final script you like the most in your typewriter in order to be able to use it efficiently later on.

(4) Box-and-Whiskers plot of "CCND3 Cyclin D3". Use the gene expressions "CCND3 Cyclin D3" of Golub et al. (1999) from row 1042 of the object golub of the multtest library.

(a) Construct the boxplot.

(b) Add text to the plot to explain the meaning of the upper and lower part of the box.

(c) Do the same for the wiskers.

(d) Export your plot to eps format.

Hint 1: locator() to find coordinates of the position of the plot.
Hint 2: lim to make the plot somewhat wider.
Hint 3: arrows to add an arrow.
Hint 4: text to add information at a certain position.

(5) Box-and-wiskers plot of persons of Golub et al. (1999) data.

   (a) Use boxplot(data.frame(golub)) to produce a box-and-wiskers plot for each column (person). Make a screen shot to save it in a word processor. Describe what you see. Are the medians of similar size? Is the inter quartile range more or less equal. Are there outliers?

   (b) Compute the mean and medians of the persons. What do you observe?

   (c) Compute the range (minimal and maximum value) of the standard deviations, the IQR and MAD of the persons. Comment of what you observe.

(6) Oncogenes of Golub et al. (1999) data.

   (a) Select the oncogens by the grep facility and produce a box-and-wiskers plot of the gene expressions of the ALL patients.

   (b) Do the same for the AML patients and use par(mfrow=c(2,1)) to combine the two plots such that the second is beneath the first. Are there genes with clear differences between the groups

(7) Descriptive statistics for the ALL gene expression values of the Golub et al. (1999) data.

   (a) Compute the mean and median for gene expression values of the ALL patients, report their range and comment on it.

   (b) Compute the SD, IQR, and MAD for gene expression values of the ALL patients, report their range and comment on it.