# MEM-264 Applied Statistics

**Department of Mathematics and Applied Mathematics, University of Crete**

Costas Smaragdakis (kesmarag@uoc.gr)

4th Lecture - 19-02-2021

**Definition**

Let $X_1, \dots, X_n$ be an i.i.d sample. The sample variance $S_n^2$ is defined as

$$S_n^2 = \frac{1}{n-1} \sum_{j=1}^{n} (X_j - \bar{X}_n)^2.$$

- $S_n^2$ is a random variable.
- We denote as $s_n^2$ a realization of the sample variance.

**Expectation of the sample variance**

$$\mathbb{E}(S_n^2) = \mathsf{Var}(X_1).$$

$$\mathbb{E}\left\{\sum_{j=1}^{n}(X_j-\bar{X}_n)^2\right\}=\mathbb{E}\left\{\sum_{j=1}^{n}X_j^2-2\bar{X}_n\sum_{j=1}^{n}X_j+n\bar{X}_n^2\right\}=\mathbb{E}\left\{\sum_j X_j^2-2n\bar{X}_n^2+n\bar{X}_n^2\right\}=\mathbb{E}\left\{\sum_j X_j^2-n\bar{X}_n^2\right\}=$$

$$=\sum_{j=1}^{n}\mathbb{E}\{X_j^2\}-n\mathbb{E}\{\bar{X}_n^2\}=n(\sigma^2+\mu^2)-n\left(\frac{\sigma^2}{n}+\mu^2\right)=n\sigma^2-\sigma^2=(n-1)\sigma^2$$

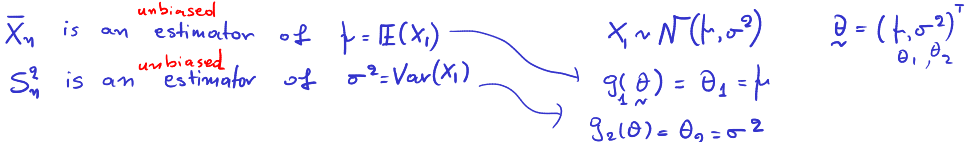
$$\mathrm{var}\{X\}=\mathbb{E}(X^2)-(\mathbb{E}(X))^2 \Leftrightarrow \boxed{\mathbb{E}\{X^2\}=\mathrm{Var}\{X\}+(\mathbb{E}(X))^2=\sigma^2+\mu^2}$$

$$\mathbb{E}\{\bar{X}_n^2\}=\mathrm{Var}\{\bar{X}_n\}+(\mathbb{E}(\bar{X}_n))^2=\frac{\sigma^2}{n}+\mu^2$$

$\Rightarrow$

$$\mathbb{E}\{S_n^2\}=\sigma^2=\mathrm{Var}\{X_1\}$$

$$\mathbb{E}\left\{\frac{1}{n}\Sigma(X_j-\bar{X}_n)^2\right\}=\frac{n-1}{n}\mathbb{E}\{S_n^2\}=\frac{n-1}{n}\sigma^2$$




**Definition**

A statistic $t$ is called **an estimator** of a function $g$ of the model parameters $\theta$ if it is used to estimate $g(\theta)$.

**Definition**

An estimator $t$ is called **unbiased (αμερόληπτη)** iff

$$\mathbb{E}(t(\mathbf{X})) = g(\theta)$$

$\frac{1}{n}\sum_{j=1}^{n}(X_j - \bar{X}_n)^2$ is an biased estimator of $\sigma^2$

**Statistical model :** $f_X(x;\theta),\ \theta \in \Theta \subseteq \mathbb{R}^d$.

$\Theta = \mathbb{R}$ → $\Theta_0 = [0, +\infty)$, $\Theta_1 = (-\infty, 0)$

- A hypothesis is a statement about an unknown state of the nature.

$$\Theta = \Theta_0 \cup \Theta_1, \quad \Theta_0 \cap \Theta_1 = \emptyset.$$

- **Null hypothesis (μηδενική υπόθεση)**

$$H_0 : \theta \in \Theta_0$$

The null hypothesis is going to be the trivial statement of the problem.

- **Alternative hypothesis (εναλλακτική υπόθεση)**

$$H_1 : \theta \in \Theta_1$$

$\Theta = \mathbb{R}$ → $\Theta_0 = \{0.5\}$ → $H_0$ is a simple hypothesis

$\Theta = \mathbb{R}$ → $\Theta_1 = \mathbb{R} - \{0.5\}$ → $H_1$ is a composite hypothesis

A hypothesis is called **a simple hypothesis** if it specifies the statistical distribution without any uncertainty. Otherwise the hypothesis is called **a composite hypothesis**.

## One-sided hypotheses

Let $\Theta = \mathbb{R}$, $H_0$ and $H_1$ are one-sided hypotheses iff we have one of the following:

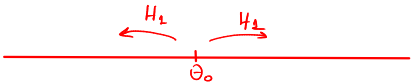

$$H_0 : \theta \geq \theta_0 \quad \text{vs} \quad H_1 : \theta < \theta_0$$

$$H_0 : \theta > \theta_0 \quad \text{vs} \quad H_1 : \theta \leq \theta_0$$

$$H_0 : \theta \leq \theta_0 \quad \text{vs} \quad H_1 : \theta > \theta_0$$

$$H_0 : \theta < \theta_0 \quad \text{vs} \quad H_1 : \theta \geq \theta_0$$

where $\theta_0 \in \mathbb{R}$.




**Two-sided hypothesis**

Let $\Theta = \mathbb{R}$, We call $H_1$ two-sided hypothesis iff:

$$H_0 : \theta = \theta_0 \quad \text{vs} \quad H_1 : \theta \neq \theta_0,$$

where $\theta_0 \in \mathbb{R}$.

**Definition**

A **test statistics** is a function that measures the inconsistency between the data and the null hypothesis.

$$\text{Random variable : } T = t(\mathbf{X}), \quad \mathbf{X} = (X_1, \dots, X_n)^T$$

$$\text{Realization : } t_{obs} = t(\mathbf{x}), \text{ where } \mathbf{x} = (x_1, \dots, x_n)^T \text{ is a realization of } \mathbf{X}$$

**Example**

Let $X_1, \dots, X_{100}$ be an i.i.d sample, and $X_1 \sim \mathcal{N}(\theta, 1),\ \theta \in \mathbb{R}$

$X_1 = 1.5$ $\quad$ $X_{100} = -0.02$

$$H_0 : \theta = 1 \quad \text{vs} \quad H_1 : \theta \neq 1$$

$\mathbb{E}(X_1) = \theta$ $\quad$ $\bar{X}_n = \frac{1}{n}\sum_{j=1}^{n} X_j$ $\quad$ $\bar{X}_n \sim \mathcal{N}\left(\mathbb{E}(X_1), \frac{Var(X_1)}{n}\right)$ $\quad$ $\theta$ $\quad$ $\sigma^2 = 1$

$T = t(\underset{\sim}{X}) = \bar{X}_n - \theta$

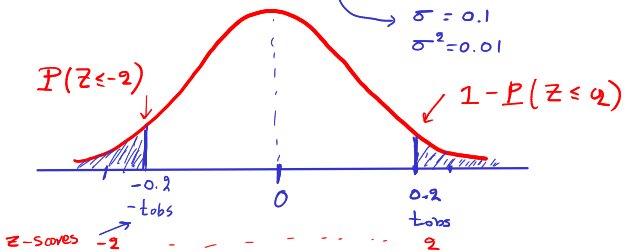
$T \sim N(0, \frac{1}{n}) = N(0, 0.01)$

$\sigma = 0.1$

$\sigma^2 = 0.01$

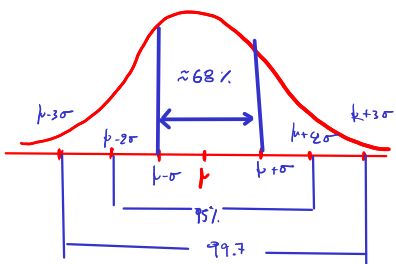
$Var(X - a) = Var X$

$t_{obs} = 0.2$




$z = \frac{t_{obs} - 0}{\sigma} = \frac{0.2}{0.1} = 2$

p.value $= 1 + P(Z \leq -2) - P(Z \leq 2) \approx 5\%$



**Definition : p-value**

To any realization $t_{obs}$ of the test statistics we associate a p-value. In the general case, the p-value is the probability of $|t(\mathbf{X})|$ lying at and beyond $|t_{obs}|$.

| p-value | amount of evidence |
|---|---|
| p-value < 0.01 | strong evidence against $H_0$ |
| 0.01 < p-value < 0.05 | evidence against $H_0$ |
| 0.05 < p-value < 0.12 | weak evedence against $H_0$ |
| p-value > 0.12 | no evidence against $H_0$ |

# Test statistics

Let $X$ be Bin$(100, \theta)$, $\theta \in [0.5, 1)$ distributed.

$$H_0 : \theta = 0.5 \quad \text{vs} \quad H_1 : \theta \in (0.5, 1)$$

Calculate the p-value if the observed value (a single realization) of $X$ is $x = 95$.