Notes of probability and inferential statistics as background to linear regression anaylsis

using python.

Iván Andrés Trujilllo Abella

Facultad de Ingenieria
Pontificia Universidad Javeriana

What is a random variable?

Introduction

In several trials when you specified same and invariant conditions always have different results!.

Expected value

is defined as the mean of a random variable

Properties

where f(x) it is the probability density function of the random varible x

$$E(x) = \int_{-\infty}^{\infty} xf(x).$$

Propertie

proof with integrals that

$$E(aX + b) = aE(x) + b \tag{1}$$

Variance

$$\begin{aligned}\sigma^2 &= E[(X-\mu)^2] \\ &= E(x^2) - \mu^2 \\ &= E(x^2) - [E(X)]^2\end{aligned} \tag{2}$$

remember that $\sum_x f(x)x = \mu$ and $\sum_x f(x) = 1$

$$\begin{aligned}\sigma^2 &= \sum_x (x-\mu)^2 f(x) \\ &= \sum (x^2 f(x) - 2x\mu f(x) + \mu^2 f(x))\end{aligned} \tag{3}$$

Applying the algebra we find $\sum x^2 f(x) - \mu^2$.

μ and σ^2 estimators

$\bar{X}$ is a estimatior of μ and S^2 is a estimatior of population variance σ^2.

$$\bar{X} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{4}$$

$$S^2 = \frac{\sum (x_i - \mu)^2}{n-1} \tag{5}$$

Why divide in S^2 by $n-1$?

Unbiasedness

This is very important to show the concept of **unbiased estimator** that it is a important property of some parameters, for instance in linear regression, we need that our parameters are **unbiased**.

Variance estimators

We can check mathematically (**See here**) that S_N^2 is a biased estimator

$$S_n^2 = \frac{\sum(x_i - \bar{x})^2}{n} \tag{6}$$

and that S_{n-1}^2 is unbiased.

$$S^2 = \frac{\sum(x_i - \bar{x})^2}{n-1} \tag{7}$$

However to see a simulation **click here**

Degree of freedom

For instance if said that $\bar{x} = 10$, and if we take for values 5 for instance:

$$\begin{aligned} 3 - 10 &= -7 \\ 10 - 10 &= 0 \\ 14 - 10 &= 4 \\ 11 - 10 &= 1 \\ \mathbf{x} - 10 &= \mathbf{2} \end{aligned} \tag{8}$$

Notice, that the first four number could be any value, however the last value is attached to fill the propertie of sum equal to zero, therefore **x** muest be equal to 12.
How many degree of fredoms we have?

Unbiased estimator

Remember that all this is very important to another concepts neccesary to understand some interesting properties of another tecnhiques as **linear regression coefficients estimation**

$E(s^2) = \sigma^2$

We are going to probe that $\frac{(\sum x_i - \bar{x})^2}{n-1}$ fill that the expected value is equal to the population parameters. We are going to use the following notation $X_n \sim (\mu, \sigma^2)$ to indicate n random and independent observations drawn from a population with mean μ and variance σ^2.

$E(s^2) = \sigma^2$

proof

To make this proof we need related the sample variance with the population variance,

$$E(x^2) = \sigma^2 + \mu^2$$

remember that $var(\bar{x}) = \frac{\sigma^2}{n}$

$$E(\bar{x}^2) = \frac{\sigma^2}{n} + \mu^2$$

now aplying the expected value definition (and properties) to variance sampling estimator we have:

$$E[s^2] = \frac{1}{n-1}\left(nE(x_i^2) - nE(\bar{x}^2)\right) \tag{9}$$

reeplacing and applyinng basic algebra operations then $E(s^2) = \sigma^2$.

Excersices

- Proof that sample mean is unbiased
- Proof why $\frac{\sum(x_i-\bar{x})^2}{n}$ is biased.

Computational lab

- Show with a normal distribution that $E(s^2) = \sigma^2$.

Theorem 1

X is a random variable with a pdf $f(x)$ then μ of $g(x)$ is

$$\mu_{g(x)} = E(g(x)) = \sum g(x)f(x) \tag{10}$$

Example of income and the probability of sell a product.

Theorem

X is a random variable with pdf $f(x)$ then then variance of $g(x)$ will be:

$$\sigma^2_{g(x)} = E((g(x) - \mu_{g(x)})^2 \tag{11}$$

this is equation is derived of the definition of variance of a random variable, remember that $g(x)$ is a random variable with mean $\mu_{g(x)}$.

Join distribution

until now we try Ω in $\mathbf{R}^1$ and we can be interested in find the probability of occurrence of two simultaneous random variables.

$$f(x, y) = P(X = x, Y = y) \tag{12}$$

Some intuitive properties are:

- $f(x, y) \geq 0$
- $\sum_x \sum_y f(x, y) = 1$

Excercise

Suppose the bag model with n balls and there there are r balls and w balls where $r + w = n$ find the probability of get x, y balls respectively.

Marginal distribution

From the joint distribution $f(x,y) = P((X = x) \cap P((Y = y))$

$$g(x) = \sum_{y} f(x,y) \tag{13}$$

$$h(y) = \sum_{x} = f(x,y) \tag{14}$$

Expected value of two random variables

let be X, Y two random variables with joint probability function distribution $f(x, y)$ the mean of $g(X, Y)$ is:

$$\mu_{g(X,Y)} = E[g(X, Y)] = \sum_x \sum_y g(x, y) f(x, y) \tag{15}$$

Covariance

Insights

$$\begin{aligned}\sigma_{X,Y} &= E[(X-\mu_x)(Y-\mu_y)] \\ &= \sum_x \sum_y (x-\mu_x)(y-\mu_y) f(x,y)\end{aligned} \tag{16}$$

is a measure of association between two variables

Propertie

Proof the following propertie $E(X+Y) = E(X) + E(Y)$.

$$P(x) = \sum_y P(x,y)$$
$$P(y) = \sum_x P(x,y) \tag{17}$$
$$E(X+Y) = \sum (x+y)P(x,y)$$

Take in mind, that we have that $\sum_{x,y} = \sum_x \sum_y$

$$E(X+Y) = \sum_{x,y}(x)P(x,y) + \sum_{x,y}(y)P(x,y) \tag{18}$$
$$E(X+Y) = E(X) + E(Y)$$

Note from before definition that x and y correspond to specific values.

proof the following propertie: $E(XY) = E(X)E(Y)$.

$$\begin{aligned} E(XY) &= \sum_{x,y} xyP(x,y) \\ &= \sum_{x,y} xyP(x)p(y) \\ &= \sum_{x,y} xP(x)yP(y) \\ &= \sum_{x} xp(x) \sum_{y} yP(y) \end{aligned} \tag{19}$$

Variance

From here we refer we refer to the population mean as μ and estimated mean as $\hat{\mu}$.
The definition of variance.

Variance

important the following properties, about variance:

- $var(x) >= 0$
- $var(c) = 0$
- $var(cx) = c^2 var(x)$
- $var(x + c) = var(x)$
- $var(x) = E(x^2) - E^2(X)$
- $Var(x + y) \neq var(x) + var(y)$

Propertie

$$\begin{aligned} var(c) &= E[c - E(c)]^2 \\ &= E(c - c)^2 = 0. \end{aligned} \tag{20}$$

Note remember that $E(c) = c$.

Propertie

$$
\begin{aligned}
var(cx) &= E[(cx - E(cx))^2] \\
&= E[(cx - cE(x))^2] \\
&= E[c^2(x - E(x))^2] \\
&= c^2 E[(x - E(x))^2] \quad = c^2 var(x)
\end{aligned} \tag{21}
$$

Propertie

Proof that $var(x + c) = var(x)$.

$$\begin{aligned} var(x + c) &= E[((x + c) - E(x + c))^2] \\ &= E[(x + c - E(x) - c)^2 \\ &= E[(x - E(x))^2] \\ &= var(x) \end{aligned} \tag{22}$$

Propertie

$$
\begin{aligned}
var(x) &= E[(x - E(x))^2] \\
&= E[x^2 - 2xE(x) + E^2(x)] \\
&= E(x^2) - 2E(x)E(x) + E^2(x) \\
&= E(x^2) - 2E(x) + E^2(x) \\
&= E(x^2) - E^2(x)
\end{aligned} \tag{23}
$$

Propertie

proof that $var(mX + nY) = m^2 var(X) + n^2 var(Y) + 2mncov(X, Y)$.

$$\begin{aligned} var(mX + nY) &= E[mX + nY - E(mX + nY)]^2 \\ &= E[m(X - E(X)) + n(Y - E(Y))]^2 \\ &= E[m^2(X - E(X))^2 + n^2(Y - E(Y))^2 + \\ &\quad 2mn(X - E(X))^2(Y - E(Y))] \\ &= m^2 var(X) + n^2 var(Y) + 2mncov(X, Y) \end{aligned} \tag{24}$$

Note that if $cov(X, Y) = 0$ and $m, n = 1$ then:

$$var(X + Y) = var(X) + var(Y). \tag{25}$$

Propertie

With all properties we can derive another important.

- $var(mX + b) = m^2 var(X)$

Propertie

$$
\begin{aligned}
var(\bar{x}) &= \frac{1}{n^2}\sum var(x_i) \\
&= \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n} \\
sdv(\bar{x}) &= \frac{\sigma}{\sqrt{n}}
\end{aligned}
\tag{26}
$$

Central limit theorem

Suppose that we draw samples from a population and get the mean of each sample for instance:

$$\begin{aligned}
\bar{x}_1 &= \frac{1}{n}\sum(x_1^1 + x_2^1 + ... + x_k^1) \\
\bar{x}_2 &= \frac{1}{n}\sum(x_1^2 + x_2^2 + ... + x_k^2) \\
&\vdots \\
\bar{x}_j &= \frac{1}{n}\sum(x_1^j + x_2^j + ... + x_k^j)
\end{aligned} \tag{27}$$

Thus $\bar{x}_j$ is the $j - th$ sample mean composed of k terms.

Insights

Think in θ as a unknown paramater, but in **Classical statistics** this parameters is **fixed** our work is find it.
We need a **estimator**, that process the data observed, in other words a **estimator** is a function $(\hat{\Theta})$ aiming to find θ. In the process $\hat{\theta}(x)$ is a estimate of a particular set(x) of X.

Think in variance that we have two possible estimators(which are?), and $\hat{\theta(x)}$ could be a specific value of them.

Estimation

Now we are interested in find a **estimator** $\hat{\Theta}$ that

$$\min(\hat{\theta} - \theta) \tag{28}$$

Namely, we are finding that the estimated value of the unkonw parameter is small in relation with the true value.

Approaches

What approaches we have to design a estimator?

Estimation of mean

A estimator of mean is $\hat{\Theta} = \frac{\sum x_i}{n}$. note that $\hat{\Theta}$ is random variable.
we are interested that a estimator have the following properties; **Unbiased** and **consistency**.

- $E(\hat{\Theta}) = \theta$
- By Weak Law Larger Numbers $\hat{\Theta} \to \theta$ when $n \to \infty$ (more data, higher accuracy in the estimation).

The error of a estimator

Remember that $\hat{\Theta}$ is a random variable, Therefore Mean squared error(MSE) $E[(\hat{\Theta} - \theta)^2]$, could be expresed as

$$E[(\hat{\Theta} - \theta)^2] = var(\hat{\Theta} - \theta) + (E[(\hat{\Theta} - \theta])^2 \tag{29}$$

Applying properties of variance (variance plus or minus a constant is the variance of the r.v).

$$var(\hat{\Theta}) + (E[(\hat{\Theta} - \theta])^2 \tag{30}$$

$$MSE = E[(\hat{\Theta} - \theta)^2] = var(\hat{\Theta}) + bias^2 \tag{31}$$

Analysis about mean estimator

We known that the

$$MSE(\hat{\Theta}) = \frac{\sigma^2}{n} + 0 \tag{32}$$

According to the consistency property then the error is lesser wich the size of data increase $n \to \infty$.
it is important note here that if bias is equal to zero, then we avoid that the error rely on about θ.

standard error

Is a important concep in estimator

$$SE = \sqrt{var(\hat{\Theta})} \tag{33}$$

a geat **SE** is

not divide by $n-1$

Here we are seeing the variance of the estimator.

$$\begin{aligned} E(\hat{\Theta}) = E(\bar{x}) &= E\left(\frac{\sum(x_i - \bar{x})^2}{n}\right) \\ &= \frac{1}{n}E\left(\sum x_i^2 - n\bar{x}^2\right) \end{aligned} \tag{34}$$

Applying some algebraic operations we have, remember as compute $E(x_i^2)$ and $E(\bar{x}^2)$.

$$E(\hat{\Theta}) = \frac{\sigma^2(n-1)}{n} \tag{35}$$

Confidence interval

For what it is useful confidence interval?
Now assume that $\alpha \in [0, 1]$

$$P(\hat{\theta}_{low} < \theta < \hat{\theta}_{upper}) = 1 - \alpha. \tag{36}$$

Note that the interval is also random.

CI

Remember that CI is aming to find θ therefore is a mistake said that $(1-\alpha)*100$ times the parameter falls inside the interval (There is a common mistake). How find $\hat{\theta}_{low}, \hat{\theta}_{upper}$
A better approximation is that the probability of the interval contain the parameter is $1-\alpha$.

Interpretation of CI

it is uses the interval in n sampling evaluations then $(1-\alpha)$ times the interval contain θ.

Excersices

- For what it is useful
- what is the relation behind CI and Hypothesis testing

Insight

We need a random variable that its definition contain θ but its probability distribution function not rely on over θ.

Mean

Unkonwn σ^2 and **known** σ^2.

CI(mean) with known σ^2

$X_n \sim N(\mu, \sigma^2)$, then

$$\frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1) \tag{37}$$

$$P(-1.96 \leq \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \leq 1.96) = 0.95 \tag{38}$$

before of some algebraic inequalities operations we have:

$$\bar{x} - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + 1.96\frac{\sigma}{\sqrt{n}} \tag{39}$$

See simulation here: CI simulation (click)
Which is the pivotal quantity?

CI(mean) with unknown σ^2

Before tackle this problem we need understand the t-student distribution.

t-student

A useful distribution, discover by William Gosset

Hypothesis testing

Hypothesis as idea or believe about a issue.

Null hypothesis

H_0 describe the current believe, and H_1 is a option if there is enough evidence to reject H_0.

Conditional probability

We can make the following lecture of conditional probability

$$P(A \mid B) \tag{40}$$

What is the probability of A occur given that B already happened.

Examples

A good lecture is **El azar en la vida cotidiana** alberto rojo, exposed basic examples for instance: *if we encounter in the street a woman that have two childrens, what is the probability of one of them is girl if the another is called* **pedro**?. The answer is $\frac{2}{3}$ how get this answer?

Disease diagnose and bayes

The sensitivity (The probability that test is positive given that the person really have the disease $P(+test \mid disease)$ for instance is 90% therefore if a person is positive test could be seen the life pass for eyes?, but we need is:

$$P(disase \mid +test) \tag{41}$$

Here appear Bayes theorem.

Bayes

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)} \tag{42}$$

Bayes in diagnose

Causality

In a causal sense the proper order is **Cause produce effect**, note here the cronological order, notice that an effect is associated with multiple causes.

Bayes in diagnose

In a medical sense we observe different sympstoms and we need diagnose could have:

$$P(cause \mid effect) = \frac{P(effect \mid cause)P(cause)}{P(effect)} \tag{43}$$

Bayes in diagnose

we can calculate $P(disease \mid +test)$ with the following formula:

$$\frac{P(+test \mid disease)}{P(+test \mid disease)P(disease) + P(positive \mid disease^c)P(disease^c)} \tag{44}$$

The denominator is calculated with total law of probability.

Prosecutor fallacy

Innocent person have a probability of $\frac{1}{10.000}$ that the evidence be damning, namely:

$$P(evidence \mid inoccent) = \frac{1}{10.0000} \tag{45}$$

Bayes theorem in prosecutor fallacy

$$P(innocent \mid evidence) = \frac{P(evidence \mid inoccent)P(inoccent)}{P(evidence)} \tag{46}$$

Sally clark's history

- in 1996 born her first son and die few months later
- in 1997 born her second son and also die few months laters
- She was the last person with stay with both childrens.

Meadow (He was a pediatrician) argument was that the probability of a child die by sudden death is:

$$P(S) = \frac{1}{8573} \tag{47}$$

Sally clarks history

- Clark was convicted in 1999
- Clark was released in 2003.
- Clak died four years later by alcohol intoxication.

Not is enough collecting data

Sally clark's history

Remember that

$$P(A \cap B) = P(A)P(B) \tag{48}$$

and therefore:

$$\begin{aligned} P(A \mid B) &= \frac{P(A \cap B)}{P(B)} \\ P(A \mid B) &= \frac{P(A)P(B)}{P(B)} = P(A) \end{aligned} \tag{49}$$

This is very important because the events are not independent.
Given that, the second child die is more probable if the first die suddently (could be appear genetic predisposition), namely: $P(S_2 \mid S_1) > P(S_1)$.

Sally clark's history

Fallacy?

Asumming that sally is innoncent then the trial uses:

$$P(evidence \mid innocent) = \frac{1}{8573} . \frac{1}{8573} \tag{50}$$

what is wrong? what formula you need?.

Bayes in ML lingo

$$P(Y \mid X) = \frac{P(X \mid Y)P(Y)}{P(X)} \tag{51}$$

we can descompose (**??**) as :

- $P(Y \mid X)$ posterior
- $P(Y)$ prior
- $P(X)$ evidence
- $P(X \mid Y)$ likelihood

Excersie

rolling dice

what is the probability of have a odd in a rolling dice if the number given that we had is equal or greater than four.

Excercises of probability

Spam detector

$$P(broma \mid spam) = b \tag{52}$$

b is the conditional probability, the probability that word **broma** is contained given belong to spam.

Diagnostico de zika

Resulta que la $P(+test \mid zika) = 0.98$, pero al año solo se reporta en el 3% de la población el contagio, y se conoce que el test da positivo el 80% de las veces independiente si se tiene o no zika.

Zika

$$P(zika \mid test+) = \frac{P(+test \mid zika)P(zika)}{P(test+)} \tag{53}$$

Naive Bayes classifier

Messages	Category
This message contain the words; a, b, c	No spam
This message contain the words; a, b	No spam
This message contain the words; a, c	No spam
This message contain the words; a, e	Spam
This message contain the words; b, d	Spam
This message contain e	Spam
This message contain d, f	Spam

Split data by category

Messages	Category
This message contain the words; a, e	Spam
This message contain the words; b, d	Spam
This message contain e	Spam
This message contain d, f	Spam

Table: Spam

Messages	Category
This message contain the words; a, b, c	No spam
This message contain the words; a, b	No spam
This message contain the words; a, c	No spam

Table: No spam

Conditional probability of the words according to the category

The **prior probability** of $word_a$ is:

$$P(word_a) = \frac{4}{7} \tag{54}$$

$$P(word_a \mid Spam) = \frac{1}{4} \tag{55}$$

$$P(word_a \mid Spam^c) = \frac{3}{3} \tag{56}$$

Notice, that the probability that the $word_a$ appear in a spam message is greater than in a not spam message. What are the probabilities for another words?.

Scoring

$$P(spam) \prod_{i}^{n} (word_i \mid spam) \tag{57}$$

according to (**??**) then we can have a problem if in the training data sets there are not a set of words that appear in another dataset then the conditional probabilities are equal to zero, to tackle this we could add α(integer) count to each category.

There are something behind

There is a bias in the order

Naive ignore the language! Whats means that Naive have low variance?

Naive Bayes in Bankruptcy

Sector	Income	Bankruptcy
Financial	High	No
Financial	Small	Yes
Agricultural	Small	Yes
Agricultural	High	Yes
Financial	Small	No
Financial	Small	No
Agricultural	High	No
Agriulcural	Small	Yes
Agricultural	Small	No

Table: Dataset Bankruptcy

Notice, that we want to uses this dataset to answer the question $P(Bankruptcy \mid Sector \cap Income) = P(Bankruptcy \mid Sector, Income)$.

Example

Naive Bayes Classifier

let's consider that we want guess the probability that a occur a **default** given that te firm belong to the **agricultural** sector and its size or income is **small**:

$$P(Banruptcy = Yes \mid Sector = Agricultural, Income = Small) \tag{58}$$

Therefore we could use **Bayes theorem** using (**??**) to estimate:

$$P(Sector = Agricultural, Income = Small \mid Bankruptcy = Yes) \tag{59}$$

Split data

Naive Bayes classifier

Sector	Income	Bankruptcy
Financial	Small	Yes
Agricultural	Small	Yes
Agricultural	High	Yes
Agricultural	Small	Yes

Table: Banruptcy = Yes

With this data we can get:

$$P(Sector = Agricultural, Income = Small \mid Bankruptcy = Yes) = \frac{2}{4} = \frac{1}{2}.$$

a lot of if!

Naive bayes classifier

Could be impractical find concrete combinations of values($X_1 = x_1, X_2 = x_2, ..., X_n = x_n$), with a great number of features(zeros could be appear). if Assume that the features are indepedent, we dont need be corncened by combinatios! and therefore:

$$P(X_1 = x_1, X_2 = x_2, ..., X_n = x_n \mid Y = y_0) = \prod_{i=1}^{n} P(X_i = x_i \mid Y = y_0)$$

arg max

Assume a function $f : X \rightarrow Y$ then $\arg\max_x$ is

$$\arg\max_x f(x) = \{x \mid \forall x' : f(x') \leq f(x)\} \tag{60}$$

in other words are set of values in domain that give us the maximun value in the function, what is the $\arg\max_x(1 - |x|)$?.

$$\max_x f(x) = \{f(x) \mid \forall f(x') : f(x') \leq f(x)\} \tag{61}$$

$$f\left(\arg\max_x f(x)\right) = \max_x f(x) \tag{62}$$

$$\arg\max_{y} P(y|X) \tag{63}$$

Pseudocode

```
Count the number of classes,
get the prior probabilities,
determine likelihoods table;
    how many times appear the feature (i) in each class.
```

Nb is the baseline model...

NB

NB could be:

- **Multinomial NB:**
  Allow us model multiple occurrences of the feauture for instance the number of times that occur a especific word.
- Binomial NB (Allow us model if the word occur or not).
- Gaussian NB