Distributions and basic exploratory data analysis

Assessment: Read in the data here. Use the `filter` function to determine how many males and females we have.

Assessment: Use filter and select to obtain a `data.frame` with just male heights. How many NAs do we have when we convert them to numeric?

Exploratory Data Analysis

An indispensable part of data wrangling is exploratory data analysis. In particular with large dataset, it is practically impossible to examine the data from potential errors by looking at tables. Here we describe some simple visualization techniques that are quite powerful for summarizing data and identifying potential errors.

Distributions

The most basic statistical summary of a list of numbers is it's distribution. The simplest way to think of a *distribution* is as a compact description of many numbers. For example, we have measured the heights of all students in a course. Imagine you need to describe these numbers to someone that has no idea what these heights are, such as an alien that has never visited Earth.

One approach to summarizing these numbers is to simply list them all out for the alien to see. Here are 10 randomly selected heights :

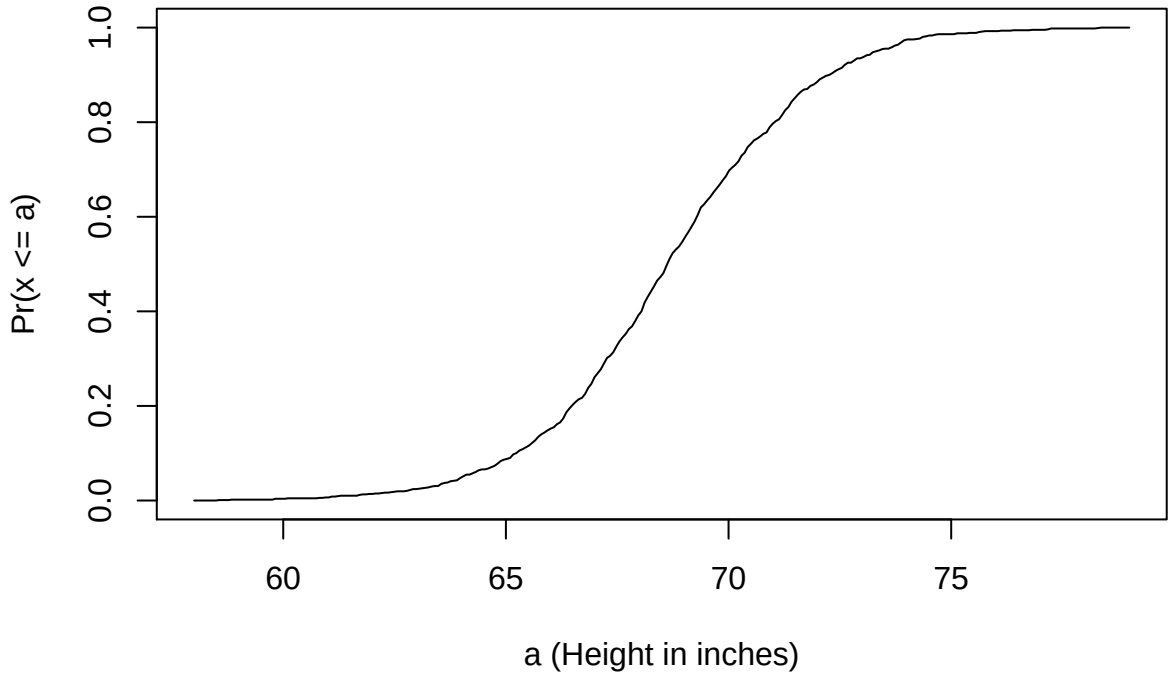
```
library(dplyr)
library(tidyr)
select(dat, height) %>% print(n=10)
```

```
## Source: local data frame [148 x 1]
##
##     height
##      (dbl)
## 1  63.0000
## 2  62.0000
## 3  69.0000
## 4  68.0000
## 5  71.6500
## 6  75.0000
## 7  68.8976
## 8  74.0000
## 9  65.0000
## 10 64.0000
## ..     ...
```

Cumulative Distribution Function Scanning through these numbers, we start to get a rough idea of what the entire list looks like, but it is certainly inefficient. We can quickly improve on this approach by defining and visualizing a *distribution*. To define a distribution we compute, for all possible values of a, the proportion of numbers in our list that are below a. We use the following notation:

$$F(a) \equiv \Pr(x \leq a)$$

This is called the cumulative distribution function (CDF). When the CDF is derived from data, as opposed to theoretically, we also call it the empirical CDF (ECDF). The ECDF for the adult male height data looks like this:



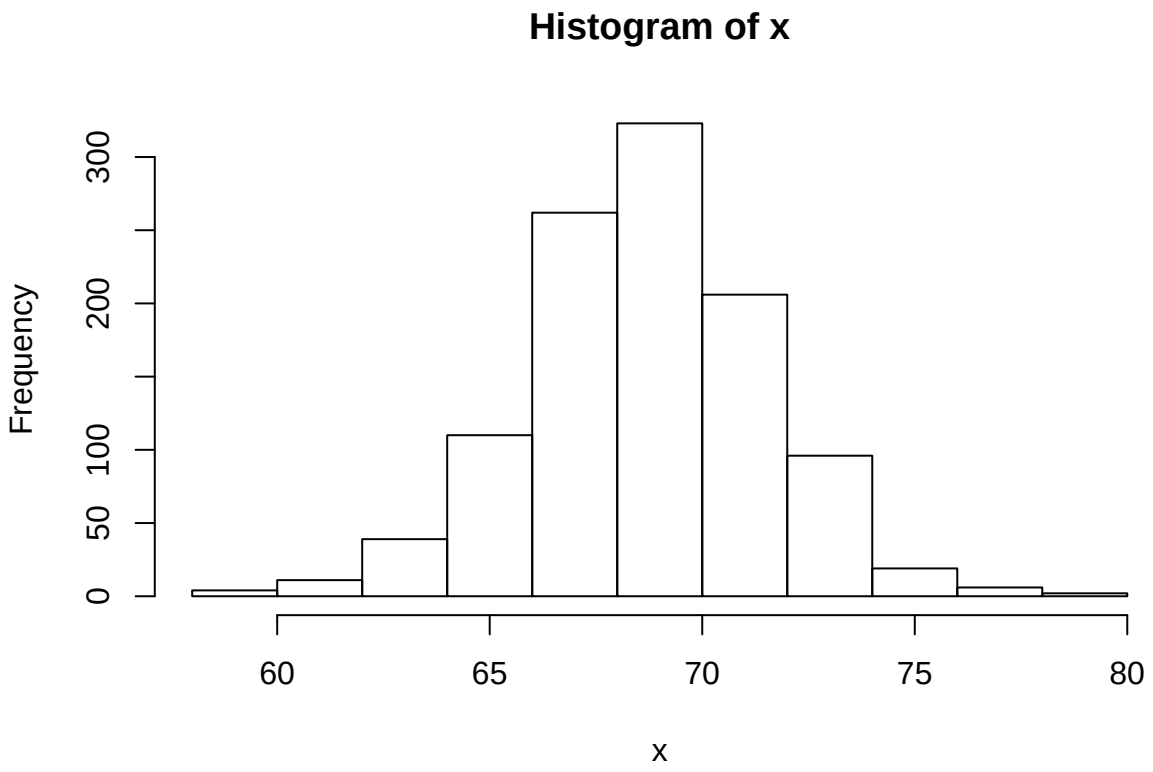
Histograms

Although the empirical CDF concept is widely discussed in statistics textbooks, the plot is actually not very popular in practice. The reason is that histograms give us the same information and are easier to interpret. Histograms show us the proportion of values in intervals:

$$\Pr(a \leq x \leq b) = F(b) - F(a)$$

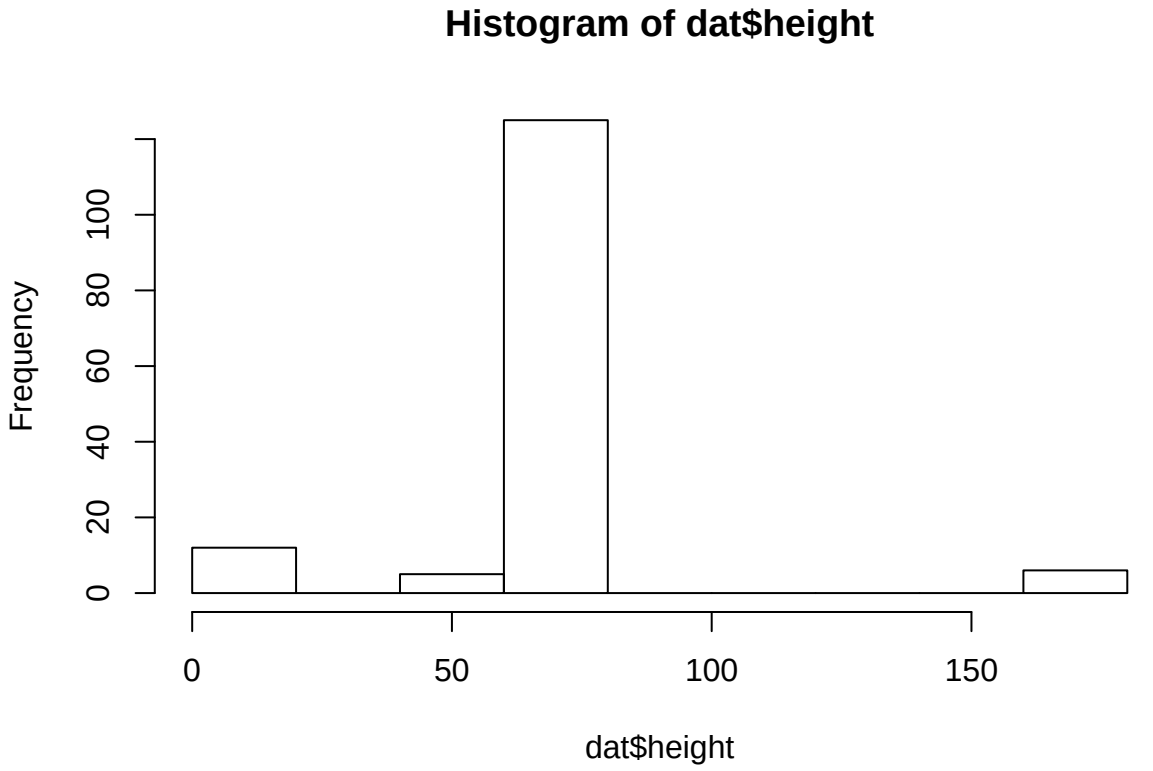
Plotting these heights as bars is what we call a *histogram*. It is a more useful plot because we are usually more interested in intervals, such and such percent are between 70 inches and 71 inches, etc., rather than the percent less than a particular height. It is also easier to distinguish different types (families) of distributions by looking at histograms. Here is a histogram for the general population:

```
data(father.son,package="UsingR")
x <- father.son$sheight
hist(x)
```



Here is a histogram for our class:

```
hist(dat$height)
```



Outliers

The plot has revealed another problem. We have a number hieghts that are larger than 96 inches and shorter than 12 inches. Let's view the data for which this is the case. To do this we introduce the or logical operator |

```
filter(dat, height>96 | height < 12) %>% select(original)
```

We see several heights that appear to be in centimeters. We will go ahead and assume this is the case and make the convertion:

```
dat <- mutate(dat, height=ifelse(height>96, height/2.54, height))
```

Now let's see what outliers remain:

```
filter(dat, height>96 | height < 12) %>% select(height)
```

```
## Source: local data frame [12 x 1]
##
##    height
##     (dbl)
## 1    5.80
## 2    5.10
## 3    5.11
## 4    5.70
## 5    5.00
## 6    5.90
## 7    5.20
## 8    5.50
## 9    5.51
## 10   5.80
## 11   5.70
## 12   6.00
```

These values appear to use the format $x.y$ with x feet and y inches. Note that these are not numbers. In particualr note that 5.11 is larger than 5.5. We also not a particularly strange entry which we will just assume is 5.5:

```
dat <- mutate(dat, height=ifelse(height==5.51, 65, height))
```

For the rest we convert to inches using the functions `floor`. This approach does not work for `5.11` so we treat that differently:

```
dat <- mutate(dat, height=ifelse(height==5.11, 71, height))
dat <- mutate(dat, height=ifelse(height>12, height, floor(height)*12+(height-floor(height))*10))
```

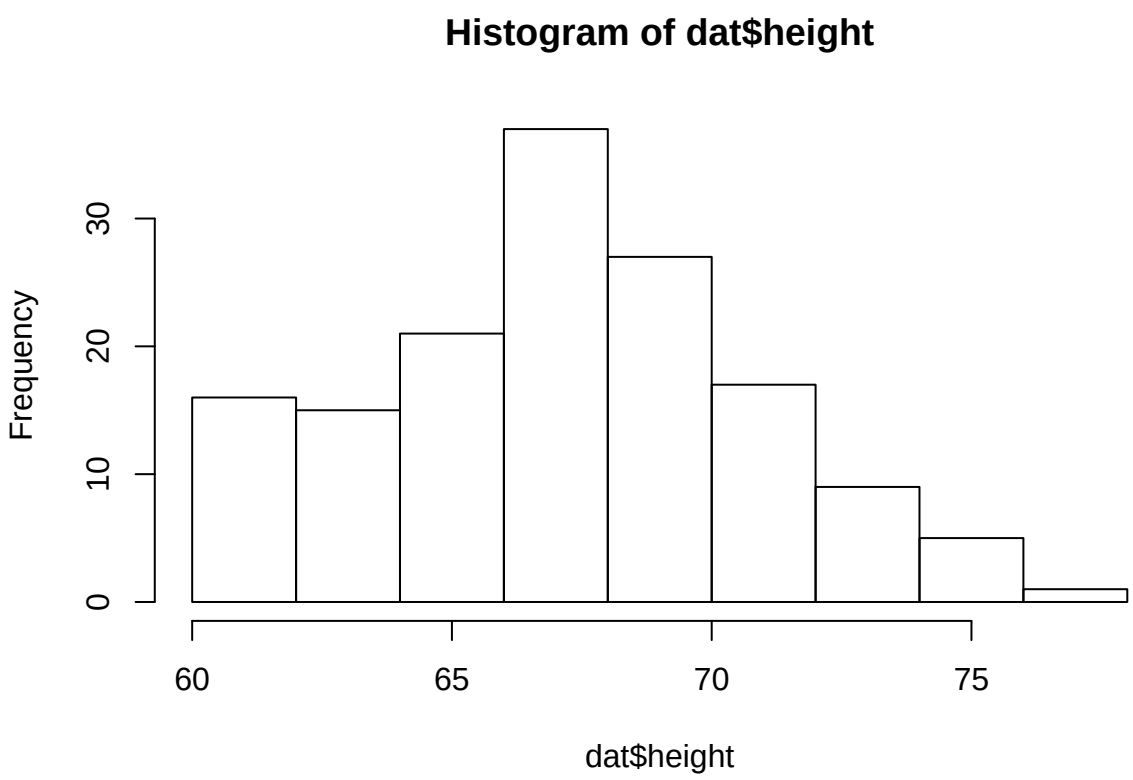
Let's confirm that we have removed all the outliers:

```
filter(dat, height>96 | height < 12) %>% select(height)
```

```
## Source: local data frame [0 x 1]
##
## Variables not shown: height (dbl)
```

The histogram looks more like the general population now

```
hist(dat$height)
```



Although not quite the same.

Assessment. Wrange the data as we did in class. You can obtain the function to fix heights from here

What is the height of the 113th entry ?

Normal Distribution

The histogram provides an excellent summary plot of a distribution. Can we summarize even further? We often see the average and standard deviation used as summary statistics. To understand why these are so widely used we need to undertand the normla distribution.

The bell curve, also known as the normal distribution or Gaussian distribution is commonly found in nature. There are reasons for this which we will explain later.

When the histogram of a list of numbers is said to be approximated by the normal distribution, it means we can use a convenient mathematical formula to approximate the proportion of values or outcomes in any given interval:

$$\Pr(a < x < b) = \int_a^b \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right) dx$$

While the formula may look intimidating, don't worry, you will never actually have to type it out, as it is stored in a more convenient form (as `pnorm` in R which sets a to $-\infty$, and takes b as an argument).

Note that if this distribution approximates our data, then we only need to know the average μ and the standard deviation σ to describe the entire population.

If we denote the values in our list as $x_1, \ldots, x_n$. The mean:

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$$

The variance:

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \mu_X)^2$$

with the standard deviation being the square root of the variance.

If we have a vector in R we can use the following functions:

```
x <- father.son$sheight
mu <- mean(x)
sigma <- sqrt( mean( (x - mu)^2) )
```

If in fact, the normal distribution is a good approximation we should be able to obtain approximations of the proportion in any range. For exmaple, how many men are taller than six feet?

We can obtain the exact answer like this:

```
sum( x> 6*12 ) / length(x)
```

```
## [1] 0.1141002
```

```
##which is equivalent to
mean( x>6*12)
```

```
## [1] 0.1141002
```

How good is the normal approximation?

```
1 - pnorm( 6*12, mu, sigma )
```

```
## [1] 0.1192743
```

We can try other values and see that we get very good approximations. Once we know the US adult men have an average height of 69 inches and standard deviation of 3 inches, we know everything: two numbers are all we need to describe the distribution.

Asssessment: If a list of numbers has a distribution that is well approximated by the normal distribution, what proportion of these numbers are within one standard deviation away from the list's average?

Standard units

Once we know that a list of numbers follow the normal distribution, a convinent way to describe a value is the number of standard deviations away from the average. We say these values are in *standard units*:

```
z <- (x - mu)/sigma
```

So, for example, a male seven footer is 5 SDs away from the average.

If the original distribtuion is approximately normal, then these values will have a *standard normal* distribution: average 0 and standard deviation 1. Notice that about 95% of the values are within two standard deviation of the average:

```
pnorm(2)-pnorm(-2)
```

```
## [1] 0.9544997
```

and most values are within 3

```
pnorm(3)-pnorm(-3)
```

```
## [1] 0.9973002
```

Assessment: Use the normal approximation to determine what proportion of US men are as tall or taller as Michael Jordan?

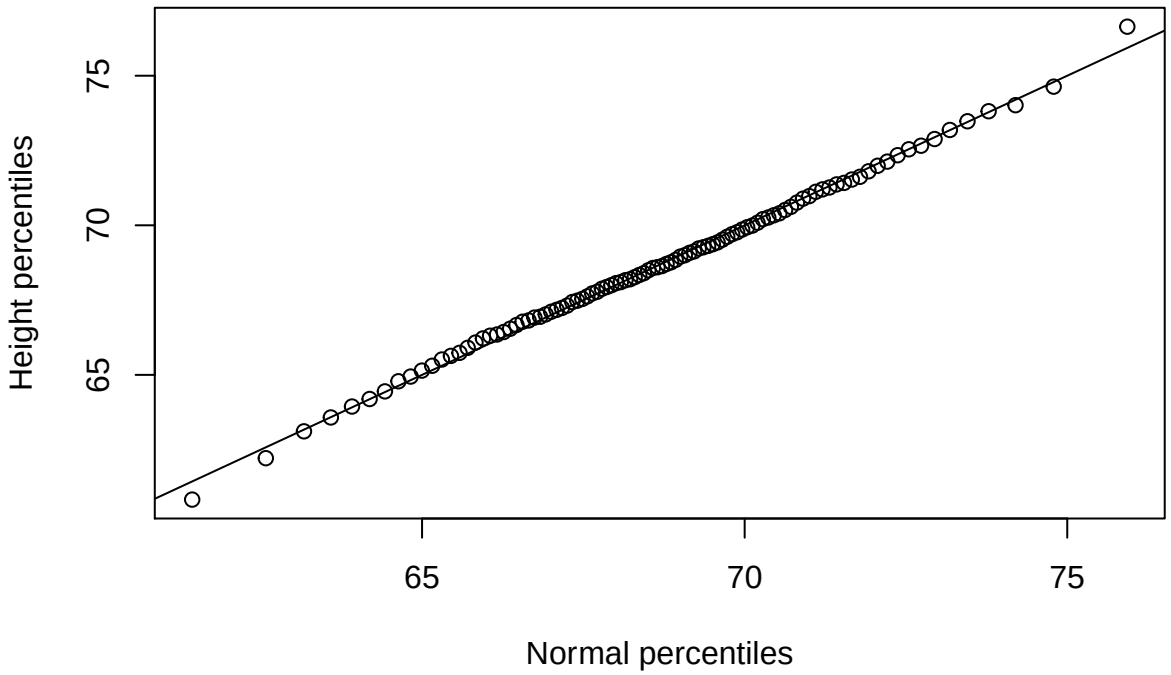
Assessment: In the US, how many more 6'6 or taller men are there than 7 foot or taller? Compute a proportion.

Assessment: What proportion of 7 footers (in the world) are in the NBA? Hint: There are 36 7 footers in the NBA. Hint 2: The average height of men 18-35 is 68.5 and SD 3 inches and is approximately normal?

Quantile Quantile Plots

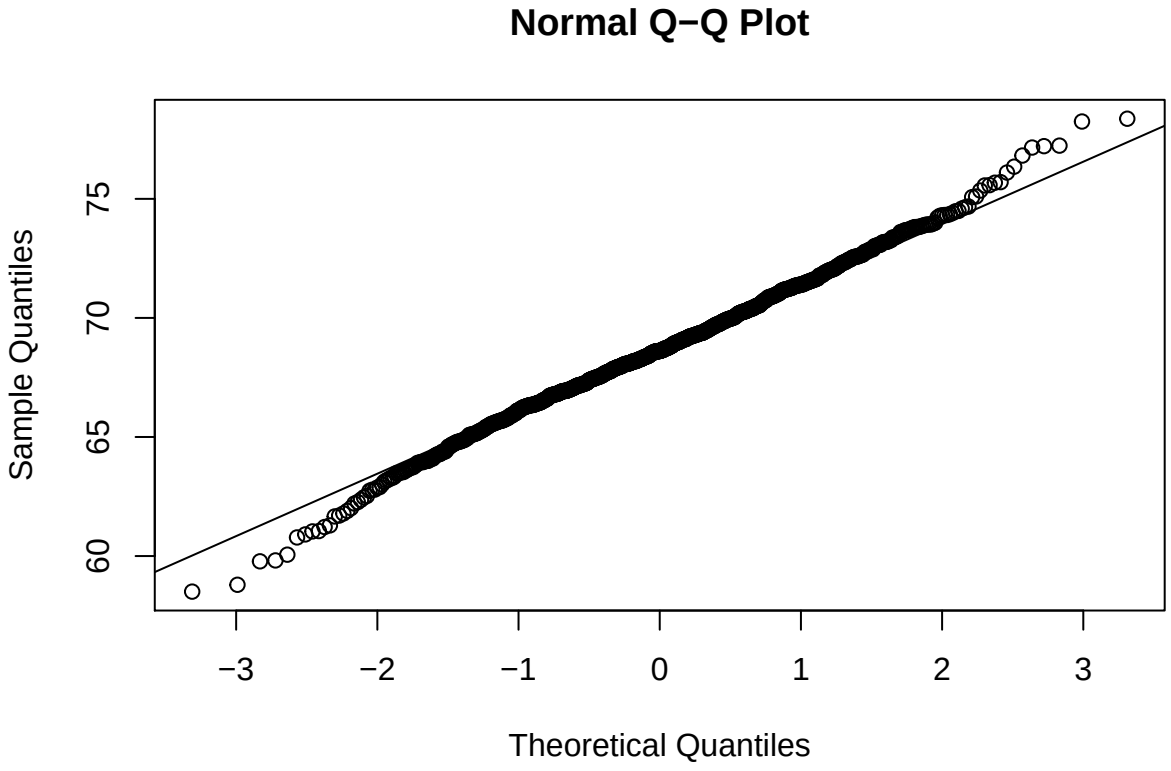
To corroborate that a theoretical distribution, for example the normal distribution, is in fact a good approximation, we can use quantile-quantile plots (qq-plots). Quantiles are best understood by considering the special case of percentiles. The p-th percentile of a list of a distribution is defined as the number q that is bigger than p% of numbers (so the inverse of the cumulative distribution function we defined earlier). For example, the median 50-th percentile is the median. We can compute the percentiles for male heights and for the normal distribution:

```
ps <- ( seq(0,99) + 0.5 )/100
qs <- quantile(x, ps)
normalqs <- qnorm(ps, mu, sigma)
plot(normalqs,qs,xlab="Normal percentiles",ylab="Height percentiles")
abline(0,1) ##identity line
```



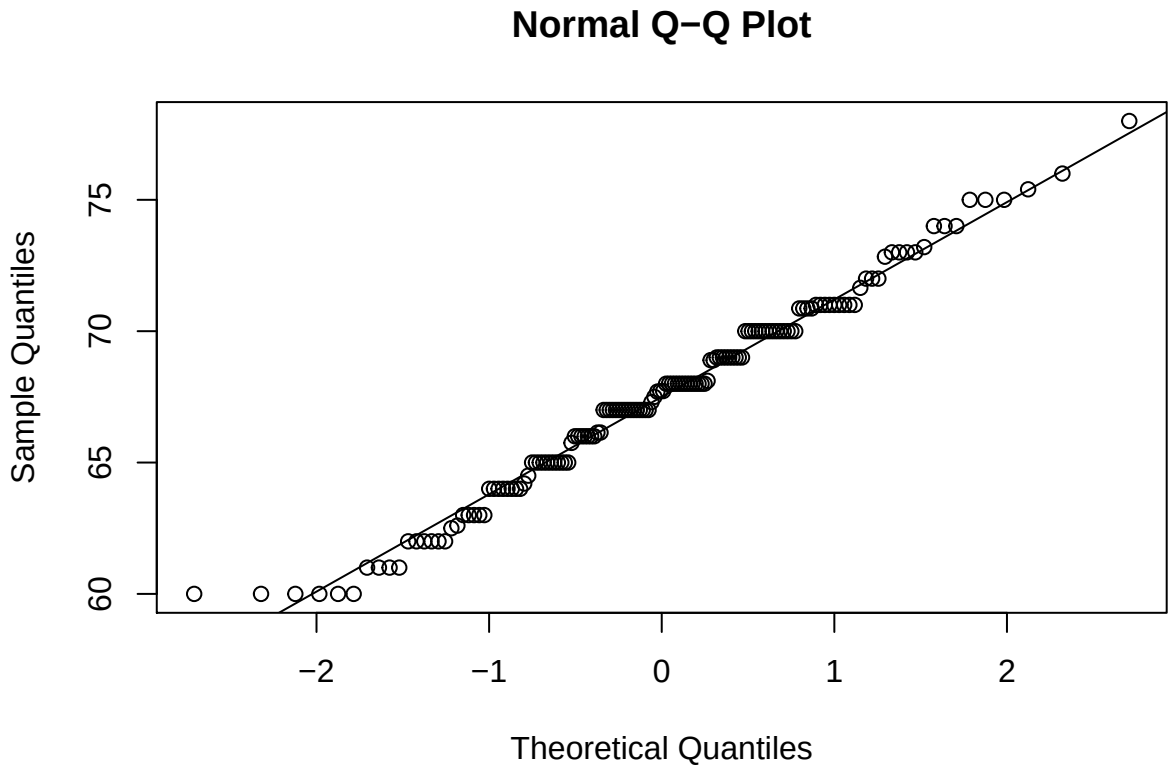
Note how close these values are. Also, note that we can see these qq-plots with less code (this plot has more points than the one we constructed manually, and so tail-behavior can be seen more clearly).

```
qqnorm(x)
qqline(x)
```



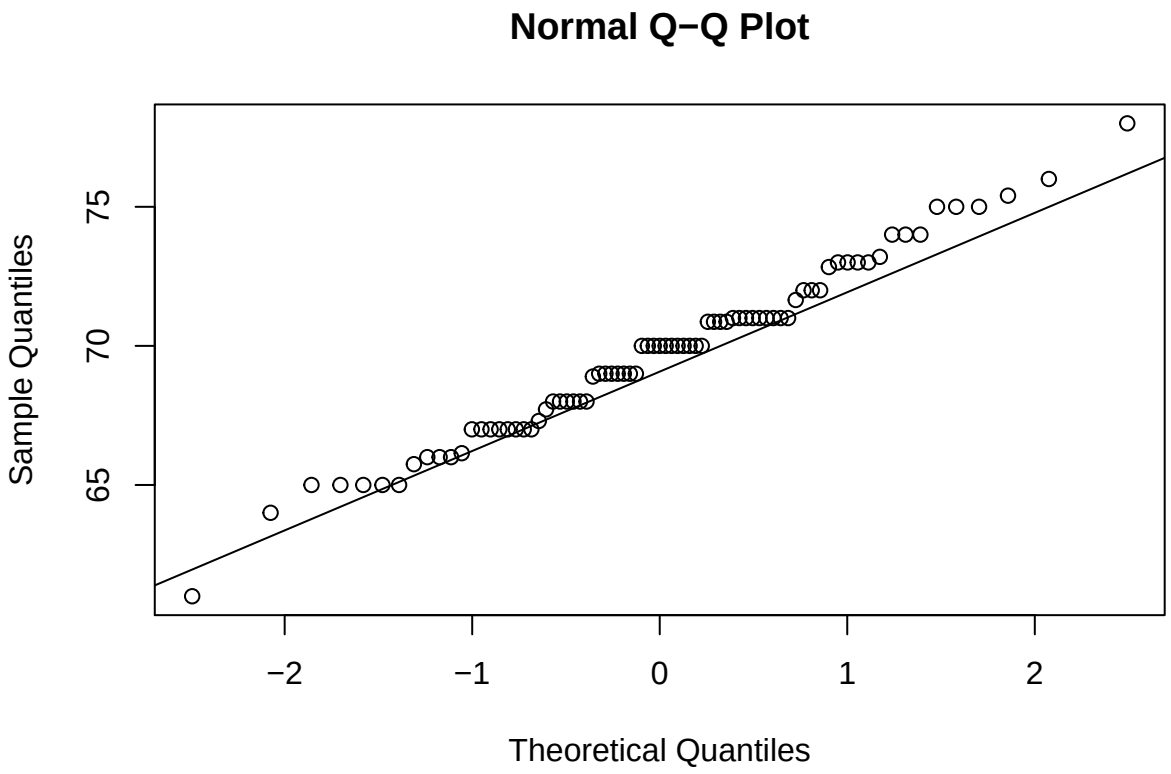
However, the `qqnorm` function plots against a standard normal distribution. This is why the line has slope σ and intercept μ.

```
qqnorm(dat$height)
qqline(dat$height)
```

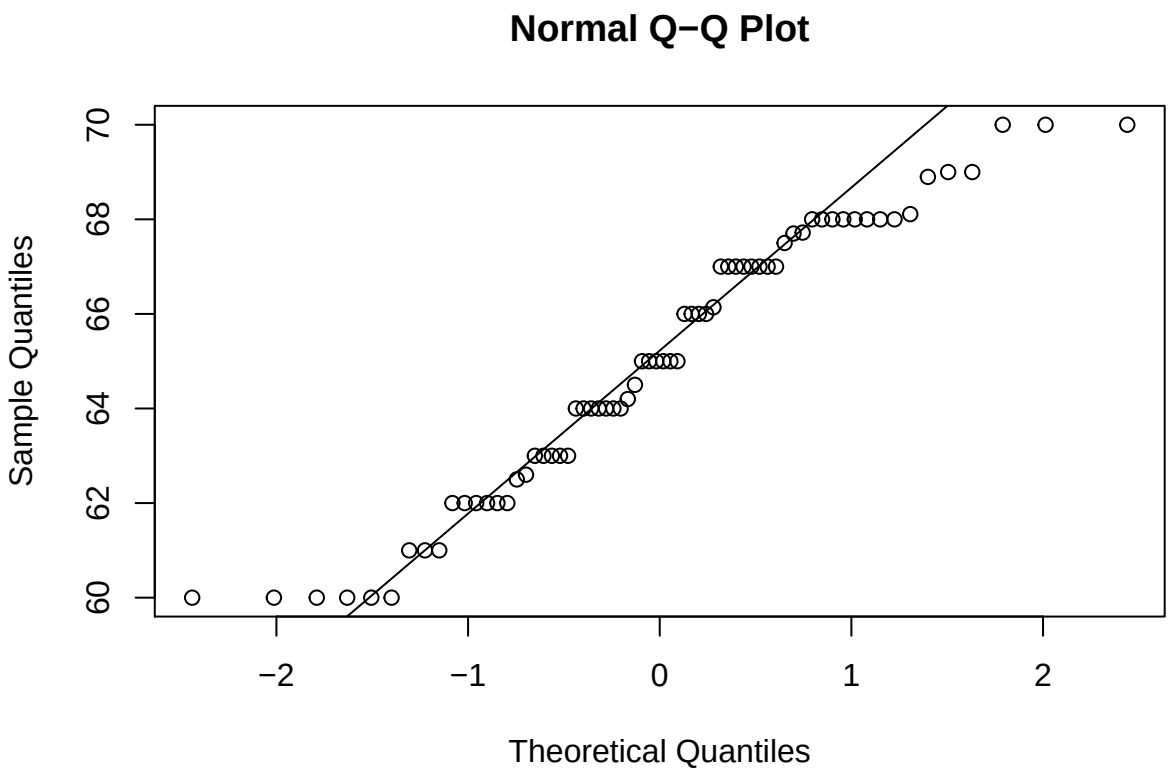


We can split it into the two genders and see slighlty better fits (excpet for the tails)

```
men <- filter(dat, gender=="Male")
qqnorm(men$height)
qqline(men$height)
```




```
women <- filter(dat, gender=="Female")
qqnorm(women$height)
qqline(women$height)
```



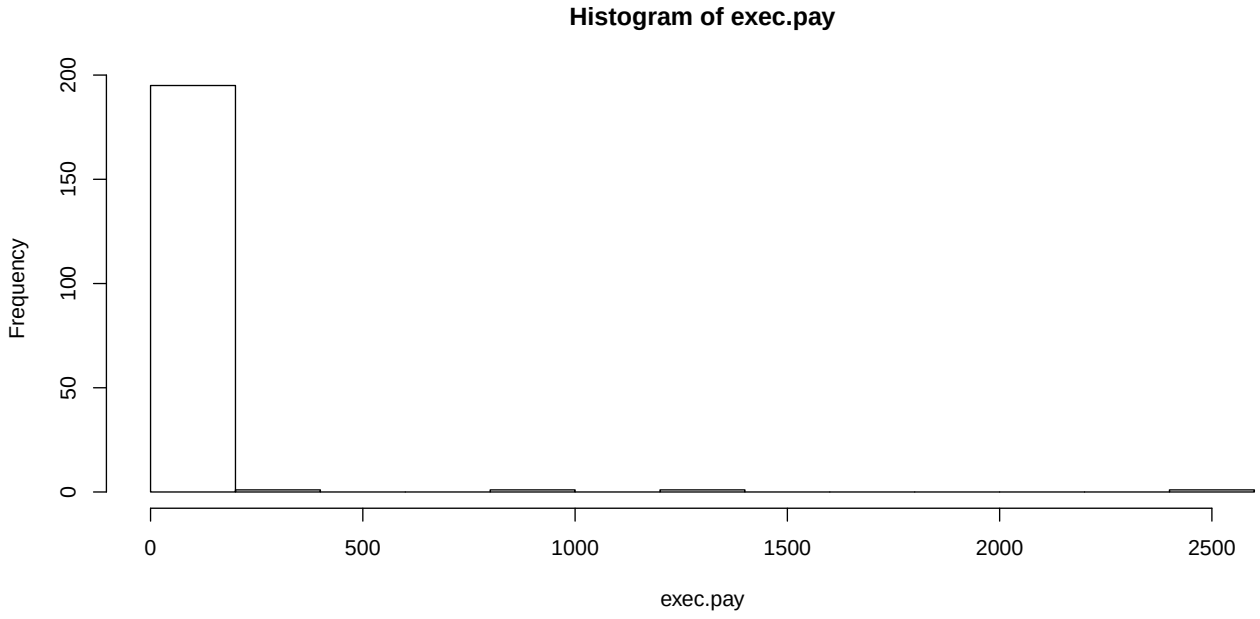
Assessment In our empirical distribution of heights, how many men are taller tha Micael Jordan? Why the

discrepency with our answer above?

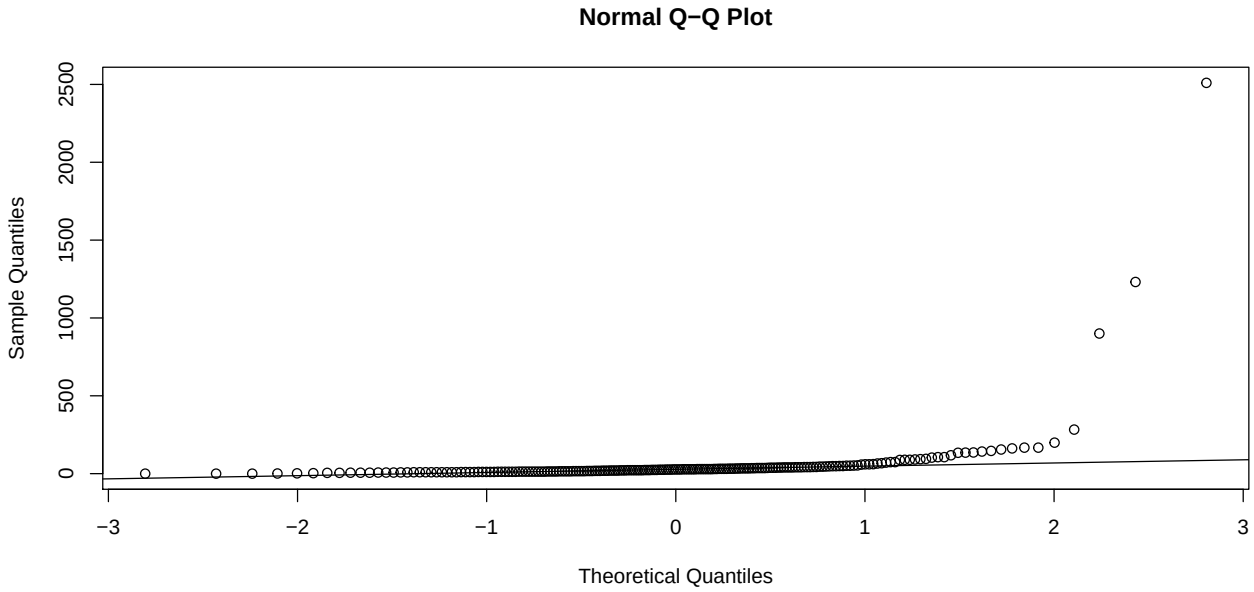
Boxplots

Data is not always normally distributed. Income is a widely cited example. In these cases, the average and standard deviation are not necessarily informative since one can't infer the distribution from just these two numbers. The properties described above are specific to the normal. For example, the normal distribution does not seem to be a good approximation for the direct compensation for 199 United States CEOs in the year 2000.

```
data(exec.pay,package="UsingR")
hist(exec.pay)
```

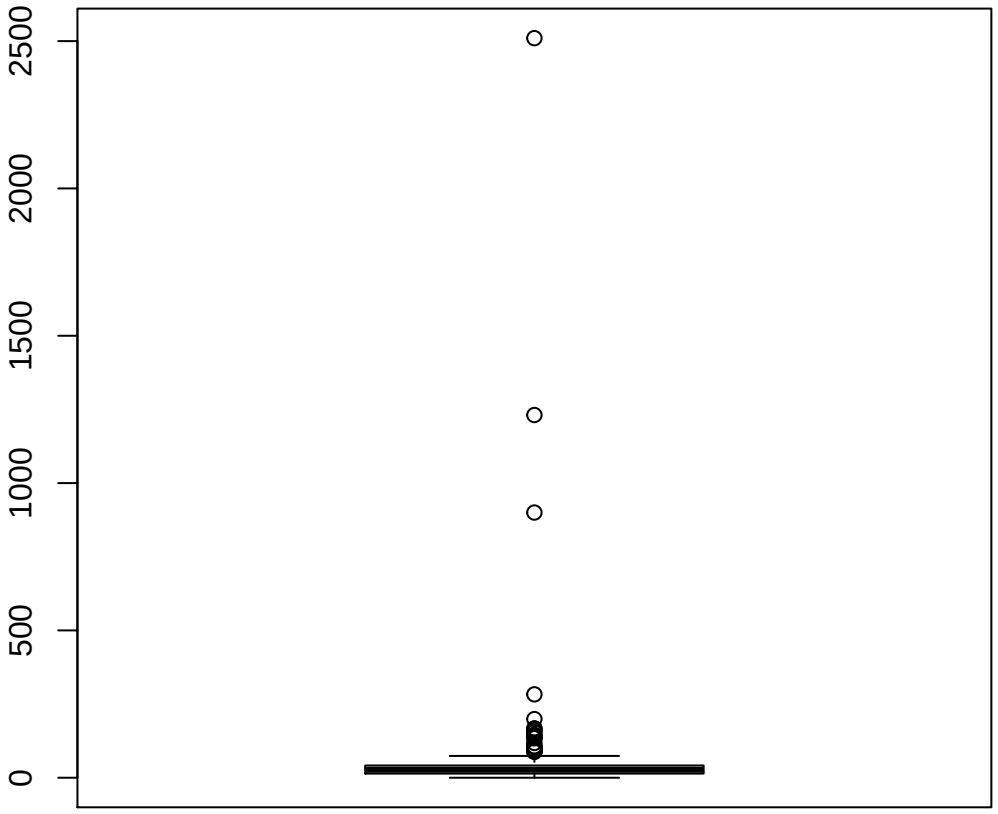



```
qqnorm(exec.pay)
qqline(exec.pay)
```



In addition to qq-plots, a practical summary of data is to compute 3 percentiles: 25-th, 50-th (the median) and the 75-th. A boxplot shows these 3 values along with a range of the points within median $\pm$ 1.5 (75-th percentile - 25th-percentile). Values outside this range are shown as points and sometimes refereed to as *outliers*.

```
boxplot(exec.pay)
```



Here we show just one boxplot. However, one of the great benefits of boxplots is that we could easily show many distributions in one plot, by lining them up, side by side.

Here are the distribution of men and women heights from our class

```
boxplot( women$height, men$height )
```