Statistics: Sampling Statistics

S. Maset

Dipartimento di Matematica e Geoscienze, Università di Trieste

PEM 2018-2019

Outline

1. Introduction

2. Sampling statistics

3. Sample mean

4. Estimating proportions

5. Sample variance
   - Sample Variance in case of the normal distribution

6. Normal data

Introduction

- One of the key concerns of Statistics is drawing conclusions from observed data (**Inferential Statistics**).

  These data usually are obtained by a representative sample of individuals of a population and then they are used to draw conclusions about the entire population.

  In the following, we assume that we have to draw conclusions about some numerical quantity associated to each individual of the population. For example, the height of the individuals in a human population.

We give a mathematical structure to this problem in the following way.

Consider the experiment, where an individual is randomly selected in the population and let X be the random variable that gives the numerical quantity of interest associated to this individual.

The sample space of this experiment is the set P of the individuals in the population.

We assume that the distribution of X belongs to a family of distributions. Any distribution in this family is determined by parameters.

The particular parameters that determine the distribution of X are unknown and they have to be estimated.

This is the mathematical form of an Inferential Statistics problem.

The experiment of the random selection of an individual from the population is like to select a card from an urn and then read the number written in the card.

Observe that the numbers in the cards are written independently of the random selection experiment.

Indeed, we are interested to have information on the manner with which these number are written in the cards.

For this reason, now we adopt another point of view.

- The numerical quantities of interest associated to each individual in the population are considered as IID random variables $Y_1, Y_2, \ldots, Y_m$, where m is the number of individuals of the population.

  As an example, we suppose $Y_1, Y_2, \ldots, Y_m$ the heights of the individuals in a human population.

  The experiment relevant to these random variables $Y_1, Y_2, \ldots, Y_m$ is the formation process as adults of the individuals of the population.

  We have seen that it reasonable to assume that the common distribution of $Y_1, Y_2, \ldots, Y_m$ belongs to the family of the normal distributions.

  The parameters of a normal distribution are the mean μ and the standard deviation σ. The particular μ and σ that determine the distribution of $Y_1, Y_2, \ldots, Y_m$ are unknown and have to be estimated.

Now, let X be the random variable height of an individual randomly selected in the population.

Observe that the experiment relevant to X (random selection of an individual) is different from the experiment relevant to $Y_1, Y_2, \ldots, Y_m$ (formation process of all individuals).

Since the population is finite, X is a discrete random variable. The range $X(\Omega)$ of X is the set of the heights of the individuals in the population, which are particular realizations of the random variables $Y_1, Y_2, \ldots, Y_m$.

Let $x \in \mathbb{R}$. Since the individual is randomly selected, we have

$$F_X(x) = \mathbb{P}(X \leq x) = \frac{\text{number of individuals with height} \leq \text{x}}{m}.$$

By thinking the process of formation of all individuals in the population as repetitions of the process of formation of one individual we have, since m is very large and the Strong Law of Large Numbers (or the frequentist interpretation) holds, that

$$\begin{aligned} F_X(x) &= \mathbb{P}(X \leq x) = \frac{\text{number of individuals with height} \leq \text{x}}{m} \\ &\approx \mathbb{P}(Y \leq x) = F_Y(x) \end{aligned}$$

where Y is the random variable height of one individual in the population: $Y_1, Y_2, \ldots, Y_m$ are IID random variables distributed as Y.

So, the distribution function of X is very close to the common distribution function of $Y_1, Y_2, \ldots, Y_m$.

Therefore, we can assume that X has the common distribution of $Y_1, Y_2, \ldots, Y_m$ by transforming X in a continuous random variable.

Of course, this is valid in general not only for the specific example of the heights. We can assume that:

- the random variable X giving the numerical quantity of interest associated to a randomly selected individual in the population has the same distribution of the IID random variables $Y_1, Y_2, \ldots, Y_m$ giving the numerical quantity for the m individuals in the population.

Therefore, when we are estimating the parameters that determine the distribution of X, we are also estimating the parameters that determine the common distribution of $Y_1, Y_2, \ldots, Y_m$.

Sampling statistics

- Let X be the random variable that gives the numerical quantity of interest associated to the individual randomly selected in the population.

  How to determine the unknown parameters of the distribution of X?

  We use another experiment, where we randomly select n times an individual in the population and each selection is independent of the others. The selected elements are called the **sample**.

  The sample space of this other experiment is the set P^n. Therefore, each selected element is not put aside but it is put back into play and can be chosen for the successive selections.

  Now, we observe, for each individual in the sample, the numerical quantity of interest. These observed numerical quantities are the **data**.

By using the data, we try to estimate the unknown parameters by means of **sampling statistics**.

In the example of the height, we randomly select from the population n individuals. The n selected individuals are the sample.

Then, we measure the heights of the individuals in the sample. The n measured heights are the data.

Finally, starting from the data, the unknown mean and standard deviation are estimated by means of suitable sampling statistics.

Example. Consider a manufacturer producing a new type of battery to be used in a particular electric-powered car. We assume that each battery will last for a random number of kilometers with an unknown normal distribution.

To estimate mean and the standard deviation of this distribution, the manufacturer randomly selects from the produced batteries some batteries and test them on the road. These selected batteries are the sample.

We record the number of kilometers of use of each selected battery. These numbers are the data.

Finally, starting from the data, the unknown mean and standard deviation of the normal distribution relevant to the duration of the new type of battery are estimated by means of suitable sampling statistics.

Exercise. In this example, what is the experiment relevant to the random variables $Y_1, Y_2, \ldots, Y_m$*?*

- In the following definition, $X_1, X_2, \ldots, X_n$ are the IID random variables distributed as X giving the numerical quantity of interest for the individuals $1, 2, \ldots, n$ selected in the sample.

  Observe that in order to have $X_1, X_2, \ldots, X_n$ IID random variables distributed as X is necessary that each selected element is put back into the play for the next selections.

  However, the probability of selecting two times a same individual is very small.

  Exercise. Compute this probability.

Definition

Let $X_1, X_2, \ldots, X_n$ be IID random variables, all with a same distribution determined by unknown parameters. The $n-$tuple $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ is called a **Sample** of size n from that distribution.

If $x_1^{\text{obs}}, x_2^{\text{obs}}, \ldots, x_n^{\text{obs}}$ are the observed values of $X_1, X_2, \ldots, X_n$, then the $n-$tuple $\boldsymbol{x}^{\text{obs}} = \left(x_1^{\text{obs}}, x_2^{\text{obs}}, \ldots, x_n^{\text{obs}}\right)$ is called a **sample** of size n from that distribution.

A **sampling statistic** based on a sample of size n is a function $\tau : \mathbb{R}^n \to \mathbb{R}$: $\tau\left(\boldsymbol{x}^{\text{obs}}\right)$ is the number provided by the statistic.

Let θ be one of the unknown parameters of the distribution. An **estimator** of θ (based on a sample of size n) is a suitable sampling statistic τ (based on a sample of size n) providing an estimate of θ: the number $\tau\left(\boldsymbol{x}^{\text{obs}}\right)$ is the estimate of θ.

At each sampling statistic τ is associated the Sampling Statistic random variable $\tau(X_1, X_2, \ldots, X_n)$.

We use the rule that a sampling statistic is denoted by a name with lower-case initial letters, whereas the corresponding Sampling Statistic random variable is denoted by the same name with upper-case initial letters.

For example, now we will see the sampling statistic called "sample mean", whose corresponding Sampling Statistic is called "Sample Mean".

Observe that the same rule is used when we call $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ "Sample" and $\boldsymbol{x}^{\mathrm{obs}} = \left(x_1^{\mathrm{obs}}, x_2^{\mathrm{obs}}, \ldots, x_n^{\mathrm{obs}}\right)$ "sample".

Note that the names "Sample" and "sample" are not exactly correct, because they denote the numerical quantities relevant to the individuals selected in the sample, not the individuals. Better names are "Data" and "data", but we do not use these names.

Sample mean

- Let us consider a Sample $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ from a distribution of mean μ and standard deviation σ.

  The sampling statistic

  $$\tau(\boldsymbol{x}) = \overline{x} := \frac{x_1 + x_2 + \cdots + x_n}{n}, \ \boldsymbol{x} \in \mathbb{R}^n,$$

  is called the **sample mean.**

  The Sampling Statistic

  $$\tau(\boldsymbol{X}) = \overline{X} := \frac{X_1 + X_2 + \cdots + X_n}{n}$$

  is then called the **Sample Mean.**

  Observe that $\overline{x}^{\mathrm{obs}} := \tau\left(\boldsymbol{x}^{\mathrm{obs}}\right) = \overline{x^{\mathrm{obs}}}$ is the mean (seen in descriptive statistics) of the data $\boldsymbol{x}^{\mathrm{obs}}$.

- Now, we determine mean and standard deviation of the Sample Mean.

We have

$$\begin{aligned}\mathbb{E}\left(\overline{X}\right) &= \mathbb{E}\left(\frac{X_1+X_2+\cdots+X_n}{n}\right)=\frac{\mathbb{E}\left(X_1+X_2+\cdots+X_n\right)}{n}\\ &= \frac{\mathbb{E}\left(X_1\right)+\mathbb{E}\left(X_2\right)+\cdots+\mathbb{E}\left(X_n\right)}{n}=\frac{\mu+\mu+\cdots+\mu}{n}\\ &= \frac{n\mu}{n}=\mu.\end{aligned}$$

Since $X_1, X_2, \ldots, X_n$ are independent, we have

$$\begin{aligned}\mathrm{Var}\left(\overline{X}\right) &= \mathrm{Var}\left(\frac{X_1+X_2+\cdots+X_n}{n}\right)=\frac{\mathrm{Var}\left(X_1+X_2+\cdots+X_n\right)}{n^2}\\ &= \frac{\mathrm{Var}\left(X_1\right)+\mathrm{Var}\left(X_2\right)+\cdots+\mathrm{Var}\left(X_n\right)}{n^2}\\ &= \frac{\sigma^2+\sigma^2+\cdots+\sigma^2}{n^2}=\frac{n\sigma^2}{n^2}=\frac{\sigma^2}{n}.\end{aligned}$$

Finally, we have

$$\mathrm{SD}\left(\overline{X}\right) = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}.$$

By summarizing:

- the mean of the Sample Mean is the mean μ of the distribution from which the Sample $\boldsymbol{X} = (X_1, X_2, \dots, X_n)$ is taken;
- the standard deviation of the Sample Mean is smaller by the factor $\frac{1}{\sqrt{n}}$ than the standard deviation σ of the distribution from which the Sample $\boldsymbol{X}$ is taken.

We can conclude that the Sample Mean is centered around μ and its spread becomes more and more reduced as the sample size n increases.

Indeed, we known from the Central Limit Theorem that, for n large, the Sample Mean is approximately distributed as $\mathrm{N}(\mu, (\frac{\sigma}{\sqrt{n}})^2)$ and so

$$\mathbb{P}\left(|\overline{X} - \mu| \leq 2\frac{\sigma}{\sqrt{n}}\right) \approx 95\% \text{ and } \mathbb{P}\left(|\overline{X} - \mu| \leq 3\frac{\sigma}{\sqrt{n}}\right) \approx 99.7\%.$$

We also know that if the distribution from which the sample is taken is normal, then , for any n, the Sample Mean is exactly distributed as $\mathrm{N}(\mu, (\frac{\sigma}{\sqrt{n}})^2)$ and so

$$\mathbb{P}\left(|\overline{X} - \mu| \leq 2\frac{\sigma}{\sqrt{n}}\right) = 95\% \text{ and } \mathbb{P}\left(|\overline{X} - \mu| \leq 3\frac{\sigma}{\sqrt{n}}\right) = 99.7\%.$$

Therefore, the sample mean can be considered as an estimator of the mean μ, if μ is one of the unknown parameters.

Examples.

- *In the case of the height of an individual of a human population, an estimate of the unknown mean μ of the normal distribution is given by the mean of the observed heights of the individuals in the sample.*
- *In the case of the new type of battery for electric-powered cars, an estimate of the unknown mean μ of the normal distribution is given by the mean of the observed durations of the batteries in the sample.*

Estimating proportions

- Before considering the estimating proportions problem, we introduce the notion of a Bernoulli distribution.

Definition

A discrete random variable Y is said to have the **Bernoulli distribution** $\mathrm{Bernoulli}\,(p)$, where $p \in [0,1]$, if $Y\,(\Omega) = \{0,1\}$ and

$$\mathbb{P}\,(Y=1)=p \quad \text{and} \quad \mathbb{P}\,(Y=0)=1-p=:q.$$

Let Y be a random variable with distribution $\mathrm{Bernoulli}(p)$, $p \in [0,1]$. The mean of Y is

$$\mu = 1\cdot\mathbb{P}(Y=1)+0\cdot\mathbb{P}(Y=0)=p$$

and the variance is

$$\sigma^2=\mathbb{E}(Y^2)-\mu^2=1^2\cdot\mathbb{P}(Y=1)+0^2\cdot\mathbb{P}(Y=0)-p^2=p-p^2=pq.$$

Example of random variables with Bernoulli distribution: given a Bernoulli process of lenght n with outcomes α and β at any trial, the independent random variables

$$X_i = \begin{cases} 1 \text{ if the outcome of the } i-\text{th trial is } \alpha \\ 0 \text{ if the outcome of the } i-\text{th trial is } \beta \end{cases}, \ i \in \{1, \ldots, n\},$$

have distribution $\mathrm{Bernoulli}(p)$*, where p is the probability of the outcome* α *at any trial.*

- Here is the problem of the estimating proportions.

  Consider the situation where we have a population whose individuals have or have not a certain characteristic.

  An example could be the left handedness in a population of human individuals.

  Another example could be the support to a given party in a population of voters for an election.

  Let $p \in [0, 1]$ be the proportion of the individuals in the population that have this characteristic. Suppose that the proportion p is unknown and we are interested in estimating it.

  Now, we give to this problem the form of an Inferential Statistics problem.

Consider the experiment where an individual is randomly chosen from the population P with m individuals.

The random variable

$$X(\omega) = \begin{cases} 1 \text{ if the individual } \omega \text{ has the characteristic} \\ 0 \text{ if the individual } \omega \text{ has not the characteristic} \end{cases}, \ \omega \in P,$$

has distribution $\mathrm{Bernoulli}(p)$. In fact

$$\mathbb{P}(X = 1) = \frac{\text{number of individuals in the population with the characteristic}}{m} = p.$$

Exercise. What are the random variables $Y_1, Y_2, \ldots, Y_m$ in this situation of estimating proportions? What is the experiment relevant to the random variables $Y_1, Y_2, \ldots, Y_m$ in case of left handedness and in case of supporting a given party in the election.

Now, in order to estimate p we select n times an individual from the population with each selection independent of the others.

We obtain a Sample $X_1, X_2, \ldots, X_n$ from the distribution $\mathrm{Bernoulli}\,(p)$, where

$$X_k = \begin{cases} 1 \text{ if the } k - \text{th selected individual has the characteristic} \\ 0 \text{ if the } k - \text{th selected individual has not the characteristic} \end{cases}$$

$$k \in \{1, 2, \ldots, n\}\,.$$

Since the mean of the distribution $\mathrm{Bernoulli}\,(p)$ is p, the sample mean is an estimator of the unknown parameter p.

So, we estimate p, the proportion of the individuals in the population with the characteristic, by the Sample Mean

$$\begin{aligned} \overline{X} &= \frac{X_1 + X_2 + \cdots + X_n}{n} \\ &= \frac{\text{"number of individuals in the sample with the characteristic"}}{n}, \end{aligned}$$

i.e. by the proportion of the individuals in the sample with the characteristic.

Exercise. Consider a shipment of $10K$ pieces. Explain how to estimate the proportion of defective pieces in this shipment. Assume that these pieces are produced by a machine. Explain why this is also an estimate of the probability that the machine will produce a defective piece.

- By considering the selection of the *n* individuals as a Bernoulli process with outcomes 1 if the individual has the characteristic and 0 otherwise, then

$$S_n = X_1 + X_2 + \cdots + X_n$$

the number of the selected individuals with the characteristic has distribution $\mathrm{Binomial}(n, p)$.

We known that, for *np* and *nq* larger than 5, the normal approximation

$$\mathrm{N}(n\mu, (\sqrt{n}\sigma)^2) = \mathrm{N}(np, (\sqrt{n} \cdot \sqrt{pq})^2)$$

of S_n is quite good.

So, for *np* and *nq* larger than 5, also the normal approximation

$$\mathrm{N}\left(\mu, \left(\frac{\sigma}{\sqrt{n}}\right)^2\right) = \mathrm{N}\left(p, \left(\frac{\sqrt{pq}}{\sqrt{n}}\right)^2\right)$$

of the Sample Mean $\overline{X} = \frac{S_n}{n}$ is quite good and we have

$$\mathbb{P}\left(|\overline{X} - p| \leq 2\frac{\sqrt{pq}}{\sqrt{n}}\right) \approx 95\% \text{ and } \mathbb{P}\left(|\overline{X} - p| \leq 3\frac{\sqrt{pq}}{\sqrt{n}}\right) \approx 99.7\%.$$

Sample variance

- Let $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ be a Sample from a distribution of mean μ and standard deviation σ.

  The sampling statistic

  $$\tau(\boldsymbol{x}) = s_x^2 := \frac{\sum_{k=1}^{n} (x_k - \overline{x})^2}{n-1}, \; \boldsymbol{x} \in \mathbb{R}^n,$$

  is called the **sample variance.**

  The Sampling Statistic

  $$\tau(\boldsymbol{X}) = S_X^2 := \frac{\sum_{k=1}^{n} \left(X_k - \overline{X}\right)^2}{n-1}$$

  is then called the **Sample Variance.**

  Observe that $\left(s^{\mathrm{obs}}\right)^2 := \tau\left(\boldsymbol{x}^{\mathrm{obs}}\right) = s^2_{x^{\mathrm{obs}}}$ is the variance (seen in descriptive statistics) of the data $\boldsymbol{x}^{\mathrm{obs}}$.

Of course, the sampling statistic

$$\tau\left(\boldsymbol{x}\right) = s_x = \sqrt{s_x^2} = \sqrt{\frac{\sum_{k=1}^{n}\left(x_k - \overline{x}\right)^2}{n-1}},\ \boldsymbol{x} \in \mathbb{R}^n,$$

is called the **sample standard deviation** and the Sampling Statistic

$$\tau\left(\boldsymbol{X}\right) = S_X = \sqrt{S_X^2} = \sqrt{\frac{\sum_{k=1}^{n}\left(X_k - \overline{X}\right)^2}{n-1}}$$

is called the **Sample Standard Deviation.**

Observe that $s^{\text{obs}} := \tau\left(\boldsymbol{x}^{\text{obs}}\right) = s_{x^{\text{obs}}}$ is the standard deviation (seen in descriptive statistics) of the data $\boldsymbol{x}^{\text{obs}}$.

- Now, we determine the mean of the Sample Variance.

  We have

$$\sum_{k=1}^{n}\left(X_k-\overline{X}\right)^2=\sum_{k=1}^{n}\left(X_k-\mu-\left(\overline{X}-\mu\right)\right)^2$$

$$=\sum_{k=1}^{n}\left(\left(X_k-\mu\right)^2-2\left(X_k-\mu\right)\left(\overline{X}-\mu\right)+\left(\overline{X}-\mu\right)^2\right)$$

$$=\sum_{k=1}^{n}\left(X_k-\mu\right)^2-2\cdot\underbrace{\sum_{k=1}^{n}\left(X_k-\mu\right)}_{=\sum_{k=1}^{n}X_k-\sum_{k=1}^{n}\mu=n\left(\overline{X}-\mu\right)}\cdot\left(\overline{X}-\mu\right)+\sum_{k=1}^{n}\left(\overline{X}-\mu\right)^2$$

$$=\sum_{k=1}^{n}\left(X_k-\mu\right)^2-2n\left(\overline{X}-\mu\right)^2+n\left(\overline{X}-\mu\right)^2$$

$$=\sum_{k=1}^{n}\left(X_k-\mu\right)^2-n\left(\overline{X}-\mu\right)^2.$$

So

$$\mathbb{E}\left(\sum_{k=1}^{n}\left(X_k-\overline{X}\right)^2\right)=\mathbb{E}\left(\sum_{k=1}^{n}\left(X_k-\mu\right)^2-n\left(\overline{X}-\mu\right)^2\right)$$
$$=\sum_{k=1}^{n}\mathbb{E}\left(\left(X_k-\mu\right)^2\right)-n\mathbb{E}\left(\left(\overline{X}-\mu\right)^2\right)=\sum_{k=1}^{n}\mathrm{Var}\left(X_k\right)-n\mathrm{Var}\left(\overline{X}\right)$$
$$=\sum_{k=1}^{n}\sigma^2-n\cdot\frac{\sigma^2}{n}=n\sigma^2-\sigma^2=\left(n-1\right)\sigma^2.$$

We conclude that

$$\begin{aligned}\mathbb{E}\left(S_X^2\right) &= \mathbb{E}\left(\frac{\sum_{k=1}^{n}\left(X_k-\overline{X}\right)^2}{n-1}\right)=\frac{\mathbb{E}\left(\sum_{k=1}^{n}\left(X_k-\overline{X}\right)^2\right)}{n-1}\\ &= \frac{\left(n-1\right)\sigma^2}{n-1}=\sigma^2.\end{aligned}$$

Is the Sample Variance concentrated around its mean σ^2?

Observe that

$$
\begin{aligned}
S_X^2 &= \frac{\sum_{k=1}^{n} \left(X_k - \overline{X}\right)^2}{n-1} = \frac{\sum_{k=1}^{n} (X_k - \mu)^2 - n\left(\overline{X} - \mu\right)^2}{n-1} \\
&= \frac{n}{n-1}\left(\underbrace{\frac{\sum_{k=1}^{n} (X_k - \mu)^2}{n}}_{=A} - \underbrace{\left(\overline{X} - \mu\right)^2}_{=B} \right) \\
&= \frac{n}{n-1}(A - B).
\end{aligned}
$$

The term

$$A = \frac{\sum_{k=1}^{n} (X_k - \mu)^2}{n}$$

is the Sample Mean of the IID random variables

$$(X_1 - \mu)^2, (X_2 - \mu)^2, \ldots, (X_n - \mu)^2 \tag{1}$$

with common mean σ^2.

So, for n large, A is concentrated around σ^2 with large probability: its distribution is close to $\mathrm{N}(\sigma^2, (\frac{\alpha}{\sqrt{n}})^2)$, where α is the common standard deviation of (1).

Now we consider the other term

$$B = \left(\overline{X} - \mu\right)^2 .$$

For n large, since $\overline{X}$ has distribution close to $\mathrm{N}(\mu, (\frac{\sigma}{\sqrt{n}})^2)$, we have

$$\mathbb{P}\left(|\overline{X} - \mu| \lessapprox 3\frac{\sigma}{\sqrt{n}} \right) \approx 99.7\%$$

and so with large probability we have

$$B = \left(\overline{X} - \mu\right)^2 \lessapprox \left(3\frac{\sigma}{\sqrt{n}}\right)^2 = \frac{9}{n}\sigma^2 \ll \sigma^2 \approx A$$

So, for n large,

$$S_X^2 = \frac{n}{n-1}(A - B) \approx A - B \approx A$$

is concentrated around σ^2X with large probability.

Since the Sample Variance has mean σ^2 and, for n large, it is concentrated around σ^2, the sample variance can be considered as an estimator of the variance σ^2, if σ^2 is one of the unknown parameters.

- Observe that if we defined the Sample Variance as

$$C = \frac{\sum_{k=1}^{n} \left(X_k - \overline{X}\right)^2}{n},$$

then

$$\mathbb{E}(C) = \mathbb{E}\left(\frac{\sum_{k=1}^{n} \left(X_k - \overline{X}\right)^2}{n}\right) = \frac{\mathbb{E}\left(\sum_{k=1}^{n} \left(X_k - \overline{X}\right)^2\right)}{n} = \frac{n-1}{n}\sigma^2 \neq \sigma^2.$$

This is the reason for which we have divided by $n-1$ instead of n when we have introduced the statistic sample variance as well as the concept of variance for data.

However, since $C = \frac{n-1}{n} S_X^2$, also the values of C are concentrated around σ^2 for n large.

- If the mean μ is known, we can use the Sampling Statistic

  $$A = \frac{\sum_{k=1}^{n} (X_k - \mu)^2}{n}$$

  as an Estimator of the variance.

  In fact,

  $$\begin{aligned}\mathbb{E}(A) &= \mathbb{E}\left(\frac{\sum_{k=1}^{n} (X_k - \mu)^2}{n}\right) = \frac{\sum_{k=1}^{n} \mathbb{E}\left((X_k - \mu)^2\right)}{n} \\ &= \frac{\sum_{k=1}^{n} \mathrm{Var}\left(X_k\right)}{n} = \frac{n\sigma^2}{n} = \sigma^2.\end{aligned}$$

  and we have previously seen that, for n large, the values of A are concentrated around σ^2.

Sample Variance in case of the the normal distribution

Now, we determine the distribution of the Sample Variance when the Sample is from a normal distribution.

Before to do this, we need to introduce the concept of a chi-squared distribution and the "Principle of Orthonormalization".

- We begin with the chi-squared distribution.

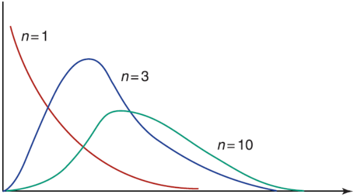
Definition

Let $Z_1, Z_2, \dots, Z_n$ be IID random variables each with the standard normal distribution. The distribution of

$$Z_1^2 + Z_2^2 + \cdots + Z_n^2$$

is said a **chi-squared distribution** with *n* degrees of freedom and it is denoted by χ_n^2.

In figure, we see the chi-squared distribution (the pdf) for some degrees of freedom. The distribution is zero before 0.



- Now, we determine the mean and the variance of the distribution χ^2_n.

  Observe that, for $k \in \{1, 2, \ldots, n\}$, we have

  $$\mathbb{E}\left(Z_k^2\right) = \operatorname{Var}(Z_k) + \mathbb{E}(Z_k)^2 = 1 + 0^2 = 1.$$

  So

  $$\begin{aligned} \mathbb{E}\left(Z_1^2 + Z_2^2 + \cdots + Z_n^2\right) &= \mathbb{E}\left(Z_1^2\right) + \mathbb{E}\left(Z_2^2\right) + \cdots + \mathbb{E}\left(Z_n^2\right) \\ &= 1 + 1 + \cdots + 1 = n. \end{aligned}$$

As for the variance, we have, since $Z_1^2, Z_2^2, \ldots, Z_n^2$ are independent,

$$\begin{aligned}\mathrm{Var}\left(Z_1^2 + Z_2^2 + \cdots + Z_n^2\right) &= \mathrm{Var}\left(Z_1^2\right) + \mathrm{Var}\left(Z_2^2\right) + \cdots + \mathrm{Var}\left(Z_n^2\right) \\ &= n \cdot \mathrm{Var}\left(Z^2\right),\end{aligned}$$

where Z is a standard normal variable.

Exercise. By using the moment generating function of Z, show that $\mathrm{Var}(Z^2) = 2$.

So

$$\mathrm{Var}\left(Z_1^2 + Z_2^2 + \cdots + Z_n^2\right) = 2n$$

and then

$$\mathrm{SD}\left(Z_1^2 + Z_2^2 + \cdots + Z_n^2\right) = \sqrt{2}\sqrt{n}.$$

- Now, we present the "Principle of Orthonormalization"

Theorem

(The Principle of Orthonormalization). *Let* $X_1, \ldots, X_n, Y_1, \ldots, Y_n : \Omega \to \mathbb{R}$ *be random variables such that*

$$\boldsymbol{X} = Q\boldsymbol{Y},$$

where $\boldsymbol{X} = (X_1, \ldots, X_n)$, $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ *and* $Q \in \mathbb{R}^{n\times n}$*. If* $X_1, \ldots, X_n$ *are independent standard normal random variables and* Q *is orthogonal, then* $Y_1, \ldots, Y_n$ *are independent standard normal random variables.*

The name "Principle of Orthonormalization" comes from the fact that the vector $\boldsymbol{Y}$ is the vector of the components of the vector $\boldsymbol{X}$ in the orthonormal basis of $\mathbb{R}^n$ given by the columns of Q.

Proof.

Assume that $X_1, \ldots, X_n$ are independent standard normal random variables and Q is orthogonal. We prove that $Y_1, \ldots, Y_n$ are independent standard normal random variables.

Since Q is orthogonal, i.e. the columns of Q constitute an orthonormal basis of $\mathbb{R}^n$ and so $Q^T Q = I_n$, we have

$$\|Q\boldsymbol{x}\|_2 = \|\boldsymbol{x}\|_2\,, \ \boldsymbol{x} \in \mathbb{R}^n,$$

and

$$|\det(Q)| = 1.$$

Exercise. Prove these two facts.

Proof.

Let V be a Borel subset of $\mathbb{R}^n$. We have

$$
\begin{aligned}
&\mathbb{P}\left(\boldsymbol{Y} \in V\right) = \mathbb{P}\left(\boldsymbol{X} = Q\boldsymbol{Y} \in Q\left(V\right)\right),\ Q(V) = \{Q\boldsymbol{y} : \boldsymbol{y} \in \mathbb{R}^n\}, \\
&= \int\limits_{\boldsymbol{x} \in Q(V)} f_{X_1}\left(x_1\right) \cdots f_{X_n}\left(x_n\right) d\boldsymbol{x},\ \text{since } X_1, \ldots, X_n \text{ are independent}, \\
&= \int\limits_{\boldsymbol{x} \in Q(V)} g\left(x_1\right) \cdots g\left(x_n\right) d\boldsymbol{x},\ g \text{ is the pdf of a standard normal random variable}, \\
&= \int\limits_{\boldsymbol{x} \in Q(V)} \frac{1}{\sqrt{2\pi}} e^{-\frac{x_1^2}{2}} \cdots \frac{1}{\sqrt{2\pi}} e^{-\frac{x_n^2}{2}} d\boldsymbol{x} = \int\limits_{\boldsymbol{x} \in Q(V)} \left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\frac{\|\boldsymbol{x}\|_2^2}{2}} d\boldsymbol{x} \\
&= \int\limits_{\boldsymbol{y} \in V} \left|\det\left(Q\right)\right| \left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\frac{\|Q\boldsymbol{y}\|_2^2}{2}} d\boldsymbol{y},\ \boldsymbol{x} = Q\boldsymbol{y}, \\
&= \int\limits_{\boldsymbol{y} \in V} \left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\frac{\|\boldsymbol{y}\|_2^2}{2}} d\boldsymbol{y},\ \ |\det(Q)| = 1 \text{ and } \|Q\boldsymbol{y}\|_2 = \|\boldsymbol{y}\|_2, \\
&\int\limits_{\boldsymbol{y} \in V} \frac{1}{\sqrt{2\pi}} e^{-\frac{y_1^2}{2}} \cdots \frac{1}{\sqrt{2\pi}} e^{-\frac{y_n^2}{2}} d\boldsymbol{y} = \int\limits_{\boldsymbol{y} \in V} g\left(y_1\right) \cdots g\left(y_n\right) d\boldsymbol{y}.
\end{aligned}
$$

Proof.

Now, for $i \in \{1, \ldots, n\}$ and $x \in \mathbb{R}$, we have

$$F_{Y_i}(x) = \mathbb{P}(Y_i \leq x) = \mathbb{P}(Y \in \mathbb{R} \times \cdots \times \mathbb{R} \times (-\infty, x] \times \mathbb{R} \times \cdots \times \mathbb{R})$$

$$= \int\limits_{\mathbb{R}\times\cdots\times\mathbb{R}\times(-\infty,x]\times\mathbb{R}\times\cdots\times\mathbb{R}} g(y_1) \cdots g(y_n)\, d\boldsymbol{y}$$

$$= \underbrace{\left(\int\limits_{\mathbb{R}} g(y_1)\, dy_1\right)}_{=1} \cdots \underbrace{\left(\int\limits_{\mathbb{R}} g(y_{i-1})\, dy_{i-1}\right)}_{=1}$$

$$\cdot \int\limits_{-\infty}^{x} g(y_i)\, dy_i \cdot \underbrace{\left(\int\limits_{\mathbb{R}} g(y_{i+1})\, dy_{i+1}\right)}_{=1} \cdots \underbrace{\left(\int\limits_{\mathbb{R}} g(y_n)\, dy_n\right)}_{=1}$$

$$= \int\limits_{-\infty}^{x} g(y_i)\, dy_i.$$

We conclude that Y_i has pdf g and so it has the standard normal distribution.

Proof.

Now, we prove that $Y_1, Y_2, \ldots, Y_n$ are independent.

For $i_1, \ldots, i_k \in \{1, \ldots, n\}$ distinct, we have

$$
\begin{aligned}
&\mathbb{P}\left(Y_{i_1} \in [a_{i_1}, b_{i_1}] \cap \cdots \cap Y_{i_k} \in [a_{i_k}, b_{i_k}]\right) \\
&= \int\limits_{y_{i_1} \in \left[a_{i_1}, b_{i_1}\right], \ldots, y_{i_k} \in \left[a_{i_k}, b_{i_k}\right]} g\left(y_1\right) \cdots g\left(y_n\right) d\boldsymbol{y} \\
&= \left(\int\limits_{y_{i_1} \in \left[a_{i_1}, b_{i_1}\right]} g\left(y_{i_1}\right) \right) \cdots \left(\int\limits_{y_{i_k} \in \left[a_{i_k}, b_{i_k}\right]} g\left(y_{i_k}\right) \right) \cdot \prod_{\substack{j=1 \\ j \notin \{i_1, \ldots, i_k\}}}^{n} \underbrace{\int\limits_{y_j \in \mathbb{R}} g\left(y_j\right) dy_j}_{=1} \\
&= \left(\int\limits_{y_{i_1} \in \left[a_{i_1}, b_{i_1}\right]} g_{i_1}\left(y_{i_1}\right) \right) \cdots \left(\int\limits_{y_{i_k} \in \left[a_{i_k}, b_{i_k}\right]} g\left(y_{i_k}\right) \right) \\
&= \mathbb{P}\left(Y_{i_1} \in [a_{i_1}, b_{i_1}]\right) \cdots \mathbb{P}\left(Y_{i_k} \in [a_{i_k}, b_{i_k}]\right).
\end{aligned}
$$

□

- We have seen that, for a Sample from a general distribution, the Sample Variance has mean the variance σ^2 of the distribution and, for *n* large, its values are concentrated around the mean σ^2 with large probability.

  In the next theorem, we give the distribution of the Sample Variance in case of a Sample from a normal distribution.

  A chi-squared distribution appears in the statement of the theorem and the "Principle of Orthonormalization" is used in the proof.

Theorem

Let $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ be a Sample from the normal distribution $\mathrm{N}\left(\mu, \sigma^2\right)$ and let S_X^2 be the Sample Variance. Then

$$Y = \frac{(n-1)\, S_X^2}{\sigma^2} = \frac{\sum\limits_{k=1}^{n} \left(X_k - \overline{X}\right)^2}{\sigma^2}$$

has the distribution χ^2_{n-1}. Moreover $\overline{X}$ and Y, and so $\overline{X}$ and S_X^2, are independent.

Proof.

For $k \in \{1, 2, \ldots, n\}$, let $Z_k = \frac{X_k - \mu}{\sigma}$ be the standardized form of X_k. Since

$$\begin{aligned} \overline{Z} &= \frac{Z_1 + Z_2 + \cdots + Z_n}{n} = \frac{\frac{X_1-\mu}{\sigma} + \frac{X_2-\mu}{\sigma} + \cdots + \frac{X_n-\mu}{\sigma}}{n} = \frac{\frac{X_1+X_2+\cdots+X_n-n\mu}{\sigma}}{n} \\ &= \frac{\frac{n\overline{X}-n\mu}{\sigma}}{n} = \frac{\overline{X}-\mu}{\sigma}, \end{aligned}$$

Proof.

we have

$$Z_k - \overline{Z} = \frac{X_k - \mu}{\sigma} - \frac{\overline{X} - \mu}{\sigma} = \frac{X_k - \overline{X}}{\sigma},\ k \in \{1, 2, \ldots, n\}\,.$$

So

$$Y \quad = \quad \frac{\sum\limits_{k=1}^{n} \left(X_k - \overline{X}\right)^2}{\sigma^2} = \sum_{k=1}^{n} \left(\frac{X_k - \overline{X}}{\sigma}\right)^2 = \sum_{k=1}^{n} \left(Z_k - \overline{Z}\right)^2.$$

Now, note that

$$\sum_{k=1}^{n} (Z_k - \overline{Z}) = \sum_{k=1}^{n} Z_k - \sum_{k=1}^{n} \overline{Z} = n\overline{Z} - n\overline{Z} = 0.$$

and so the vector random variable $\mathbf{\Delta} := (Z_1 - \overline{Z}, Z_2 - \overline{Z}, \ldots, Z_n - \overline{Z})$ lies on the hyperplane H of equation $x_1 + x_2 + \cdots + x_n = 0$, which is a subspace di $\mathbb{R}^n$ of dimension $n - 1$.

Proof.

Let $\boldsymbol{u}^{(1)}, \dots, \boldsymbol{u}^{(n-1)}$ be an orthonormal basis of H and let $C_1, \dots, C_{n-1}$ be the random variables components of $\boldsymbol{\Delta}$ in such basis, i.e. $C_1, \dots, C_{n-1}$ are the scalars such that

$$\boldsymbol{\Delta} = \sum_{i=1}^{n-1} C_i \boldsymbol{u}^{(i)} = Q_{n-1} \boldsymbol{C},$$

where $Q_{n-1} \in \mathbb{R}^{n \times (n-1)}$ is the matrix of columns $\boldsymbol{u}^{(1)}, \dots, \boldsymbol{u}^{(n-1)}$ and $\boldsymbol{C} = (C_1, \dots, C_{n-1})$. The coefficients $C_1, \dots, C_{n-1}$ are such that

$$\sum_{k=1}^{n} \left(Z_k - \overline{Z}\right)^2 = \|\boldsymbol{\Delta}\|_2^2 = \sum_{i=1}^{n-1} C_i^2$$

Exercise. Prove this fact by considering the matrix $Q_{n-1}^T Q_{n-1}$ and observing that $Q_{n-1}^T Q_{n-1} = I_{n-1}$.

Proof.

Now, by setting $\mathbf{1} = (1, 1, \ldots, 1) \in \mathbb{R}^n$, observe that

$$\begin{aligned}(Z_1, Z_2, \ldots, Z_n) &= (\overline{Z}, \overline{Z}, \ldots, \overline{Z}) + \boldsymbol{\Delta} \\ &= \overline{Z} \cdot \mathbf{1} + \sum_{i=1}^{n-1} C_i \boldsymbol{u}^{(i)} = (\sqrt{n} \cdot \overline{Z}) \, \frac{\mathbf{1}}{\sqrt{n}} + \sum_{i=1}^{n-1} C_i \boldsymbol{u}^{(i)} \\ &= Q(\sqrt{n} \cdot \overline{Z}, C_1, \ldots, C_{n-1}),\end{aligned}$$

where $Q \in \mathbb{R}^{n \times n}$ is the matrix of columns $\frac{\mathbf{1}}{\sqrt{n}}, \boldsymbol{u}^{(1)}, \ldots, \boldsymbol{u}^{(n-1)}$.

Exercise. Prove that the vector $\frac{\mathbf{1}}{\sqrt{n}}$ has unit length and, by recalling that $\boldsymbol{u}^{(1)}, \ldots, \boldsymbol{u}^{(n-1)}$ belong to the hyperplane H, prove that $\frac{\mathbf{1}}{\sqrt{n}}$ is orthogonal to $\boldsymbol{u}^{(1)}, \ldots, \boldsymbol{u}^{(n-1)}$.

The previous exercise show that $\frac{\mathbf{1}}{\sqrt{n}}, \boldsymbol{u}^{(1)}, \ldots, \boldsymbol{u}^{(n-1)}$ is an orthonormal basis of $\mathbb{R}^n$ and so Q is an orthogonal matrix.

Proof.

We have

$$(Z_1, Z_2, \ldots, Z_n) = Q(\sqrt{n} \cdot \overline{Z}, C_1, \ldots, C_{n-1}).$$

with $Z_1, Z_2, \ldots, Z_n$ independent standard normal variables (they are the standardized forms of the independent normal variables $X_1, X_2, \ldots, X_n$) and Q orthogonal. By the Principle of Orthonormalization, it follows that $\sqrt{n} \cdot \overline{Z}, C_1, \ldots, C_{n-1}$ are independent standard normal variables.

Since $C_1, \ldots, C_{n-1}$ are independent standard normal variables and

$$Y = \sum_{k=1}^{n} \left(Z_k - \overline{Z}\right)^2 = \sum_{i=1}^{n-1} C_i^2,$$

we have that Y has the distribution χ^2_{n-1}.

Moreover, since $\sqrt{n} \cdot \overline{Z}, C_1, \ldots, C_{n-1}$ are independent, we have that $\overline{X} = \mu + \sigma\overline{Z} = \mu + \frac{\sigma}{\sqrt{n}}\sqrt{n} \cdot \overline{Z}$ and $Y = \sum_{i=1}^{n-1} C_i^2$ are independent and so also $\overline{X}$ and $S_X^2 = \frac{\sigma^2 Y}{n-1}$ are independent. □

- By using the previous theorem, we can determine the standard deviation of the Sample Variance in case of a Sample $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ from a normal distribution $\mathrm{N}(\mu, \sigma^2)$.

  The previous theorem says that the Sample Variance S_X^2 is such that
  $$Y = \frac{(n-1)\, S_X^2}{\sigma^2}$$
  has distribution χ^2_{n-1} and so
  $$\begin{aligned}\mathrm{SD}\left(S_X^2\right) &= \mathrm{SD}\left(\frac{\sigma^2 Y}{n-1}\right) = \frac{\sigma^2}{n-1}\mathrm{SD}\,(Y) \\ &= \frac{\sigma^2}{n-1}\sqrt{2}\sqrt{n-1} = \frac{\sqrt{2}\sigma^2}{\sqrt{n-1}}.\end{aligned}$$

  Observe that
  $$\mathrm{SD}\left(S_X^2\right) \to 0,\ n \to \infty.$$
  and so S_X^2 concentrates its values around its mean σ^2, as n increases, as we have already seen in the general case without determining the distribution of S_X^2.

Exercise. By using the Chebyshev's inequality find a lower bound for the probability

$$\mathbb{P}\left(\left|\frac{S_X^2-\sigma^2}{\sigma^2}\right|<\varepsilon\right),$$

where $\varepsilon>0$ and $\frac{S_X^2-\sigma^2}{\sigma^2}$ is the relative error when we estimate σ^2 by S_X^2.

Exercise. What is the normal distribution close to the distribution of $Y=\sum_{i=1}^{n-1}C_i^2$, when n is large? The C_is are in the proof of the previous theorem. Then, what is the normal distribution close to the distribution of S_X^2 , when n is large?

- Exercise. In case of a Sample from a normal distribution $\mathrm{N}\left(\mu, \sigma^2\right)$, where μ is known, the Sampling Statistic

$$A = \frac{\sum\limits_{k=1}^{n} (X_k - \mu)^2}{n}$$

is used as an Estimator of the variance. Prove that

$$\frac{nA}{\sigma^2} = \frac{\sum\limits_{k=1}^{n} (X_k - \mu)^2}{\sigma^2}$$

has distribution χ_n^2 and then compute the standard deviation of A.

Normal data

- Consider a Sample $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ and the corresponding sample $\boldsymbol{x}^{\text{obs}} = \left(x_1^{\text{obs}}, \ldots, x_n^{\text{obs}}\right)$, where n is large, from a normal distribution $\mathrm{N}\left(\mu, \sigma^2\right)$.

  We recall that a data $\boldsymbol{x}$ is called normal if there is an histogram for the data such that:

  - the histogram is symmetric with respect to an interval called the middle interval, i.e. $\boldsymbol{x}^{\text{hist}}$ is symmetric about the middle point c of the middle interval and c is the mean of $\boldsymbol{x}^{\text{hist}}$.
  - the middle interval has the highest frequency, i.e. $\boldsymbol{x}^{\text{hist}}$ has mode c;
  - the frequencies decrease from the middle interval in a bell-shaped fashion, i.e. $68\%, 95\%, 99.7\%$ of the components of $\boldsymbol{x}^{\text{hist}}$ lie in the intervals between $c - ks$ and $c + ks$ with $k = 1, 2, 3$, respectively, s being the standard deviation of $\boldsymbol{x}^{\text{hist}}$.

- Now we prove that $\boldsymbol{x}^{\text{obs}}$ is an approximately normal data.

  Let $h > 0$ be a small number. Consider the points

  $$y_i := \mu + ih\sigma, \ i \in \mathbb{Z}.$$

  and then the histogram based on the class intervals

  $$I_i = (y_i, y_{i+1}], \ i \in \mathbb{Z}.$$

  Since n is large, by the Strong Law of the Large Numbers (or the frequentist interpretation), we obtain, for $i \in \mathbb{Z}$,

  $$\begin{aligned} & \text{relative frequency of the } i\text{th interval} \\ = & \frac{\text{number of components of } \boldsymbol{x}^{\text{obs}} \text{ in } I_i}{n} \\ \approx & \, \mathbb{P}\left(Y \in I_i\right) = F_Y\left(y_{i+1}\right) - F_Y\left(y_i\right), \end{aligned}$$

  where Y has distribution $\mathrm{N}\left(\mu, \sigma^2\right)$.

Now

$$\begin{aligned} F_Y(y_{i+1}) - F_Y(y_i) &= \Phi\left(\frac{y_{i+1}-\mu}{\sigma}\right) - \Phi\left(\frac{y_i-\mu}{\sigma}\right) \\ &= \Phi((i+1)h) - \Phi(ih) \approx \Phi'(ih)\,h = f_Z(ih)\,h \quad \text{since } h \text{ is small,} \end{aligned}$$

where f_Z is the pdf of a standard normal variable Z.

Therefore

$$\text{relative frequency of the } i\text{th interval} \approx f_Z(ih)\,h,\ i \in \mathbb{Z}.$$

This means that the histogram has approximately the maximum frequency in the middle interval I_0 and the frequencies are approximately symmetric with respect to I_0: this follows from the fact f_Z has the maximum at 0 and it is symmetric with respect to 0.

Since n is large, we can assume that $\overline{x}^{\text{obs}} \approx \mu$ e $(\overline{s}^{\text{obs}})^2 \approx \sigma^2$ and so $\overline{s}^{\text{obs}} \approx \sigma$.

Let c and s be mean and standard deviation, respectively, of the histogram version $\boldsymbol{x}^{\text{hist}}$ of $\boldsymbol{x}^{\text{obs}}$. Since h is small, we have $c \approx \overline{x}^{\text{obs}} \approx \mu$ and $s \approx \overline{s}^{\text{obs}} \approx \sigma$.

The histogram has approximately the bell-shaped fashion decrease: since n is large and h is small, we have, for $k > 0$,

$$\begin{aligned}
&\frac{\text{number of components of } \boldsymbol{x}^{\text{hist}} \text{ in the intervals between } c-ks \text{ and } c+ks}{n}\\
&= \frac{\text{number of components of } \boldsymbol{x}^{\text{obs}} \text{ in the intervals between } c-ks \text{ and } c+ks}{n}\\
&\approx \frac{\text{number of components of } \boldsymbol{x}^{\text{obs}} \text{ in } (\mu-k\sigma,\mu+k\sigma]}{n}\\
&\approx \mathbb{P}\left(Y \in (\mu-k\sigma,\mu+k\sigma]\right) = \Phi(k)-\Phi(-k) = 2\Phi(k)-1 =: c(k)\%.
\end{aligned}$$

and

$$c(k)\% = \begin{cases} 68\% \text{ if } k=1\\ 95\% \text{ if } k=2\\ 99.7\% \text{ if } k=3. \end{cases}$$

- Now, we can understand our previous observation that:
  - if the data represents some biological characteristic (for example heights, weights, blood pressure,...) of a sample taken from an homogeneus population of human beings, or other living beings, and its size n is large, then it is approximately normal, and it becomes normal as $n \to \infty$.

The data is normal because it is a sample from a normal distribution: the biological characteristics of the individuals in the population are normally distributed random variables.