Chapter 6: Two sample tests

1 Session 1

Objective: association of a continuous variable with a binary variable; Student's T-test, Fisher's (variance) F-test, Kolmogorov-Smirnov test.

Steps of the session:

1. Individual work: Practice with the instructions given on the web site.
2. Share and work: discuss what you understand.
3. Individual work: Practice with the exercises of Session 1.
4. Share and work: discuss your scripts together.

Exercise 1.1. Upload `titanic.csv`, read data description. Assign columns `pclass`, `survived`, `gender`, `age`, to variables `P`, `S`, `G`, `A`. Conclusions are to be given at threshold 5%.

1. Display summaries of `A` by values of `G`, and a box plot of `A` against `G`. Does the one-way anova conclude that the mean ages of women and men were different? Does the T-test conclude that the mean ages of women and men were different? Does the T-test conclude that women were younger than men on average? Compare the p-values of the three tests.
2. Plot the ecdf of ages of women in red, add the ecdf of ages of men in blue. On the graphic, where does the main difference between the two curves occur? Why is this so? Does the Kolmogorov-Smirnov test confirm that the ecdf of women ages is above the ecdf of men ages?
3. Are the age variances among men and women significantly different?
4. Repeat questions 1. to 3., replacing `G` by `S`. Were survivors younger than non survivors?
5. Display summaries of `A` by values of `P`, and a box plot of `A` against `P`. Does the one-way anova conclude there that there were differences in mean among the three classes of passengers? Were first class passengers older on average, than the other passengers? than second class passengers? than third class passengers? Were second class passengers different in age from the others, were they older than third class passengers?

6. Plot the ecdf of ages for the three passenger classes with three different colors. Where are the main differences? How can you explain them? Is the curve for second class significantly below the curve for third class? above the curve for first class?

7. Is the variance of ages in first class different from the variance of ages in second class? is it larger than the variance of ages in third class? Can you explain this result?

8. Display the values, and a table of `G:S`. What does `G:S` represent? Display a box plot of `A` against `G:S`. Does a one-way anova detect significant differences? Where do differences come from? Were women who did not survive, younger than other passengers? Were surviving men younger than other passengers? Can you explain these results?

Exercise 1.2. Upload `bosson.csv`, read data description. Assign columns `country`, `gender`, `aneurysm`, `bmi`, to variables `C`, `G`, `A`, `B`. Assign column `risk` to variable `R`; group classes `3,4,5` into a single class, interpreted as "`3` or more".

1. Display a box plot of `A` against `G`. Does the one-way anova conclude that the mean aneurysm size of women and men are different? Does the T-test conclude that the mean ages of women and men are different? Does the T-test conclude that women have smaller aneurysms than men on average? Compare the p-values of the three tests.

2. Plot the ecdf of aneurysm sizes of women in red, add the ecdf of aneurysm sizes of men in blue. On the graphic, where do the main differences between the two curves occur? Does the Kolmogorov-Smirnov test confirm that the ecdf of women aneurysms is different from the ecdf of men aneurysms?

3. Are the aneurysm variances among men and women significantly different? Can this be seen on the ecdf plots? On the box plots?

4. Repeat questions 1. to 3., replacing `G` by `C`. Do the Vietnamese get smaller aneurysms than the French?

5. Display a box plot of `A` against `R`. Does the one-way anova conclude there that there are differences in mean among the four risk classes? Is the variance for 0 risk factors different from the variance for 1 risk factor?

6. Display a box plot of `A` against `G:C`. Does a one-way anova detect significant differences? Where do differences come from? Do French men get bigger aneurysms than the rest of the population?

7. Repeat the same study for body mass index.

Exercise 1.3. Upload `ferretti.csv`, read data description. Assign columns `height`, `diameter`, `density`, `invasive`, to variables `H`, `Di`, `De`, `I`.

1. Display a box plot of `H` against `I`. Does the one-way anova conclude that the mean height of invasive and non invasive tumors are different? Does the T-test conclude that invasive tumors are higher than non invasive tumors on average?

2. Plot the ecdf of heights of invasive tumors in red, add the ecdf of heights of non invasive tumors in blue. Does the Kolmogorov-Smirnov test confirm that the heights of invasive tumors are larger?

3. Is the variance of heights of invasive tumors larger than that of non invasive tumors? Can this be seen on the ecdf plots? On the box plots?

4. Display a box plot of `H` against `De`. Does the one-way anova conclude that there are differences in mean among the three density classes? Are positive density tumors higher on average?

5. Repeat the same study for diameter.

2 Session 2

Objective: association of two discrete variables; chi-squared test of independence, proportion test, Fisher's exact test, odds-ratio, correlation test between two continuous variables.

Steps of the session:

1. Individual work: Practice with the instructions given on the web site.
2. Share and work: discuss what you understand.
3. Individual work: Practice with the exercises of Session 2 (at least exercises 1, 2 and 4).
4. Share and work: discuss your scripts together.
5. Individual work at home: Read the chapter *Two Sample tests*, sections *Continuous against continuous*.

Exercise 2.1. Assume 23 patients were exposed to a certain factor and got the disease, 56 were exposed and did not get the disease, 67 were not exposed and got the disease, 136 were not exposed and did not get the disease.

1. What are the odds of getting the disease for exposed patients? for unexposed patients? What is the odds-ratio? Does exposure favor the disease, or the contrary?
2. Make the contingency table of the two variables disease and exposure. Display the bar plot of conditional probabilities of disease given the exposure. Does the chi-square test conclude that the two variables are dependent or independent? Does the proportion test conclude that the proportion of diseased among exposed patients is smaller than the proportion of diseased among unexposed patients? Does Fisher's exact test conclude that the odds-ratio is significantly smaller than 1?
3. Repeat, multiplying all frequencies by 10.
4. Repeat, swapping the frequencies of exposed and unexposed. How does the odds-ratio change? How do the alternatives change? Do the p-values change?

Exercise 2.2. Upload `titanic.csv`, read data description. Assign columns `pclass`, `survived`, `gender`, `age`, to variables `P`, `S`, `G`, `A`. Conclusions are to be given at threshold 5%.

1. Display the contingency table of `G` and `S`. How many passengers were women? men? How many women survived? How many men survived? Display the bar plots of conditional distributions, of `S` given values of `G`. According to the chi-squared test, are the two variables dependent or independent? According to the proportion test, was the proportion of survivors among women greater than among men? According to Fisher's exact test, what was the odds-ratio of survival with respect to gender? what was the odds-ratio of death with respect to gender? Was it significantly different from 1? larger than 1?

2. Display the contingency table of `A<10` and `S`. How many passengers were below 10? above 10? How many of them survived? According to the chi-squared test, are the two variables dependent or independent? According to the proportion test, was the proportion of survivors among children below 10, greater than among other passengers? According to Fisher's exact test, what was the odds-ratio of survival? Was it significantly different from 1? larger than 1?

3. In the previous question, replace `A<10` by `A<21`. What do you conclude?

4. Display the contingency table of `P` and `S`. How many passengers was there in each class? How many survived in each class? According to the chi-squared test, are the two variables dependent or independent? According to the proportion test, were the proportions of survivors among the three classes different? According to Fisher's exact test, did first class passengers have better chances of survival than second class passengers? than third class passengers ? Did second class passengers have better chances of survival than third class passengers?

Exercise 2.3. Upload `bosson.csv`, read data description. Assign columns `country`, `gender`, to variables `C`, `G`. Assign column `risk` to variable `R`; group classes `3,4,5` into a single class, interpreted as "`3` or more".

1. Display the contingency table of `R` and `G`, then the bar plots of conditional distributions, of `R` given values of `G`. According to the chi-squared test, are the two variables dependent or independent?

2. Is the proportion of patients having zero risk factor, greater among women than among men? Is the proportion of patients having 2 risk factors or more, smaller among women than among men?

3. What is the odds-ratio of having zero risk factor, according to gender? What does Fisher's exact test conclude?

4. Repeat questions 1. to 3., replacing `G` by `C`, `women` by `Vietnam`, `men` by `France`.

Exercise 2.4. Upload `tauber.csv`, read data description. Assign columns `age`, `height`, `weight` to variables `A`, `H`, `W`. Assign to variable `AHW` the last three columns.

1. Display paired scatter plots of `AHW`. Which scatter plot is closest to a straight line? Display the correlation matrix of `AHW`. Is it coherent with the scatterplots? Explain why the correlation of height and weight is higher.

2. Does the correlation test conclude that the three correlations are significantly positive?

3 Session 3

Objectives: study the link between couples of variables on your dataset.

Steps of the session:

1. Team work: Compute two sample tests on your project dataset. Test the correlations between two continuous variables (what do you observe on the transcriptome variables ?), the independence between binary variables, test the link between the clinical variables and the transcriptome variables.

2. Team work: Write the report corresponding to your results

3. Individual work at home: Read the chapter *Regression models*, section *Linear regression*.