Assignment 1 - Statistics for molecular biomedicine

February 24, 2021

- Please read the assignment instructions in the assignment module on Absalon.
- In this assignment you can use the R package gplots but otherwise only use base R.
- In the following questions use a significance threshold of 0.05 unless otherwise stated.
- This is an individual assignment. You cannot share code and answers with someone else.
- Use the discussion board for question about the assignment - however DO NOT post your suggested solution (see pinned post in the discussion board)

Assignment part 1 of 1

Type 2 diabetes is a common disease that affect many individuals in most of the world including Greenland where 10% of the Inuit population have type 2 diabetes. In a recent genetic study of the Greenlandic population a variant(mutation) was identified on chromosome 13 that affects the risk of developing type 2 diabetes. Individuals who are homozygous (carries two copies) of the risk variant have a 50% risk of developing type 2 diabetes – in this assignment we will refer to these individuals as homozygous. Likewise we will refer to the individuals carrying zero or one risk allele as non-homozygous. Many individuals in Greenland have ancestry that is both Inuit and European. For this assignment we will focus only on individuals with only Inuit ancestry. Additionally the alleles are in Hardy Weinberg equilibrium meaning that the presence of the variant is independent between an individual's two alleles.

For the curious the study is briefly described here (you do not need to read it)

They identified a stop-gain variant (mutation) with an allele frequency of 23% in individuals with only Inuit ancestry (based on thousands of individuals). However, 0nitial discovery of the mutation was made using exome sequencing of 18 unrelated individuals with only Inuit ancestry.

1A What was the probability of them observing the stop-gain variant at least one time in the 18 sequenced samples?

1B When sequencing individuals using a next generation sequencing platform there can be a considerable amounts of errors. If a mutation is observed in multiple individuals the probability of it being an error is much lower. What was the probability of observing the mutation in at least 2 individuals?

1C They could have saved money by sequencing fewer individuals. What is the minimum number of individuals they would have needed in order to have at least 90% probability of observing the mutation in at least two individuals

1D In the study the variant was present in almost half of the 18 sequenced individuals and in total 14 copies was observed. Give the estimated allele frequency as well as the 99% confidence interval.

1E As mentioned the homozygous carriers of the variant have a much larger risk of type 2 diabetes. In Greenlandic Inuit what is probability of being homozygous if you have type 2 diabetes?

1F What is the probability of having type 2 diabetes if you are not homozygous for the variant(non-homozygous)?

1G The study also investigated 2-hour glucose levels from an oral glucose test. Due to privacy laws we cannot use the individual data so instead simulate the study data using the following function based on your KUid and the below function

```
seeder<-function(name){
    n<-sapply(strsplit(name,""),function(x)
     sum(match(x,c(LETTERS,letters)),na.rm=T))
    set.seed(n)
    N<-2700
    genotype<-rbinom(N,2,0.23)
    glucose <- rnorm(N,5.7,sd=1.4) +
     ifelse(genotype==2,rnorm(N,4,sd=1),0)
    data.frame(stopGain=genotype,glucose=glucose)
}
```

for example my data would be generated as

```
myData<-seeder("bcn627")
```

- the stopGain column contains the number of stop-gain copies for each individual
- the glucose column contains the glucose levels

make a boxplot of glucose levels stratified on homozygous and non-homozygous individuals using the standard R boxplot function with default options. Remember to make the plot understandable with informative labels, title and/or legend.

1H For each of the 3 genotype levels calculate the mean, the standard error (of the mean) and the 99% confidence interval (assume each category is normal distributed). Print the results as a table, data.frame or matrix

1I Make a barplot of the mean glucose levels for all three genotype levels and include the 99% confidence (use barplot2 in the gplots package)

1J Test if there is a difference in glucose levels between homozygous and non-homozygous individuals. Write your two hypotheses and provide a p-value as well as your test statistic

1K Investigate if the assumptions in your test hold. Provide text and/or appropriate figures.

1L From a much larger study you are told that the standard deviation of glucose levels are equal to 1.4 for the non-homozygous and 1.7 in the homozygous. Using that information test if there is a difference in glucose levels between the homozygous and non-homozygous. Provide the test statistic and state whether the p-value if smaller (more significant) than without using this information.