Introduction to statistics

TODO subtitle

Caio Volpato (caioau)
caioau.keybase.pub → caioauheyuuiavlc.onion
210B C5A4 14FD 9274 6B6A 250E **EFF5 B2E1 80F2 94CE**

Summary:

- Probability concepts
- Discrete distributions.
- Continuous distributions.
- Calculations on the Normal distribution.
- Convergence
- Inference

Motivation

Figure 1: Dados apontam ... (data shows ...)

Basic concepts of probability:

Sample space Ω

It's the set of all the possible outcomes of a experiment, denoted by S or Ω

Event

It's a subset of the sample space.

Basic concepts of probability:

Probability (Definition):

Given a experiment with a sample space Ω and a class of events $\mathcal{A}$, the probability denoted by $\mathbb{P}$ is a function which has $\mathcal{A}$ as domain and associate a numerical value between $[0, 1]$ as image.

Probability properties:

1. $\mathbb{P}(\Omega) = 1$ and $\mathbb{P}(\emptyset) = 0$
2. $0 \leq \mathbb{P}(A) \leq 1$, for every event A
3. For any sequence of mutually exclusive events $A_1, A_2, \ldots$ that's events that $A_i \bigcap A_j$ when $i \neq j$ we have that:

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$$

Basic concepts of probability:

Event independence:

Two events are independent when the occurrence of the first does not affect the probability of ocurrence of the second.
Two events A and B are independent if:

$$\mathbb{P}(A \bigcap B) = \mathbb{P}(A)\mathbb{P}(B)$$

Conditional Events:

The probability of a event A to occur given that the event B occurred is:

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

Basic concepts of probability:

Bayes theorem:

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

General case:

$$\mathbb{P}(A_i|B) = \frac{\mathbb{P}(B|A_i)\mathbb{P}(A_i)}{\sum_{j=1}^{n} \mathbb{P}(B|A_j)\mathbb{P}(A_j)}$$

Bayes example (from Veritasium):

You are felling sick, so you go to the doctor, there you run a battery of tests. After getting the results you tested positive for a rare disease (affects 0.1% of the population), the test will correctly identify that you have it 99% of the times.
What's the chances that you actually have the disease? 99%?

Bayes example Solution

Let's denote the event of you have the disease H (stands for hypothesis, the prior) and the test been positive denoted by E (stands for evidence), so we have: $\mathbb{P}(H) = 0.001$ and $\mathbb{P}(E|H) = 0.99$

$$\mathbb{P}(H|E) = \frac{\mathbb{P}(E|H)\mathbb{P}(H)}{\mathbb{P}(E)} = \frac{\mathbb{P}(E|H)\mathbb{P}(H)}{\mathbb{P}(H)\mathbb{P}(E|H) + \mathbb{P}(H^C)\mathbb{P}(E|H^C)} =$$

$$= \frac{0.99 \cdot 0.001}{0.001 \cdot 0.99 + 0.999 \cdot 0.01} = 0.09 = 9\%$$

What if you test again and it's also positive? You can just take the posterior probability we just calculated and use as a prior:

$$= \frac{0.99 \cdot 0.09}{0.09 \cdot 0.99 + 0.91 \cdot 0.01} = 0.907 \approx 91\%$$

- Awesome video: A visual guide to Bayesian thinking

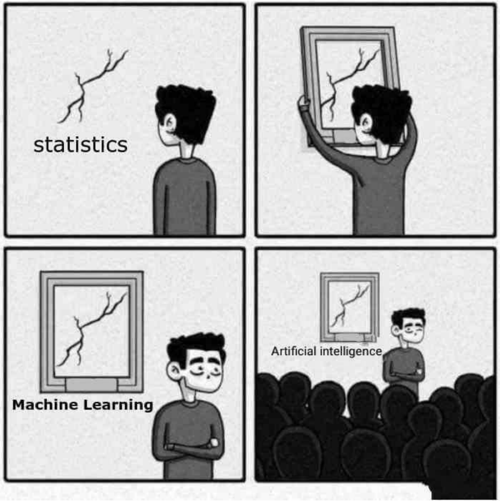

Figure 2: Credits: sandserifcomics

Random Variable (RV)

Consider a experiment with a sample space Ω associated with it. A function that maps each element $\omega \in \Omega$ to a Real number such that $[w \leq X]$ it's called random variable (RV) $(X : \Omega \rightarrow \mathbb{R})$

- Example: Imagine a experiment that consist of 3 consecutive fair coin tosses, so the sample space of this experiment is: S = {(H,H,H), (H,H,T) , . . . (T,T,T)} . Now we want to create a random variable X that counts the number of heads in each outcome, so X((H,H,H)) = 3 and X((H,H,T)) = 2.

Random Variable:

Probability Mass Function (PMF):

$$f_X(x) = \mathbb{P}[X = x] = \mathbb{P}[\{\omega \in \Omega : X(\omega) = x\}]$$

Probability Density Function (PDF)

$$\mathbb{P}[a \leq X \leq b] = \int_a^b f(x)dx$$

Cumulative Distribution Function (CDF)

$$F_X(x) = \mathbb{P}[X \leq x]$$

Expectation:

- Discrete : $\mathbb{E}[X] = \sum x\mathbb{P}(X = x)$
- Continuous: $\mathbb{E}[X] = \int_{-\infty}^{\infty} xf(x)dx$

Variance:

$$\mathbb{V}[X] = \sigma_X^2 = \mathbb{E}[X^2] - \mathbb{E}^2[X]$$

Sample mean:

$$\overline{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Sample variance and standard deviation:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

Standard deviation $= s$

Discrete distributions

Bernoulli:

Consider a experiment with has two possible outcomes: success (X=1, with probability p) or failure (X=0), this random variable is called Bernoulli, the PMF is:

$$\mathbb{P}(X = k) = p^k (1 - p)^{1-k}$$

Binomial:

Now consider a Bernoulli experiment conducted n times, let X be the random variable that represents the number of successes, X is called Binomial, the PMF is:

$$\mathbb{P}(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$$

Discrete distributions

Geometric:

Again consider a Bernoulli experiment conducted n times, but the first n-1 are failures and the last nth is a success. Let X be number of tries , which is called Geometric, the PMF is:

$$\mathbb{P}(X = k) = (1 - p)^k p$$

- A important property is that Geometric distribution is the only discrete distribution that is **memoryless**.

Poisson:

A random variable which value can assume 0,1,2 ... is called Poisson with $\lambda > 0$ parameter if your PMF is:

$$\mathbb{P}(X = k) = \frac{e^{-\lambda}\lambda^k}{k!}$$

graficos distribuicoes discretas

Continuous distributions

Normal (or Gaussian, bell curve):

A continuous real random variable is called Normal with $\sigma^2 > 0$ (squared scale), $\mu \in \mathbb{R}$ (location) parameters if your PDF is:

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

Continuous distributions

Exponential

A continuous **positive** random variable is called Exponential with $\lambda > 0$ (rate or inverse scale) parameter if your PDF is:

$$f(x) = \lambda e^{-\lambda x}$$

Important property: Exponential and Geometric (discrete) distribution are the only distributions that are **memoryless**.

Memoryless property:

$$\mathbb{P}[X > x + y \mid X > y] = \mathbb{P}[X > x]$$

So no matter how much time has passed it's like the process is starting from beginning.

Continuous distributions

Pareto

A continuous $x \in [x_m, \infty)$ random variable is called pareto with $x_m > 0$ (scale) and $\alpha > 0$ (shape) parameters if your PDF is:

$$f(x) = \frac{\alpha x_m^\alpha}{x^{\alpha+1}}$$

Zipf is the discrete distribution of pareto
Pareto is a **heavy tailed** distribution: It means it goes to zero slower.

Pareto principle (80-20 law):

The pareto principle states that 80% of results is caused by 20% of the effects, for example Wealth distribution, software bugs and so on.
It's a particular pareto distributed values when $\alpha = 1.161$

graficos distribuicoes continuas

meme know distributions

Calculations on the Normal distribution

tabela e calcular python, excel, normalizar (z score), regra 65,95,99

Assumptions on distribution choice

antes se eh discreta ou continua
descrever a natureza da normal, exp, pareto
Discretas:
Bernouli: resultado dicotonimo , exemplo: moeda, homem ou mulher, sim ou nao. Voto em 2o turno
Binom: quantidade de sucesso dado um numero fixo de experimento independente; Dados 20 dispositivos independentes, depois de muitas horas, qual a prob de 15 apresentarem defeito.
Geometrica: Numero de falhas ate primeiro sucesso. Exemplo loteria: Dado p ser 1 em 1 milhao qual a prob de ganhar depois de 3 tentativas?
Poisson: Contagem de pessoas inscritas em algum programa que desistem.
Continuas
Normal : Sem restricao de valores (pode ser positivo ou negativo). Exemplo: Altura de criancas do mesmo sexo e idade (funciona bem pra qui quadrado).

order statitics

defs, min, max, median, q1,q3, IQR, pq? estat robusta, boxplot

Convergence

defs, lei dos grandes numeros, teorema do limite central

Inference

metodo da maxima verosimilhanca, e grafico qxq (scipy.stats.probplot) (colocar exemplo com distribuicao errada: t student e fitar com a normal)

max veross: find the most likely parameter value, given data. That is, given a prob description of data, find the optimum value for that data (derivatives).

Further reading:

- Portal action (pt)
- stat cookbook
- havard youtube (https://www.youtube.com/playlist?list=PL2SOU6wwxB0uwwH80KT
- ross, barry james, meyer
- khan academy
- http://www.randomservices.org/random/
- divulgacao: pizza de dados, senhora toma cha, andar do bebado, /r/dataisbeautiful