

Statistics  
Semester 4

# Contents

<b>Chapter 1</b>	<b>Revision of Probability</b>	<b>Page 2</b>
1.1	Discrete Distributions	2
1.2	Continuous Distributions	3
1.3	Convergence	4
1.4	Approximations	4
1.5	Further laws	5
<b>Chapter 2</b>	<b>Estimators</b>	<b>Page 7</b>
2.1	Point estimation of the mean	7
2.2	Point estimator of the variance Suppose $\mu$ is unknown — 8 • Suppose $\mu$ is known — 8	8
2.3	Point estimation of a proportion (percentage)	9
2.4	Confidence interval Confidence interval for the mean — 9 • Confidence interval for a proportion (percentage) — 11 • Confidence interval for the variance — 11	9
2.5	Mean Squared Error	12
<b>Chapter 3</b>	<b>Hypothesis Testing</b>	<b>Page 13</b>
3.1	Introduction	13
3.2	Comparison between a mean and a reference value Two sided test — 13 • One sided test — 14	13
3.3	Comparison between a proportion and a reference value Two sided test — 15 • One sided test — 15	15
3.4	Comparison between a variance and a reference value Two sided test — 16 • One sided test — 16	16
3.5	Critical Probability	16
3.6	Comparison between two means Two sided test — 16 • One sided test — 17	16
3.7	Comparison between 2 proportions Two sided test — 17 • One sided test — 18	17
3.8	Comparison between two variances Two sided test — 18 • One sided test — 18	18
<b>Chapter 4</b>	<b>Regression</b>	<b>Page 19</b>
4.1	Linear Regression	19
4.2	Coefficient of linear correlation	19

# Chapter 1

## Revision of Probability

I'm simply gonna list rules.

$$\mathbb{E}(X) = \mu = \sum_{i \in \Omega} X_i \Pr(X_i)$$

$$\mathbb{E}(g(X)) = \sum_{i \in \Omega} g(X_i) \Pr(X_i)$$

$$\mathbb{E}(aX + b) = a\mathbb{E}(X) + b$$

$$\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y) \quad \text{if both variables are independent}$$

$$\text{Var}(X) = \sigma^2 = \mathbb{E}(X^2) - \mu^2$$

$$\text{Var}(aX + bY) = a^2 \text{Var}(X) + b^2 \text{Var}(Y) + 2ab \text{cov}(X, Y)$$

where

$$\text{cov}(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X) \cdot \mathbb{E}(Y).$$

### 1.1 Discrete Distributions

#### 1. Uniform discrete law

$$X(\Omega) = \{1, 2, 3, \dots, n\}$$

$$\Pr(X = k) = \frac{1}{n} \quad \forall k = 1, 2, 3, \dots, n$$

$$\begin{cases} \mathbb{E}(X) = \frac{n+1}{2} \\ \text{Var}(X) = \frac{n^2-1}{12} \end{cases}$$

#### 2. Bernoulli law of parameters $p$ ( $0 < p < 1$ )

$$X \sim B(p)$$

$$X(\Omega) = \{0, 1\}$$

$$\Pr(X = 1) = p \quad \Pr(X = 0) = 1 - p$$

$$\begin{cases} \mathbb{E}(X) = p \\ \text{Var}(X) = p(1 - p) \end{cases}$$

### 3. Binomial law of parameters $n$ and $p$

$$\begin{aligned} X &\sim \text{Bin}(n, p) \\ X(\Omega) &= \{1, 2, \dots, n\} \\ \Pr(X = k) &= C_n^k p^k q^{n-k} \quad \forall k \in \{0, 1, 2, \dots, n\} \\ \begin{cases} \mathbb{E}(X) = np \\ \text{Var}(X) = np(1-p) \end{cases} \end{aligned}$$

### 4. Hypergeometric law

$$\begin{aligned} X &\sim \mathcal{H}(N, n, p) \\ X(\Omega) &= [\max\{0, n - N + M\}, \min\{M, n\}] \\ \Pr(X = k) &= \frac{C_M^k \cdot C_{N-M}^{n-k}}{C_N^n} \quad \forall k \in X(\Omega) \\ \begin{cases} \mathbb{E}(X) = np \\ \text{Var}(X) = np(1-p) \left(\frac{N-n}{N-1}\right) \end{cases} \end{aligned}$$

### 5. Geometric law

$$\begin{aligned} X &\sim G(p) \\ X(\Omega) &= \mathbb{N}^* \\ \Pr(X = k) &= p(1-p)^{k-1} \quad \forall k \in \mathbb{N}^* \\ \begin{cases} \mathbb{E}(X) = \frac{1}{p} \\ \text{Var}(X) = \frac{1-p}{p^2} \end{cases} \end{aligned}$$

### 6. Poisson's law of parameter $\lambda$ ( $\lambda \in \mathbb{R}_+^*$ )

$$\begin{aligned} X &\sim \mathcal{P}(\lambda) \\ X(\Omega) &= \mathbb{N} \\ \Pr(X = k) &= e^{-\lambda} \frac{\lambda^k}{k!} \quad \forall k \in \mathbb{N} \\ \begin{cases} \mathbb{E}(X) = \lambda \\ \text{Var}(X) = \lambda \end{cases} \end{aligned}$$

## 1.2 Continuous Distributions

#### 1. Uniform law

$$\begin{aligned} f(x) &= \begin{cases} \frac{1}{b-a} & \text{if } x \in [a, b] \\ 0 & \text{else} \end{cases} \\ \begin{cases} \mathbb{E}(x) = \frac{a+b}{2} \\ \text{Var}(x) = \frac{(b-a)^2}{12} \end{cases} \end{aligned}$$

#### 2. Exponential law

$$\begin{aligned} x &\sim \xi(\lambda) \\ f(x) &= \begin{cases} \lambda e^{-\lambda x} & \text{if } x > 0 \\ 0 & \text{else} \end{cases} \\ \begin{cases} \mathbb{E}(x) = \frac{1}{\lambda} \\ \text{Var}(x) = \frac{1}{\lambda^2} \end{cases} \end{aligned}$$

### 3. Normal law

$$x \sim \mathcal{N}(\mu, \sigma)$$

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$\begin{cases} \mathbb{E}(x) = \mu \\ \text{Var}(x) = \sigma^2 \end{cases}$$

For  $\mathcal{N}(0, 1)$

$$\Phi(z) = \int_{-\infty}^z \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx.$$

$$\pi(z) = \Phi(z) - 0.5 = \int_0^z \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx.$$

## 1.3 Convergence

### Theorem 1.3.1 Chebyshev's inequality

Let  $X$  be a random variable of expectation  $\mathbb{E}(X)$  and variance  $\text{Var}(X)$ . Then  $\forall \varepsilon$

$$\Pr(|X - \mathbb{E}(X)| \geq \varepsilon) \leq \frac{\text{Var}(X)}{\varepsilon^2}.$$

it can also be stated as

$$\Pr(|X - \mathbb{E}(X)| < \varepsilon) \geq 1 - \frac{\text{Var}(X)}{\varepsilon^2}.$$

We say a sequence of random variables  $X_n$  converges to  $a$  ( $X_n \xrightarrow{\Pr} a$ ) if  $\forall \varepsilon$

$$\lim_{n \rightarrow +\infty} \Pr(|X_n - a| > \varepsilon) = 0.$$

or

$$\lim_{n \rightarrow +\infty} \Pr(|X_n - a| \leq \varepsilon) = 1.$$

### Theorem 1.3.2 Weak law of large numbers

Consider a random variable  $(X_n)$  of mean  $\mu$  and variance  $\sigma^2$ . Consider the random variable  $\tilde{X}_n = \frac{X_1 + X_2 + \dots + X_n}{n}$ . It can be shown that  $\tilde{X}_n$  converges to  $\mu$  meaning  $\forall \varepsilon$

$$\lim_{n \rightarrow +\infty} \Pr(|\tilde{X}_n - \mu| > \varepsilon) = 0.$$

## 1.4 Approximations

### Theorem 1.4.1 Binomial by a Poisson

$$\text{Bin}(n, p) \sim \mathcal{P}(np) \quad \text{if} \quad \begin{cases} n \geq 30 \\ p \leq 0.1 \\ np < 15 \end{cases}.$$

**Theorem 1.4.2** Hypergeometric by a Binomial

$$\mathcal{H}(N, n, p) \sim \text{Bin}(n, p) \quad \text{if } n \leq 0.05N.$$

**Theorem 1.4.3** De Moivre–Laplace theorem

$$\text{Bin}(n, p) \sim \mathcal{N}\left(np, \sqrt{np(1-p)}\right) \quad \text{if } \begin{cases} n \geq 30 \\ np \geq 5 \\ n(1-p) \geq 5 \end{cases}.$$

In this case the event  $X = k$  can be replaced by  $k - 0.5 < X < l + 0.5$

**Theorem 1.4.4** Central limit theorem

Let  $(X_n)$  be a sequence of independent random variables following the same law of expectation  $\mu$  and of standard deviation  $\sigma$ . Let  $S_n = \sum_{i=1}^n X_i$  and  $S_n^* = \frac{S_n - n\mu}{\sigma\sqrt{n}}$ . It can be shown that  $S_n^*$  converges in law to  $\mathcal{N}(0, 1)$ .

$$\begin{aligned} \mathbb{E}(S_n) &= n\mu \\ \text{Var}(S_n) &= n\sigma^2 \end{aligned}$$

## 1.5 Further laws

**Theorem 1.5.1** Chi square law

Let  $X_1, X_2, \dots, X_n$  be  $n$  independent random variables following the standard normal law  $\mathcal{N}(0, 1)$ . Let  $Y = X_1^2 + X_2^2 + \dots + X_n^2$ . We say that  $Y$  follows a chi-square law with  $n$  degrees of freedom.  $Y \sim \chi_n^2$ .

$$\begin{aligned} \mathbb{E}(Y) &= n \\ \text{Var}(Y) &= 2n \end{aligned}$$

It can be shown that the density function of  $Y$  is

$$f(x) = \begin{cases} \frac{1}{2^{\frac{n}{2}} \Gamma(\frac{n}{2})} x^{\frac{n}{2}-1} e^{-\frac{x}{2}} & \text{if } x > 0 \\ 0 & \text{else} \end{cases}.$$

where  $\Gamma$  is the gamma function

$$\Gamma(x) = \int_0^{+\infty} t^{x-1} e^{-t} dt \quad \forall x > 0.$$

**Theorem 1.5.2** Student law(t-distribution)

Let  $X, Z$  be two independent random variables such that  $X \sim \mathcal{N}(0, 1)$  and  $Z \sim \chi_n^2$ . Hence the random variable

$$T = \frac{X}{\sqrt{\frac{Z}{n}}}.$$

is said to be following a student law.  $T \sim \mathcal{T}_n$

$$f(t) = \frac{1}{\sqrt{n\pi}} \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)} \left(1 + \frac{t^2}{n}\right)^{-\frac{n+1}{2}}.$$

# Chapter 2

## Estimators

Let  $\theta$  be a certain characteristic of a population  $P$  of  $N$  individuals, for example letting  $\theta$  be the expectation of a certain random variable  $X$  defined over the population. We take a sample of size  $n < N$  of the population to estimate the value of  $\theta$ .

Let  $Y_n$  be a function of the random variables  $X_1, X_2, \dots, X_n$ .  $Y_n$  is called an estimator of  $\theta$  if

$$\lim_{n \rightarrow +\infty} \mathbb{E}(Y_n) = \theta.$$

a consistent estimator if

$$\lim_{n \rightarrow +\infty} \text{Var}(Y_n) = 0.$$

and an unbiased estimator if

$$\mathbb{E}(Y_n) = \theta \quad \forall n \in \mathbb{N}^*.$$

the value  $y_n$  of  $Y_n$  computed from any observed sample is called point estimation of  $\theta$

### 2.1 Point estimation of the mean

Let  $X$  be a random variable defined over the population  $P$  of the expected value  $\mu$  and standard deviation  $\sigma$ . Consider a sample  $(X_1, X_2, \dots, X_n)$  of size  $n$ , randomly selected from  $P$  such that  $X_i$  are independent and follow the same law.

Consider the statistic  $\bar{X}_n = \frac{X_1 + X_2 + \dots + X_n}{n}$ , it is a random variable whose distribution is called the sample distribution of the mean.

$$\begin{aligned} \mathbb{E}(\bar{X}_n) &= \mu \\ \text{Var}(\bar{X}_n) &= \frac{\sigma^2}{n} \end{aligned}$$

Since  $\text{Var}(\bar{X}_n) \xrightarrow{n \rightarrow +\infty} 0$  then  $\bar{X}_n$  is a consistent unbiased estimator of the mean  $\mu$ .

**Note:-**

The standard deviation of  $\bar{X}_n$  is called standard error of the mean

$$\sigma(\bar{X}_n) = \frac{\sigma}{\sqrt{n}}.$$

Due to the central limit theorem, as the sample size gets larger and larger  $\bar{X}_n$  approaches a normal distribution  $\bar{X}_n \sim \mathcal{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$ .



## 2.2 Point estimator of the variance

### 2.2.1 Suppose $\mu$ is unknown

Consider the random variable  $S^2$  (estimator of  $\sigma^2$ )

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

The expectation of  $S^2$  can be proved to be

$$\mathbb{E}(S^2) = \frac{n-1}{n} \sigma^2.$$

Since  $\mathbb{E}(S^2) \xrightarrow{n \rightarrow +\infty} \sigma^2$  then  $S^2$  is a biased estimator of  $\sigma^2$ .

Consider the random variable  $S'^2$

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

Since  $\mathbb{E}(S'^2) = \sigma^2$  then  $S'^2$  is an unbiased estimator of  $\sigma^2$ .

Hence  $\sigma$  can be estimated by

$$S' = \sqrt{\frac{n}{n-1}} S.$$

and

$$\sigma(\bar{X}_n) = \frac{S}{\sqrt{n-1}}.$$

where

$\sigma^2$  variance of the population.

$S^2$  variance of the sample.

$\sigma^2(\bar{X}_n)$  variance of the distribution of the sample mean.

$S'^2$  corrected variance of the sample.

#### Note:-

It is better to estimate  $\sigma^2$  using  $S'^2$  than  $S^2$  since  $S^2$  is a biased estimator. However, if  $n$  (sample size) is big enough ( $\frac{n}{n-1} \approx 1$ ), then  $\sigma^2$  can be estimated by  $S^2$

### 2.2.2 Suppose $\mu$ is known

Consider the random variable  $Z^2$  (not the variance of the sample)

$$Z^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2.$$

Since  $\mathbb{E}(Z^2) = \sigma^2$  then  $Z^2$  is an unbiased estimator of  $\sigma^2$  thus the value  $z^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2$  is a point estimation of the variance  $\sigma^2$  of the population.

#### Note:-

If  $n > 0.05N$  and if the sample is selected without replacement then the value of the variance changes to become

$$\text{Var}(\bar{X}_n) = \left( \frac{N-n}{N-1} \right) \frac{\sigma^2}{n}.$$

and the standard error becomes

$$\sigma(\bar{X}_n) = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}.$$

If the variance of the population is not known then we can use  $S^2$  or  $Z^2$  to estimate  $\text{Var}(\bar{X}_n)$

$$\text{Var}(\bar{X}_n) = \left( \frac{N-n}{N-1} \right) \frac{S^2}{n-1}.$$

and the standard error with

$$\sigma(\bar{X}_n) = \frac{S}{\sqrt{n-1}} \sqrt{\frac{N-n}{N-1}}.$$

## 2.3 Point estimation of a proportion (percentage)

Consider a population  $P$  of individuals with a proportion  $p$  if individuals having a certain characteristic  $\theta$ . Let  $(a_1, a_2, \dots, a_n)$  be a sample randomly selected  $P$ . We define for each individual  $a_i$  the Bernoulli random variable  $X_i$  as follows

$$\begin{cases} X_i = 1 & \text{if } a_i \text{ has the characteristic } \theta \text{ with probability } p \\ X_i = 0 & \text{else} \end{cases}.$$

Let  $Y_n = \frac{X_1 + X_2 + \dots + X_n}{n} = \frac{1}{n} \sum_{i=1}^n X_i$ .  $Y_n$  is the random variable giving the proportion of individuals of the sample that have the characteristic  $\theta$ .

$$\begin{aligned} \Pr(X_i = 1) &= \frac{\text{number of individuals of the population having } \theta}{\text{total number of individuals}} = p \\ \Pr(X_i = 0) &= 1 - p \end{aligned}$$

Thus  $X_1 + X_2 + \dots + X_n \sim \text{Bin}(n, p)$

$$\begin{aligned} \mathbb{E}(X_1 + X_2 + \dots + X_n) &= np \\ \text{Var}(X_1 + X_2 + \dots + X_n) &= np(1 - p) \end{aligned}$$

$$\begin{aligned} \mathbb{E}(Y_n) &= p \\ \text{Var}(Y_n) &= \frac{p(1 - p)}{n} \end{aligned}$$

Hence  $Y_n$  is a consistent unbiased estimator of  $p$ . Therefore any observed value  $y_n$  of  $Y_n$  is a point estimator of  $P$ , meaning  $p$  is estimated by the proportion of the sample.

## 2.4 Confidence interval

### 2.4.1 Confidence interval for the mean

1. **Suppose that  $n \geq 30$ , the population is normally distributed, and  $\sigma$  is known**

Let  $X$  be a random variable defined over a population  $P$  of mean  $\mathbb{E}(X) = \mu$  and of variance  $\text{Var}(X) = \sigma^2$ .

Here we consider that  $\bar{X}_n \sim \mathcal{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$ . Hence  $\frac{\bar{X}_n - \mu}{\sigma_{\bar{X}_n}} \sim \mathcal{N}(0, 1)$ .

Given the probability  $\gamma$  (level of confidence), we can find  $t$  such that

$$\begin{aligned} \Pr\left(-t \leq \frac{\bar{X}_n - \mu}{\sigma_{\bar{X}_n}} \leq t\right) &= \gamma \\ \Pr(\bar{X}_n - t\sigma_{\bar{X}_n} \leq \mu \leq \bar{X}_n + t\sigma_{\bar{X}_n}) &= \gamma \end{aligned}$$

where  $\pi(t) = \frac{\gamma}{2}$ . Knowing  $\gamma$  we get  $t$ . Therefore a  $\gamma\%$  confidence interval for the mean  $\mu$  is given by

$$\text{IC}_\gamma(\mu) = [\bar{x}_n - t\sigma_{\bar{X}_n}, \bar{x}_n + t\sigma_{\bar{X}_n}].$$

where

$$\sigma_{\bar{X}_n} = \begin{cases} \frac{\sigma}{\sqrt{n}} & \text{if } \sigma \text{ is known} \\ \frac{S}{\sqrt{n-1}} & \text{if } \sigma \text{ is unknown (estimated by } S' = \sqrt{\frac{n}{n-1}}S) \end{cases}.$$

**2. Suppose that  $n < 30$ , the population is normally distributed, and  $\sigma$  is unknown:**

Using the table of student distributed knowing  $\gamma$ , we determine  $t$  such that

$$\Pr\left(\bar{X}_n - t\frac{S}{\sqrt{n-1}} \leq \mu \leq \bar{X}_n + t\frac{S}{\sqrt{n-1}}\right) = \gamma.$$

hence the confidence interval for the mean  $\mu$  is

$$\text{IC}_\gamma(\mu) = \left[\bar{X}_n - t\frac{S}{\sqrt{n-1}}, \bar{X}_n + t\frac{S}{\sqrt{n-1}}\right].$$

**Theorem 2.4.1**

- (a)  $\bar{X}_n$  and  $S^2$  are two independent random variance.
- (b) The random variable  $n\frac{S^2}{\sigma^2}$  follows a chi-square law with  $n - 1$  degrees of freedom.

**Theorem 2.4.2**

The random variable

$$\tilde{T} = \frac{\bar{X}_n - \mu}{\frac{S'}{\sqrt{n}}} = \frac{\bar{X}_n - \mu}{\frac{S}{\sqrt{n-1}}}.$$

follows a student law (t-distribution) with  $n - 1$  degrees of freedom

**3. Suppose that  $n < 30$ , the population is not normally distributed:**

In this case we cannot use the normal distributed nor the student distribution. However we can use Chebyshev's inequality.

$$\Pr(|\bar{X}_n - \mu| \leq \varepsilon) \geq 1 - \frac{\sigma_{\bar{X}_n}^2}{\varepsilon^2}.$$

Take  $\varepsilon = t\sigma_{\bar{X}_n}$

$$\Pr(\bar{X}_n - t\sigma_{\bar{X}_n} \leq \mu \leq \bar{X}_n + t\sigma_{\bar{X}_n}) \geq 1 - \frac{1}{t^2}.$$

Then we set  $1 - \frac{1}{t^2}$  equal to  $\gamma$  solve for  $t$  and find the interval as follows

$$\text{IC}_\gamma = [\bar{x}_n - t\sigma_{\bar{X}_n}, \bar{x}_n + t\sigma_{\bar{X}_n}].$$

**Note:-**

- if  $\sigma$  is known then  $\sigma_{\bar{X}_n} = \frac{\sigma}{\sqrt{n}}$
- if  $\sigma$  is unknown then we replace  $\sigma_{\bar{X}_n}$  by its point estimator  $\frac{S'}{\sqrt{n}} = \frac{S}{\sqrt{n-1}}$

### 2.4.2 Confidence interval for a proportion (percentage)

same setup as last time. If we assume this time that  $\text{Bin}(n, p) \approx \mathcal{N}(np, \sqrt{np(1-p)})$  if  $(n \geq 30, np, n(1-p) \geq 5)$  then we can say  $Y_n \sim \mathcal{N}\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$ . Knowing  $\gamma$  we can determine  $t$  such that

$$\Pr\left(-t \leq \frac{Y_n - p}{\sigma_{Y_n}} \leq t\right) = \gamma.$$

The confidence interval becomes

$$[y_n - t\sigma_{Y_n}, y_n + t\sigma_{Y_n}].$$

where  $\sigma_{Y_n} = \sqrt{\frac{p(1-p)}{n}}$  estimated by

$$\sqrt{\frac{n}{n-1}} \sqrt{\frac{y_n(1-y_n)}{n}} = \sqrt{\frac{y_n(1-y_n)}{n-1}}.$$

Therefore the confidence interval becomes

$$\text{IC}_\gamma(p) = \left[ y_n - t \sqrt{\frac{y_n(1-y_n)}{n-1}}, y_n + t \sqrt{\frac{y_n(1-y_n)}{n-1}} \right].$$

#### Note:-

If  $n \geq 100$  then  $\frac{n}{n-1} \approx 1$ , then the confidence interval is

$$\left[ y_n - t \sqrt{\frac{y_n(1-y_n)}{n}}, y_n + t \sqrt{\frac{y_n(1-y_n)}{n}} \right].$$

#### Note:-

If the sample is selected without replace and if  $n > 0.05N$  then we shall put a correcting factor  $\frac{N-n}{N-1}$  to  $\sigma_{Y_n} = \sqrt{\frac{p(1-p)}{n}}$ , thus the confidence interval for proportion  $p$  becomes

$$\left[ y_n - t \sqrt{\frac{N-n}{N-1}} \sqrt{\frac{y_n(1-y_n)}{n-1}}, y_n + t \sqrt{\frac{N-n}{N-1}} \sqrt{\frac{y_n(1-y_n)}{n-1}} \right].$$

### 2.4.3 Confidence interval for the variance

Assume  $X \sim \mathcal{N}(\mu, \sigma)$  and  $X_1, X_2, \dots, X_n$   $n$  independent random variables and identically distributed as  $X$ . We set the variables

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

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

$$S'^2 = \frac{n}{n-1} S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$$

$$Z^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2$$

we have

$$\begin{aligned}\bar{X}_n &\sim \mathcal{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\right) \\ n \frac{S^2}{\sigma^2} &\sim \chi_{n-1}^2 \\ n \frac{Z^2}{\sigma^2} &\sim \chi_n^2\end{aligned}$$

1. Suppose  $\mu$  is unknown

Since  $n \frac{S^2}{\sigma^2} \sim \chi_{n-1}^2$ , then we determine the values  $v_{\alpha/2}$  and  $v_{1-\alpha/2}$  from the chi-square table such that

$$\Pr\left(v_{\alpha/2} \leq \frac{nS^2}{\sigma^2} \leq v_{1-\alpha/2}\right) = \gamma = 1 - \alpha.$$

therefore a confidence interval of level  $\gamma$  (risk  $\alpha$ ) is given by

$$IC_{\gamma}(\sigma^2) = \left[ \frac{nS^2}{v_{1-\alpha/2}}, \frac{nS^2}{v_{\alpha/2}} \right] = \left[ \frac{(n-1)S'^2}{v_{1-\alpha/2}}, \frac{(n-1)S'^2}{v_{\alpha/2}} \right].$$

2. Suppose  $\mu$  is known

From the chi-square table, we determine the values of the quantities  $v_{\alpha/2}$  and  $v_{1-\alpha/2}$  for the law  $\chi_n^2$  such that

$$\Pr\left(v_{\alpha/2} \leq \frac{nZ^2}{\sigma^2} \leq v_{1-\alpha/2}\right) = \gamma.$$

Therefore the confidence interval of level  $\gamma$  is given by

$$IC_{\gamma}(\sigma^2) = \left[ \frac{nz^2}{v_{1-\alpha/2}}, \frac{nz^2}{v_{\alpha/2}} \right].$$

## 2.5 Mean Squared Error

Consider the estimator  $y_n = f(x_1, x_2, \dots, x_n)$  of  $\theta$ . The bias of  $y_n$  relative to  $\theta$  is

$$\text{Bias}(y_n) = \mathbb{E}(y_n) - \theta.$$

The mean squared error of  $y_n$  with respect to  $\theta$  is

$$\text{MSE}(y_n) = \mathbb{E}[(y_n - \theta)^2].$$

It can also be shown that

$$\text{MSE}(y_n) = \text{Var}(y_n) + \text{Bias}(y_n)^2.$$

If  $y_n$  is an unbiased estimator of  $\theta$  then  $\text{Bias}(y_n) = \mathbb{E}(y_n) - \theta = 0 \Rightarrow$

$$\text{MSE}(y_n) = \text{Var}(y_n).$$

Assume  $y_n$  and  $z_n$  are two estimators of the same parameter  $\theta$ . We say  $y_n$  is more efficient than  $z_n$  if

$$\text{MSE}(y_n) < \text{MSE}(z_n).$$

Assume  $y_n$  and  $z_n$  are two unbiased estimators of  $\theta$ , then  $y_n$  is more efficient than  $z_n$  if and only if  $\text{Var}(y_n) < \text{Var}(z_n)$

# Chapter 3

## Hypothesis Testing

### 3.1 Introduction

In hypothesis testing, we have a null-hypothesis  $H_0$  on a sample space  $\Omega$  and an alternative hypothesis  $H_1$  on  $\Omega$ . We want to test  $H_0$  against  $H_1$ .

To do so we consider a random sample size  $n$  and calculate the probability that  $H_0$  is within a certain *significance level* and deduce if we can accept  $H_0$ .

As an example consider  $H_0 = \mu \neq m$ , then  $H_1$  is

$$\begin{aligned} H_1 = \mu > m \\ H_1 = \mu < m \end{aligned} \Bigg\} \text{One sided test}$$
$$H_1 = \mu \neq m \Bigg\} \text{Two sided test}$$

Type I error

$$\alpha = \Pr(\text{Type I error}) = \Pr(\text{Reject } H_0 | H_0 \text{ is true}).$$

Type II error

$$\beta = \Pr(\text{Type II error}) = \Pr(\text{Accept } H_0 | H_0 \text{ is false}).$$

Power of the test

$$\pi = \begin{cases} \alpha & \text{if } H_0 \text{ is true} \\ 1 - \beta & \text{if } H_1 \text{ is true} \end{cases}$$

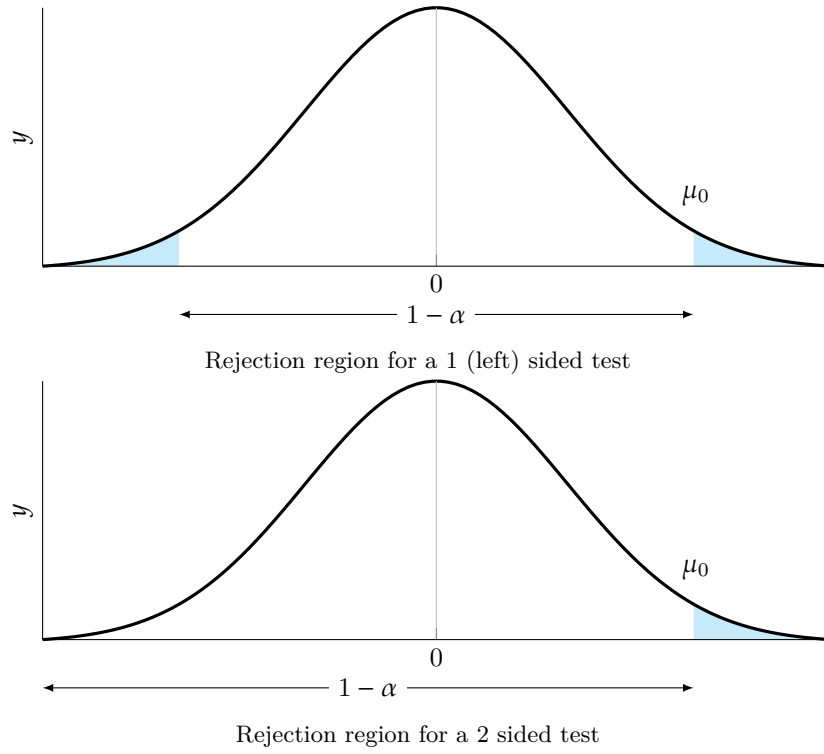
### 3.2 Comparison between a mean and a reference value

#### 3.2.1 Two sided test

$$H_0 : \mu_0 = \mu$$

$$H_1 : \mu_0 \neq \mu$$

	$H_0$ is true	$H_0$ is false
Accept $H_0$	Correct decision	Type II error
Reject $H_0$	Type I error	Correct decision



- Assume  $\sigma$  is known. Test statistic  $T = \frac{\bar{x}_n - \mu_0}{\frac{\sigma}{\sqrt{n}}}$ .

Rule of rejection

$$|T| > t \quad \begin{cases} \Pr(|T| > t) = \alpha \\ T \sim \mathcal{N}(0, 1) \end{cases}.$$

- Assume  $\sigma$  is unknown. Test statistic  $T = \frac{\bar{x}_n - \mu_0}{\frac{S'}{\sqrt{n}}} = \frac{\bar{x}_n - \mu_0}{\frac{S}{\sqrt{n-1}}}$ .

Rule of rejection

$$|T| > t \quad \begin{cases} \Pr(|T| > t) = \alpha \\ T \sim \mathcal{T}_{n-1} \end{cases}.$$

### 3.2.2 One sided test

**Left sided test**

$$\begin{aligned} H_0 : \mu_0 &= \mu \\ H_1 : \mu_0 &< \mu \end{aligned}$$

- Assume  $\sigma$  is known. Test statistic  $T = \frac{\bar{x}_n - \mu_0}{\frac{\sigma}{\sqrt{n}}}$ .

Rule of rejection

$$T > t \quad \begin{cases} \Pr(T > t) = \alpha \\ T \sim \mathcal{N}(0, 1) \end{cases}.$$

- Assume  $\sigma$  is unknown. Test statistic  $T = \frac{\bar{x}_n - \mu_0}{\frac{S'}{\sqrt{n}}} = \frac{\bar{x}_n - \mu_0}{\frac{S}{\sqrt{n-1}}}$ .

Rule of rejection

$$T > t \quad \begin{cases} \Pr(T > t) = \alpha \\ T \sim \mathcal{T}_{n-1} \end{cases}.$$

### Right sided test

Same as left sided test but with  $T < -t$ .

## 3.3 Comparison between a proportion and a reference value

### 3.3.1 Two sided test

$$\begin{aligned} H_0 : p_0 &= p \\ H_1 : p_0 &\neq p \end{aligned}$$

Test statistic  $X \sim \text{Bin}(n, p)$ .

Rule of rejection

$$\begin{cases} \Pr(X > b_{n,p_0,1-\alpha/2}) = \Pr(X > b_{n,p_0,\alpha/2}) = \frac{\alpha}{2} \\ X \sim \text{Bin}(n, p_0) \end{cases}.$$

Acceptance region  $[b_{n,p_0,\alpha/2}, b_{n,p_0,1-\alpha/2}]$ .

$$\Pr(X \in [b_{n,p_0,\alpha/2}, b_{n,p_0,1-\alpha/2}]) = 1 - \alpha.$$

However if  $n \geq 30$  we can use the normal approximation.

$$\text{Test statistic } T = \frac{X - np_0}{\sqrt{np_0(1-p_0)}}.$$

Rule of rejection

$$|T| > t \quad \begin{cases} \Pr(|T| > t) = \alpha \\ T \sim \mathcal{N}(0, 1) \end{cases}.$$

### 3.3.2 One sided test

$$\begin{aligned} H_0 : p_0 &= p \\ H_1 : p_0 &< p \end{aligned}$$

Test statistic  $X \sim \text{Bin}(n, p)$ .

Rule of rejection

$$\Pr(X \geq b_{n,p_0,1-\alpha}) = \alpha.$$

Acceptance region  $[-\infty, b_{n,p_0,1-\alpha}]$ .

$$\text{Normal approximation Test statistic } T = \frac{X - np_0}{\sqrt{np_0(1-p_0)}}.$$

Rule of rejection

$$T > t \quad \begin{cases} \Pr(T > t) = \alpha \\ T \sim \mathcal{N}(0, 1) \end{cases}.$$



## 3.4 Comparison between a variance and a reference value

### 3.4.1 Two sided test

$$\begin{aligned}H_0 : \sigma_0^2 &= \sigma^2 \\H_1 : \sigma_0^2 &\neq \sigma^2\end{aligned}$$

Test statistic

$$T = \begin{cases} n \frac{Z^2}{\sigma_0^2} \sim \chi_n^2 & \text{if } \mu \text{ is known} \\ n \frac{S^2}{\sigma_0^2} \sim \chi_{n-1}^2 & \text{if } \mu \text{ is unknown} \end{cases}.$$

We reject  $H_0$  if  $T \notin [v_{\alpha/2}, v_{1-\alpha/2}]$ .

$$\Pr(T \notin [v_{\alpha/2}, v_{1-\alpha/2}]) = \frac{\alpha}{2}.$$

### 3.4.2 One sided test

$$\begin{aligned}H_0 : \sigma^2 &= \sigma_0^2 \\H_1 : \sigma^2 &> \sigma_0^2\end{aligned}$$

Test statistic

$$T = \begin{cases} n \frac{Z^2}{\sigma_0^2} \sim \chi_n^2 & \text{if } \mu \text{ is known} \\ n \frac{S^2}{\sigma_0^2} \sim \chi_{n-1}^2 & \text{if } \mu \text{ is unknown} \end{cases}.$$

We reject  $H_0$  if  $T > v_{1-\alpha}$ .

$$\Pr(T > v_{1-\alpha}) = \alpha.$$

## 3.5 Critical Probability

$$P_c = \begin{cases} \Pr(|T| \geq |t_0|/\theta = \theta_0) & \text{for one sided tests} \\ \Pr(T \geq t_0/\theta = \theta_0) & \text{for } H_1 : \theta > \theta_0 \\ \Pr(T \leq t_0/\theta = \theta_0) & \text{for } H_1 : \theta < \theta_0 \end{cases}.$$

## 3.6 Comparison between two means

### 3.6.1 Two sided test

$$\begin{aligned}H_0 : \mu_1 &= \mu_2 \\H_1 : \mu_1 &\neq \mu_2\end{aligned}$$

Test statistic  $T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}.$

Rule of rejection

$$\begin{cases} \Pr(|T| > t) = \alpha \\ T \sim \mathcal{N}(0, 1) \end{cases}.$$

**Note:-**

$$\begin{aligned}\bar{X}_1 - \bar{X}_2 &\sim \mathcal{N}\left(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right) \\ \mathbb{E}(\bar{X}_1 - \bar{X}_2) &= \mu_1 - \mu_2 \\ \text{Var}(\bar{X}_1 - \bar{X}_2) &= \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\end{aligned}$$

### 3.6.2 One sided test

$$H_0 : \mu_1 = \mu_2$$

$$H_1 : \mu_1 > \mu_2$$

Test statistic same as before.

Rule of rejection

$$\begin{cases} \Pr(T > t) = \alpha \\ T \sim \mathcal{N}(0, 1) \end{cases}.$$

**Note:-**

If  $\sigma_1$  and  $\sigma_2$  are unknown, and  $\sigma_1 = \sigma_2$  we perform all the same preceding tests but this a new test statistic

$$T = \frac{\bar{X}_1 - \bar{X}_2}{S \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim \mathcal{T}_{n_1+n_2-2}.$$

if  $\sigma_1 \neq \sigma_2$  we use the preceding tests only if  $n_1$  and  $n_2$  are sufficiently large ( $> 30$ ).

## 3.7 Comparison between 2 proportions

Consider the following 2 random variables

$$P = \frac{X + Y}{n_1 + n_2} \quad S_d^2 = P(1 - P) \left( \frac{1}{n_1} + \frac{1}{n_2} \right).$$

### 3.7.1 Two sided test

$$H_0 : p_1 = p_2$$

$$H_1 : p_1 \neq p_2$$

$$\text{Test statistic } T = \frac{\frac{X}{n_1} - \frac{Y}{n_2}}{S_d}.$$

Rule of rejection

$$\begin{cases} \Pr(|T| > t) = \alpha \\ T \sim \mathcal{N}(0, 1) \end{cases}.$$

### 3.7.2 One sided test

$$H_0 : p_1 = p_2$$

$$H_1 : p_1 > p_2$$

Same job as before.

## 3.8 Comparison between two variances

### 3.8.1 Two sided test

$$H_0 : \sigma_1^2 = \sigma_2^2$$

$$H_1 : \sigma_1^2 \neq \sigma_2^2$$

Test statistic  $F = \frac{S_1'^2}{S_2'^2}$ .

$$F \sim \mathcal{F}_{n_1-1, n_2-1}.$$

where  $\mathcal{F}$  is the Fisher distribution.

Rule of rejection

$$\begin{cases} \Pr(F < f_{n_1-1, n_2-1, 1-\alpha/2}) = \Pr(F > f_{n_1-1, n_2-1, \alpha/2}) = \frac{\alpha}{2} \\ F \sim \mathcal{F}_{n_1-1, n_2-1} \end{cases}.$$

### 3.8.2 One sided test

$$H_0 : \sigma_1^2 = \sigma_2^2$$

$$H_1 : \sigma_1^2 < \sigma_2^2$$

Rule of rejection

$$\begin{cases} \Pr(F > f_{n_1-1, n_2-1, 1-\alpha}) = \frac{\alpha}{2} \\ F \sim \mathcal{F}_{n_1-1, n_2-1} \end{cases}.$$

#### Note:-

Chi-squared test

$H_0$  : the population follows law M

$H_1$  : the population does not follow law M

Test statistic

$$Y = \sum_{i=1}^k \frac{(n_i - n_{p_i})^2}{n_{p_i}} \sim \chi^2_{k-1}.$$

Rule of rejection

$$\begin{cases} \Pr(Y > t) = \alpha \\ Y \sim \chi^2_{k-1} \end{cases}.$$

# Chapter 4

## Regression

$$\text{cov}(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y).$$

### 4.1 Linear Regression

The objective of linear regression is to find the best linear approximation  $D : \alpha x + \beta$  of a set of data points.

The least squares method consists in minimizing the sum  $S(\alpha, \beta)$  of the squares of the residuals.

$$S(\alpha, \beta) = \frac{1}{n} \sum_{i=1}^p \sum_{j=1}^q n_{ij} (y_j - \alpha x_i - \beta)^2.$$

It can be shown that the couple  $(a, b)$  that minimizes  $S$  is given by

$$\begin{aligned} a &= \frac{\text{cov}(X, Y)}{\text{Var}(X)} \\ b &= \mathbb{E}(Y) - a\mathbb{E}(X) \end{aligned}$$

$$S_{\min} = \text{Var}(Y) \cdot \left[ 1 - \frac{\text{cov}(X, Y)^2}{\text{Var}(X) \cdot \text{Var}(Y)} \right].$$

Similarly, we can construct a regression line  $D' : a'y + b'$  of  $y$  on  $x$  where  $a'$  and  $b'$  are given by

$$\begin{aligned} a' &= \frac{\text{cov}(X, Y)}{\text{Var}(Y)} \\ b' &= \mathbb{E}(X) - a'\mathbb{E}(Y) \end{aligned}$$

### 4.2 Coefficient of linear correlation

$$\rho(X, Y) = \frac{\text{cov}(X, Y)}{\sqrt{\text{Var}(X) \text{Var}(Y)}} = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y}.$$

Some properties of  $\rho(X, Y)$

$$\rho(X, Y)^2 \leq 1$$

$$\rho(X, Y)^2 = a \cdot a'$$

$$\rho(X, Y) = \begin{cases} \sqrt{a \cdot a'} & \text{if } a, a' > 0 \text{ (cov}(X, Y) > 0) \\ -\sqrt{a \cdot a'} & \text{if } a, a' < 0 \text{ (cov}(X, Y) < 0) \\ 0 & \text{if } a = a' = 0 \text{ (cov}(X, Y) = 0) \end{cases}$$

$$\rho(X, Y) \begin{cases} > 0 & \text{if } X \text{ and } Y \text{ are positively correlated} \\ < 0 & \text{if } X \text{ and } Y \text{ are negatively correlated} \\ = 0 & \text{if } X \text{ and } Y \text{ are not correlated} \end{cases}$$

$$\rho(X, Y)^2 = 1 \text{ if and only if } X \text{ and } Y \text{ are perfectly correlated}$$