Practical 2

Student t-test

Exercise 1

Here we have data on the genome size (measured in picograms of DNA per haploid cell) in two large groups of crustaceans. The cause of variation in genome size has been a puzzle for a long time; weâ??ll use these data to answer the biological question of whether some groups of crustaceans have different genome sizes than others.

1. First we should observe the data, load the file into R and graphically explore the dispersion and normality of the whole dataset. Looking at the histograms, do you think the data is normal?

```
genome_size <- read.table("genome_size_long_format.txt")
boxplot(genome_size[,2])
```

```
# different dispersion

hist(genome_size[,2])
```

A better graphical way to look at data normality is to perform a QQ plot. A histogram shows the frequencies of different values in the variable (counts). Depending on how the histogram looks it can be misleading. It's better to use the QQ plot. A Q-Q plot shows the mapping between the distribution of the data and the ideal distribution (the normal distribution in this case). Usually a line is plotted through the quartiles. When the dots follow the line closely, the data has a normal distribution.

```
# make a QQ plot
qqnorm(genome_size[,2], main = "QQ plot")
# add a QQ line
qqline(genome_size[,2], col=2)
```

2. Calculate mean and variance of each group. Compare them.

```
mean(genome_size[genome_size[,1]=="Decapods",2])
```

```
## [1] 4.133
```

```
mean(genome_size[genome_size[,1]=="Isopods",2])
```

```
## [1] 1.377
```

```
var(genome_size[genome_size[,1]=="Decapods",2])
```

```
## [1] 10.08
```

```
var(genome_size[genome_size[,1]=="Isopods",2])
```

```
## [1] 2.734

# both measures are very different
```

3. Using the statistical test called Kolmogorov-Smirnov check if the dataset is normally distributed.

```
ks.test(genome_size[,2],"pnorm", mean(genome_size[,2]), sd(genome_size[,2]))

##
##  One-sample Kolmogorov-Smirnov test
##
## data:  genome_size[, 2]
## D = 0.2, p-value = 0.02
## alternative hypothesis: two-sided
```

4. If the variances are not similar between the groups and the variable is not normally distributed, we cannot use the Student t test directly. To do it we try to transform the data to fit it into a normal distribution. We will apply the log10 transformation to our data. Use the *log10()* function. Calculate the mean and the variance of the newly transformed data. Are the variances more similar now?

```
genome_size$log10 <- log10(genome_size[,2])

mean(genome_size[genome_size[,1]=="Decapods",3])

## [1] 0.402

mean(genome_size[genome_size[,1]=="Isopods",3])

## [1] -0.1067

var(genome_size[genome_size[,1]=="Decapods",3])

## [1] 0.3807

var(genome_size[genome_size[,1]=="Isopods",3])

## [1] 0.2129

# the means are different but the variances are more similar
```

5. Now plot the histogram and the Q-Q plot of the transformed data. Do they look nearly normal?

```
hist(genome_size[,3])
```



```
qqnorm(genome_size[,3])
qqline(genome_size[,3], col=2)
```


6. Check if the transformed data follows a normal distribution with Kolmogorov-Smirnov test.

```
ks.test(genome_size[,3],"pnorm", mean(genome_size[,3]), sd(genome_size[,3]))

##
##  One-sample Kolmogorov-Smirnov test
##
## data:  genome_size[, 3]
```

```
## D = 0.12, p-value = 0.3
## alternative hypothesis: two-sided

# now the variable is normal
```

7. After transforming the data, now we can apply the Student t test to answer the question: Do both groups have the same mean genome size? What is the value of the t statistic? And the p-value?

```
result_StudentT <- t.test(genome_size[genome_size[,1]=="Decapods",3],
                          genome_size[genome_size[,1]=="Isopods",3],var.equal =
TRUE)
result_StudentT

##
##  Two Sample t-test
##
## data:  genome_size[genome_size[, 1] == "Decapods", 3] and
genome_size[genome_size[, 1] == "Isopods", 3]
## t = 3.4, df = 52, p-value = 0.001
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.2112 0.8063
## sample estimates:
## mean of x mean of y
##    0.4020   -0.1067

tvalue <- result_StudentT$statistic
pvalue <- result_StudentT$p.value
```

Exercise 2

1. The p-value obtained in Exercise 1 is based on the theoretical t-distribution. However, we can also calculate the p-value by simulating the null hypothesis of the groups having the same mean genome size. Then, the p-value would be the probability of obtaining a t-value equal or higher than the one observed in Exercise 1.

Letâ??s create a random normal distribution with the mean and variance of the whole dataset.

```
popmean <- mean(genome_size[,3])
popsd <- sd(genome_size[,3])
population <- rnorm(n = 1e6, mean = popmean, sd=popsd)
```

2. Now we will start our sampling experiment, which will consist of drawing 10.000 times two samples of size 27 each and calculate the t-value. The distribution of the t-values obtained is an â??empiricalâ?? t-distribution. There are several ways of doing this calculation, a useful function is *replicate()*.

```
nReplicates <- 1e4

t.samples <- replicate(nReplicates,
  t.test(sample(population,length(genome_size[genome_size[,1]=="Decapods",3])),
         sample(population,length(genome_size[genome_size[,1]=="Isopods",3])),
         var.equal = TRUE)$statistic)
```

```r
hist(t.samples)
points(tvalue,0, col="red")
```

3. Now calculate the p-value using the distribution of t-values obtained.

```r
# How many values are higher or > than the observed tvalue?
signif <- t.samples[t.samples >= tvalue]
(length(signif) / nReplicates) *2
```

Exercise 3 -

You have to reproduce the same simulation from Exercise 2, but instead of using a sampling from a normal distribution with the same mean and variance, you will apply a randomization test. The randomization test gives us a way to measure the variability in the difference of two sample means. Our goal is to compare the observed difference between Decapods and Isopods to the expected difference due to chance.

To perform a randomization test you have to take your initial dataset of 54 individuals and randomly divide the sample into two groups. There is a huge number of different possible ways to divide the 54 observations into two groups, each of size 27 (could you calculate the number?). We will do this resampling process only 10.000 times. For each time you resample, you will have to perform the t-test on the two samples and save the t-value in a vector (as you already did in Exercise2). The result will be a distribution of t-values to estimate the significance of your result.

Do the experiment on both the raw data and the transformed one. Compare the results that you obtained in the previous section and give your opinion of the whole experiment.

```r
nReplicates <- 1e4

# raw data
randomTest <- function(genome_size){
g1 <- sample(genome_size,size = 27)
g2 <- genome_size[which(!genome_size %in% g1)]
tval <- t.test(g1,g2,var.equal = TRUE)$statistic
}

t.samples.original <- replicate(nReplicates, randomTest(genome_size[,2]))
t.samples.transformed <- replicate(nReplicates, randomTest(genome_size[,3]))


hist(t.samples.original)
points(tvalue,0, col="red")
```

```r
hist(t.samples.transformed)
points(tvalue,0, col="red")
```

```r
signif <- t.samples.original[t.samples.original >= tvalue]
(length(signif) / nReplicates)*2
```

```
signif <- t.samples.transformed[t.samples.transformed >= tvalue]
(length(signif) / nReplicates)*2
```

Exercise 4: Paired t-test -

To investigate the effects of lighting conditions on the orb-spinning spider webs, researchers measured the horizontal (width) and vertical (height) dimensions of the webs made by 17 spiders under light and dim conditions. Accepting that the webs of individual spiders vary considerably, they employed a paired design in which each individual spider acts as its own control. A paired t-test performs a one sample t-test on the differences between dimensions under light and dim conditions.

You can find the data in the *spider_web.txt* file. Note the format of this data set. Rather than organizing the data into the usual long format in which variables are represented in columns and rows represent individual replicates, these data have been organized in wide format. Wide format is often used for data containing repeated measures from individual or other sampling units. Even though this is not necessary (as paired t-tests can be performed on long format data), traditionally it did allow more compact data management as well as making it easier to calculate the differences between repeated measurements on each individual.

Perform two separate paired t-tests to test the following null hypotheses:

- No effect of lighting on web width
- No effect of lighting on web height