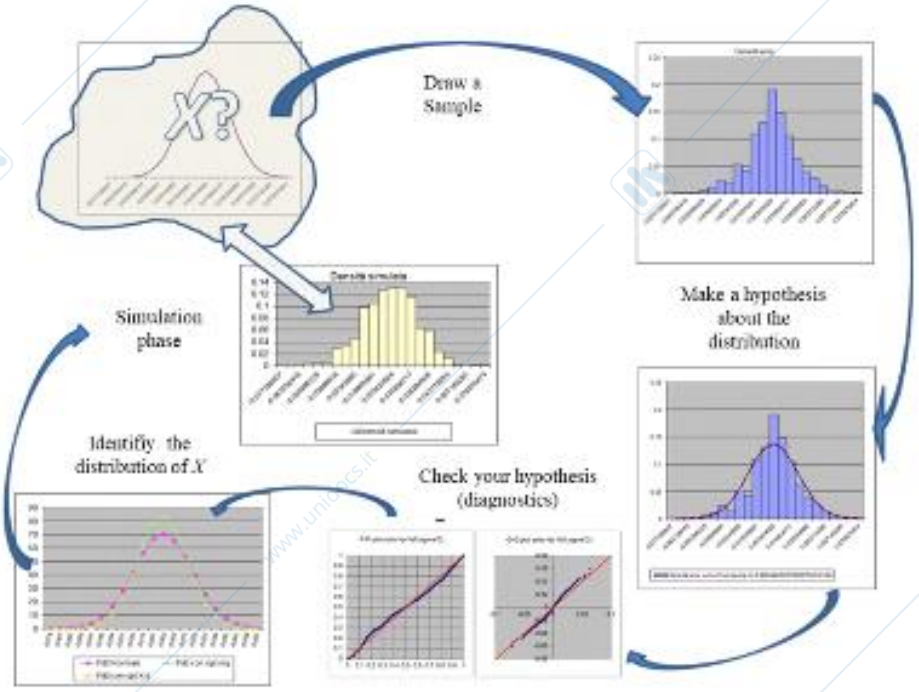
Theory

We'll try to understand the aim of making a good analysis process in reading a database

Q. Why are we willing to make data analyis? What are the aims ?

Q. Are basic descriptive statistics miniumally useful to take decisions?



Q1? It's such a philosophical topic but it's very important.
Q2? = is it possible to manage a company just computing an average? No!

We are interested in the interpretation of the random variable X (that represents for instance the number of accidents). The difficult point is that it does really exist in reality, but we want to study it ex ante (before its realizations) because we need that in order to consider the risk that the company will probably have to bear, in order to compute premiums (ex ante, at inception!)
Typically, so we can try to exploit some past information: we will use a sample trough a database we have (for instance ISTAT database)
Starting from our database we can compute some basic descriptive statistics (avg, ...), and draw an picture/histogram (for now, with a database we are in the discrete world, because we have some recordings)
According to the result we graphically obtain the aim is to understand what is the stochastic mechanism that generates those data (for instance: when u collect the data of the returns of an investment, drawing the histogram u expect a sort of Gaussian distribution, not exactly it but quite close). So we won't get exactly a distribution as the Gaussian, Gamma and so on but something that remind them: we have to "chose", to make an hypothesis about the distribution we think that can descript the data (maybe others use another hypothesis/idea) ➔ The Poisson distribution in general is good to describe the number of accidents occurred... in fact it can be shown that the Poisson distribution fit the case of occurrences in a time period assuming that: Consider the binomial random variable $X \sim Bin(x; n, p)$. Let $\lambda = np$. If for $n \to \infty$, λ remains constant, then the pdf of X converges to $$p(x;\lambda) = \frac{e^{-\lambda}\lambda^x}{x!}$$ $p(x;\lambda)$ is the pdf of the Poisson r.v. with x=0,1,... 1. the probability of a single arrival in a very short time interval is constant 2. the probability of more than one arrival during time interval is negligible 3. the probability on an arrival during the time interval is independent of any occurrences in previous periods

The occurrences of accidents in time and space is often (not always) well suited by a Poisson distribution
It is the basic distribution to fit (e.g.) the number of accidents in non-life LoBs
The unique parameter (λ) corresponds to the average and the variance
Its interpretation is the average (rate) of accidents of a portfolio or, in case of premium personalization, the propensity to get an accident of each customer (each customer has a different risk... so if I want to reduce the risk of accidents I will select only students that in probability are expected to return 0 accidents or few few, but I'm not sure they don't really have them).

➔ The lognormal distribution in general is good to describe the costs of the accidents

Consider the normal (Gaussian) random variable $X \sim N(\mu, \sigma^2)$.
The transformation $Y = \exp(X)$ returns the Lognormal distribution

$$Y \sim N(\mu_L, \sigma_L^2)$$

where $\mu_L = \exp\left(\mu + \frac{1}{2}\sigma^2\right), \sigma_L^2 = \exp(2\mu + \sigma^2)\left(e^{\sigma^2} - 1\right)$

On the log (normal) scale $\mu = \ln\left(\frac{\mu_L^2}{\sqrt{\sigma_L^2 + \mu_L^2}}\right)$, $\sigma^2 = \ln(1 + \mathrm{cv}_L^2)$

$X \sim Gamma(\alpha, \beta)$.
The pdf of X is $f(x; \alpha, \beta) = \frac{1}{\beta^\alpha \Gamma(\alpha)} e^{-\frac{x}{\beta}} x^{\alpha-1}$

E(X)= $\alpha\beta$, Var(X)= $\alpha\beta^2$ ($\alpha = shape, \beta = scale$)
Then β = Var(X)/E(X), and α= E(X)^2/Var(X)

The Gaussian in insurance studies is not often used because of its -oo;+oo domain and because of its symmetry.

You do an hypothesis of the distribution, but then you work with unknown parameters... When we are willing to identify unknown parameters, you need to apply a basic strategy (a theoretical method) in order to understand them: the **likelihood principle** is in general the most used.

Let $X_1, \ldots, X_n$ *be n* i.i.d. r.v. with pdf $f(x, \theta)$. Let $x_1, \ldots, x_n$ be a sample.
The likelihood is

$$L(\theta) = \prod_{i=1}^{n} f(x_i, \theta)$$

or the log likelihood is

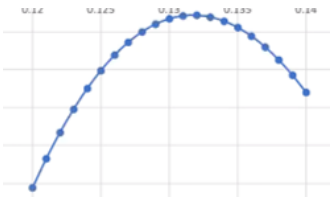
$$l(\theta) = \sum_{i=1}^{n} \log\left(f(x_i, \theta)\right)$$

In case of a discrete r.v. $L(\theta)$ is the probability of the sample *given* θ
The maximum likelihood estimate of θ is the solution to

$$\hat{\theta}: L(\hat{\theta}) = \sup_{\theta \in \Theta} L(\theta)$$

*Iid = independent and identically distributed random variable. So…
->independent = so when you collect data you are assuming that what happens to one unit is independent to the observation on the second statistical unit and on the other ones. …let's think about the insurance context. You are collecting data and typically even in the real contexts the basic assumptions is that each customers is supposed to be independent with respect to the others. In reality - thinking about the example of drivers, they are not really independent: They are in a quite similar context, with same signals and rules, but, in order to avoid higher complexity, we assume that there is in general independence.
->Identically distributed = our propensity to get an accident is depicted by the same random variable (for instance a Poisson with same parameters, supposed to be the same for each of us!)
**

AIM: if I have those data the probability of observing that data is the higher that I can have with respect to selecting any other lambda. So maximizing the overall likelihood of our sample we can compute the best parameter(s) that corresponds the maximum probability to have those data = If we collected again the data probably we would collect the same data! So the database observed there contains the maximum information for our data => so we want also the maximum likelihood of our data. We can substitute "likelihood" with "probability": the maximum probability of collecting that data.
HOW CAN WE DO THAT? we need to chose a "starting point/value" - subjective selection - for the parameter that we want to estimate, then we have to do some assumptions and then to maximize that function.



Now that we have an hypothesis, we need to check/validate it.
In fact now we have a theoretical distribution (our hp, not an histogram 'cause once we have chosen a distribution we are able to substitute our data with theoretical data from the distribution chosen) and an empirical one (our data sample, represented by an histogram) and we have to verify if our choice is really good for our data.
In order to do that we apply some ***diagnostics***

- ➔ (If we chose the Poisson distribution an easy check is to see – after the application of the likelihood principle – if the mean and the variance obtained are the same because so they are if the distribution is a Poisson!
- ➔ Q plot or P plot

So we can have done a good choice (next step) or a bad one (try again…)

Once you have checked your hypothesis you have identified a distribution that fits the data sample we have. So we can now say that the stochastic process that generates the data works as a … distribution.

The power of this process is that so X can be used also to simulate the future!!!
So, with the strong assumption that the future is roughly or strongly related to the mechanisms observed into the past, you can project the behavior of the past also for the future.
So it's possible from this to compute the basic amount of premium that we have to receive from our customers!

Fair premium: amount that covers the expected losses, E(X)
Premium

For the insurance company is really important the pricing, so the computation of the premium because a wrong premium can affect the stability of the result of the year, and in the worst cases of the insurance company itself.
We can compute…

A premium equal to everyone

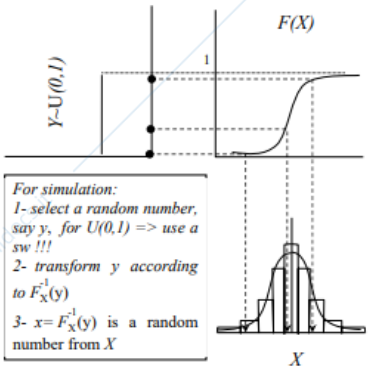
A different premium for every customer/group of customers applying the **PERSONALIZATION PRINCIPLE -> PERSONALIZATION OF THE PREMIUM**,
=The insurance company collect information about its customers defining profiles (using some covariates) trough which is possible to specify in a more precise way the risk related to the customer.

RANDOM NUMBER GENERATION

Teo. If X is a r.v. with $F_X(x)$ and $Y=F_X(x)$, then the cdf of $Y=F_X(x)$ is

$$F_Y(y)=\begin{cases}0 & y<0\\ y & 0\le y<1\\ 1 & y\ge 1\end{cases}$$

If Y has pdf $f_Y(y)=1$ for $0\le y<1$ and 0 elsewhere then $X = F_X^{-1}(y)$ has cdf $F_X(x)$.



The rationale is:

1. choose a probability, say p
2. given F_X, from

$$F_X(x_p) = p$$

x_p is the quantile of order p

3. if

$$F_X^{-1}(p) = x_p$$

exists, then

4. If p is *random* then x_p is a *random* number from X

What's the meaning from a theoretical point of view? How can be generated

In general, random numbers are generated from a specific stochastic mechanism and the corresponding cdf is an invertible function. Supposing for instance that that is a Gaussian distribution (continuous pdf), next we have his cdf F(x). The inverse of F(x) can be graphically visualized imagining rotating the picture by 90° (now x axis is y and vice versa) and it's possible to show it corresponds to a U(0,1).

With any pc (and in any kind of languages) it's possible to generate random numbers from a U(0,1) and also from any type of distribution, if we have a good knowledge of the cdf (and inverse). Suppose to pick up a random number from a U(0,1) as 0,65... we would have a value of the cdf of the U(0,1) calculated in 0,65. Taking the corresponding value on the inverse of the cdf (we look at the paper rotated, now we rotate it again) we would find also the corresponding value to the random number of a specific distribution (using the scale of the distribution you have selected).

So if you repeat a lot of times this procedure selecting a lot of uniform random numbers you are able to generate, at best with a simulation, the distribution into which you are interested.

What can be the application?

Supposing that a problem corresponds to a specific stochastic mechanism (=the generating process of accidents is a Poisson with a specific lambda), you can simulate the expected evolution of the accidents in the future. These studies are fundaments to have an idea of the scenarios to face (if bad, good and so on)

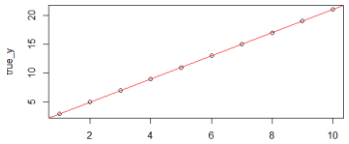
MODELS

In general, when we deal with a model we have two possible situations:

We collect real data -> we would understand the relationship between a couple of variables: as the income with respect to the age or the economic system as the accidents recorded by my customers with respect to the power of the car -> we can set up a model with a **dependent variable** and one/more **independent / explanatory variable** = here we arrive to a model = technical point of view	Simulation approach (the major property is that we are ready to control almost everything of the model… its dgp for instance and to understand in a much better way the output reported by our system) = here we start from a model =applied point of view

Simple linear model

- *STEP 1: "theoretical/true model"* that here is linear (so it's a linear combination, it's a straight line -> linear regression)



$$Y = a + bX$$

Y is the dependent / response variable *cost of an accident*
a is the intercept
b is the sloap coefficient/ multiplier
they are the coefficient, that here are known (in reality they are not!)
X is the independent / explanatory variable *power of the car*
!! it's intuitive to think that the greater is the power of the car, the greater will be the cost of accidents, but it's not always so clear neither easy to understand in what proportions
!!!! this true model is in reality unknown…we will try to find it out, trying to be the closest possible at the true solution.

- *STEP 2*: We start from some data collected from the real world = so a sample from Y
!!!! typically the underlying relationship between y and x is masked by some *randomness* -> so collecting data we will obtain something close to the black points, but not exact
So the model is better represented as

$$Y=a+bx + \varepsilon$$

a and b are the unknown coefficients
ε is the error component / disturbance term: it's a random number, and so we have to do some hypothesis about its stochastic mechanism
In many application we have $\varepsilon \sim N(0, \sigma^2)$, but not in insurance studies (Gamma or beta i.e)!

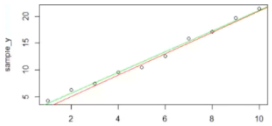
⇨ You have true data and you are willing to know, to find out the "true" mechanism that generates them.
Obviously, the larger is the sample size the greater is the probability to be very very close to the true model (! Not exaclty)
Obviously, when sigma increases we have a much more evident dispersive effect (we can see a relation between them – for instance a positive correlation – but the link between them is less clear)

- *STEP 3*: We want to get $\hat{a}$ and $\hat{b}$, the estimates of the true unknown a and b
The estimation exploit the OLS approach
Then we can compute also the estimated value of y for each value of the sample size

$$\hat{y} = \hat{a} + \hat{b}\,x$$

And the related basic statistics (mean, variance and so on)

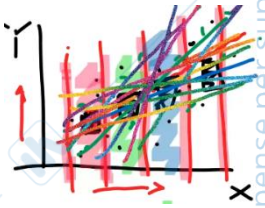
- *STEP 4*: We can plot the true model and the estimated one and comparing them. We would have not found exactly the correct model, but a close one.



!!!! For now, we had based our comments on only ONE sample of data, so on ONE selection of observations from the real population y + there is an error component !!!
= "A competitor" with another pc would have found other results (because he starts from a different sample). How can we be sure my estimates are the best?
We really have to find a strategy because we really want to have a good knowledge of the expected relationship between Y and X: according to my results I'm going to ask my customers specific amounts of money!

SIMULATE A MODEL A LOT OF TIMES

When we want to be ready/conscious of all the possible scenarios we have to deal with different samples collected from the population.
Considering another sample may be compared to considering another customer's profile
So, have to replicate a lot of times the process proposed above;
Every sample would have a related sigma (we can compute it simply using the data of the sample) and related estimates = we would obtain lots of different straight lines representing the estimated models ($\hat{y}$)
-> They are all *good* models because they are based on real data affected by different stochastic mechanism. We expect them all to be at least close to the "straight line corresponding to the "real" model.



We will use a strategy (=a standard mechanism to simplify the scenarios that you have generated; we will use statistics - principle of consistency, of asymptotical unbiased property) in order to be able to reproduce reality *on average*
on average = suppose splitting the domain of X into the pink those vertical regions and compute the corresponding average (of Y). You would find results into the black boxes, that contain also the value from the true model.

!!! ... from statistics:
If the error component of a linear model is supposed to be Gaussian, the distribution of the random variable which corresponds to the estimators of the model is supposed to be as well Gaussian

SOLVING AN OPTIMIZATION PROBLEM
But...how can I solve this problem numerically?

TRIAL AND ERROR APPROACH =>iterate procedure in which we hypothese some possible solutions and then plot them to see – through a comparison – which can be the best direction to follow. At the end we can so chose the best solution Pro: simple and intuitive Drawback: long procedure and not precise
OPTIMIZE() → =>the starting points is the implementation of a function. Then using the optimize I can find out the optimum. !!! Here I have to do a subjective choice: the domain to consider Pro: Simple, rich and powerful Drawback: subjectivity of the choice
OPTIM() → typically used when you have more than one unknown coefficient to be found and estimated =>the starting points is the implementation of a function. Then using the optime I can find out the optimum. !!! Here I have to do a subjective choice: the starting point from which starting the research Pro: Simple, powerful Drawback: no choice of a interval, but still subjectivity of the choice

CENTRAL LIMIT THEOREM

If you have a population with mean μ and standard deviation σ and take sufficiently large random samples from the population with replacement, then the distribution of the sample means will be approximately normally distributed

Let $X_1 \dots X_n$ a sequence of i.i.d r.v. If $E(X_i) = \mu$ and $Var(X_i) = \sigma^2$ $\forall i$, then

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \xrightarrow{D} N(0,1)$$

So we want to show that the distribution of the standardized x bar (→empirical data) converges asymptotically to the standard gaussian distribution (theoretical model).

- ➔ we have to use the sample mean and the sample deviation because we have got the data of the sample, not of the entire population.
- ➔ So $E(\bar{X}) = \mu$ $Var(\bar{X}) = \sigma^2/n$ is what we have asymptotically

A possible interpretation of the random variables...

$X_1, \dots, X_n$ can be interpreted as the behaviors of n customers, as the specific risk to get an accident of each of us. We have not the opportunity to have all the distribution of those variables, but only a sample from each of them: for instance X_1 corresponds to Mr Zappa, I have some observation about him (for instance a sample size of 10000) and so I can have an idea of his risk.

When we want a general view, a look at portfolio level we are not focusing only on "Zappa" (X1) but on all those random variables...we want so to sum up the number of accidents recorded by each customers.

Multiple linear model

$$Y = a + bX1 + cX2 + \dots$$

$$Y = a + bX1 + cX2 + \dots + \varepsilon$$

Suppose to have only 2 explanatory variables: **X1** and **X2**
You have two ways to write your model: one by one or merging the explanatory variables into a matrix

$Y = 1a + b \cdot X1 + cX2 + \varepsilon$

$\equiv$

$Y = X \cdot \vec{\beta} + \varepsilon$

$(n \times 3)$ (3×1)

in this second notation we will have:

- The matrix **X** of the explanatory variables (n x 3)
  !!! why "3"? implicitly there is also the explanatory variable related to the a coefficient, that is a vector of 1s
  (! It's useful for the linear algebra computation of the product a*1 + bx1+x2c of course for each statistical unit you have)
- The vector β of the unknwons of the model (3 x 1)
- The vector ε of the disturbance terms (3 x 1)

$$Y = \begin{bmatrix} 1 & x_1 & x_2 \\ 1 & x_1 & x_2 \\ \vdots & \vdots & \vdots \\ 1 & x_1 & x_2 \end{bmatrix} \cdot \begin{bmatrix} a \\ b \\ c \end{bmatrix} + \varepsilon$$

the difference between fit and predict

- ***Fit a model***: it consists in using our "framing dataset" to compute the fitted values of the model, so y hat
  framing dataset = dataset of information well known by us
  Suppose to have a portfolio of (1000) customers. To compute the risk of getting an accidents of each of them we can use a model to fit the corresponding risk to get an accident of each of them basing us to the information we have collected in the years about them

$\hat{Y} = X\hat{\beta}$

- ***Predict a model:*** it consists in using a "texting dataset" to try to predict something
  texting dataset = (new/other) data to text the capability of our model to grasp details on something not yet know
  Suppose to have a brand new customer: you are some basic information available (birth, age, geographical area, ...) but little knowledge about its history. We could use some text data to try to predict its behavior, its risk

Dummy variables

Read linear and multiple regression hand notes – page 26

In statistics and econometrics, particularly in regression analysis, a dummy variable[a] is one that takes only the value 0 or 1 to indicate the absence or presence of some categorical effect that may be expected to shift the outcome. They can be thought of as numeric stand-ins for qualitative facts in a regression model, sorting data into mutually exclusive categories (such as smoker and non-smoker).

Why is this topic so relevant?

In a database – for instance one used to compute premiums – we have the presence of a lot of different variables: some numerical (integers or continuous), some are characters. It's not possible to treat them all together and to use them in a model, unless you do some transformation.

The logical transformation consists in using a "Boolean" or T/F logic in order to have now "factors", whose categories will be called in R "levels".