❈ Appendix B: Hypothesis Tests

Each test is made with a significance level α. The null hypothesis H_0 is specified together with the decision variable D and its **distribution under the hypothesis** H_0. We denote by $\hat{d}$ the observation of D on the sample.

The shape of the critical region A depends on the alternative hypothesis H_1. It is defined with respect to some d which is defined by:

- $d = \hat{d}$ in the p-value method;
- $d = d_c$, the critical value of the test, in the critical value method.

In the p-value method, the expected answer is the

$$p - \text{value} = \mathbb{P}\left[A |\, H_0\right]$$

together with the result of the test with a sentence of explanation adapted to the context:

- If p-value $< \alpha$ we reject H_0 at the significance level α;
- if p-value $\geq \alpha$ we do not reject H_0, the data bringing no evidence of H_0 being not true.

In the critical value method, the expected answer is the dritical value d_c defined by

$$\mathbb{P}\left[A |\, H_0\right] = \alpha,$$

together with the result of the test with a sentence of explanation adapted to the context:

- If $\hat{d} \in A$, we reject H_0 at the significance level α:
- if $\hat{d} \notin A$ we do not reject H_0, the data bringing no evidence of H_0 being not true.

B.1 Parametric statistical tests for one sample

We assume that $(X_1, ..., X_n)$ is a sample of a random variable X. The null hypothesis deals with the value of some parameter θ which is compared to some fixed value θ_0.

The possibilities for the hypotheses and the related reject regions are:

H_0	$\theta = \theta_0$ or $\theta \geq \theta_0$	$\theta = \theta_0$ or $\theta \leq \theta_0$	$\theta = \theta_0$
H_1	$[\theta < \theta_0]$	$[\theta > \theta_0]$	$[\theta \neq \theta_0]$
A	left-sided	right-sided	two-sided

We recall the definitions:

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 \quad \text{and} \quad T_n = \frac{1}{n}\sum_{i=1}^{n}(X_i - m)^2.$$

B.1.1 Tests on the mean and variance of a Gaussian variable

We assume that $X \rightsquigarrow \mathcal{N}(m, \sigma^2)$.

Tests on the mean with known variance or Z-test

$$\theta = m,\ \theta_0 = m_0, \quad D = \frac{\overline{X}_n - m_0}{\sigma/\sqrt{n}} \rightsquigarrow \mathcal{N}(0,1) \quad \text{and} \quad \hat{d} = \frac{\overline{x}_n - m_0}{\sigma/\sqrt{n}}$$

Tests on the mean with unknown variance or T-test

$$\theta = m,\ \theta_0 = m_0, \quad D = \frac{\overline{X}_n - m_0}{S_n/\sqrt{n}} \rightsquigarrow \mathcal{T}_{n-1} \quad \text{and} \quad \hat{d} = \frac{\overline{x}_n - m_0}{s_n/\sqrt{n}}$$

Tests on the variance with known mean

$$\theta = \sigma,\ \theta_0 = \sigma_0, \quad D = n\frac{T_n}{\sigma_0^2} \rightsquigarrow \chi^2_n \quad \text{and} \quad \hat{d} = n\frac{t_n}{\sigma_0^2}$$

Tests on the variance with unknown mean

$$\theta = \sigma,\ \theta_0 = \sigma_0, \quad D = (n-1)\frac{S_n^2}{\sigma_0^2} \rightsquigarrow \chi^2_{n-1} \quad \text{and} \quad \hat{d} = (n-1)\frac{s_n^2}{\sigma_0^2}$$

B.1.2 Test on a proportion of a Bernoulli variable

We assume that $X \rightsquigarrow \mathcal{B}(p)$ and that we have a large sample.

$$\theta = p,\ \theta_0 = p_0, \quad D = \frac{\overline{X}_n - p_0}{\sqrt{p_0(1-p_0)}/\sqrt{n}} \rightsquigarrow \mathcal{N}(0,1) \quad \text{and} \quad \hat{d} = \frac{\overline{x}_n - p_0}{\sqrt{p_0(1-p_0)}/\sqrt{n}}$$

B.1.3 Test on a mean for orther square integrable variables

When X is square integrable (it has a variance and an expectation) and the sample size n is large we can use both Z and T- tests but only with the $\mathcal{N}(0,1)$ distribution (not the Student distribution).

Remark B.1. The explanations on how we compute the p-value or the critical value in the different cases of tests and reject regions are given on a separate document.

B.2 Parametric statistical tests for comparing two independent homogeneous samples

Let $(X_1^1, ..., X_{n_1}^1)$ and $(X_1^2, ..., X_{n_2}^2)$ be two indenpendent hommogeneous samples of parent variables X^1 and X^2 respectively. We assume their distributions depend on some parameters of same nature (means, variances,...) denoted θ_1 and θ_2.

The possibilities for the hypotheses and the related reject regions are

H_0	$\theta_1 = \theta_2$ or $\theta_1 \geq \theta_2$	$\theta_1 = \theta_2$ or $\theta 1 \leq \theta_2$	$\theta_1 = \theta_2$
H_1	$[\theta_1 < \theta_2]$	$[\theta_1 > \theta_2]$	$[\theta_1 \neq \theta_2]$
A	one-sided	one-sided	two-sided

B.2.1 Gaussian samples

We assume that $X^1 \rightsquigarrow \mathcal{N}(m_1, \sigma_1^2)$ and $X^2 \rightsquigarrow \mathcal{N}(m_2, \sigma_2^2)$.

Test on the variances

$$\theta_1 = \sigma_1,\ \theta_2 = \sigma_2, \quad D = \frac{S_1^2}{S_2^2} \rightsquigarrow F_{n_1-1,n_2-1} \quad \text{and} \quad \hat{d} = \frac{s_1^2}{s_2^2}.$$

In practice, we chose the indices such that $s_1 > s_2$ and we use only the right tail of the F-table.

Tests on the means with known variances

$$\theta_1 = m_1,\ \theta_2 = m_2, \quad D = \frac{\overline{X^1}_{n_1} - \overline{X^2}_{n_2}}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} \rightsquigarrow \mathcal{N}(0,1) \text{ and } \hat{d} = \frac{\overline{x^1}_{n_1} - \overline{x^2}_{n_2}}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$$

Tests on the means with unknown but equal variances

$$\theta_1 = m_1,\ \theta_2 = m_2, \quad D = \frac{\overline{X^1}_{n_1} - \overline{X^2}_{n_2}}{\sqrt{(n_1-1)S_1^2 + (n_2-1)S_2^2}} \frac{\sqrt{n_1+n_2-2}}{\sqrt{1/n_1 + 1/n_2}} \rightsquigarrow T_{n_1+n_2-2}$$

$$\text{and } \hat{d} = \frac{\overline{x^1}_{n_1} - \overline{x^2}_{n_2}}{\sqrt{(n_1-1)S_1^2 + (n_2-1)S_2^2}} \frac{\sqrt{n_1+n_2-2}}{\sqrt{1/n_1 + 1/n_2}}$$

B.2.2 Tests on the parameters of Bernoulli samples

We assume $X^1 \rightsquigarrow \mathcal{B}(p_1)$ et $X^2 \rightsquigarrow \mathcal{B}(p_2)$ and that the samples are large. We set

$$\hat{p} = \frac{n_1\overline{X}_1 + n_2\overline{X}_2}{n_1 + n_2}, \hat{d} = \frac{\overline{x^1}_{n_1} - \overline{x^2}_{n_2}}{\sqrt{\hat{p}(1-\hat{p})}\sqrt{1/n_1 + 1/n_2}}$$

and

$$\theta_1 = p_1, \theta_2 = p2, \quad \frac{\overline{X^1}_{n_1} - \overline{X^2}_{n_2}}{\sqrt{\hat{p}(1-\hat{p})}\sqrt{1/n_1 + 1/n_2}} \rightsquigarrow \mathcal{N}(0,1) \text{ approximatively.}$$

B.2.3 Tests on mean for square integrable parent variables.

When the samples are not Gaussian but eh variable is square integrable, and if the samples are large enough, we can still use the decision variable of the above tests on average but by using the $\mathcal{N}(0,1)$ distribution.

B.3 Nonparametric chi-squared tests

In each test, the decision variable D^2 has a χ^2 distribution under H_0, the constant d^2 being its actual value.

1. We compute either the p-value of the test

$$p\text{-value} = \mathbb{P}\left[D^2 \geq d^2 \mid H_0\right]$$

or d_c^2 the critical value of the test such that

$$\mathbb{P}\left[D^2 \geq d_c^2 \mid H_0\right] = \alpha.$$

2. We conclude the test as usual, rejecting the hypothesis H_0 whenever

$$p - \text{value} < \alpha \text{ or } d^2 > d_c^2.$$

B.3.1 Chi-square goodness of fit test

Let $(X_1, ..., X_n)$ be an n sample of some parent variable X.

With no estimated parameter. $\mathcal{L}$ being a distribution completely specified, the hypothesis are:

$$H_0 : [X \rightsquigarrow \mathcal{L}], \qquad H_1 : \overline{H_0}.$$

Let C_1,...,C_J be a partition of the range of both distributions, sample and target. For all $j \in \{1, \ldots, J\}$, we denote

$$p_j = \mathbb{P}\left[X \in C_j | H_0\right] \text{ and } N_j = \text{Card}\ \{i \in \{1, \ldots, n\} /\ X_i \in C_j\}$$

Then, under hypothesis H_0,

$$D^2 = \sum_{j=1}^{J} \frac{(N_j - np_j)^2}{np_j} \rightsquigarrow \chi^2_{J-1} \text{ approximatively.}$$

With estimated parameters. Here, we assume that the distribution $\mathcal{L}$ depends on k parameters θ_1, ..., θ_k. We denote by Θ_1, ..., Θ_k some estimators of these parameters and $\hat{\theta}_1$, ..., $\hat{\theta}_k$ their actual values caclulated with the sample data. Then

$$H_0\ :\ [X \rightsquigarrow \mathcal{L}(\hat{\theta}_1, \ldots, \hat{\theta}_k)], \qquad H_1\ :\ \overline{H_0}.$$

The test is similar to the previous one except that, under hypothesis H_0, we have:

$$D^2 \rightsquigarrow \chi^2_{J-1-k} \text{ approximatively.}$$

B.3.2 Comparaison of the distribution of samples of two or more groups

Let $(X^1_1, \ldots, X^1_{n_1})$,..., $(X^m_1, \ldots, X^m_{n_m})$ be m independent samples of parent variables X^1,...,X^m.

$$H_0\ :\ [\text{"the samples have the same distribution"}], \quad H_1\ :\ \overline{H_0}.$$

For each sample k, we denote $N_{kj} = \text{Card}\{i \in \{1, \ldots, n_k\} /\ X^k_i \in C_j\}$ and

$$N_{k\cdot} = \sum_{j=1}^{J} N_{kj}, \quad N_{\cdot j} = \sum_{k=1}^{m} N_{kj}, \quad N = \sum_{k=1}^{m} \sum_{j=1}^{J} N_{kj}.$$

$$\text{Under } H_0,\ D^2_0 = \sum_{k=1}^{m} \sum_{j=1}^{J} \frac{(N_{kj} - N_{k\cdot}N_{\cdot j}/N)^2}{N_{k\cdot}N_{\cdot j}/N} \rightsquigarrow \chi^2_{(J-1)(m-1)}.$$

B.3.3 Chi-squared independence test

We test the independence of two characteristics with a sample $((X_1, Y_1), \ldots, (X_n, Y_n))$ of parent variable (X, Y) reprensenting the characteristics. We denote $\{C_i,\ i \in \{1, \ldots, I\}\}$ and $\{C^j,\ j \in \{1, \ldots, J\}\}$ partitions of the ranges of X and Y respectively and

$$N_{ij} = \text{Card}\{l \in \{1, \ldots, n\} /\ X_l \in C_i \text{ et } Y_l \in C^j\}.$$

Finally,

$$H_0\ :\ [\text{"}X \text{ and } Y \text{are independent"}], \quad H_1\ :\ \overline{H_0}.$$

We proceeed as in the previous test.

References

[1] Textbook: B. Illowsky and S. Dean *Introductory Statistics*. OpenStax 2018, https://openstax.org/details/books/introductory-statistics

[2] In French: G. Saporta *Probabilités, analyse des données et statistique*. Edition Technip (1990, 2006 or 2011 editions are all sufficient for this course, the last one includes machine learning)