Masco; Probability and statistics

Academic year 2025-2026

November 2025

Instructions

This practical is composed of 4 parts:

- **Part 1**: Perform a two-sample test with both Frequentist and Bayesian approaches, and discover the problem of multiple hypothesis testing
- **Part 2**: About multiple tests.
- **Part 3**: Create a permutation test function in R.
- **Part 4**: Perform a full linear regression analysis with Jasp.

Form groups of 2-3 students.

Part 1: Comparing the means of two normal populations

We consider a between-subjects experiment to test the effect of oxygenated drink on cognitive abilities. The experiment consists of a set of memory scores for 20 subjects who took plain water (C), and a set of memory scores for 20 subjects who took oxygenated water (O).

```
O = c(70,80,79,83,77,75,84,78,75,75,78,82,74,81,72,70,75,72,76,77)
C = c(56,80,63,62,67,71,68,76,79,67,76,74,67,70,62,65,72,72,69,71)
```

Questions

You want to perform both Bayesian and Frequentist two-sample test.

1. Provide graphical representation using Boxplot. Comment.
2. Discuss all stages for setting-up a proper hypothesis test of equality of the means of these two groups (modeling assumptions, choice of the test...).
3. Import the data in Jasp.
4. Interpret the results, emphasize the advantages/drawbacks/eventual differences of both approaches.

Part 2: About multiple tests (Frequentist approach)

In many applications, we have to perform simultaneously many tests. For example, Statistical Parametric Mapping aims at identifying task-related areas of the brain from fMRI images. At each voxel (3D pixels), we perform a test statistic with null hypothesis: the i-th voxel is 'not task related' (understand the underlying neural mass did not specifically reacts to the stimulus). We assume that the voxels are independently reacting from each other (very strong assumption).

In other words, that consists in setting B hypothesis testing problems, based on B, n-samples $\left(X_1^{(i)}, \cdots, (X_n^{(i)}\right) \overset{i.i.d}{\sim} \mathcal{N}(\mu_i, \sigma^2), 1 \leq i \leq B$. Each hypothesis test has the form:

$$H_{0,i}: \ \mu_i = \mu_0,$$
$$H_{1,i}: \ \mu_i \neq \mu_0.$$

Imagine that we want to perform such a test for $B = 5000$ voxels, at the level $\alpha = 0.05$. From previous studies we know that about 150 voxels should be identified as task-related.

Questions

1. On average how many false positives shall we get ?
2. Discuss this result and illustrate it with R.
3. Compute the probability under the null hypothesis of committing at least one false positive among B tests. Deduce an adjusted level for testing your hypotheses.
4. To tackle this problem, Bonferroni suggested to use tests of level α/B. Comment on this choice.

Part 3: Permutation test

We have at our disposal $x = (x_1, \cdots, x_{n_1})$ and $y = (y_1, \cdots, y_{n_2})$ two observed samples of sizes n_1 and n_2 respectively. We are interested in the statistic $T(x, y)$ given by:

$$T(x,y) = \frac{\bar{x} - \bar{y}}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) V_{x,y}^2}}, \text{ with } V_{x,y}^2 = \frac{(n_1 - 1)s_{c,x}^2 + (n_2 - 1)s_{c,y}^2}{n_1 + n_2 - 2},$$

where $\bar{x}$ and $\bar{y}$ are empirical means; $s_{c,x}^2$ and $s_{c,y}^2$ are corrected empirical variances.

For a two independent sample location test problem, the permutation test algorithm consists in the following steps:

- Compute the observed test statistic $T(x, y)$, and set $N = 0$.
- For i from 1 to M:
  - Generate $\left(u_1^{(i)}, \cdots, u_{n_1+n_2}^{(i)}\right)$ by permuting the component of the data vector $(x_1, \cdots, x_{n_1}, y_1, \cdots, y_{n_2})$.
  - Define $x^{(i)} = \left(u_1^{(i)}, \cdots, u_{n_1}^{(i)}\right)$ and $y^{(i)} = \left(u_{n_1+1}^{(i)}, \cdots, u_{n_1+n_2}^{(i)}\right)$.
  - Check whether $T\left(x^{(i)}, y^{(i)}\right) \geq T(x, y)$. If it is the case $N \leftarrow N + 1$.
- Compute $p = \frac{N}{M}$.

The resulting p is a p-value based on the permutations of the statistic T.

Questions

1. Create a commented R function with four input arguments: `x`, the data vector of size n_1 for the first sample; `y`, the data vector of size n_2 for the second sample; `M`, the number of iterations; and `seed`, an integer with a default value of `NULL`. If `seed` is not specified, the random seed is not fixed; otherwise, it is set to the given value. The function's output is the p-value p.
2. Give your 'intuition' on how this kind of permutation tests work.
3. Apply your function to the data vectors of Part 1 exercise with $M = 10^5$. Conclude.

Part 4: Bayesian linear regression

Following the article *Van den Bergh et al. (2021) A tutorial on Bayesian mlti-model lienar regression with BAS and JASP*, perform a linear regression analysis on the dataset `'attitude.csv'` (provide descriptive statistics, check the underlying assumptions, comment the Jasp output, robustness check, interaction terms,...)

```
data <- read.csv("attitude.csv", row.names = 1) # Load the attitude data
head(data) # Display the first 6 rows of the data
```

```
##   rating complaints privileges learning raises critical advance
## 1     43         51         30       39     61       92      45
## 2     63         64         51       54     63       73      47
## 3     71         70         68       69     76       86      48
## 4     61         63         45       47     54       84      35
## 5     81         78         56       66     71       83      47
## 6     43         55         49       44     54       49      34
```

Data description

The dataset `'attitude.csv'` arises from a study in industrial psychology. It is a survey of clerical employees of a large financial organization with questions related to the employee satisfaction with their supervisors. There was a question designed to measure the overall performance of a supervisor, as well as questions that were related to specific activities involving interaction between supervisor and employees. The data are aggregated from the questionnaires of the approximately 35 employees for each of 30 (randomly selected) departments. The numbers give the percentage of favorable responses to seven questions in each department.

The data consists in 30 observations and 7 variables. The first column are the short names from the reference, the second one the variable names in the data frame:

- Y `rating` *numeric* Overall rating,
- $X[1]$ `complaints` *numeric* Handling of employee complaints,
- $X[2]$ `privileges` *numeric* Does not allow special privileges,
- $X[3]$ `learning` *numeric* Opportunity to learn,
- $X[4]$ `raises` *numeric* Raises based on performance,
- $X[5]$ `critical` *numeric* Too critical,
- $X[6]$ `advance` *numeric* Advancement.