**Lab 3:***Basic Concepts II*

***3.1. The t-test***

The t-test is used to test for differences in means between two groups. Recall that there are several versions of the t-test depending on the properties of the data collected. The most common t-test is the “*two-sample t-test for independent samples*” which assumes that the individual observations of the data are independently collected and that the variance in the two samples that are being compared is equal. Let’s start from there.

You want to compare the number of mutations passed on through the maternal and paternal lineage to offspring (you are interested in knowing whether there are differences in germline mutation rates between males and females). You have access to 100 sequenced “trios” consisting of the mother, father, and one of their children. By deep DNA-sequencing you can with fair confidence identify the number of de novo mutations passed by the mother and father respectively by comparing DNA sequences in the three. The trios come from different studies and were not sequenced for the same portion of the genome, so the data you have access to is scaled up to give the number of de novo mutations in offspring, attributed to either the mother or the father, per 6.4 giga bases (the size of the human diploid genome). Read in the file “two.sample” into Rstudio.

read.csv("two.sample.csv") -> name\_of\_my\_data #I will call my dataset “two.sample”

Make sure the file looks as it should and that you understand what it contains by clicking on it in the top right menu in Rstudio. First, you might also want to check that the data is approximately normally distributed, since this is what the t-test assumes. Remember the “*ass-principle*”?! That’s right, if you plot data with a histogram from two groups which means truly differ, the distribution might look a bit like… In this case, therefore, you can plot the data on the number of mutations for each group separately:

hist(two.sample$mutations[two.sample$parent.sex=="Mother"])

hist(two.sample$mutations[two.sample$parent.sex=="Father"])

It does not look too bad. You can now try to run the two-sample t-test by entering the code:

t.test(mutations ~ parent.sex, var.equal = T, data = two.sample) -> my.test

my.test

Note that I first saved my test as an object (called “*my.test*”) before I looked at it. This was not necessary in this case, but often this is a preferred way of working in R. You read the first line of code above as: the number of mutations *depends on* (~) the parent sex. Hence, you put the dependent variable to the left, and the independent variable to the right.

**3.1a.** ***Now look at the output. Was your test significant? What was the t-value and degrees of freedom? What is the mean number of mutations passed to offspring from mothers and fathers respectively?***

It turns out that you know the trio from which each mother and father came from. It seems likely that there are differences among studies in the overall number of mutations found. Such differences could easily be imagined since sequencing technology, bioinformatic pipelines and age of parental couples, and genomic regions that were sequenced (etc.) likely vary between trios. This will cause a lot of variation in the data that could obscure the difference between mothers and fathers. One way to control for such variation is to perform a *“paired” t-test*. This test compares paired data-points with each other. These paired observations have something more in common with each other compared to the other observations in the dataset. In our case, this thing is that the mother and father comes from the same trio and hence conceived their child together. It would then seem obvious that the best analysis would be to compare the parents from a trio with each other directly. The *paired t-test* does this by calculating a difference between the paired observations (mutations mother – mutation father) so that we get a total of 100 differences from our 100 trios. It then tests if the mean of this difference is significantly different from 0. Let’s run a paired t-test by first reading in the new file “paired.csv” that contains the exact same data, but rearranged so that mothers and fathers from the same trio occur on the same row (click on the data and inspect it once read into Rstudio). We then perform *the paired test*:

t.test(paired$Mother , paired$Father, paired = T)

Note that the syntax now is slightly different. We do not have a dependent variable that is a function of the independent variable, as we had for the two-sample t-test. Instead, we have two vectors containing paired observations that occur on the same row.

***3.1b. Inspect the output carefully. How does it compare to the first test you performed that did not consider the pairing of the data? What has happened here?***

An intuitive way to think about the two alternative tests is to plot the data in a way that corresponds to the two hypotheses tested. We can use the **boxplot**function to look at the data. This function plots the spread in the data, but does not plot confidence intervals, so does not directly correspond to the statistical significance of the test(s). Nevertheless, it is an easy-to-use function to get a general idea of how the data looks (type ***?boxplot*** to find out more).

For the *two-sample t-test*, we tested for differences in means between the two groups, so this corresponds to the following plot:

boxplot(mutations ~ parent.sex, data = two.sample)

In the *paired t-test* we tested whether the difference between the number of mutations passed through mothers and fathers was significantly different from 0:

boxplot(paired$Mother-paired$Father)

***3.1c. Look at the figures – do they make sense and correspond to the corresponding test?***

***3.2. Distributions***

Try typing **?Distributions**. This gives you an overview of all the available distributions in R. The accompanying functions allow you to randomly sample from each distribution (e.g. **rnorm**), or calculate the cumulative probability (**pnorm**), the density (**dnorm**), or critical value for a given probability (**qnorm**). ***You can try to play around a bit with these functions for the normal distribution to see if you understand them*.** Use **?pnorm** to find out more about the function**.**

**3.2a.** For example, using **pnorm** and **qnorm**, and assuming a normal distribution with a mean of 10 and a standard deviation of 2, try to find:

* ***The probability to observe an individual with a trait value of 13 or higher***
* ***The trait value below which the 20% of the population with smallest trait values are***

Some other distributions are more or less common in statistics, and some are related to each other. For example, the Chi-square distribution is a squared (*f*(x^2)) normal distribution (*f*(x)), and also equal to the gamma distribution for a specific shape parameter of the gamma (the gamma distribution is a very general distribution and can be modified by two parameters; shape and scale). You do not need to know all these distributions, or be able to use all the functions, but note that one of these distributions is always assumed for the response variable when you model your data, and you choose which one by picking a specific test and distribution. Choosing an inappropriate distribution for your response will generate erroneous conclusions from your statistical test.

Gene expression from RNA-seq data is measured as the number of reads of a specific RNA transcript – the more reads in the sample, the greater expression of the gene. There is of course also some chance to the process of whether a specific gene product gets picked up and read in the sequencing process, and the number of reads for a specific product should essentially follow a *Poisson* process. (Note that sometimes, due to sequencing errors or genome features, several processes act simultaneously to determine gene expression. For example, a gene may not be expressed in some damaged tissue of the sample. This can make the number of reads follow a *negative binomial* distribution. This distribution can arise because there is a binomial process (expressed or not) and a Poisson process (if expressed – how many times?) acting simultaneously).

***3.2b. Use the Poisson distribution to generate 1000 biological samples of transcript reads for a specific gene with mean read count = 5 and plot the sampled data. How does it look*? *In what way is it different from a normal distribution*?**If you need help, first type: **?rpois**

The data does not look normal, because you sampled from the Poisson distribution (note that it is asymmetrical – count variables cannot take on negative values – so the distribution has a lower bound at zero), and one can analyze this with a generalized linear model (GLM) using a Poisson distribution for the response variable (number of reads). However, you don’t know about this yet, you only know about the t-test, which requires approximately normal data. One could try to transform your data to make it a better fit to a normal distribution, but for certain types of distributions this can be tricky and you are better of using GLMs. You can go back to the notes from lecture 3 to remind yourself about common transformations that often work for specific data-types. ***Now generate poisson data for another gene which mean expression = 100. Notice any difference in the shape of the distribution compared to the previous data with mean = 5?***

***3. Analysis of frequencies***

Individuals with Sickle cell anemia are homozygous for a recessive allele (*aa*) coding for dysfunctional hemoglobin and get very ill, but heterozygotes (*Aa*) do OK, and are more resistant to malaria. This causes balancing selection on this locus (the heterozygote *Aa* has an advantage over both homozygotes *aa* and *AA*) if there is malaria in the region), and the dysfunctional allele is maintained in the population. You are studying African populations where malaria is present or absent. Based on a sample of 1000 adults from each region, you establish that, in African populations where malaria is present, the genotype frequencies are: 0.40*AA*; 0.55*Aa*; 0.05*aa*. In populations where malaria is absent, the corresponding frequencies are: 0.80*AA*; 0.19*Aa*; 0.01*aa*.

**3.3a. Test if the presence/absence of malaria is consistent with selection acting differently on the locus in the geographic regions** – Use the chi-square test in R: **chisq.test().** To figure out the input you need to feed the function, type: **?chisq.test.** A hint is that you can feed a matrix containing your genotype counts to the function, such that your code will be: **chisq.test(name\_of\_my\_matrix)**. Create a matrix with your genotype counts using **matrix().** *Make sure you understand the output.*

***3.3b. ADVANCED (TOTALLY OPTIONAL) BONUS QUESTION: Test if there indeed is selection on the locus in each geographic region separately using Hardy-Weinberg’s equilibrium (HWE).*** You have the observed frequencies in each population – *But what are the expected values under the null hypothesis?* Hint, HWE gives the expected genotype frequencies at a locus with the alternative alleles *A* and *a* occurring at frequencies p and q, respectively, in an infinite population with no selection, migration or drift: Exp(*AA*) = p2; Exp(*Aa*) = 2pq; Exp(*aa*) = q2. For solving this question, use the observed and expected values to calculate a χ2-statistic. Then use the cumulative distribution function for the chi-square distribution **pchisc()** and set “lower.tail=FALSE” to find the p-value. See last slides of Lecture 3 for more hints.

***What do you find? Are the first results compatible with this last result? This is a hard one, can you solve it together?***