Statistical hypothesis testing in Biomedical Engineering: Tutorial with Python

Preparation:

1) Go to the website https://trinket.io or use your own installation of Python (with matplotlib, pandas and Scipy).
2) Upload the data files (dataA.txt, etc) and the initial script example to load data (starting_code.py).
3) Go through the tutorial below. Make a new Python script for each exercise (excerciseA.py, etc.)
4) Write a report (max 3 pages) with your results (figures and quantitative results) and your comments. Add your code in appendix as text (the code does not count in the 3 pages).
5) The **deadline** for the submission of report+code is indicated on Moodle. DO NOT WAIT FOR THE LAST MINUTE.

A. Galton's stature data: descriptive statistics and Shapiro-Wilk's normality test

In his seminal paper about regression [1], Francis Galton studied the relationship between the mean heights of 205 couples and those of their children. Here, we are concerned with the parents only. The file "dataA.txt" provides the parents' heights in inches.

1) Load the height data using pandas.
2) Transform the data in cm
3) Plot the histogram of the father and mothers heights.
4) Plot the cumulative distribution function for the fathers and mothers
5) Plot the probability plot against the theoretical Gaussian distribution. Comment on these plots.
6) Perform a Shapiro-Wilk test on both heights to determine if they follow a Gaussian distribution.
7) Compute the average, standard deviation, kurtosis, skewness of both heights.
8) Compare the two distributions with the appropriate statistical test. Comment on the result.

B. Material properties of polymerized NDGA–collagen composite fibers: development of biologically based tendon constructs

Koob and Hernandez [2] aimed to develop a novel biomaterial to repair ruptured tendon. They produced synthetic fibers from tendon type I collagen, which was polymerized with di-catechol nordihydroguaiaretic acid (NDGA). In particular, the authors studied the effects of increasing the concentration of NDGA on the stiffness of the fibers.

---

[1] Galton F. (1886). Regression towards mediocrity in hereditary stature. Journal of the Anthropogical Institute of Great Britain and Ireland 15, p. 246-263.

[2] Koob and Hernandez (2002). Material properties of polymerized NDGA–collagen composite fibers: development of biologically based tendon constructs. Biomaterials, p. 203-212. Data is *loosely* based on their results.

The data in dataB.txt file represents the stiffness of the fibers (in terms of elastic modulus) measured at different concentrations of NDGA, between 0 and 3 mg/ml, as well as the effects of a second treatment of NDGA ("3+0.1mg/ml") and a comparison with the stiffness of native tendon fibers.

1) Create a boxplot to visually compare all data series. Make a table with the appropriate descriptive statistics for each data series.
2) We want to compare all data series with each other. Check the required statistical assumptions (normality, variance) and then apply the correct test to compare the series. What can you conclude?
3) If possible, perform a post-hoc test to determine if groups "2 mg/ml" and "3 mg/ml" differ from each other. Also check if "3 mg/ml" and "3+1 mg/ml" are different.
4) Find the concentration of NDGA that best approximates the stiffness of native tendon fibers, and check with a statistical test if the two are significantly different. What can you conclude?
5) Do you have any comments on the dataset?

C. Muscle fitness-based predictors of bone quality

Osteoporosis is considered a pediatric disease with geriatric consequences. However, measuring bone strength in children is complex. V.Yngling et al. [3] hypothesized that muscle fitness could be used to predict bone quality. Those authors acquired data on hand grip strength, using a hand dynamometer, and vertical jump height, as well as a proxy for bone strength: trabecular bone mineral content, which was measured with peripheral Quantitative Computed Tomography.

1) We need to test the correlation between trabecular bone mineral content and all the other variables. Check the hypotheses required by the correlation test (i.e., normality, outliers) and apply the correct correlation test to assess which variable is the best predictor of bone quality.
2) Plot the scatter plot between mineral density and the best predictor.
3) Males and females have different bone quality and risk of osteoporosis. Check if the correlation between bone mineral content the best predictor remains significant in the two sexes, and if one sex presents a better prediction.
4) Plot a scatter between plot mineral density and the best predictor separating males and females. Can you infer (and how) if males tend to have higher or lower mineral content than females?
5) Bone mineral content should decrease with age, but it does not in this dataset. Looking at the scatter plot between age and mineral content, can you hypothesize why?

[3] V.Yngling et al. (2021). Peak vertical jump power predicts radial bone strength better than hand grip strength in healthy individuals. Communications in Kinesiology 1(2).