AN INTERESTING EXAMPLE

Il mese prossimo in Italia verra indetto un referendum per stabilire se la pizza con l'ananas debba essere dichiarata illegale.

Ogni votante avrà due scelte: rispondere si (con probabilità μ) oppure no (con probabilità $1 - \mu$). Inoltre, ogni votante darà una risposta che non dipende da quella degli altri.

La popolazione (il corpo elettorale) è composto da 40 milioni di votanti, questo numero è così grande che possiamo assumere che la frazione di votanti che daranno una risposta positiva coincida a tutti gli effetti con il valore atteso μ della variabile stocastica *a priori* che regola il voto di ogni singolo individuo.

Si vorrebbe sapere in anticipo cosa sarà deciso dal referendum, a questo scopo si faun sondaggio scegliendo in modo rappresentativo un campione di $n = 100$ aventi diritto al voto.

Possiamo formalizzare il problema assumendo che la variabile stocastica X_i associata ad ogni elettore possa assumere il valore 1 con probabilità μ (favorevole alla dichiarazione di illegalità) oppure il valore 0 con probabilità $1 - \mu$ (contrario) in modo che $\mathbb{E}(X_i) = \mu$.

Il risultato del sondaggio è di 80 pareri favorevoli alla proibizione e 20 contrari. In simboli:

$$\bar{x} = \frac{1}{100} \sum_{i=1}^{100} x_i = \frac{80}{100} = 0.8,$$

$$s^2 = \frac{1}{99} \sum_{i=1}^{100} (x_i - 0.8)^2 = \frac{1}{99} \left[80(1 - 0.8)^2 + 20(0 - 0.8)^2 \right] = 0.16$$

In definitiva abbiamo la nostra stima:

$$\mu = \bar{x} \pm \frac{s}{\sqrt{n}} = 0.8 \pm \frac{0.4}{\sqrt{100}} = 0.8 \pm 0.04$$

Ne deduciamo che la pizza all'ananas con quasi certezza sarà dichiarata fuorilegge in Italia perché la forchetta va da 0.76 a 0.84, che è quindi ben lontana dai valori inferiori a 0.5 che segnerebbero la sopravvivenza della **tradizionale pizza all'ananas italiana**.

Abbiamo detto quasi certezza perché anche se piccola esiste sempre una probabilità che il campione non rappresenti la popolazione.

Le cose sarebbero state diverse se solo 51 votanti intervistati si fossero espressi a favore dell'abolizione. In tal caso avremmo $\bar{x} = 0.51$ e $s^2 \simeq 0.25$, quindi

$$\mu = \bar{\mathbf{x}} \pm \frac{\mathbf{s}}{\sqrt{\mathbf{n}}} = \mathbf{0.51 \pm 0.05}$$

che implica una forchetta da 0.46 a 0.56 e quindi il sondaggio avrebbe avuto poco valore inferenziale.

This qualitative argument can be done much more precise introducing the notion of **confidence interval**, but this is beyond the scope of this short course.

Si noti che se la distribuzione è binomiale se $\mu = 0.8$ si deve necessariamente avere $\sigma^2 = (1-\mu)\mu = 0.16$ e se $\mu = 0.51$ si deve necessariamente avere $\sigma^2 = (1-\mu)\mu = 0.25$ e quindi il nostro risultato è coerente con il modello.

Questo non è un caso, in realtà c'è sempre coerenza(a meno di un fattore $\frac{n}{n-1}$). Infatti tenendo conto che $x_i^2 = x_i$ si ha

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{n}{n-1}\bar{x}(1-\bar{x}).$$

Si noti che in conseguenza $\bar{x}(1-\bar{x})$ non è uno stimatore corretto della varianza, tuttavia nel nostro caso il fattore $\frac{100}{99}$ non altera in modo significativo il risultato.

EFFICIENT ESTIMATORS

In general, the estimator of a parameter is not unique: to decide which of two correct estimators is the best, we use the criterion of choosing the most efficient one, that is, the one for which the variance is smaller.

Esercise 1. Let us consider a sample x_1, ...,x_n ($n \geq 2$) of independent and discrete random variables X_i which may assume one of the three values 0, -1, 1 with probabilities

$$p(-1) = \frac{a}{2}, \qquad p(0) = 1 - a, \qquad p(1) = \frac{a}{2},$$

we can ask ourselves various questions:

(1) for which values of the parameter a the function $p(i)$ is a probability distribution?

Answer: $0 \leq p(i) \leq 1$ must old or any $i = 0, -1, 1$, therefore, it is necessary and sufficient that

$$a \in [0, 1].$$

Moreover, $p(-1) + p(0) + p(1) = 1$ must also hold, but this is true for any a.

(2) what are the expected value and the variance?

Answer:

$$\mu = \mathbb{E}(X_i) = (-1) \cdot \frac{a}{2} + 0 \cdot (1 - a) + 1 \cdot \frac{a}{2} = 0.$$

$$\sigma^2 = \text{var}(X_1) = \mathbb{E}(X_1^2) = (-1)^2 \cdot \frac{a}{2} + 0^2 \cdot (1 - a) + 1^2 \cdot \frac{a}{2} = a$$

(3). Let us consider the following two estimators of a:

$$\Gamma = \frac{1}{n}\sum_{i=1}^{n}|X_i|, \qquad \Lambda = |X_n|$$

are they correct?

Answer:

$$\mathbb{E}(\Lambda) = \mathbb{E}(|X_n|) = |-1| \cdot p(-1) + 0 \cdot p(0) + |1|p(1) = a$$

Moreover since the X_i have the same distribution:

$$\mathbb{E}(\Gamma) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(|X_i|) = \frac{1}{n}\sum_{i=1}^{n} a = a.$$

Therefore, both Γ and Λ are correct estimators.

(4) What is the variance of the two estimators?

Answer:

$\text{var}(\Lambda) = \mathbb{E}\Big((|X_n| - a)^2\Big) = \mathbb{E}\big(|X_n|^2\big) - a^2$, therefore

$$\text{var}(\Lambda) = (-1)^2 p(-1) + 0^2 p(0) + 1^2 p(1) - a^2 = a(1-a).$$

Moreover, given that the random variables **are independent and they have identical distribution:**

$$\text{var}(\Gamma) = \text{var}(\frac{1}{n}\sum_{i=1}^{n}|X_i|) = \frac{1}{n^2}\sum_{i=1}^{n}\text{var}(|X_i|)$$

$$= \frac{1}{n^2}\sum_{i=1}^{n}\text{var}(|X_n|) = \frac{1}{n^2}\sum_{i=1}^{n}a(1-a) = \frac{1}{n}a(1-a)$$

(5) Which of the two estimators is more efficient?

Answer: one has

$$\frac{\mathrm{var}(\boldsymbol{\Gamma})}{\mathrm{var}(\boldsymbol{\Lambda})} = \frac{\mathbf{1}}{\mathbf{n}} < \mathbf{1},$$

where the inequality holds because $n \geq 2$.

In conclusion:

$\boldsymbol{\Gamma}$ is more efficient than $\boldsymbol{\Lambda}$.

Esercise 2. Let us consider a sample x_1, ...,x_n ($n \geq 2$) of independent and discrete random variables X_i which may assume the two values 0 e 1 with probability $1 - p$ e p respectively:
We have already seen that

$$\mathbb{E}(X_i) = p \quad \text{and} \quad Var(X_i) = p(1-p)$$

(1) For which values of a

$$aX_1$$

is a correct estimator of p?

Answer: given that

$$\mathbb{E}(aX_1)) = a\mathbb{E}(X_1) = ap$$

it is correct only if $a = 1$.

(2) For which values of a and b

$$\Gamma = aX_1 + b\sum_{i=2}^{n} X_i$$

is a correct estimator of p?

Answer: given that the X_i have the same distribution, one has

$$\mathbb{E}(\Gamma) = a\mathbb{E}(X_1) + b(n-1)\mathbb{E}(X_1) = [a + b(n-1)]p$$

therefore, Γ is correct only if $a + b(n-1) = 1$, in this case, in fact, the expected value of Γ equals p. In particular it is a correct estimator if $a = b = \frac{1}{n}$, for which Γ reduces to the ordinary average, but also $a = 0$, $b = \frac{1}{n-1}$.

(3) It is quite always possible to obtain a correct estimator from a distorted one. Assume that $a + b(n-1) \neq 0$, and find the constant c such that $\Lambda = c\Gamma$ is a correct estimator of p.

Answer: given that $\mathbb{E}(\Gamma) = [a + b(n-1)]p$, one can chose

$$c = \frac{1}{a + b(n-1)},$$

one has, in fact,

$$\mathbb{E}(\Lambda) = \frac{1}{a + b(n-1)}\mathbb{E}(\Gamma) = p,$$

therefore, Λ is a correct estimator of p.

(4) Consider a generic correct estimator of μ (and, therefore, of $p = \mu$)

$$\Gamma = aX_1 + b\sum_{i=2}^{n} X_i$$

with $a + b(n-1) = 1$ and find the values of a and b which correspond to the highest efficiency.

Answer: first of all, notice that $a + b(n-1) = 1$ is the same as to say $b = \frac{1-a}{n-1}$ so that we can write

$$\Gamma = aX_1 + \frac{1-a}{n-1}\sum_{i=2}^{n} X_i,$$

which automatically ensures that $\mathbb{E}(\Gamma) = p$ for any real number a.

In fact:

$$\mathbb{E}(\Gamma) = a\mathbb{E}(X_1) + \frac{1-a}{n-1}\sum_{i=2}^{n}\mathbb{E}(X_i) = ap + (1-a)p = p.$$

Moreover, given the independence of the variables and the properties of variance one has

$$Var(\Gamma) = a^2 Var(X_1) + \left(\frac{1-a}{n-1}\right)^2 \sum_{i=2}^{n} Var(X_i)$$

and given that that

$$Var(X_j) = p(1-p)$$

for $j = 1, 2, \ldots, n$) one finally obtains

$$\text{var}(\Gamma) = \mathbb{E}(\Gamma^2) - \mathbb{E}(\Gamma)^2 = \left(a^2 + \frac{(1-a)^2}{n-1}\right) p(1-p).$$

The highest efficiency can be than obtained by minimizing this variance with respect to a, which is the same to say minimizing

$$a^2(n-1)+(1-a)^2.$$

The minimum is in

$$a=\frac{1}{n}$$

which also implies

$$b=\frac{1-a}{n-1}=\frac{1}{n}=a.$$

In sum, the highest efficiency corresponds to $a=b=\frac{1}{n}$, for which

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

which means that the flat average $\bar{X}$ is the most efficient amog all possible Γ.

Esercise 2.
Consider two centered independent gaussian variables X_1 and X_2 of same variance σ^2, *i.e.*, variables with probability density

$$\rho(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(-\frac{x^2}{2\sigma^2}\Big).$$

Given that for any function f one has

$$\mathbb{E}(f(X_i)) = \int_{-\infty}^{\infty} f(x)\rho(x)dx.$$

simple integration gives the following averages (student exercise)

$$\mathbb{E}(|X_i|) = \sqrt{\frac{2}{\pi}}\sigma, \qquad \mathbb{E}(X_i^2) = \sigma^2, \qquad \mathbb{E}(X_i^4) = 3\sigma^4.$$

1) show that both

$$\Gamma_1 = \frac{\pi}{2}|X_1||X_2| \quad \text{and.} \quad \Gamma_2 = \frac{X_1^2 + X_2^2}{2}$$

are correct estimators of the variance σ^2.

Answer:

$$\mathbb{E}(\Gamma_1) = \frac{\pi}{2}\mathbb{E}(|X_1||X_2|) = \frac{\pi}{2}\mathbb{E}(|X_1|)\mathbb{E}(|X_2|) = \sigma^2,$$

where we have used the independence. Moreover,

$$\mathbb{E}(\Gamma_2) = \frac{\mathbb{E}(X_1^2) + \mathbb{E}(X_2^2)}{2} = \sigma^2,$$

therefore, they are both correct.

2) say which of the two estimators is more efficient.

Answer:

First of all we compute the following two averages:

$$\mathbb{E}(\Gamma_1^2) = \frac{\pi^2}{4}\mathbb{E}(X_1^2 X_2^2) = \frac{\pi^2}{4}\mathbb{E}(X_1^2)\mathbb{E}(X_2^2) = \frac{\pi^2}{4}\sigma^4,$$

where for the second equality we have used the independence of the two variables,

$$\mathbb{E}(\Gamma_2^2) = \frac{\mathbb{E}(X_1^4) + \mathbb{E}(X_2^4) + 2\mathbb{E}(X_1^2)\mathbb{E}(X_2^2)}{4} = 2\sigma^4$$

where for the first equality we have used again the independence of the two variables.

From these averages we immediately find

$$\text{var}(\Gamma_1) = \frac{\pi^2}{4}\sigma^4 - \sigma^4 = \frac{\pi^2 - 4}{4}\sigma^4$$

and

$$\text{var}(\Gamma_2) = 2\sigma^4 - \sigma^4 = \sigma^4,$$

therefore

$$\frac{\text{var}(\Gamma_1)}{\text{var}(\Gamma_2)} = \frac{\pi^2 - 4}{4} > 1$$

which means that **the estimator Γ_2 is more efficient than Γ_1**.

Notice that if we define e third estimator as $\Gamma_3 = X_1^2$, we find that it is correct (its expected value is σ^2). Moreover, its variance is $2\sigma^4$, larger than both var(Γ_1) and var(Γ_1). **Therefore, Γ_3 is the less efficient of the three estimators.**