Scientific Method

Class 2 - Part 1

MASTER INFORMATIQUE - SIIA

G. Coppin

2021-2022

According to the last polls, 47% of statistics are wrong.

A little joke

A statistician and a biologist are sentenced to death. They are given one last favor.

> *- I would like to give a big lecture on statistics in front of everyone, said the statistician.*
>
> *- Granted, said the judge*
>
> *The biologist expresses no hesitation :*
>
> *- I would like to be executed first.*

from "Comprendre et réaliser les tests statistiques à l'aide de R", Gael Millot

A question of money

You are about to be hired in the Macheprot company. Marcel-Benoit Schblurb, your joker colleague from the previous promotion, tells you that the average salary of those hired is 35 keuros. After a little investigation, you get the following data :

Emb1	Emb2	Emb3	Emb4	Emb5
34.5	36	35.2	33	34.3

Is this bloody damned Marcel-Benoit pulling your leg again?

Example developed: effect of the moon on births

We want to study the effects of the moon on births (more precisely the supposed effect of the full moon on the increase in births). We note in a maternity the following data:

Phase	New moon	First quarter	Full moon	Last quarter	Total
Births	76	88	100	96	360
Frequency	0,211	0,244	0,278	0,267	1

Can we validate the hypothesis from these data?

Effect of the moon (II)

The case is more difficult than if we had:

Phase	New moon	First quarter	Full moon	Last quarter	Total
Births	89	88	92	91	360

Phase	New moon	First quarter	Full moon	Last quarter	Total
Births	10	20	300	30	360

Effect of the moon (III)

- We define the null hypothesis H_0: births are equiprobable with respect to the phases of the moon.
- It can be rephrased with:

Phase	New moon	First quarter	Full moon	Last quarter	Total
Observed	89	88	92	91	360
Theoretical	90	90	90	90	360

Moon Effect (IV)

The statistical approach unveiled :

- We decide on a reference hypothesis (called *null*) H_0 relating to a given population,
- We get a sample of the population in question,
- We "compare" the distribution of the population to that which would be obtained from the null hypothesis,
- If the difference is too large, it cannot be *due to chance* only and we refute the null hypothesis.

Effect of the moon (V)

- In our case, we can "compare" the distributions using a global measure $M = \sum (Obs. - Theo)^2 / Theo$,
- This measurement is then compared with a reference value defined according to the number of degrees of freedom of the data (ν equal to (number of classes - 1)) and a margin of error (typically 5 %),
- If the measure is less than the threshold, we do not reject the null hypothesis.

Here, $\nu = 3$, we can show that the critical point is 7.82 and our measure is 3.83: *the difference is small enough to be justified by chance, we will keep the null hypothesis here*.

Classic statistical approach

- Data collection
  - Polls
  - Design of experiments
- Exploratory statistics: synthesis of the information related to the data
  - Descriptive statistics
  - Data analysis: classification, principal component analysis, correspondence analysis
- Inferential statistic
  - Construction of estimators
  - Hypothesis test
  - Modeling and statistical forecasting

In practice

- Identify the target population
- Determine the suitable encoding (if not natural)
- Identify and characterise the sample (descriptive statistics)
- Statistical modeling (identification of the law or model)
- Work the parameters of interest (estimate, confidence intervals)

Quality of a sample / collection procedure

The overall population is generally inaccessible and requires sampling; The samples must consist of **independent** and **identically distributed** observations in the population. Beware of collection bias :

- Bias related to telephone surveys
- Bias related to collection hours and days
- Motivation bias
- Manipulation bias

We can rely on **quota methods**.

Which global properties?

What quantities can we refer to to describe the sample - and therefore the entire population?

- mean
- variance / standard deviation
- regression line
- median
- but also ...
- extreme value
- higher order moments (kurtosis, *skewness* for example)
- more generally **parametric identification** (statistical model)

What serves you (or serves the end customer of statistical analysis).

Statistics and probabilities

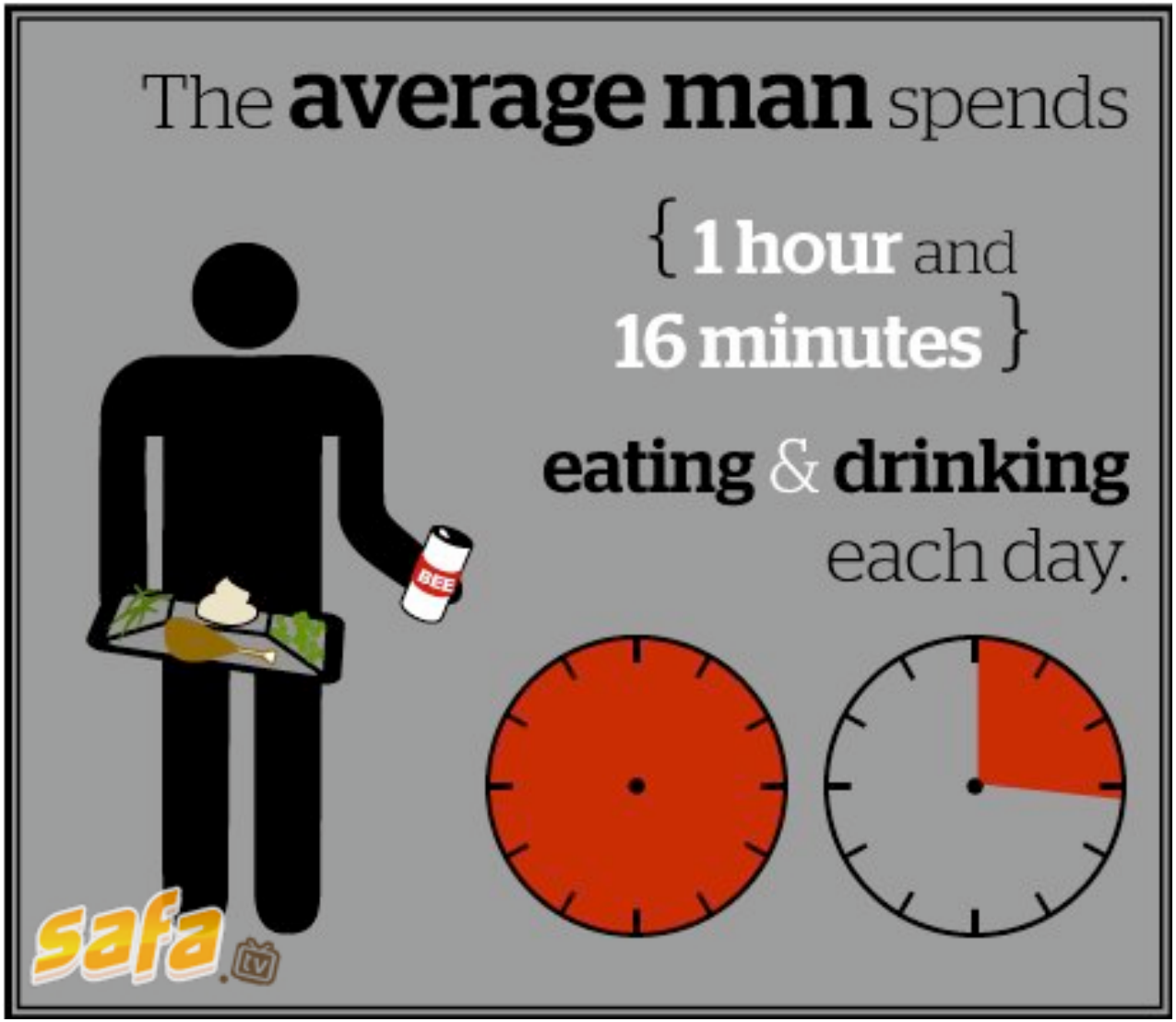
A key question (which we will find with the central limit theorem) : why do we often come down to a normal distribution?
3 very different reasons :

- The observations are imprecise data, therefore tainted with errors
- The statistical distribution of a variable is close to a theoretical model
- The method of constituting a sample makes it possible to align the statistical distribution on models

Basic concepts of descriptive statistics

You must master the concepts of :

- coding
- histogram
- mean, median, modal class
- variance
- quartiles
- mustache box
- quantiles


The average man spends
{ 1 hour and
16 minutes }
eating & drinking
each day.
BEE
safa

What are statistical laws REALLY good for?

We associate with an observed phenomenon a random variable which represents this phenomenon or the measure of the phenomenon, and of which we are supposed to master the law. Mastering the law means, for example, knowing its mean and variance (or any other time).
We can then play with :

- the probabilities (I know the law and I calculate the probability of getting a given value)
- the statistics (I know a value and I estimate if it is possible to consider it as "normal")

Binomial law (I)

- 20 flipping coins, X "tails" obtained
- 36 dice, Y "1" obtained
- 10 births, U daughters
- 45 % of the population in favor of a bill; in a sample of 100 people, W declared themselves in favor of the project

Binomial distribution $B(n,p)$ where n is the number of trials and p is the probability (on each trial) of success.

$$p(x) = C_n^x p^x (1-p)^{n-x}$$

Binomial law (II) - Example

In a factory, lots of mass-produced items are inspected using sampling methods. In each lot, 10 items are randomly selected and the lot will be rejected if 2 (or more) items are defective. If the lot contains 5% of defective items, what is the probability that the lot will be accepted or rejected?

$$X \text{ is } B(10; 0.05) \text{ and the lot is accepted if } X = 0 \text{ or } X = 1.$$

$$P(\textit{lot accept}) = p(0) + p(1) = C_{10}^{0}(0.05)^{0}(0.95)^{10} + C_{10}^{1}(0.05)^{1}(0.95)^{9}$$

$$P(\textit{lot accepts}) = 0.91386$$

Binomial law (III) - Hope and Variance

For a binomial distribution $B(n,p)$,

$$E(X) = \mu = np$$

and

$$Var(X) = \sigma^2 = npq$$

Binomial distribution (IV) - example

In the Groland presipacy 30% of the people are supporters of Vice-President Marcel. In a survey of 1000 people, X people say they are in favor of Marcel. In view of the relative size of the sample and the city, we can consider X as binomial $B(1000\ ;\ 0.3)$.

$$E(X) = np = 1000.(0.3) = 300$$

$$Var(X) = npq = 1000.(0.3).(0.7) = 210$$

$$\sigma = \sqrt{(Var(X))} = 14.49$$

If you get 700 supporters, it's fishy ...

Why is this fishy?

Because of Bienaymé-Chebychev!

$$P(|X - \mu| \geq \alpha) \leq \frac{\sigma^2}{\alpha^2}$$

Poisson's Law (I)

Poisson's law expresses the arrival of events "in the temporal sense". More precisely, X is the number of events that occur during a certain time interval. If X follows a Poisson distribution of mean λ, we have:

$$p(x) = \frac{\exp^{(-\lambda)} \lambda^x}{x!}$$

If X is a Poisson distribution

$$E(X) = \lambda$$

$$Var(X) = \lambda$$

The equality of expectation and variance constitutes an empirical indicator of the presence of a Poisson distribution.

Poisson's Law (II)

A teacher-researcher makes an haverage of 17 typos / claass. Hee has just tiped 50 pwpt slides, what is the probabillity of having less than 2 typing errors? The average number of errors expected is $50.17 = 850$ and therefore

$$P(X < 2) = p(0) + p(1) = \exp^{-850}(\frac{850^0}{0!} + \frac{850^1}{1!}) = ...$$

Poisson's Law (III)

The binomial law can be approximated by the Poisson law:

- n tends to infinity
- p tends to 0

Ex: number of calls to a switchboard between 10 a.m. and 11 a.m. To define X binomial, we divide the interval into 3600 seconds, thus giving 3600 tries. The probability of having a success per test is very low and we can shorten the interval and reduce the probability at will ... Convergence towards Poisson's law (if independence conditions and binary result validated).

Continuous laws : exponential law (I)

A random variable X has exponential distribution with mean $\theta > 0$ if its density function is :

$$f(x) = \frac{1}{\theta} \exp^{-\frac{x}{\theta}}$$

These laws model the waiting times before the arrival of an event. The link with the Poisson distribution is immediate : if the number of events occurring during an interval of time t is governed by a Poisson distribution of mean $\lambda = ct$, the waiting time between two arrivals will follow an exponential distribution with $\theta = \frac{1}{c}$

Continuous laws : exponential law (II)

- Application to reliability : if a teacher-researcher goes crazy on average every 40 hours of class, what is the probability that he will remain normal over a semester with 80 hours of class?
- We assume the operating time in exponential law of mean 40
- $P(X > 80) = \exp \frac{-80}{40} = \exp -2 = 0.1353$

Continuous laws : normal law (or Gaussian law) (I)

The variable X has a normal distribution if its density function is:

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

where μ and σ^2 are the expectation and the variance of the distribution. We denote the distribution $N(\mu, \sigma^2)$. Reduced law : for a law $N(\mu, \sigma^2)$, we classically define the reduced law $Z = \frac{X-\mu}{\sigma}$ which has law $N(0,1)$.

$$P(a < N(\mu, \sigma^2) < b) = P(\frac{a-\mu}{\sigma} < N(0,1) < \frac{b-\mu}{\sigma})$$

Continuous laws : normal law (or Gauss's law) (II)

- If X is a normal distribution $N(\mu, \sigma^2)$, $a + bX$ is a normal distribution $N(a + b\mu, b^2\sigma^2)$
- if $X_1, X_2, .., X_n$ are independent normal distributions of respective distributions $N(\mu_i, \sigma_i^2)$, then their sum is normal of distribution $N(\mu_1 + \mu_2 + .. + \mu_n, \sigma_1^2 + \sigma_2^2 + .. + \sigma_n^2)$

Continuous laws : normal law (or Gaussian law) (III)

- We assume that the weight in kilograms of a teacher-researcher is distributed according to a $N(70,5)$ law. What is the probability that the administrative vehicle of IMT Atlantique (which admits a load of 1500 kg) cannot transport the 20 researchers of the Computer Science department?
- $E(Lussi) = E(X_1 + .. + X_{20}) = 1400$
- $Var(X) = Var(X_1 + .. + X_{20}) = 500$
- $P(X > 1500) \Leftrightarrow P(N(1400,500) > 1500)$
- either $P(N(0,1)) > \frac{1500-1400}{\sqrt{500}}$
- let $P(N(0,1)) > 4.47$
- we are therefore "normally" saved ...

Law of χ^2

Let $U_1, U_2, .., U_p$ p normal variables $N(0,1)$ independent.
We call the law of χ^2 at p degrees of freedom χ^2_p the law of the variable $\sum_{i=1}^{p} U_i^2$.

The distribution of χ^2 can be approximated by a normal distribution.
When $p > 30$, we can effectively consider that $\sqrt{(2\chi^2)} - \sqrt{(ep-1)}$ is a distribution of type $N(0,1)$.

Student's law

Let U be a random variable following a normal distribution $N(0,1)$ and X independent of U following a χ^2_n distribution. We define the Student variable T at n degrees of freedom by:

$$T_n = \frac{U}{\sqrt{\frac{X}{n}}}$$

Fisher-Snedecor Law

If X and Y are respectively laws of type χ^2_n and χ^2_p, then we can define the following law:

$$F(n;p) : \frac{X/n}{Y/p}$$

This law serves as a reference for variance analyzes.

Central limit theorem (I)

- Central limit theorem (variables with the same distribution): Let n be a large number of independent variables $X_1, X_2, .., X_n$ with the same distribution. Then their sum $X = X_1 + X_2 + .. + X_n$ approximately follows a normal distribution, even if these variables are not normal
- Generalized central limit theorem: Let n be a large number of independent variables $X_1, X_2, .., X_n$. Then under certain conditions, their sum $X = X_1 + X_2 + .. + X_n$ approximately follows a normal distribution, even if these variables are not normal.

For some authors, n large for $n > 30$, for others $n > 100$.

Generalized central limit theorem (II)

Conditions : essentially that of Lindeberg, which indicates that the reduced variables $\frac{X_i-\mu_i}{S_n}$ are "uniformly small" with a high probability.

$$\lim_{n\to\infty} \frac{1}{S_n^2} \sum_{i=1}^{n} \int_{|x|>S_n} x^2 dF_i(x) = 0$$

with F_i distribution function of $X_i - \mu_i$ and

$$S_n^2 = \sum_{i=1}^{n} \sigma_i^2$$

Elements of proof (variables of the same law)

The characteristic function $\phi_x(t) = E[\exp^{itX}]$ of a variable with expectation 0 and variance 1 Y can be approximated by :

$$\phi_Y(t) = 1 - \frac{t^2}{2} + o(t^2)$$

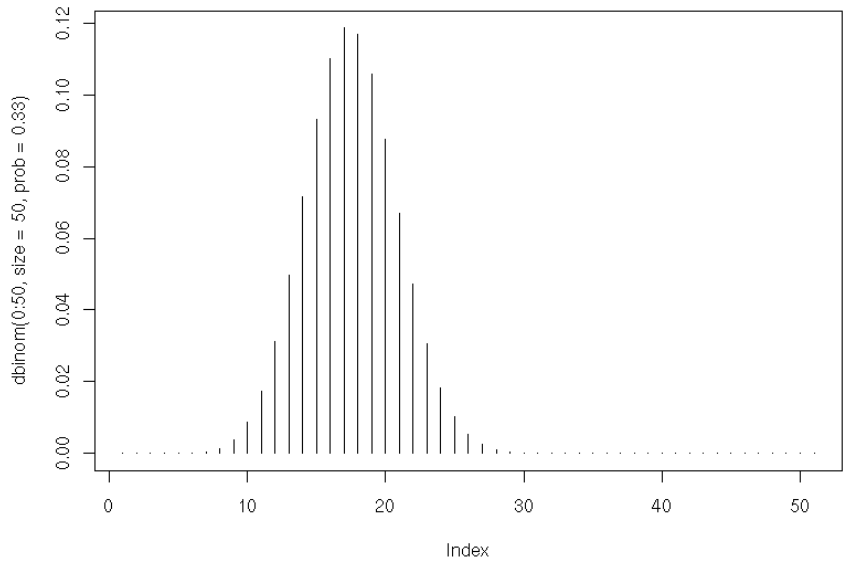
Then if the reduced centered mean of observations is :

$$Z_n = \frac{\overline{X_n} - \mu}{\frac{\sigma}{\sqrt{n}}} = \Sigma_{i=1}^{n} \frac{Y_i}{\sqrt{n}}$$

the characteristic function of Z_n is $(\phi_Y(\frac{t}{\sqrt{n}}))^n = [1 - \frac{t^{2n}}{2} + o(\frac{t^2}{n})]^n$ which converges to $\exp^{-\frac{t^2}{n}}$ ... which is the characteristic function of the normal distribution.

Application to binomial law approximation

By repeating a binomial-type experiment, several independent components - all of the same distribution - are added, so that their sum is close to a normal distribution. A binomial distribution $B(n,p)$ will thus be approximated by a normal distribution $N(\mu = np, \sigma = \sqrt{(npq)})$.



Binomial law approximation (II)

We throw 16 coins. We want to calculate the probability of obtaining a number of "faces" between 5 and 10 inclusive. With $B(16, 1/2)$, we obtain by taking the exact binomial values ?? :

$$P(5 \leq X \leq 10) = p(5)+p(6)+p(7)+p(8)+p(9)+p(10) = 0.85654$$

To use a normal approximation, we have to switch to a continuous law, and we adjust the bounds of the interval, here at 4.5 and 10.5. We obtain:

$$\begin{aligned} P(5 \leq X \leq 10) &\simeq P(4.5 < N(8.4) < 10.5) \\ &= P(-1.75 < N(0.1) < 1.25) \\ &= 1 - (P(N(0.1) > 1.75) - P(N(0.1) > 1.25) \\ &= 0.8543 \end{aligned}$$

Binomial law approximation (III)

The greater the approximation will be the greater the n, we estimate the approximation valid for $npq > 5$. Without the correction for continuity, we would have obtained for the previous example a response of 0.7745, therefore much less precise.

Binomial law approximation (IV) - Elements of proof

We start from the characteristic function of the random variable (Fourier transform of its density), i.e.

$$\varphi_X(t) = \int_{\mathbf{R}} \exp^{itx} f(x)dx = (p \exp^{it} + 1 - p)^n$$

then the characteristic function of $\frac{X_n - np}{\sqrt{npq}}$ can be approximated by:

$$\varphi(t) = (p \exp^{\frac{it}{\sqrt{npq}}} + 1 - p)^n \exp^{-\frac{itnp}{\sqrt{npq}}}$$

Binomial law approximation (V) - Elements of proof - continued

By going to the log, we get:

$$\ln \varphi = n \ln(p(\exp^{\frac{it}{\sqrt{npq}}}) - 1)) - \frac{itnp}{\sqrt{npq}}$$

By successively expanding to second order the exponential $(1 + px + px^2)$ then the logarithm $(x - \frac{x^2}{2})$, we obtain:

$$\ln \varphi \simeq \frac{-t^2}{2q} + \frac{pt^2}{2q} = \frac{t^2}{2q}(p-1) = \frac{-t^2}{2}$$

This corresponds to the characteristic function of the reduced centered normal distribution $N(0,1)$. Phew.

Return to the χ^2

If we start from a partitioning of the event space into n classes $A_1, A_2, ..A_n$, and if we measure their rate of realization $N_1, N_2, .., N_n$, the vector follows a law multinomial (extension of the binomial distribution from 2 to n classes) of total size n and parameters $p_1, p_2, ..., p_n$. We can define the conditional distribution of N_i knowing $N_j = n_j$ as $B(n - n_j, \frac{p_i}{1-p_j})$. The central limit theorem (applied to each of the components of the vector) indicates that each of these components $N_i - np_i$ tends towards a normal distribution $N(0, 1)$. So:

$$\lim_{n\to\infty} \sum_{i=1}^{k} \frac{(N_i - np_i)^2}{np_i} \to \chi^2_{k-1}$$

Here we find the result of the previous example.

Distribution of an average

Thanks to the central limit theorem, we have the following result:

- If $\overline{X}$ is the mean of n independent observations $X_1, X_2, .., X_n$ where $E(X_i) = \mu$ and $Var(X_i) = \sigma^2$, then, for n large, $\overline{X}$ has approximately the distribution of $N(\mu, \frac{\sigma^2}{n})$.
- The larger n, the closer the estimate $\overline{X}$ will be to the true value μ and the smaller the variance of the "estimated mean" variable.