Practise questions for the exam in Introductory Statistics for Bioinformatics Students. 2021.10.

1. Give a correct interpretation of the P-value? (2P)

2. What happens with the type-I error rate when you increase the number of tests? What can you do about it? (2P)

3. What happens with the type-I and type-II error rate when you increase sample size? Why does this happen? (2P)

4. You want to test if there are differences among men and women belonging to three different age classes (16-30 years old, 31-45 y.o., 46-60 y.o., and 61-90 y.o.) in the probability of having been the driver involved in a serious car crash. What type of data do you collect to answer this hypothesis? What test do you perform, and what is your null hypothesis? (2P)

5. You want to test if there is a difference in the number of de novo somatic mutations in liver tissue samples between cancer patients and a control group. What test and distribution do you choose, and what is your null hypothesis? (2P)

6. You have performed an experiment measuring the effect of a chemical treatment used in pesticide management on the cell growth of 8 different strains of yeast used in beer brewing. You grew the strains on agar plates, with 10 replicate plates per strain and treatment (added chemical or control). Below is an ANOVA table with the variation (sums of squares) explained by each factor in your model. We assume that each agar plate serves as an independent observation for the test.

Calculate the mean squared variance (MS) for each effect. Use R to calculate P-values from F-values for each effect. (4P).

Effect	SS	df	MS	F	P
Treatment	250	1			
Strain	500	7			
Treatment:Strain	500	7			
Residual error	500	144			

7. You want to see if the amount of human growth hormone (HGH) among people going to the gym affects their monthly muscle growth. You measure the baseline HGH levels of 18 people over a month and then approximate their thigh muscle gain in grams. You have saved your data in the file "muscle1.txt".

Devise a statistical test to evaluate the hypothesis. What is the null hypothesis? What is the evidence for your hypothesis in terms of the P-value, and how much variation in muscle growth is explained by the level of HGH? Make sure you deal with your data in an appropriate way before interpreting your final analysis, and motivate and document all analysis steps clearly. (4P).

8. You investigate color-blindness in mean and women. In your sample of 200 men and 300 women, 14 men and 5 women are color bind. You want to test if men and women differ in this trait. Set up a null hypothesis and calculate the expected frequency of color blind men and women given the null. Chose a proper statistical test and carry it out. Use R if needed. (3P)

9. An experiment was performed to see how genome size was affected by the number of transposable elements. Partitioning variation in genome size gave SSreg = 1000 and SSerror = 4000 across 22 different plants studied. Can you calculate the F-value, P-value, as well as the R^2 value for the effect of transposable elements on genome size in these plants? (3P)

10. You are going to work with morphological measurements of crabs in order to graphically explore differences in size between two species as well as between males and females. (4P)

```
install.packages("MASS")
install.packages("ggfortify ")

library(MASS)
library(ggfortify)
```

Load the data

```
data("crabs")
```

Have a look at the data help file in order to better understand it

```
?crabs
```

Run a PCA on all five morphological measurments for both sexes and both species

```
pca_all <- prcomp(crabs[, 4:8], scale=TRUE)
```

Look at the output

```
summary(pca_all)
print(pca_all)
```

Q1A: How much variance does PC3 explain and what is the loading for BD on PC2?

Plot and color based on species

```
autoplot(pca_all, data = crabs, colour = 'sp', frame=FALSE,
      loadings = TRUE, loadings.colour = 'blue',
      loadings.label = TRUE, loadings.label.size = 3)
```

Plot and color based on sex

```
autoplot(pca_all, data = crabs, colour = 'sex', frame=FALSE,
      loadings = TRUE, loadings.colour = 'blue',
      loadings.label = TRUE, loadings.label.size = 3)
```

Q1B: Can you detect any differences in morphological measurements between species and sexes?

In order to get a better look at the data we subset each sex and look at them seperately

```
dat_male <- crabs[crabs$sex=="M",]
pca_male <- prcomp(dat_male[, 4:8], scale=TRUE)

autoplot(pca_male, data = dat_male, colour = 'sp', frame=FALSE,
      loadings = TRUE, loadings.colour = 'blue',
```

```
        loadings.label = TRUE, loadings.label.size = 3)


dat_female <- crabs[crabs$sex=="F",]
pca_female <- prcomp(dat_female[, 4:8], scale=TRUE)

autoplot(pca_female, data = dat_female, colour = 'sp', frame=FALSE,
        loadings = TRUE, loadings.colour = 'blue',
        loadings.label = TRUE, loadings.label.size = 3)
```

Q1C: Can you detect any differences between the two species when looking at the sexes separately? And what would be your overall conclusion regarding morphological differences in these crabs?

11. One of the differences between PCA and multidimensional scaling (MDS/PCoA) is the form of the data you give as input to the method. What characterizes this form of data you would use as input in an MDS/PCoA? (1P)

12. Why is multicollinearity between explanatory variables a potential problem when running a multiple regression, and can you give two suggestions for a solution to this problem? (2P)

13. In a study of the effects of marijuana on the risk of cancer in oral squamous cells, Rosenblatt et al. (2004) examined 407 recent cases of the cancer from Washington state. They also randomly sampled 615 healthy subjects from the same region having similar frequency distribution of age and sex as the cancer cases. They found that a similar proportion of the cancer cases (25.6%) and healthy subjects (24.4%) reported having used marijuana (odds ratio = 0.9; 95% confidence interval 0.6-1.3). In this type of study, the disease itself is the "treatment" and the potential causal factor – marijuana use – is the "response". (4P)

A. Is this an experimental or observational study? Explain.

B. Does this study include a control group?

C. What was the purpose of ensuring that the healthy subjects were similar in age and sex to the cancer cases?

D. Can we conclude that marijuana does not cause cancer in oral squamous cells in this population?

14. In population-based association studies, we generally aim to test for the presence of association between the trait and each of multiple genotypes across many SNPs and gene loci, something that can result in inflation of the error rate. Describe some methods to adjust for multiple testing in the context of a genome wide association study on human height where 400K people have been genotyped on a 500K SNP array. (2P)