Statistic Stuffs

Roberto ANTONIELLO

October 8, 2022

I will drop here some statistic concepts I liked during the course I had at University of Milan.

1 Gini index

There are two types of Gini indexes. One has the purpose of telling how much a sample is heterogeneous. The other one is used for concentration. We will talk about the first one we have mentioned and is defined as follow:

$$I = 1 - \sum_{j=1}^{k} f_j^2$$

This index is always less than 1 and always more than 0.

$$0 \leq I < 1$$

So we are subtracting the squared relative frequencies from one.
we know that:

$$\forall j f_j \geq 0 \longrightarrow \sum_{j=1}^{k} f_j = 1$$

.

This implies that $\exists \acute{j} : \acute{f}_j > 0 \longrightarrow f_j^2 > 0$.
So:

$$\sum_{j=1}^{k} f_j^2 > 0 \longrightarrow I < 1$$

as said in before.

We can consider also that:

$$\sum_{j=1}^{k} f_j^2 \le \sum_{j=1}^{k} f_j$$

.

This implies that

$$1 - \sum_{j=1}^{k} f_j^2 \ge 0$$

.

We have just demonstrated what we said a few moments ago. $0 \le I < 1$.

2 Probability calculus

Before we can talk about probability, we have to define a probability function. So we need to define as first Ω as the set of all possibles outcomes. Then we can say what is an event, which is just a subset of Ω. Now let's define what is an Event algebra:

$$\mathcal{A} = \{E_i \subseteq \Omega\}$$

E_i are all the subsets events in Ω. This algebra has to respect three rules:
    1. $\Omega \in \mathcal{A}$
    2. $\forall E \subseteq \Omega \longrightarrow E \in \Omega \to \bar{E} \in \Omega$
    3. $\forall E, F \subseteq \Omega \longrightarrow E \in \mathcal{A}, F \in \mathcal{A} \to E \cup F \in \mathcal{A}$
    The third rule can be extended to this:

$$\forall i = 1, 2, .., n E_i \subseteq \Omega, E_i \in \mathcal{A} \to \bigcup_{i=1}^{n} E_i \in \mathcal{A}$$

Finally we can define the probability function.
$P : \mathcal{A} \to R$

2.1 Kolmogorov axioms

Kolmogorov axioms are the fundamental rules of the probability calculus.
    A1. $\forall E \in \mathcal{A} \quad 0 \le P(E) \le 1$
    A2. $P(\Omega) = 1$
    A3. $\forall E_1 ... E_n \quad \forall i \ E_i \in \mathcal{A}$ mutually exclusive $\to P(\bigcup_{i=1}^{n} E_i) = \sum_{i=1}^{n} P(E_i)$

2.2 Elemental theorems of Probability

1. First theorem
    $\forall E \in \mathcal{A} \quad P(\bar{E}) = 1 - P(E)$

    Demonstration:
    We know that $1 = P(\Omega)$ from the second kolmogorov axiom.

We know also that $E \cap \bar{E} = \emptyset$ and that $E \cup \bar{E} = \Omega$

So:

$1 = P(\Omega) \rightarrow 1 = P(E \cup \bar{E})$

Now we can use the third kolmogorov axiom and:

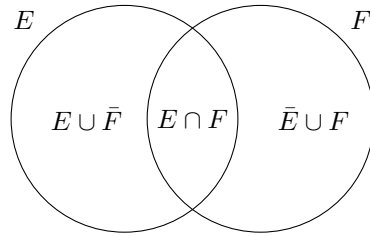$1 = P(E) + P(\bar{E}) \rightarrow P(\bar{E}) = 1 - P(E)$

Demonstrated.

2. Second theorem

$\forall E, F \in \mathcal{A} \quad P(E \cup F) = P(E) + P(F) - P(E \cap F)$

Demonstration:

Let's draw a Venn diagram to make it easier.

We can get that $E \cup F = (E \cap \bar{F}) \cup (E \cap F) \cup (\bar{E} \cap F)$

So: $P(E \cup F) = P((E \cap \bar{F}) \cup (E \cap F) \cup (\bar{E} \cap F))$

Now we use again the third kolmogorov axiom:

$P(E \cup F) = P(E \cap \bar{F}) + P(E \cap F) + P(\bar{E} \cap F)$

But now we can transform $P(E \cap \bar{F}) + P(E \cap F)$ in $P(E)$

$P(E \cup F) = P(E) + P(\bar{E} \cap F)$

We need also $P(F)$, so I add and substract $P(E \cap F)$ which doesn't change anything.

$P(E \cup F) = P(E) + P(\bar{E} \cap F) + P(E \cap F) - P(E \cap F)$

And so we finished it because the result from getting $P(F)$ is:

$P(E \cup F) = P(E) + P(F) - P(E \cap F)$

Demonstrated.

3. Third theorem

$P(\emptyset) = 0$

Demonstration:

$P(\Omega) = 1$

we know from the first theorem that: $P(\bar{E}) = 1 - P(E)$

$P(\emptyset) = P(\bar{\Omega}) = 1 - P(\Omega) = 0$

Demonstrated.

3   Total probability theorem and Bayes Theorem

these two theorems can be useful to calculate complex probability by calculating more probabilities that are more simple.

3.1 Total probability theorem(General form)

Hypothesis:

A partition of $\Omega \longrightarrow F_1, F_2, ..., F_n \subseteq \Omega$
$\forall i P(F_i) \neq 0$

Thesis:

$$P(E) = P(\bigcup_{i=1}^{n} (E \cap F_i)) = \sum_{i=1}^{n} P(E \cap F_i) = \sum_{i=1}^{n} P(E|F_i) \cdot P(F_i)$$

Where $P(A|B)$ is a conditional probability defined as follow:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

And from this last definition we say $P(A \cap B) = P(A|B) \cdot P(B)$

3.2 Bayes Theorem

The hypothesis are the same of the Total probability theorem(General form).

Thesis:

$$P(F_j|E) = \frac{P(F_j \cap E)}{P(E)} = \frac{P(E|F_j) \cdot P(F_j)}{\sum_{i=1}^{n} P(E|F_i) \cdot P(F_i)}$$

From this theorem it's possible to build a classifier that is called Naive Bayes, which allow to simplify the calculus of some probability by using a naive hypothesis. We will not define this classifier here.

4 Random variable

A random variable is a function defined as follow:

$$X : \Omega \to R$$

Or, if we want to be more precise:

$$\{X = \alpha\} \equiv \{\omega \in \Omega : X(\omega) = \alpha\}$$

Basically it's a function that associates a number to each outcome of a random experiment. We define the support as the set of all possible values that a random variable X can assume.

In this document we will see discreet and continuous random variables.

4.1 Cumulative Distribution Function

This function can be defined as the DNA of a random variable, the ID card we can say. It tells us how the variable is distributed.

Definition: $F_x(x) = P(X \leq x)$

Soon we will understand what it means when a random variable is distributed like a certain F_x and we will write $X \sim F_x$.

Properties of F_x:
$F_x : R \to [0,1]$
$F_x > 0$
$\lim_{x \to +\infty} F_x(x) = 1$

4.2 Probability Mass Function

Only a disreet random variable has this type of function. Discreets random variables assume an enumerable set of specifications. This function is always greater than zero. It's defined as follows:

$p_x : R \to [0,1]$
$p_x(i) = P(X = i)$

if we do $\sum_i p_x(i)$ we obtain 1. Basically we are doing the sum of all specifications that a random variables X can assume, so we are obtaining Ω and then we are obtaining 1.

4.3 Probability Density Function

Only a continous random variable has this type of function. It's defined as follows:
$f_x : R \to R^+$

$$\forall B \subseteq R, \ P(X \in B) = \int f_x(x)dx$$

$$P(a \leq X \leq b) \int_a^b f_x(x)dx$$

Now we'll explaining why this can't work using the Probability Mass Function. Basically it would be like calculating:

$$P(X = a) = P(a \leq X \leq a) = \int_a^a f_x(x)dx = 0$$

Properties of a Probability Density Function:

$$First: \quad \int_{-\infty}^{+\infty} f_x(x)dx = \int_R f_x(x)dx = P(X \in R) = 1$$

$$Second: \quad \int_{-\infty}^{a} f_x(x)dx = P(X \le a) = F_x(a) \longrightarrow \frac{d}{dx}F_x(x) = f_x(x)$$

5 Markov and Tchebyshev inequalities

These inequalities are interesting theoretic results we can use to give a esteem of a probability.

5.1 Markov inequality

Theorem:

$$X \ge 0 \quad \forall a > 0 \to P(X \ge a) \le \frac{\mathcal{E}(X)}{a}$$

Demonstration(For continuous):

$$\mathcal{E}(X) = \int_0^{+\infty} x \cdot f_x(x)dx = \int_0^a x \cdot f_x(x)dx + \int_a^{+\infty} x \cdot f_x(x)dx \ge \int_a^{+\infty} x \cdot f_x(x)dx$$

$$\ge \int_a^{+\infty} a \cdot f_x(x)dx = a \cdot \int_a^{+\infty} f_x(x)dx = a \cdot P(X \ge a) \longrightarrow \mathcal{E}(X) \ge a \cdot P(X \ge a)$$

Now we divide both by a and it's demonstrated. The demonstration for discreet variables is almost the same, we use summations instead of integrals.

5.2 Tchebyshev inequality

We consider a random variable X with $\mathcal{E}(X) = \mu$ and $Var(X) = \sigma^2$
Thesis:
$$\forall r > 0 \quad P(|X - \mu| \ge r) \le \frac{\sigma^2}{r^2}$$

Demonstration:
$|X - \mu| \ge r \leftrightarrow (X - \mu)^2 \ge r^2$

$$P(|X - \mu| \geq r) = P((X - \mu)^2 \geq r^2)$$
Let' define

$$Y := (X - \mu)^2 \rightarrow using\ markov\ P(Y \geq r^2) \leq \frac{\mathcal{E}(Y)}{r^2}$$

$$= \frac{\mathcal{E}((X - \mu)^2)}{r^2} = \frac{Var(X)}{r^2} = \frac{\sigma^2}{r^2}$$

Demonstrated.

6   Random variables Models

Models are useful to solve common problems with a common plan.

6.1   Bernoulli model

The bernoulli model describes the most simple experiment such as coin launch and similar. Every experiments have only two outcomes: success or failure.

$$X \sim B(p) \quad D_x = \{0, 1\} \quad p \in [0, 1]$$

$p_x(x) = P(X = x) = p^x \cdot (1 - p)^{1-x} \cdot I_{\{0,1\}^{(x)}}$
$= 1 - p$ if $x = 0$
$= p$ if $x = 1$
$= 0$ otherwise

$\mathcal{E}(X) = p = \sum_i x_i P(X = xi)$
$Var(X) = p(1 - p)$

6.2   Binomial model

This model is obtained from the Bernoulli by doing n bernoulli experiments with the same p parameter and with independence from each other. The variable counts the number of success.
$X \sim B(n, p) D_x = \{0, 1, ..., n\}$

$$p_x(i) = P(X = i) = \binom{n}{i} \cdot p^i \cdot (1 - p)^{n-i} \cdot I_{\{D_x\}^{(x)}}$$

$$\mathcal{E}(X) = (\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} \mathcal{E}(X) = \sum_{i=1}^{n} p = np$$

$$Var(X) = Var(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} Var(X_i) = \sum i = 1^n p(1 - p) = np(1 - p)$$

If we have two binomial indipendent random variables with the same p but different n and we sum them we obtain this:

$$X_1 \sim B(n, p) \ and \ X_2 \sim B(m, p) \rightarrow X_1 + X_2 = Y \sim B(n + m, p)$$

6.3 Discreet uniform model

This model is used for experiments such as dice launch.
$X \sim U(n) D_x = N \cup \{0\}$

$$p_x(i) = P(X = i) = \frac{1}{n} \cdot I_{\{1,...,n\}}(i)$$

$$\mathcal{E}(X) = \frac{n + 1}{2}$$

$$Var(X) = \frac{n^2 + 1}{12}$$

The dispersion around the $\mathcal{E}(X)$ increase with the number of equal-probability events.

6.4 Geometric Model

This model is obtained from the Bernoulli by doing experiments and stop only when it comes the first success. The variable counts the number of failures before the first success.

$X \sim G(p) \ \ p \in (0, 1]$
$p_x(i) = P(X = i)(1 - p)^i p \cdot I_{N \cup 0}(i)$

$$\mathcal{E}(X) = \frac{1 - p}{p}$$

$$Var(X) = \frac{1 - p}{p^2}$$

This model has the property of Memorylessness, which is defined as follow:
$P(X \geq i + j | X \geq i) = P(X \geq j)$
The number of failures is saying nothing about the next experiment, the probability is the same.

6.5 Poisson model

Before talking about the possion model, let's write directly the Probability Mass Function.

$X \sim P(\lambda) D_x = N \cup \{0\}$

$$p_x(i) = P(X = i) = e^{-\lambda} \cdot \frac{\lambda^i}{i!} \cdot I_{N \cup \{0\}}$$

$\mathcal{E}(X) = \lambda$
$Var(X) = \lambda$

The really importance of the Poisson model is that we can approximate a Binomial variable with it during certain conditions. If we have a random variable X distributed as a Binomial with a "big" n and a "little" p (we can say a costant product np), then we can approximate well this variable as a Poisson of parameter $np = \lambda$. This is possible only when the big n is balanced from a decrease of p.

Demonstration:
$X \sim B(n, p)$ n "big"   p "little"
I set $np = \lambda$ with $\lambda > 0 \longrightarrow p = \frac{\lambda}{n}$

$$P(X = i) = \left(\begin{array}{c} n \\ i \end{array} \right) \cdot p^i \cdot (1 - p)^{n-i}$$

$$= \left(\begin{array}{c} n \\ i \end{array} \right) \cdot (\frac{\lambda}{n})^i \cdot (1 - \frac{\lambda}{n})^{n-i}$$

$$\frac{n(n-1)...(n-i+1)}{i!} \cdot (\frac{\lambda}{n})^i \cdot (1 - \frac{\lambda}{n})^{n-i}$$

I switch the denominator $i!$ with n and I do a limit for $n \to +\infty$

$$\lim_{n+\infty} \left(\frac{n}{n} \cdot \frac{n-1}{n} \cdot ... \cdot \frac{n-i+1}{n} \cdot \frac{\lambda^i}{i!} \cdot \frac{(1-\frac{\lambda}{n})^n}{(1-\frac{\lambda}{n})^i} \right)$$

the first part tends to 1, the fraction with λ doesnt depend from n and we ignore it, we have to determine the last one:

we know that:
$$\lim_{n \to +\infty} (1 + \frac{1}{n})^n \to e^1$$

So the numerator tends to $e^{-\lambda}$ and the denominator tends to 1.
The result of the limit is:
$$\frac{\lambda^i}{i!} \cdot e^{-\lambda}$$

This is literally the probablity mass function of a poisson random variable, so it's demonstrated.

6.6 Exponential model

This model is used for the time that pass between two events in the case the events exist based of some hypothesis.

$X \sim E(\lambda)\lambda \in R^+$

$f_x(x) = \lambda \cdot e^{-\lambda x} \cdot I_{\{R^+\}}(x)$

$\mathcal{E}(X) = \frac{1}{\lambda}$
$Var(X) = \frac{1}{\lambda^2}$

The exponential model has something in common with the Geometric one, such as the Probability mass function graph of the geometric with the probability density function graph of the exponential. Both models has the Memorylessness property.

Demonstration:

$$F_x(x) = 1 - e^{-\lambda x} \rightarrow P(X > x) = 1 - F_x(x) = e^{-\lambda x}$$

$$P(X > s + t) = e^{-\lambda(s+t)} = e^{-\lambda s} \cdot e^{-\lambda t}$$

$$P(X > s + t) = P(X > s)P(X > t)$$

$$P(X > s) = \frac{P(X > s + t)}{P(X > t)} \rightarrow Memorylessness \frac{P(X > s \cap X > t)}{p(X > t)}$$

And so: $P(X > s + t | x > t) = P(X > s)$
Demonstrated.

7 The Gaussian/Normal model

$X \sim G(\mu, \sigma) \quad \mu \in R \quad \sigma \in R^+$

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

There's no need of a index set because D_x is defined on all R.

$\mathcal{E}(X) = \mu$
$Var(X) = \sigma^2$

If I calculate $aX + b := Y$
$X \sim G(\mu, \sigma) \rightarrow Y \sim G(aX + b, \sigma)$

If I consider two gaussian variables with independence:
$X_1 \sim G(\mu_1, \sigma_1)$ and $X_2 \sim G(\mu_2, \sigma_2)$
$X_1 + X_2 \sim G(\mu_1 + \mu_2, \sqrt{\sigma_1^2 + \sigma_2^2})$

7.1 Standardization

The standardization is defined as follow:

$$Y = \frac{X - \mu}{\sigma}$$

The result is again a Gaussian variable with $\mathcal{E}(X) = 0$ and $Var(X) = 1$.
This type of gaussian has a particolar name: standard normal.
$X \sim N(0, 1)$

summarizing the Standard Normal:

$$Z := \frac{X - \mu}{\sigma} \sim N(0, 1)$$

$$P(Z > x) = 1 - P(Z \leq x) = 1 - \Phi(x)$$

7.2 Reproduction

If we have $X_1, ..., X_n$ gaussian variables with independence

$$X_i \sim G(\mu_i, \sigma_i)$$

we define

$$Y = \sum_{i=1}^{n} X_i \longrightarrow \mathcal{E}(Y) = \sum_{i=1}^{n} \mu_i \quad Var(Y) = \sum_{i=1}^{n} \sigma_i^2$$

$$Y \sim G\left(\sum_{i=1}^{n} \mu_i, \sqrt{\sum_{i=1}^{n} \sigma_i^2}\right)$$

This behavior is similar to what we have seen with the binomial variable.
The Poisson model has this exact property too.

8 Central Limit Theorem

Hypothesis:
$\quad X_1, ..., X_n$ random variables with independence and with same distribution
$\quad \forall i \quad \mathcal{E}(X_i) = \mu \quad Var(X_i) = \sigma^2$

If I sum all of these variables:

$$\sum_{i=1}^{n} X_i \dot\sim N(n\mu, \sqrt{n}\sigma)$$

If I do the standardization:

$$\frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n}\sigma} \dot\sim N(0,1)$$

This approximation is better when the number of random variables is high. Now, if we want to be more formal:

$$\lim_{n \to +\infty} P\left(\frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n}\sigma} \leq x\right) = \Phi(x)$$

Where $\Phi(x)$ is the cumulative ditribution function of the Standard Normal.

9 Law of larger number

This is the only notion of Inferential statistic parametric you will read in this document. There are other concepts that you can read on the Jupyter Notebook.

This law has two forms: strong and weak.

The strong one:

$$P\left(\lim_{n \to +\infty} \bar{X}_n = \mu\right) = 1$$

More the number of elements in the sample we have, less the $\bar{X}$ is a random quantity, it becomes a constant which is exactly what I want to esteem.

The weak one:

$$\lim_{n \to +\infty} P\left(|\bar{X}_n - \mu| > \epsilon\right) = 0$$

I pin an epsilon that tell me when the error is too much. This is the probability that the error in the esteem is too much and with n that tends to infinity this probability tends to zero $\forall \epsilon > 0$.

10 Conclusion

Other concepts will be inside the jupyter notebook I put in this repository. I ignored many details such as the demonstration of how some $\mathcal{E}(X)$ and $Var(X)$ have those values and why.