Chapter 5: One sample tests

1 Session 1

Objective: understand tests on a Gaussian sample; use R functions to test values of a mean, a standard deviation.

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 (at least exercises 1 and 3).
4. Share and work: discuss your scripts together.

Exercise 1.1. A packaging machine is supposed to produce 1 kg packs. The actual weight of a pack is modeled by a random variable following a normal distribution with standard deviation 20 g. It is possible to tune the mean weight of the packs. In order to check that the tuning is correct, a sample of 10 packs is weighed.

1. Let H_0 be the hypothesis: "the mean weight is 1 kg". Give the decision rule of a test at threshold 1%, of H_0 against H_1: "the mean weight is different from 1 kg". Find the p-value of that test, for a sample of average weight 1011 grams.
2. Same question for hypothesis H_1 : "the mean weight is larger than 1 kg".
3. Answer again the two previous questions for a sample of 100 packs, with mean weight 1005 g.

Exercise 1.2. A paracetamol concentration of more than 150 mg per kilogram body weight is considered as dangerous. The mesurements of paracetamol in blood tests are modelled by a random variable with normal distribution $\mathcal{N}(\mu, \sigma)$. The standard deviation, linked to the testing method, is supposed to be known and equal to 5 mg. For better assessment, 4 blood tests are usually made. The results are assumed to be independent realisations of the same normal distribution $\mathcal{N}(\mu, \sigma)$.

1. Give the hypotheses and the decision rule for the test deciding, at threshold 5%, whether a patient is at risk, on view of 4 blood tests (you are a cautious doctor).
2. On a given patient, the 4 blood tests gave concentrations of 140, 133, 148, 144. Give the p-value for the test of the previous question. Is he at risk?

3. From now on, the standard-deviation is supposed *unknown*. Give the test statistic and the decision rule for the test deciding, at threshold 5%, whether a patient is at risk, on view of 4 blood tests.

4. For the patient of question 2, give the p-value for the test of the previous question. What is your conclusion?

Exercise 1.3. 1. Upload `bosson.csv`, read data description. Display a boxplot of bmi against gender, then against country. What do you conclude?

2. Compute the mean bmi. Is it significantly smaller than 23? larger than 22?

3. Would you say that on average,

   (a) French people have a bmi above 25?

   (b) Vietnamese people have a bmi below 22?

   (c) women have a bmi below 22?

   (d) men have a bmi above 23?

   (e) Vietnamese women have a bmi below 20?

   (f) French men have a bmi above 25?

2 Session 2

Objectives: chi-squared and Kolmogorov-Smirnov goodness-of-fit tests; interpretation of results; give statistical answers to questions on a data set.

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 (at least exercises 1, 2 and 3).
4. Share and work: discuss your scripts together.
5. Individual work at home: Read the chapter *Two sample tests*, section *Continuous against discrete*.

Exercise 2.1. Answer the questions using the chi-squared test, at threshold 5%.

1. Consider a diploid population with allele frequencies 0.4 for A, 0.6 for a. At the Hardy-Weinberg equilibrium, the probabilities of the three genotypes AA, Aa, aa, are $(0.16, 0.48, 0.36)$. Frequency tables of the three genotypes are given. For each table, answer the question: is the theoretical model acceptable?
   (a) $(1600, 4900, 3500)$, $(16000, 49000, 35000)$,
   (b) $(150, 490, 360)$, $(1500, 4900, 3600)$.
2. In a Mendelian genetic cross, starting with 2 parents of genotypes AABB (double dominant) and aabb (double recessive), the second generation may have one of 16 different genotypes, hence 4 different phenotypes, with respective probabilities 9/16 for "double dominant", 3/16 for "single dominant A" and "single dominant B", 1/16 for "double recessive". Frequency tables of the four phenotypes are given. For each table, answer the question: is the Mendelian model acceptable?
   (a) $(9, 3, 3, 1)$, $(90, 30, 30, 10)$,
   (b) $(900, 300, 310, 90)$, $(9000, 3000, 3100, 900)$,
   (c) $(40, 20, 16, 4)$, $(100, 50, 40, 10)$, $(200, 100, 80, 20)$,
   (d) $(219, 81, 69, 31)$, $(2190, 810, 690, 310)$,
   (e) $(271, 73, 63, 26)$, $(270, 73, 63, 27)$.

Exercise 2.2. Over 5000 families of three children, there were 687 families with zero girl, 1986 with one girl, 1762 with two girls, 565 with three girls.

1. Assuming that girls and boys are equally likely, to what theoretical distribution should the table of frequencies be compared? Does the chi-squared test accept the goodness-of-fit?

2. The actual sex ratio (number of males divided by the number of females) is not exactly one, but rather 1.05. What is the probability of a child being a girl? To what probability distribution should the table of frequencies be compared? Does the chi-squared test accept the goodness-of-fit?

3. Over the same table of frequencies, what is the total number of children? Compute the relative frequency of girls, denoted by $\hat{p}$. Use the chi-squared test to compare the table of frequencies to the binomial distribution with parameters 3 and $\hat{p}$. What parameter should be used for the chi-squared distribution? What is the p-value of the test? What do you conclude?

4. Which p-value does the `chisq.test` function return? Why is it wrong?

Exercise 2.3.

1. Upload `bosson.csv`, read data description. Assign to variable `B` the bmi of the whole sample, to `Bv` the bmi of the Vietnamese, to `Bf` the bmi of the French, to `Bfw` the bmi of French women, to `Bfm` the bmi of French men, to `R` the number of risk factors. Display summaries of the variables.

2. Does the $\mathcal{N}(26, 3)$ fit the distribution of `Bf`? Des the $\mathcal{N}(20, 2.7)$ fit the distribution of `Bv`?

3. Are `B`, `Bv`, `Bf` normally distributed?

4. Among the French, do women have a mean bmi significantly smaller than 25? Do men have a mean bmi significantly larger than 25?

5. Does the binomial $\mathcal{B}(5, 0.3)$ fit the distribution of

Exercise 2.4. Upload `ferretti.csv`, read data description.

1. Are the proportions of tumors with negative, null, and positive density, equal?

2. Is the diameter of invasive tumors normally distributed?

3. Does the $\mathcal{N}(12, 5)$ fit the diameter of non invasive tumors?

4. Are invasive tumors higher than 10 mm on average?

5. Are tumors with positive density higher than 10 mm on average?

3 Session 3

Objectives: compute one sample tests on your dataset.

Steps of the session:

1. Team work: Compute one sample tests on your project dataset. Test the normality, the goodness of fit of some variables, test some values known from the litterature.

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

3. Individual work at home: Read the chapter *Two Sample tests*, sections *Continuous against discrete* and *Discrete against discrete*.