Chapter 1

Statistical methods part 1

The course is not actually a statistical course and we are not going through the theory of the statistical methods that will be demonstrated. However, some explanations of what the methods do will be given. So even if you are not familiar with the methods you should still be able to follow and learn how to use R for these statistical analyses.

1.1 Statistical measures

1.1.1 Data set

Let us first import the data set norsjo86.

agegrp:	Age group	(30, 40, 50 ,60 years)
health:	Health status	(0=good, 1=not quite good/bad)
sex:	Sex	(1=man, 2=woman)
height:	Body height	(cm)
weight:	Body weight	(kg)
sbp:	Systolic blood pressure	
dbp:	Diastolic blood pressure	
kolester:	Cholesterol	
smoker:	Smoking status	(0=non-smoker, 1=smoker)
bmi:	Body mass index	(kg/m^2)

```
# import spss file norsjo86
library(haven)
norsjo86 <- read_sav("../data/norsjo86.sav")  # this becomes a tibble
norsjo86<-as.data.frame(norsjo86)
head(norsjo86)
```

```
  agegrp health sex height weight sbp dbp kolester smoker      bmi
1     60      0   2    157     61 110  70      6.7      0 24.74745
2     60      1   2    157     97 150 100      6.6      0 39.35251
3     60      0   1    170     74 136  96      8.2      0 25.60554
4     60      0   2    163     66 156  76      7.5      0 24.84098
5     60      0   2    166     66 110  70     10.2      0 23.95123
6     60      0   2    168     61 130  78      7.3      0 21.61281
```

1.1.2 Mean, median and variance

We start by calculating the most common statistical measures. The formulas and common notations are as follows: mean $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$, variance $s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$, standard deviation $s = \sqrt{s^2}$, correlation $\rho = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})^2}{s_x s_y}$ where x_i and y_i are the sample observations and n the sample size.

There are often missing in data. Arguments like **na.rm** or **use** has to be used in some of the functions to ensure the result to be calculated. If no such option exist for a function it is possible to create an own function with this option. Some of the measures can be calculated using **summary**

```
x<-norsjo86$height
summary(x)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  145.0   165.0   171.0   170.9   178.0   191.0       4
```

```
mean(x) # what happens here?
```

```
[1] NA
```

```
mean(x,na.rm=T) # na.rm=T helps
```

```
[1] 170.9141
```

```
ownmean<-function(x){mean(x,na.rm=T)} # we can define an own function that removes NA
ownmean(x)
```

```
[1] 170.9141
```

```
median(x,na.rm=T)
```

```
[1] 171
```

```
range(x,na.rm=T)
```

```
[1] 145 191
```

```
quantile(x,na.rm=T)
```

```
  0%  25%  50%  75% 100%
 145  165  171  178  191
```

The **quantile** function gives as default the same result as **summary**. It also give both results in **range** as well as **median**. While mean and median gives information of the location of data the latter gives information on the variation in data . The most common measure of variation is however standard deviation which is the square root of the variance. It gives rise to some confusion because they actually measure the same thing although in different units/scales. Standard deviation is measured in the same scale as the variable, e.g. cm for height. Standard deviation can approximately be interpreted as the average absolute deviation from the mean.

```
var(x,na.rm=T) # variance
```

```
[1] 75.85141
```

```
# alternative calculation of variance
xnm<-x[!is.na(x)]          # first remove the missing observations
v<-sum((xnm-mean(xnm))^2)/(length(xnm)-1)
v
```

```
[1] 75.85141
```

```
sd(x,na.rm=T)

[1] 8.709271

sqrt(v)                    # alternative calculation

[1] 8.709271
```

1.1.3 Correlation

Correlation is a measure of relationship with values in $[-1, 1]$ and is calculated pairwise between variables. If its value is zero there is no relationship. The most common correlation is Pearson's correlation (ρ) which measures the `linear` relationship.

Correlation can be calculated for many variables at a time but always parweise. The same function is used independent of the number of variables. Observe the argument na.rm=T does not work with the correlation function.

Pearson's correlation

```
cor(norsjo86$sbp,norsjo86$dbp)     # correlation

[1] NA

cor(norsjo86$sbp,norsjo86$dbp,use="complete.obs")

[1] 0.7582536

cor(norsjo86[,c("height","weight","sbp","dbp")],use="complete.obs")

           height    weight        sbp        dbp
height 1.00000000 0.5049999 0.01716543 0.03305511
weight 0.50499987 1.0000000 0.26780039 0.40531170
sbp    0.01716543 0.2678004 1.00000000 0.75608662
dbp    0.03305511 0.4053117 0.75608662 1.00000000

# or similarly
cor(norsjo86[,4:7],use="complete.obs")

           height    weight        sbp        dbp
height 1.00000000 0.5049999 0.01716543 0.03305511
weight 0.50499987 1.0000000 0.26780039 0.40531170
sbp    0.01716543 0.2678004 1.00000000 0.75608662
dbp    0.03305511 0.4053117 0.75608662 1.00000000
```

Rank based correlation

The same function is also used to calculate Spearman's correlation which is based on the correlation of the ranks.

```
cor(norsjo86$sbp,norsjo86$dbp,use="complete.obs",method="spearman")

[1] 0.7748761
```

Test

We can also test if there is correlation in the population. Formally we test the hypothesis $H_0 : \rho = 0$. We can also use a confidence interval. More about test and confidence intervals later.

```
cor.test(norsjo86$sbp,norsjo86$dbp) # by default NAs are removed

Pearson's product-moment correlation

data:  norsjo86$sbp and norsjo86$dbp
t = 18.572, df = 255, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.7009239 0.8058481
sample estimates:
      cor
0.7582536
```

A scatterplot can be used to visually study the pairwise relationship. In the case with sbp and dbp the correlation is high and we can clearly see it in the plot.

```
plot(norsjo86$sbp,norsjo86$dbp,cex=0.7)
```

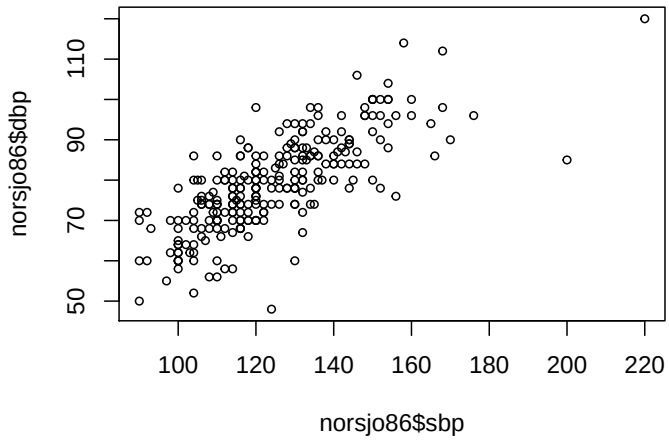

Figure 1.1:

Own experimentation

If you compare the estimated correlation between sbp and dbp you can see that it differs when it was made for these two variables only and when it was made for the four variables above. How come? Calculate Pearson correlation for other variables or subsets and test it. Can you test the correlation using Spearman's correlation -check help?

1.2 Test and confidence intervals

What is a statistical test? First of all we need to set up hypotheses; first a so called null hypothesis. It is often about the expected value (theoretical mean value or population mean). Your data can be thought of as a sample from a larger population. If you for example have measured systolic blood pressure in 30 randomly chosen individuals you can calculate the sample mean ($\bar{x}$). However, if someone else would do the same thing he/she would end up with a different sample and will also get a different sample mean. The sample mean can be thought of as an estimation of the true theoretical mean in the full population. We cannot observe the population mean because we don't have data on all individuals.We can theoretically have an infinite population. Think of experiments. If 100 experiments are performed this is the sample but the number of experiments that could have been made is high maybe unlimited (which can be thought of as the population). Thus the sample include observations that are randomly chosen from the population. If the sample includes the full population we don't need statistics, we can than just calculate the population mean.

The null hypothesis, called H_0 is usually what we don't believe is true. We actually want a decision: can we conclude that H_0 is false? If H_0 is false then the alternative hypothesis called H_1 must be true. This is the first step of a test. Assume we want to test if the population mean of systolic blood pressure are equal between men and women, then the null hypothesis may be $H_0 : \mu_M = \mu_W$ where μ_M and μ_W are the population means in men and women. The alternative hypothesis may be $H_1 : \mu_M \neq \mu_W$. Thus if we reject H_0 then H_1 is true. Otherwise we may keep on to H_0 although we cannot be sure it is true.

To test H_0 we try to find a test statistic, let us call it TS (it can for example be based on the difference of the sample means $\bar{x}_M - \bar{x}_W$) for which the distribution is known if H_0 is true. We would expect the difference of the sample means $\bar{x}_M - \bar{x}_W$) to be close to zero if H_0 is true but if H_0 is false the difference is expected to be further away from zero (either negative or positive). The question is how far from zero should it be for us to decide that H_0 is false. This is where the known distrubiution of TS comes in. We can calculate the probability to get the resulting TS we got or even more extreame given H_0 is true. If this probability is low, i.e. this is a rare event if H_0 is true, we will not belive in H_0 and reject it. The probability is called **p-value**. The limit for when a p-value is regarded as statistically significant is called **significance level** and is most often chosen as 5%

The method to calculate a so called **confidence interval** is similar, utilizing that the distrubution of TS is known if H_0 is true. In the example above a 95% confidence interval for $\mu_M - \mu_W$ is an estimated interval that with 95% probability cover the true $\mu_M - \mu_W$

We may have a one-sided alternative hypotheses saying for example $H_1 : \mu_M > \mu_W$. Then we are still interested if we can reject H_0 but we are only interested if the mean of men is larger then for women. For a confidence interval it will then be one-sided so that the upper limit is infinity.

1.2.1 T-test

T-test is perhaps the most common statistical test. However it can be performed in different ways; independent two-sample, one-sample or pairwise test. They can also be two-sided or one-sided which has to do with the alternative hypothesis. For two-sample test we can choose between assuming the standard deviations to be equal (t-test) or not (Welch's test) in the two samples. The nice thing in R is that we can use the same function for all different tests just by varying the arguments of the function. The result also include confidence interval. The confidence level is 95% by default but can be changed using the argument `conf.level = 0.95`.

There are however some requirements for the test to be correctly performed.

- The observations have to be independent

- The observations have to be normally distributed or the sample size should exceed 30 per group
- Standard deviations in the two groups are equal

1.2.2 Two-sample t-test

We start by creating two groups; smokers and non-smokers as two vectors. We then test the null hypothesis: the population means of systolic blood pressure are the same for smokers and non-smokers.

There are two ways to use the **t.test** function; using two vectors or using the formula method. In the last method the groups are identified by a variable. Observe that you need to give the argument `var.equal=T` to get the two-sample t-test.

```
table(norsjo86$smoker)

  0   1
206  54

# contruct two groups, smokers(1) and non-smokers(0)
sbp.smoker<-norsjo86$sbp[norsjo86$smoker==1]
sbp.nonsmoker<-norsjo86$sbp[norsjo86$smoker==0]
summary(sbp.nonsmoker)

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
   90.0   113.0   122.0   125.6   136.0   220.0       3

summary(sbp.smoker)

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   92.0   104.5   118.0   120.2   131.0   176.0

res.t<-t.test(sbp.nonsmoker,sbp.smoker,var.equal=T)
res.t

Two Sample t-test

data:  sbp.nonsmoker and sbp.smoker
t = 1.8528, df = 255, p-value = 0.06506
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.3373127 11.0676539
sample estimates:
mean of x mean of y
 125.6059  120.2407
```

The result shows that we don't reject the null hypothesis because the p-value is 0.065 which is over 5%. Furthermore we seee that the 95% confidence interval for the difference of the population means is $(-0.34, 11.07)$

The other alternative is the formula method (using a formula object). Let us take a closer look to the resulting object from **t.test** which is a list. The formula method also makes it easier to analyse subgroups.

```
res.t<-t.test(sbp~smoker,data=norsjo86,var.equal=T) #  here the groups are separated by variable smo
res.t
```

```
Two Sample t-test

data:  sbp by smoker
t = 1.8528, df = 255, p-value = 0.06506
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.3373127 11.0676539
sample estimates:
mean in group 0 mean in group 1
       125.6059        120.2407

class(res.t)

[1] "htest"

str(res.t)

List of 10
 $ statistic  : Named num 1.85
  ..- attr(*, "names")= chr "t"
 $ parameter  : Named num 255
  ..- attr(*, "names")= chr "df"
 $ p.value    : num 0.0651
 $ conf.int   : num [1:2] -0.337 11.068
  ..- attr(*, "conf.level")= num 0.95
 $ estimate   : Named num [1:2] 126 120
  ..- attr(*, "names")= chr [1:2] "mean in group 0" "mean in group 1"
 $ null.value : Named num 0
  ..- attr(*, "names")= chr "difference in means"
 $ stderr     : num 2.9
 $ alternative: chr "two.sided"
 $ method     : chr " Two Sample t-test"
 $ data.name  : chr "sbp by smoker"
 - attr(*, "class")= chr "htest"

t.test(sbp~smoker,data=subset(norsjo86,agegrp==50),var.equal=T) # we can select a subset

Two Sample t-test

data:  sbp by smoker
t = -0.44741, df = 60, p-value = 0.6562
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -18.20692  11.55094
sample estimates:
mean in group 0 mean in group 1
       129.4902        132.8182
```

1.2.3 One-sided t-test

For a one sided test the alternative hypothesis is either $H_1 : \mu_{NS} < \mu_S$ or $H_1 : \mu_{NS} > \mu_S$. Alternatively we may think of the hypothesis as $H_1 : \mu_{NS} - \mu_S < 0$ or $H_1 : \mu_{NS} - \mu_S > 0$. You can probably guess what NS and S stands for. You can choose alternative by argument `alternative="greater"` or `alternative="less"`. Now check the difference compared to the two-sided test result.

```
res.t1<-t.test(sbp~smoker,data=norsjo86,var.equal=T,alternative="greater")
res.t1

Two Sample t-test

data:  sbp by smoker
t = 1.8528, df = 255, p-value = 0.03253
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 0.5848424       Inf
sample estimates:
mean in group 0 mean in group 1
       125.6059        120.2407

res.t   # compare with two-sided alternative

Two Sample t-test

data:  sbp by smoker
t = 1.8528, df = 255, p-value = 0.06506
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.3373127 11.0676539
sample estimates:
mean in group 0 mean in group 1
       125.6059        120.2407
```

The result of the one-sided test shows that we reject the null hypothesis because the p-value is 0.033 which is over 5%. Furthermore we seee that the 95% confidence interval for the difference of the population means is $(0.58, \infty)$

1.2.4 Welch's test

This test is calculated similarly as a two-sample t-test but without assumption of equal standard deviations in the two groups. The default is "var.equal=F" so we don't need to give this argument.

```
t.test(sbp~smoker,data=norsjo86)

Welch Two Sample t-test

data:  sbp by smoker
t = 1.8184, df = 81.404, p-value = 0.07268
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.5049962 11.2353374
sample estimates:
mean in group 0 mean in group 1
       125.6059        120.2407
```

Own experimentation

Try to do a t-test using a different variable or a subset of the data. Try out one-sided test. Study the p-values and the confidence intervals. Compare the result of two-sample t-test and Welch's test.

1.2.5 One-sample t-test

For this test we only have one sample. Instead our hypotheses are based on the true mean in the population. $H_0 : \mu = \mu_0$ where μ_0 is the mean we want to test. In the examples below we start by the default nullhypothesis $H_0 : \mu = 0$ which may not be a realistic null hypothesis and then $H_0 : \mu = 120$. When we use one-sample test the argument var.equal=T is not relevant.

```
t.test(sbp.smoker) # not realistic null hypothesis (default mu=0)

One Sample t-test

data:  sbp.smoker
t = 45.552, df = 53, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 114.9463 125.5352
sample estimates:
mean of x
 120.2407

t.test(sbp.smoker,mu=120)

One Sample t-test

data:  sbp.smoker
t = 0.091201, df = 53, p-value = 0.9277
alternative hypothesis: true mean is not equal to 120
95 percent confidence interval:
 114.9463 125.5352
sample estimates:
mean of x
 120.2407

res.t.one<-t.test(sbp~1,mu=120,data=subset(norsjo86,smoker==1)) # alternative using formula
```

The test was not statistically significant, p=0.928 so we cannot conclude that systolic blood pressure among smokers in the population is not 120.

1.2.6 Paired t-test

So far we have dealt with independent observations. Now we are going to look at a situation with paired data. In can be for example observations before and after a treatment so there are two observations for each individual. In such situations we cannot assume that the observations are independent although we have independence between individuals. The null hypothesis of interest is most often: there is no difference before and after treatment. In such situations we should then use a paired test.

We use the Subliminal data for an example.

In an intervention study 18 students were randomized to receive either of two messages with the intention to see if this would affect their performance on the mathematics exam. The control group received neutral messages whereas the intervention group received messages confirming their learning process. All students participated in a summer school in mathematics and were tested at the beginning and end of the intervention. We are going to test the hypothesis that the test result did nott differ at the beginning and end for the neutral group.

The dataset contains the following variables:

Message:	If the student received neutral or confirmatory (positive) messages
Before:	Test result at the beginning of the study
After:	Test result at the end of the study
Improvement:	Improvement of their results (AfterBefore)

```
# paired t-test
library(haven)
Subliminal <- read_sav("../data/Subliminal.sav") ## find data set
Subliminal

# A tibble: 18 x 4
   Message  Before After Improvement
   <chr>     <dbl> <dbl>       <dbl>
 1 positive     18    24           6
 2 positive     18    25           7
 3 positive     21    33          12
 4 positive     18    29          11
 5 positive     18    33          15
 6 positive     20    36          16
 7 positive     23    34          11
 8 positive     23    36          13
 9 positive     21    34          13
10 positive     17    27          10
11 neutral      18    29          11
12 neutral      24    29           5
13 neutral      20    24           4
14 neutral      18    26           8
15 neutral      24    38          14
16 neutral      22    27           5
17 neutral      15    22           7
18 neutral      19    31          12

Sub.n<-subset(Subliminal,Message=="neutral")
t.test(Sub.n$After,Sub.n$Before,paired=T)

Paired t-test

data:  Sub.n$After and Sub.n$Before
t = 6.3175, df = 7, p-value = 0.0003974
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  5.162053 11.337947
sample estimates:
mean of the differences
                   8.25

res.t.p<-t.test(Sub.n$After-Sub.n$Before,var.equal=T)
```

The conclusion is that there is a significant difference in the neutral group with p= 4×10^{-4}. Thus we reject the null hypothesis of no difference. What the paired test actually does is however a one-sample t-test of the null hypothesis: the population mean of the individual differences is zero (see the last example).

1.3 Nonparametric tests

Nonparametric alternatives to t-test like Wilcoxon rank sum test (Mann Witney U test) and Wilcoxon signed rank test are shown below. They use the ranks instead of the observed values to calculate the test statistic.

One reason for doing non-parametric test is e.g. when the distribution is not normal and we have less than 30 observations in each group. Observe that the hypothesis is not about mean values in the population. The null hypothesis in a two sample test is: the two groups have the same distribution.

1.3.1 Wilcoxon rank sum test (Mann-Whitney test)

This is the non-parametric vesion of two sample t-test. Also here we have two alternativs, with or without using formula. We also have similar options as for t-tests using the argument `alternative` to get one-sided tests.

```
wilcox.test(sbp.nonsmoker,sbp.smoker) # alternative with two vectors

Wilcoxon rank sum test with continuity correction

data:  sbp.nonsmoker and sbp.smoker
W = 6505, p-value = 0.03486
alternative hypothesis: true location shift is not equal to 0

wilcox.test(sbp~smoker,data=norsjo86) #  alternative using formula

Wilcoxon rank sum test with continuity correction

data:  sbp by smoker
W = 6505, p-value = 0.03486
alternative hypothesis: true location shift is not equal to 0
```

1.3.2 Wilcoxon signed rank test

This is the non-parametric version of a paired t-test.

```
wilcox.test(Sub.n$After,Sub.n$Before,paired=T,exact=F)

Wilcoxon signed rank test with continuity correction

data:  Sub.n$After and Sub.n$Before
V = 36, p-value = 0.01415
alternative hypothesis: true location shift is not equal to 0
```

Own experimentation

For example you can compare confidence intervals and p-values for two-sample t-test. You can compare two-sample t-test and Mann-Whitney test for another variable or a subset .

1.4 Chi-2 test and Fishers exact test, tests for count data

Here the null hypothesis is that there is no relationship between categorical column and row variables.

1.4.1 Aggregated data

We start with an example where the relationship between back pain and weight increase during pregnancy was studied. In total 180 pregnant women were asked whether they experienced pain in the back during pregnancy or not. Further their weight increase during pregnancy were measured. The aggregated data are given below.

Weight increase	Pain	No pain
< 15	22	12
> 15	131	15

The data are prepared in aggregated form in a two by two table.

```
pregpain<-as.table(matrix(c(22,131,12,15),nrow=2))
pregpain

    A   B
A  22  12
B 131  15

dimnames(pregpain)<-list(weight=c("increase <15","increase >15"),pain=c("pain","no pain"))
pregpain

              pain
weight         pain no pain
  increase <15   22      12
  increase >15  131      15

round(prop.table(pregpain,margin=1),2)

              pain
weight         pain no pain
  increase <15 0.65    0.35
  increase >15 0.90    0.10

# it works with pregpain as a data frame or matrix too
res.c<-chisq.test(pregpain)
res.c

Pearson's Chi-squared test with Yates' continuity correction

data:  pregpain
X-squared = 11.649, df = 1, p-value = 0.0006424
```

The risk of pain seems to be higher with weight increase > 15. The p-value of the test is lower than 5% so the null hypothesis (independence) is rejected. Thus we can conclude that weight increase and pain in the back is related.

```
class(res.c)

[1] "htest"
```

```
str(res.c)

List of 9
 $ statistic: Named num 11.6
  ..- attr(*, "names")= chr "X-squared"
 $ parameter: Named int 1
  ..- attr(*, "names")= chr "df"
 $ p.value  : num 0.000642
 $ method   : chr "Pearson's Chi-squared test with Yates' continuity correction"
 $ data.name: chr "pregpain"
 $ observed : 'table' num [1:2, 1:2] 22 131 12 15
  ..- attr(*, "dimnames")=List of 2
  .. ..$ weight: chr [1:2] "increase <15" "increase >15"
  .. ..$ pain  : chr [1:2] "pain" "no pain"
 $ expected : num [1:2, 1:2] 28.9 124.1 5.1 21.9
  ..- attr(*, "dimnames")=List of 2
  .. ..$ weight: chr [1:2] "increase <15" "increase >15"
  .. ..$ pain  : chr [1:2] "pain" "no pain"
 $ residuals: 'table' num [1:2, 1:2] -1.284 0.619 3.055 -1.474
  ..- attr(*, "dimnames")=List of 2
  .. ..$ weight: chr [1:2] "increase <15" "increase >15"
  .. ..$ pain  : chr [1:2] "pain" "no pain"
 $ stdres   : 'table' num [1:2, 1:2] -3.68 3.68 3.68 -3.68
  ..- attr(*, "dimnames")=List of 2
  .. ..$ weight: chr [1:2] "increase <15" "increase >15"
  .. ..$ pain  : chr [1:2] "pain" "no pain"
 - attr(*, "class")= chr "htest"

res.c$expected

              pain
weight          pain no pain
  increase <15  28.9     5.1
  increase >15 124.1    21.9

res.c$observed

              pain
weight         pain no pain
  increase <15   22      12
  increase >15  131      15

res.c$observed-res.c$expected

              pain
weight         pain no pain
  increase <15 -6.9     6.9
  increase >15  6.9    -6.9
```

The test is based on the differences between observed and expected numbers.

1.4.2 Individual data

Here we use the norsjo86 data set for another example to study the relationship between BMI (split into low and high) and smoking.

We start by aggregating the individual data to a 2 by 2 table and do the test on aggregatedf data as

above and then we use individual data directly in the same function. The results are the same.

```
norsjo86$bmic<-cut(norsjo86$bmi,breaks=c(0,25,Inf),labels=c("<25",">25"))
norsjo86$smokerf<-factor(norsjo86$smoker,labels=c("No smoker","Smoker"))  # optional

aggtab1<-table(norsjo86$smokerf,norsjo86$bmic)
aggtab1

            <25 >25
  No smoker 112  90
  Smoker     32  22

aggtab2<-xtabs(~smokerf+bmic,data=norsjo86)     # alternative table
aggtab2

           bmic
smokerf     <25 >25
  No smoker 112  90
  Smoker     32  22

res.ca<-chisq.test(aggtab2)                     # chi-square test
res.ca

Pearson's Chi-squared test with Yates' continuity correction

data:  aggtab2
X-squared = 0.1207, df = 1, p-value = 0.7283

# We need not do this procedure - using individual data works as well
res.ci<-chisq.test(norsjo86$smokerf,norsjo86$bmic)
res.ci

Pearson's Chi-squared test with Yates' continuity correction

data:  norsjo86$smokerf and norsjo86$bmic
X-squared = 0.1207, df = 1, p-value = 0.7283

res.ci$observed # compare with aggtab1 (aggtab2)

                norsjo86$bmic
norsjo86$smokerf <25 >25
       No smoker 112  90
       Smoker     32  22
```

Thus we can use either aggregated or individual data in the same function.

1.4.3 Fishers exact test

This test can be useful especially with small samples. While the p-value from a chi-square test is an approxmation (asymptotic result based on normal approximation) Fishers exact test is always correct. Here we need another R function **fisher.test**. The function also gives odds ratio and its confidence interval.

```
res.fa<-fisher.test(aggtab2)                          # aggregated data
res.fa

Fisher's Exact Test for Count Data
```

```
data:  aggtab2
p-value = 0.6461
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.4409047 1.6395816
sample estimates:
odds ratio
  0.856075

res.fi<-fisher.test(norsjo86$smokerf,norsjo86$bmic)  # individual data
res.fi

Fisher's Exact Test for Count Data

data:  norsjo86$smokerf and norsjo86$bmic
p-value = 0.6461
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.4409047 1.6395816
sample estimates:
odds ratio
  0.856075
```

1.4.4 Tests in n by k tables

As an example we test if there is a relationship between age (4 categories) and BMI (2 categories).

```
t22<-table(norsjo86$agegrp,norsjo86$bmic)
round(prop.table(t22,margin=1),2)

      <25  >25
  30 0.69 0.31
  40 0.59 0.41
  50 0.48 0.52
  60 0.48 0.52

chisq.test(norsjo86$agegrp,norsjo86$bmic)

Pearson's Chi-squared test

data:  norsjo86$agegrp and norsjo86$bmic
X-squared = 7.6546, df = 3, p-value = 0.05372
```

Close but no significant diffence of BMI between the age groups.

Alternative

We can also use **summary** on a `table` to calculate chi-square test. Here demonstrated with `xtabs` but it also works for `table`.

```
summary(xtabs(~agegrp+bmic, data=norsjo86))

Call: xtabs(formula = ~agegrp + bmic, data = norsjo86)
Number of cases in table: 256
```

```
Number of factors: 2
Test for independence of all factors:
Chisq = 7.655, df = 3, p-value = 0.05372
```

1.4.5 Count data in one dimension

It is also possible to do chi-square tests on one variable. Then the aim is not to test relationship but to test if the observations of the different categories all have the same expected number i.e the observed variation is not systematic, only random. Assume we analyse the number of sold houses per month January to June in a municipality a certain year. Is the expected frequency of sold houses same for all months (null hypothesis) or is the observed variation systematic?

```
n.sold.houses<-c(17,14,18,26,19,24)
n.sold.houses

[1] 17 14 18 26 19 24

chisq.test(n.sold.houses)

Chi-squared test for given probabilities

data:  n.sold.houses
X-squared = 5.1525, df = 5, p-value = 0.3975
```

The test is not significant so the conclusion is that there is no systematic pattern.

1.5 Test of distribution, Kolmogorov-Smirnov and Shapiro Wilk test

The purpose of this test is to check if a distribution follows a certain distribution. In this example we test if systolic blood pressure in age group 50 years follows a normal distribution. Thus it is important to give the correct parameters as arguments. The Shapiro-Wilk only test the hypothesis: normal distribution.

```
# KS test
sbp50<-norsjo86$sbp[norsjo86$agegrp==50]

# standardise the observations
m<-mean(sbp50,na.rm=T)
s<-sd(sbp50,na.rm=T)
sbp.st<-(sbp50-m)/s

ks.test(sbp50,"pnorm")  # null hypothesis is standard normal

One-sample Kolmogorov-Smirnov test

data:  sbp50
D = 1, p-value < 2.2e-16
alternative hypothesis: two-sided

table(duplicated(sbp50)) # We should avoid ties for this method.

FALSE  TRUE
   37    25
```

```
        # There are a lot of ties. They are removed below

        # compare standardised observations to standard normal
ks.test(sbp.st[!duplicated(sbp.st)],"pnorm")

One-sample Kolmogorov-Smirnov test

data:  sbp.st[!duplicated(sbp.st)]
D = 0.077041, p-value = 0.9685
alternative hypothesis: two-sided

        # compare original observations to normal(mean=m,sd=s)
ks.test(sbp50[!duplicated(sbp50)],"pnorm",m,s)

One-sample Kolmogorov-Smirnov test

data:  sbp50[!duplicated(sbp50)]
D = 0.077041, p-value = 0.9685
alternative hypothesis: two-sided

shapiro.test(sbp50)     # Shapiro-Wilk test

Shapiro-Wilk normality test

data:  sbp50
W = 0.92256, p-value = 0.0007866
```

There was a lot of ties and the p-value was quite different when the ties were removed. However, the conclusion was the same: we don't reject the hypothesis that the observations follow a normal distribution. The function `pnorm()` can be found in the help. We can choose any of these functions for the test.

1.6 Q-Q plot

In a Q-Q plot the quantiles of the observed variable is plotted on the y-axis and the corresponding expected quantiles had the distribution been normal are plotted on the x-axis.

```
qqnorm(sbp50,cex=0.7)
qqline(sbp50,col=2)
```

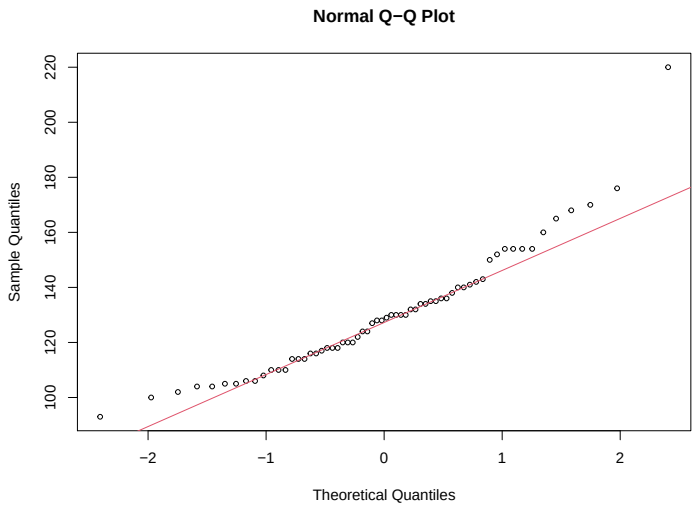

Figure 1.2: Q-Q plot