Chapter 1

Introduction to Hypothesis Testing Theory

Introduction

Statistical tests are methods of inferential statistics which, like estimation, allow for the analysis of data obtained through random sampling. They consist of generalizing properties observed in samples to the population from which they were drawn, and answering questions regarding, for example, the nature of a probability distribution, the value of a parameter, or the independence of two random variables.

While it is important to present all statistical tests in detail, the literature on the subject is very abundant. Therefore, in this chapter, we will limit ourselves to the simplest and most common classical tests used in practice. Specifically, the tests presented concern single-sample tests, goodness-of-fit tests, two-sample comparison tests, and finally, one-way and two-way analysis of variance (ANOVA).

1.1 Empirical Estimator of the Mean and Variance

1.1.1 Construction of the Estimator

Given a sample $(X_1, \ldots, X_n)$ of X, we define the empirical probability distribution P_n such that: $P_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}$ where δ_{X_i} is the Dirac mass (indicator function) at point X_i. This probability distribution admits an empirical distribution function, denoted F_n:

$$F_n(x) = \frac{\text{number of } x_i \text{ less than or equal to } x}{n} = P_n(]-\infty, x]). \tag{1.1}$$

Definition 1.1 (Empirical Mean)
The empirical mean of an n-sample (i.i.d) is the mean of the empirical distribution denoted:

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

this random variable (r.v.) admits expectation and variance.

Definition 1.2 (Empirical Variance) The empirical variance of the empirical distribution is:

$$\begin{cases} S_n^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \mu)^2, & \text{if } \mu \text{ is known;} \\ S_n^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X}_n)^2, & \text{if } \mu \text{ is unknown.} \end{cases}$$

Remark 1.1 The average value of the empirical variance is not exactly equal to the theoretical variance σ^2 (it is a biased estimator), which is why we introduce the corrected empirical variance defined by:

$$S_{n,c}^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2$$

1.1.2 Distribution of the Estimator of a Mean and Variance

In general, a simple estimation of a characteristic (mean, variance, median, ...) of a sample is not enough: it is necessary to know its degree of imprecision. The fundamental tool for evaluating an estimator and comparing it to others is its sampling distribution.

Consider a quantitative character represented by a random variable X with mathematical expectation μ, variance σ^2, and a sample $X_1, X_2, \ldots, X_n$ of X of size n.

1. For each sampling, we can calculate the observed mean of the character:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$$

   It is demonstrated that $E(\bar{X}) = \mu$ (an unbiased estimator of μ) and $Var(\bar{X}) = \frac{\sigma^2}{n}$.

2. If the mean μ is **known**, then we consider the sample variance:

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \mu)^2 = \left(\frac{1}{n}\sum_{i=1}^n X_i^2\right) - \mu^2$$

3. If the mean μ is **unknown**, then the unbiased estimator of the sample variance is defined as:

$$\hat{\sigma}_c^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2 = \frac{n}{n-1}\hat{S}^2$$

   with $\hat{S}^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2$.

The probability distributions of the mean and variance estimators can be summarized as follows:

1. **Case of a small Gaussian sample ($n \leq 30$ and $X \sim N(\mu, \sigma^2)$):**
   - If σ is known, $U = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ follows the normal distribution $N(0,1)$.
   - If σ is unknown, $T = \frac{\bar{X}-\mu}{\hat{\sigma}_c/\sqrt{n}}$ follows the Student's t-distribution with $n-1$ degrees of freedom.
2. **Case of a large sample ($n > 30$) and X from any distribution:**
   - In this case, $U = \frac{\bar{X}-\mu}{\hat{\sigma}_c/\sqrt{n}}$ approximately follows the normal distribution $N(0,1)$.

1.2 Conformity Tests for a Mean

We consider a quantitative character represented by a random variable X with mean μ and standard deviation σ.

1.2.1 Required Conditions

- The sample must be drawn from a population following a Normal distribution $N(\mu, \sigma^2)$, OR the sample size must be large ($n > 30$).
- Observations must be independent.

1.2.2 Case 1: Known Variance (σ known), or unknown variance and $n > 30$

The test statistic used is:

$$U = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$$

Under the null hypothesis $H_0 : \mu = \mu_0$, U follows the standard normal distribution $N(0, 1)$.

Three Hypothesis Scenarios and Decisions

1. **Bilateral Test (Two-tailed):**
   - **Hypotheses:** $H_0 : \mu = \mu_0$ vs $H_1 : \mu \neq \mu_0$
   - **Condition:** We look for values extremely far from μ_0 in both directions.
   - **Decision:** Reject H_0 if $|u| > u_{\alpha/2}$.
2. **Right Unilateral Test (Upper-tailed):**
   - **Hypotheses:** $H_0 : \mu = \mu_0$ vs $H_1 : \mu > \mu_0$
   - **Condition:** We check if the observed mean is significantly larger than μ_0.
   - **Decision:** Reject H_0 if $u > u_\alpha$.
3. **Left Unilateral Test (Lower-tailed):**
   - **Hypotheses:** $H_0 : \mu = \mu_0$ vs $H_1 : \mu < \mu_0$
   - **Condition:** We check if the observed mean is significantly smaller than μ_0.
   - **Decision:** Reject H_0 if $u < -u_\alpha$.

1.2.3 Case 2: Unknown Variance (σ unknown, $n \leq 30$)

The test statistic follows a Student's t-distribution with $n - 1$ degrees of freedom:

$$T = \frac{\bar{X} - \mu_0}{\hat{\sigma}_c/\sqrt{n}}$$

Decision: Similar to the normal case, but using critical values t_α from the Student's table with $(n - 1)$ d.f.

1.3 Homogeneity Tests

In the previous sections, we only considered a single sample to see if one of its characteristics (mean, variance, distribution) conformed to an arbitrarily fixed value (usually representing a standard or known norm). However, in practice, we often have two populations P_1 and P_2 (or more) where we study the same characteristic and wish to compare the populations to determine if they are homogeneous or not. In this section, we limit ourselves to testing the homogeneity of variance and means for two independent populations.

1.3.1 Required Conditions

- Both samples must be independent.
- Populations follow Normal distributions $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$.
- For the homogeneity of the means, variances must be equal.

1.3.2 Comparison of Two Variances

Let X and Y be two independent random variables representing the same quantitative characteristic in populations P_1 and P_2. We assume that X and Y follow normal distributions $N(\mu_1; \sigma_1^2)$ and $N(\mu_2; \sigma_2^2)$, respectively.

From P_1, we extract a sample $X_1, X_2, \ldots, X_{n_1}$ of size n_1. From P_2, we extract a sample $Y_1, Y_2, \ldots, Y_{n_2}$ of size n_2. The empirical means of the two samples are:

$$\bar{X} = \frac{1}{n_1}\sum_{i=1}^{n_1} X_i, \quad \bar{Y} = \frac{1}{n_2}\sum_{i=1}^{n_2} Y_i$$

And their corrected variances are:

$$\hat{\sigma}_{c,1}^2 = \frac{n_1}{n_1 - 1}\hat{\sigma}_1^2 \text{ where } \hat{\sigma}_1^2 = \frac{1}{n_1}\sum_{i=1}^{n_1} X_i^2 - \bar{X}^2$$

$$\hat{\sigma}_{c,2}^2 = \frac{n_2}{n_2 - 1}\hat{\sigma}_2^2 \text{ where } \hat{\sigma}_2^2 = \frac{1}{n_2}\sum_{i=1}^{n_2} Y_i^2 - \bar{Y}^2$$

We want to perform the following two-tailed test:

$$H_0 : \sigma_1^2 = \sigma_2^2 \quad \text{against} \quad H_1 : \sigma_1^2 \neq \sigma_2^2$$

The steps for this test are summarized as follows:

1. Calculate the realized value $f_c = \frac{\hat{\sigma}_{c,1}^2}{\hat{\sigma}_{c,2}^2}$. If necessary, permute the samples so that $f_c \geq 1$ (i.e., $f_c = \frac{\max(\hat{\sigma}_{c,1}^2, \hat{\sigma}_{c,2}^2)}{\min(\hat{\sigma}_{c,1}^2, \hat{\sigma}_{c,2}^2)}$).

2. Under the null hypothesis H_0, the statistic $F = \frac{\hat{\sigma}_{c,1}^2}{\hat{\sigma}_{c,2}^2}$ follows a Fisher distribution with $(n_1 - 1; n_2 - 1)$ degrees of freedom. From the Fisher table, determine f_α such that $P(F \geq f_\alpha) = \alpha/2$.

3. The decision rule is:

   - If $f_c < f_\alpha$, we cannot reject H_0 (H_0 is assumed true).
   - If $f_c \geq f_\alpha$, we reject H_0 with a probability α of being wrong.

Summary of Non-Rejection Zones for H_0:

Hypothesis	**Non-rejection Zone for H_0**
$H_0 : \sigma_1^2 = \sigma_2^2$ vs $H_1 : \sigma_1^2 \neq \sigma_2^2$	$[1; f(n_1 - 1, n_2 - 1, 1 - \alpha/2)]$
$H_0 : \sigma_1^2 = \sigma_2^2$ vs $H_1 : \sigma_1^2 > \sigma_2^2$	$[1; f(n_1 - 1, n_2 - 1, 1 - \alpha)]$
$H_0 : \sigma_1^2 = \sigma_2^2$ vs $H_1 : \sigma_1^2 < \sigma_2^2$	$[1; f(n_2 - 1, n_1 - 1, 1 - \alpha)]$, with $f_c = \frac{\hat{\sigma}_{c,2}^2}{\hat{\sigma}_{c,1}^2}$

1.3.3 Comparison of Two Means

In this section, we focus on the homogeneity of two populations regarding their mean. Note that the comparison test depends on the distribution and size of the samples. We focus on large samples from any distribution and small Gaussian samples.

Case of Large Samples ($n > 30$)

Let X and Y be independent random variables with means μ_1, μ_2 and standard deviations σ_1, σ_2. We extract samples of size $n_1 > 30$ and $n_2 > 30$. The test statistic is:

$$U = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \tag{1.2}$$

Two-tailed Test: $H_0 : \mu_1 = \mu_2$ vs $H_1 : \mu_1 \neq \mu_2$
Under H_0, U approximately follows $N(0, 1)$. Calculate $u = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{\hat{\sigma}_{c,1}^2}{n_1} + \frac{\hat{\sigma}_{c,2}^2}{n_2}}}$. Determine u_α such that $P(-u_\alpha < U < u_\alpha) = 1 - \alpha$.

- Do not reject H_0 if $u \in]-u_\alpha, u_\alpha[$.
- Reject H_0 if $u \notin]-u_\alpha, u_\alpha[$.

One-tailed Test (Upper): $H_0 : \mu_1 = \mu_2$ vs $H_1 : \mu_1 > \mu_2$
Reject H_0 if $u \geq u_\alpha$ where $P(U \geq u_\alpha) = 1 - \alpha$.

Case of Small Samples ($n \leq 30$)

Assume $X \sim N(\mu_1, \sigma_1^2)$ and $Y \sim N(\mu_2, \sigma_2^2)$ with $n_1, n_2 \leq 30$. The statistic is:

$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}}} \tag{1.3}$$

If we assume $\sigma_1^2 = \sigma_2^2 = \sigma^2$ (after performing the variance comparison test), the common variance is estimated by:

$$\hat{\sigma}_c^2 = \frac{(n_1 - 1)\hat{\sigma}_{c,1}^2 + (n_2 - 1)\hat{\sigma}_{c,2}^2}{n_1 + n_2 - 2}$$

Then calculate:

$$t = \frac{\bar{x} - \bar{y}}{\hat{\sigma}_c \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Under H_0, T follows a Student's t-distribution with $n_1 + n_2 - 2$ degrees of freedom.

Decision for Two-tailed Test ($H_1 : \mu_1 \neq \mu_2$):

- Do not reject H_0 if $t \in]-t_\alpha, t_\alpha[$.
- Reject H_0 if $t \notin]-t_\alpha, t_\alpha[$.

1.4 Analysis of variance with one factor (ANOVA 1)

In this section, we will focus on a more general case for comparing means, specifically when the number of samples is strictly greater than two. More precisely, we will examine the single-factor analysis of variance technique, which is most appropriate for this situation.

1.4.1 Problem Statement

Suppose we have 3 databases containing a specific type of information, and we want to determine whether these databases influence the storage space occupied on the hard disk by this information. To this end, we have collected the storage capacities occupied by six (06) files from each of the three databases, the measurements of which are arranged in the following table:

$N°$	Base 1	Base 2	Base 3
1	23.3	18.9	22.5
2	24.4	21.1	22.9
3	24.6	21.1	23.7
4	24.9	22.1	24.0
5	25.0	22.5	24.0
6	26.2	23.5	24.5

Table 1.1: Storage space occupied according to the database

Let us define the following concepts and notations:

- **The Database:** Qualitative variable containing three modalities, referred to as the factor.
- **Occupied memory space:** Response variable, denoted by X, and μ_i represents the average memory space occupied in the i-th database ($i = 1, 3$).

Answering our objective consists of performing the following test:

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu \quad \text{against} \quad H_1 : \exists\, i, j \in \{1, 2, 3\} \text{ such that } \mu_i \neq \mu_j$$

To perform this test, we could decompose it into three sub-tests comparing the average memory space two-by-two according to the databases. However, order to avoid the problem of an inflated α error, we use the statistical technique known as **Analysis of Variance (ANOVA)** rather than multiple Student t-tests (see section 1.3.2). Note that ANOVA can also be used when $p = 2$, as it returns the same conclusion as a t-test.

1.4.2 Single-factor analysis of variance

The identifying description of ANOVA 1 in a literal sense can be summarized in the following definition:

Definition 1.3 (ANOVA 1)
Single-factor analysis of variance tests the effect of a controlled factor A having p modalities (groups) on the means of a quantitative variable X.

The problems concerned by the ANOVA 1 technique are presented in their general form as follows:

N°	group 1	group 2	...	group p
1	$X_{1,1}$	$X_{1,2}$	...	$X_{1,p}$
2	$X_{2,1}$	$X_{2,2}$	...	$X_{2,p}$
3	$X_{3,1}$	$X_{3,2}$	...	$X_{3,p}$
4	$X_{4,1}$	$X_{4,2}$	...	$X_{4,p}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
n_j	$X_{n_1,1}$	$X_{n_2,2}$	...	$X_{n_p,p}$

And the associated mathematical model is given by:

$$X_{ij} = \mu_i + \epsilon_{ij}, \quad \text{with } i = \overline{1,n}, j = \overline{1,p} \text{ and } \epsilon_{ij} \sim N(0, \sigma^2), \tag{1.4}$$

where X_{ij} is the j-th realization of the quantitative variable X in the i-th sample and ϵ_{ij} are the measurement errors.

If this model is retained, then the test to be performed is defined by:

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_p = \mu \quad \text{against} \quad H_1 : \exists\, i, j \in \{1, 2, \ldots, p\} \text{ such that } \mu_i \neq \mu_j. \tag{1.5}$$

In what follows, we will enumerate the steps for implementing ANOVA 1 which allow us to perform this test.

1.4.3 Steps of ANOVA 1

To carry out the test defined in (1.5), three conditions must primarily be verified beforehand:

- The p compared samples are independent.
- The studied quantitative variable follows a normal distribution in the p compared populations.
- The p compared populations have the same variance: **Homogeneity** of variances or **homoscedasticity**.

If these last three conditions are verified, then we can use the ANOVA 1 technique for the test (1.5). We need the following statistics:

- The mean of all observations: $\bar{X} = \frac{1}{n} \sum_{j=1}^{p} \sum_{i=1}^{n_j} X_{ij}$ with $n = \sum_{j=1}^{p} n_j$
- Mean of each sample: $\bar{X}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} X_{ij}$, for $j = \overline{1,p}$

- Variance of each sample: $\hat{\sigma}_i^2 = \frac{1}{n_j}\sum_{i=1}^{n_j}(X_{ij} - \bar{X}_j)^2$, for $j = \overline{1,p}$
- The variance of all observations: $\hat{\sigma}^2 = \frac{1}{n}\sum_{j=1}^{p}\sum_{i=1}^{n_j}(X_{ij} - \bar{X})^2$ with $n = \sum_{j=1}^{p} n_j$

It can be easily demonstrated that the variance of all observations is the sum of the variance of the means and the mean of the variances of the p samples:

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{j=1}^{p}\sum_{i=1}^{n_j}(X_{ij} - \bar{X})^2 = \frac{1}{p}\sum_{j=1}^{p}\sigma_i^2 + \frac{1}{p}\sum_{j=1}^{p}(\bar{X}_j - \bar{X})^2, \tag{1.6}$$

or even:

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{j=1}^{p}\sum_{i=1}^{n_j}(X_{ij} - \bar{X})^2 = \frac{1}{n}\sum_{j=1}^{p}\sum_{i=1}^{n_j}(X_{ij} - \bar{X}_j)^2 + \frac{1}{p}\sum_{j=1}^{p}(\bar{X}_j - \bar{X})^2. \tag{1.7}$$

By multiplying (1.7) by n, we obtain:

$$\underbrace{\sum_{j=1}^{p}\sum_{i=1}^{n_j}(X_{ij} - \bar{X})^2}_{SC_{Tot}} = \underbrace{\sum_{j=1}^{p}\sum_{i=1}^{n_j}(X_{ij} - \bar{X}_j)^2}_{SC_{Res}} + \underbrace{\sum_{j=1}^{p}\sum_{i=1}^{n_j}(\bar{X}_j - \bar{X})^2}_{SC_{Fac}}, \tag{1.8}$$

where:

- SC_{Tot}: Total variation representing the dispersion of data around the general mean.
- SC_{Fac}: Variation due to the factor (inter-group variation) representing the dispersion of means around the general mean.
- SC_{Res}: Residual variation (intra-group variation) representing the dispersion of data within each sample around its mean.

The most natural hypothesis is that the factor has no impact if the total variation is primarily generated by intra-group variation:

- If H_0 is true, the variation SC_{Fac} due to the factor should be small compared to SC_{Res}.
- If H_1 is true, the variation SC_{Fac} due to the factor should be large compared to SC_{Res}.

To compare these, Fisher considered the ratio of the Mean Squares (CM):

- Mean square for the factor: $CM_{Fac} = \frac{SC_{Fac}}{p-1}$
- Mean residual square: $CM_{Res} = \frac{SC_{Res}}{n-p}$

If the conditions are met and H_0 is true, then:

$$F_{obs} = \frac{CM_{Fac}}{SC_{Res}} \sim f_{(p-1,n-p)}$$

Decision: For a given risk threshold α, Fisher tables provide a critical value f_α such that:

$$P\left(\frac{CM_{Fac}}{SC_{Res}} < f_\alpha\right) = 1 - \alpha$$

- if $f_{obs} < f_\alpha \implies$ we cannot reject H_0 (Factor has no influence).
- if $f_{obs} \geq f_\alpha \implies$ we reject H_0 (Factor influences the variable).

Source of variation	Sum of Squares (SC)	Degrees of Freedom (df)	Mean Square (CM)	Ratio F_{obs}	Fisher c
Inter-group (Fac)	SC_{Fac}	$p-1$	CM_{Fac}	$\frac{CM_{Fac}}{CM_{Res}}$	c
Intra-group (Res)	SC_{Res}	$n-p$	CM_{Res}		
Total	SC_{Tot}	$n-1$			

1.3.4 Example

Problem: An agronomist wants to know if three different types of fertilizers (Soil A, Soil B, Soil C) influence the growth height of bean plants. Six plants are tested for each soil type.

Table 1.2: Plant Height (cm) after 4 weeks

Plant №	**Soil A**	**Soil B**	**Soil C**
1	23.3	18.9	22.5
2	24.4	21.1	22.9
3	24.6	21.1	23.7
4	24.9	22.1	24.0
5	25.0	22.5	24.0
6	26.2	23.5	24.5

Hypotheses:

- H_0: $\mu_A = \mu_B = \mu_C$ (Fertilizer has no effect on growth)
- H_1: At least one group mean is different.

Step 1: Calculate Group Means

- $\bar{X}_1 = 24.73$
- $\bar{X}_2 = 21.53$
- $\bar{X}_3 = 23.60$
- **Global Mean** ($\bar{X}_{total}$): 23.29

Step 2: ANOVA Table Results

Source	**Sum of Squares (SS)**	**df**	**Mean Square (MS)**	**F-ratio**	**F-crit (5%)**
Between Groups	31.59	2	15.79	12.02	3.68
Within Groups	19.71	15	1.31		
Total	51.30	17			

Decision: Since $F_{obs} = 12.02$ is **greater** than $F_{crit} = 3.68$, we **reject** H_0.

Conclusion

Based on the concepts and tests discussed in this chapter, we conclude that:

A hypothesis test is an inference process for controlling (accepting or rejecting) the validity of hypotheses regarding one or more populations based on random samples.

Statistical inference methods allow us to determine, with a given probability, whether observed differences are due to chance or if they signify that samples come from different populations.

Tests require a set of underlying assumptions (normality, equality of variances, independence, etc.) which must be verified beforehand.