Scientific Method

Class 2 - Part 2

MASTER INFORMATIQUE - SIIA

G. Coppin

2021-2022

Class objectives

Objectives:

1. Understand the mechanisms of estimation
2. Know how to estimate a mean from a sample
3. Know how to estimate a percentage from a sample

Basic hypothesis

Definition

We suppose the target population to be infinite

Estimation et statistiques

The statistician's first job is to establish one or more estimators that describe the population from the observation sample. By definition, the estimators should not depend on the actual parameters of the distribution, but only on observables present in the sample which are assumed to be generated from random variables of the same law P_{θ_0}.

Definition

An estimator is any random variable constructed from the observations $X_1, X_2, .., X_n$. In particular, it should not depend on unknown quantities, such as θ_0 ou P_{θ_0}.

Example : punctual estimation of proportion

The practice of sampling is essential. If the sample is correctly formed, the experimental proportion should be close to the theoretical proportion. By noting p the real (unknown) proportion, n the sample size, and X the number of individuals possessing the characteristic that interests us (ie of **success**), we can evaluate :

$$\widehat{p} = \frac{X}{n}$$

$\widehat{p}$ is therefore an estimator of p.

Toy example

Two surveys have been carried out on Guinness addiction among UBO students. With a small sample of 5 individuals (met at 9 a.m. on Monday), we get $X = 1$ positive responses, or $\widehat{p} = 20\%$. The same poll of 300 students on Thursday night at the students' party gives $X = 150$, a value of $\widehat{p} = 50\%$ but this time ... maybe more reliable!

Qualities of an estimator

The estimator must be unbiased (or well centered) : its expectation is equal to the value of the parameter to be estimated.

Definition

$$E[\widehat{\theta}] = \theta_0$$

The estimator must be consistent

Definition

$$\theta_n \Rightarrow_{n\to\infty} \theta_0$$

The estimator should have *variance as small as possible* to be as *precise* as possible.

Estimation of a proportion

$\widehat{p}$ is an unbiased estimator for the binomial distribution

$$E(\widehat{p}) = E(\frac{1}{n}X) = \frac{1}{n}E(X) = \frac{1}{n}np = p$$

$$Var(\widehat{p}) = Var(\frac{1}{n}X) = \frac{1}{n^2}Var(X) = \frac{npq}{n^2} = \frac{pq}{n}$$

We can clearly see the importance of the sample size for the quality (precision) of the estimate.

Estimation of a mean

In the same way, we can estimate without bias the mean μ of a normal distribution ... by the mean of the observations in the sample

$$E(\overline{X}) = E(\frac{1}{n}\sum_i X_i) = \frac{1}{n}\sum_i E(X_i) = \mu$$

et

$$Var(\overline{X}) = \frac{1}{n^2} Var(\sum_i X_i) = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}$$

Example

We assume that the average duration in minutes of the consumption of a pint follows a normal distribution N, of unknown mean μ and variance 2. We observe the following durations :

7,3	5,7	6,4	6,7	8,2	6,0	5,8	8,3

The average of these observations is $\overline{X} = 6.8$ and the variance of $\overline{X}$ is $\frac{\sigma^2}{n} = \frac{1}{4}$ so its standard deviation of $\sigma = 0.5$. With 1000 observations, the standard deviation of $\overline{X}$ would have been 0.004 ...

Example

We assume that the average duration in seconds of the consumption of a pint follows a normal distribution N, of unknown mean μ and variance 2. We observe the following durations :

7,3	5,7	6,4	6,7	8,2	6,0	5,8	8,3

The average of these observations is $\overline{X} = 6.8$ and the variance of $\overline{X}$ is $\frac{\sigma^2}{n} = \frac{1}{4}$ so its standard deviation of $\sigma = 0.5$. With 1000 observations, the standard deviation of $\overline{X}$ would have been 0.004 ...

Confidence interval

- We try to set the estimate in an confidence interval which allows the quality of the estimate to be assessed. If n is large enough, the estimation error ($\overline{X} - \ mu$ or $\widehat{p} - p$) will be smaller than a given deviation and therefore within of an interval.
- In practice, we define a risk that we accept to run, α, which represents the probability that the interval does not contain the true value of the parameter. $(1 - \alpha)$ is the alert confidence level of the interval

$$P(Y_1 < \theta < Y_2) = 1 - \alpha$$

Estimation of a proportion (I)

$\widehat{p} = \frac{X}{n}$ is an unbiased variance estimator $\sigma^2_{\widehat{p}} = \frac{pq}{n}$. If n large, thanks to the central limit theorem, $\widehat{p}$ is approximately Gaussian with distribution $N(E(\widehat{p}) = p, \sigma^2_{\widehat{p}})$, so $\frac{\widehat{p}-p}{\sigma_{\widehat{p}}}$ is a reduced distribution $N(0,1)$. We can therefore find a threshold c_α in the reduced centered normal distribution table such as:

$$P(-c_\alpha < \frac{\widehat{p}-p}{\sigma_{\widehat{p}}} < c_\alpha) \simeq 1-\alpha$$

which amounts to:

$$P(p - c_\alpha \sigma_{\widehat{p}} < \widehat{p} < p + c_\alpha \sigma_{\widehat{p}}) \simeq 1-\alpha$$

or better

$$P(\widehat{p} - c_\alpha \sigma_{\widehat{p}} < p < \widehat{p} + c_\alpha \sigma_{\widehat{p}}) \simeq 1-\alpha$$

Estimation of a proportion (II)

It remains to estimate $\sigma_{\widehat{p}} = \sqrt{\frac{pq}{n}}$, which is done naturally by $\widehat{\sigma}_{\widehat{p}} = \sqrt{\frac{\widehat{p}\widehat{q}}{n}}$. This makes it possible to determine the interval of confidence :

$$(\widehat{p} \pm c_\alpha \widehat{\sigma}_{\widehat{p}}) = (\widehat{p} \pm c_\alpha \sqrt{\frac{\widehat{p}\widehat{q}}{n}})$$

And there, you will tell me ...

That's enough ! An example !

Example

- In a Grolandish survey of 500 people, 180 people declared themselves positive to the wearing of the ponpon cap. What is the theoretical proportion p of people in favor of the ponpon cap (with 90% confidence interval)?
- $\widehat{p} = \frac{X}{n} = \frac{180}{500} = 0,360$
- to have $\alpha = 10\%$, we must take $c_\alpha = 1,644854$ (2-tail)
- $(\widehat{p} \pm c_\alpha \sqrt{\frac{\widehat{pq}}{n}}) = (0,360 \pm 1,645 \sqrt{\frac{0,36.0,64}{500}}) = (0,325; 0,395)$

Normal distribution table

TABLE 3 **Loi Normale $N(0, 1)$: Valeur de $P(N(0,1) > x)$ en fonction de x**

x	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,4960	0,4920	0,4880	0,4840	0,4801	0,4761	0,4721	0,4681	0,4641
0,1	0,4602	0,4562	0,4522	0,4483	0,4443	0,4404	0,4364	0,4325	0,4286	0,4247
0,2	0,4207	0,4168	0,4129	0,4090	0,4052	0,4013	0,3974	0,3936	0,3897	0,3859
0,3	0,3821	0,3783	0,3745	0,3707	0,3669	0,3632	0,3594	0,3557	0,3520	0,3483
0,4	0,3446	0,3409	0,3372	0,3336	0,3300	0,3264	0,3228	0,3192	0,3156	0,3121
0,5	0,3085	0,3050	0,3015	0,2981	0,2946	0,2912	0,2877	0,2843	0,2810	0,2776
0,6	0,2743	0,2709	0,2676	0,2643	0,2611	0,2578	0,2546	0,2514	0,2483	0,2451
0,7	0,2420	0,2389	0,2358	0,2327	0,2296	0,2266	0,2236	0,2206	0,2177	0,2148
0,8	0,2119	0,2090	0,2061	0,2033	0,2005	0,1977	0,1949	0,1922	0,1894	0,1867
0,9	0,1841	0,1814	0,1788	0,1762	0,1736	0,1711	0,1685	0,1660	0,1635	0,1611
1,0	0,1587	0,1562	0,1539	0,1515	0,1492	0,1469	0,1446	0,1423	0,1401	0,1379
1,1	0,1357	0,1335	0,1314	0,1292	0,1971	0,1251	0,1230	0,1210	0,1190	0,1170
1,2	0,1151	0,1131	0,1112	0,1093	0,1075	0,1056	0,1038	0,1020	0,1003	0,0985
1,3	0,0968	0,0951	0,0934	0,0918	0,0901	0,0885	0,0869	0,0853	0,0838	0,0823
1,4	0,0808	0,0793	0,0778	0,0764	0,0749	0,0735	0,0721	0,0708	0,0694	0,0681
1,5	0,0668	0,0655	0,0643	0,0630	0,0618	0,0606	0,0594	0,0582	0,0571	0,0559
1,6	0,0548	0,0537	0,0526	0,0516	0,0505	0,0495	0,0485	0,0475	0,0465	0,0455
1,7	0,0446	0,0436	0,0427	0,0418	0,0409	0,0401	0,0392	0,0384	0,0375	0,0367
1,8	0,0359	0,0351	0,0344	0,0336	0,0329	0,0322	0,0314	0,0307	0,0301	0,0294
1,9	0,0287	0,0281	0,0274	0,0268	0,0262	0,0256	0,0250	0,0244	0,0239	0,0233
2,0	0,0228	0,0222	0,0217	0,0212	0,0207	0,0202	0,0197	0,0192	0,0188	0,0183
2,1	0,0179	0,0174	0,0170	0,0166	0,0162	0,0158	0,0154	0,0150	0,0146	0,0143

Methodological and terminological remarks

- it is necessary to balance precision and security : the smaller the interval, the greater the risk and vice versa. Precision must be "paid for" by a greater risk of error. To do better ... we must increase the value of n, therefore the number of pollers and the data collection time.
- it is often said that p has a 9 in 10 chance (10 % risk) of falling within the confidence interval. This is a misnomer ... since p is not ... (up to you to fill in).

Interval vs. risk

With $n = 100$ et $\widehat{p} = 0,21$, we can have the following intervals :

α	c_α	Interval	Length
50%	0,674	(0,18 - 0,24)	0,06
10%	1,645	(0,14 - 0,28)	0,14
5%	1,960	(0,13 - 0,29)	0,16
1%	2,576	(0,11 - 0,31)	0,20
0,1%	3,291	(0,08 - 0,34)	0,26

Size of the sample

How many individuals must be taken so that whatever p the confidence interval has radius r_{ref} at most 0.05 (resp. 0.03 / 0.02 / 0.01)?

- the radius of the 95% confidence interval is equal to $1,960\sqrt{\frac{\widehat{p}_n\widehat{q}_n}{n}}$
- the maximum value of $\widehat{p}_n\widehat{q}_n$ is 0.25 so $r_{max} = \frac{1,960}{\sqrt{4n}}$
- so to have $r < r_{max} < r_{ref}$, we must take $n \geq (\frac{0,98}{r_{ref}})^2$
- which gives respectively $n \geq 385$, $n \geq 1068$, $n \geq 2041$, $n \geq 9604$

Indeed, most of the surveys given at 3% are done on samples of ... more than 1000 participants.

Application to an election poll (real but here with fictitious data)

At 8:45 p.m., in the municipal elections of 2008, the media announced the victory of Lyne Cohen-Solal as mayor of the fifth arrondissement of Paris. At 9.15 p.m., they reverse and announce that of Jean Tiberi. In fact, at 8:45 p.m., statisticians had access to a partial sample of votes, such as LCS 473, JT 418 and PM 108. By applying the estimate to each of the three candidates, we obtain (from the empirical percentage measured on the first 978 samples, with a risk of 5%) :

- LCS - $[0,473 \pm \frac{1,96\sqrt{0.473*0.527}}{\sqrt{978}}] = [44,2\% - 50,5\%]$
- JT - $[0,418 \pm \frac{1.96\sqrt{0.418*0.582}}{\sqrt{978}}] = [38,7\% - 44,9\%]$
- PM - $[0,108 \pm \frac{1.96\sqrt{0.108*0.892}}{\sqrt{978}}] = [8,8\% - 12,8\%]$

Difficult to assert anything at 8:45 p.m. ... .

Estimation of a mean

We have seen that if n large, the mean $\overline{X}$ has approximately the distribution of $N(\mu, \sigma^2_{\overline{X}})$. Remember that $\sigma^2_{\overline{X}} = \frac{\sigma^2}{n}$. As for the proportions, we will fix a risk α and get

$$P(\mu - c_\alpha \sigma_{\overline{X}} < \overline{X} < \mu + c_\alpha \sigma_{\overline{X}}) \simeq 1 - \alpha$$

ans similarly

$$P(\overline{X} - c_\alpha \sigma_{\overline{X}} < \mu < \overline{X} + c_\alpha \sigma_{\overline{X}}) \simeq 1 - \alpha$$

Estimation of a mean (II)

As for the proportions, it remains to evaluate the standard deviation $\sigma_{\overline{X}}$, a priori unknown. As the mean is unknown, we have to approach the variance by

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

but ... this estimator is biased! he has an expectation of $\frac{n-1}{n}\sigma^2$. To have an unbiased estimator, we must divide by $n-1$ and not n

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

The level $1-\alpha$ confidence interval for μ is then

$$\overline{X} \pm c_\alpha \frac{\widehat{\sigma}}{\sqrt{n}}$$

We may be cautious about getting alcool issues by mean, but certainly by extreme values.

Phew, another example ...

We observe the number of beverages drunk in the evening by each student customer during the weekly party and we obtain the following results:

Number of bev.	1	2	3	4	5	6	Total
Number of students	230	248	117	76	14	3	688

The following ...

The total of observations is 688, the number of swallowed drinks is 1469, which leads to $\overline{X} = 2.135$ and $\widehat{\sigma}^2 = 1.183$ or $\widehat{\sigma} = 1.088$. To have a risk of 5 %, we must take $c_\alpha = 1.960$ and the confidence interval is therefore :

$$(\overline{X} + c_\alpha \frac{\widehat{\sigma}}{\sqrt{n}}) = (2,135 \pm 1,960.1,088/26,23) = (2,054; 2,216)$$

The following ...

The total of observations is 688, the number of swallowed drinks is 1469, which leads to $\overline{X} = 2.135$ and $\widehat{\sigma}^2 = 1.183$ or $\widehat{\sigma} = 1.088$. To have a risk of 5 %, we must take $c_\alpha = 1.960$ and the confidence interval is therefore :

$$(\overline{X} + c_\alpha \frac{\widehat{\sigma}}{\sqrt{n}}) = (2,135 \pm 1,960.1,088/26,23) = (2,054; 2,216)$$

Sampling normal law

If the laws $X_1, X_2, .., X_n$ are normal, their mean is also normal and $\frac{\overline{X}-\mu}{\sigma_{\overline{X}}}$ is then exactly law $N(0.1)$. We still have to deal with the case $\sigma_{\overline{X}}$ (which remains unknown). We replace it (as usual) by $\widehat{\sigma}_{\overline{X}}$ and we get the variable $\frac{\overline{X}-\mu}{\widehat{\sigma}_{\overline{X}}}$ as follows ...

Sampling normal law

If the laws $X_1, X_2, .., X_n$ are normal, their mean is also normal and $\frac{\overline{X}-\mu}{\sigma_{\overline{X}}}$ is then exactly law $N(0.1)$. We still have to deal with the case $\sigma_{\overline{X}}$ (which remains unknown). We replace it (as usual) by $\widehat{\sigma}_{\overline{X}}$ and we get the variable $\frac{\overline{X}-\mu}{\widehat{\sigma}_{\overline{X}}}$ as follows ... a Student's law!!

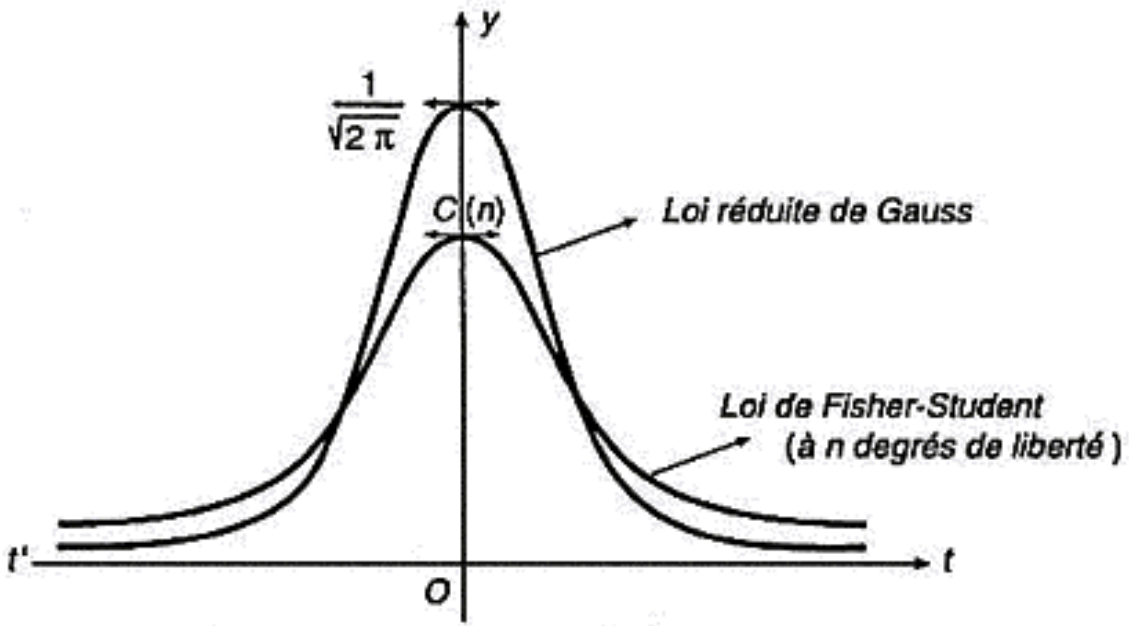
Reminder : Student's law

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:

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

In our case, U corresponds to the normal variable $\overline{X} - \mu$ and the variable X in the denominator is indeed a variable of χ^2 since it is calculated from the variance (therefore sum of Gaussians). Student's law will be parameterized by a number of degrees of freedom equal to $\nu = n - 1$.

Normal law, Student's law



An example again : the problem

We want to estimate the average duration of a 33 rpm record face. By measuring the faces of 5 disks, we get the vector

$$(17,5-22,4-18,6-24,3-19,5-21,6-15,9-20,4-18,7-20,3)$$

We assume that these laws are normal. What is the 90 % confidence interval for μ?

An example again : the solution

The observations give $\sum X = 199.2$ and $\sum X^2 = 4022.2$. So we have

$$\overline{X} = 19,92$$

et

$$\widehat{\sigma}^2 = \frac{\sum X_i^2 - n\overline{X}^2}{n-1} = 5,9951$$

An example again : the solution

The observations give $\sum X = 199.2$ and $\sum X^2 = 4022.2$. So we have

$$\overline{X} = 19,92$$

et

$$\widehat{\sigma}^2 = \frac{\sum X_i^2 - n\overline{X}^2}{n-1} = 5,9951$$

Avec $\alpha = 10\%$ et $\nu = 9$, on obtient $t_\alpha = 1,833$ and the confidence interval :

$$(\overline{X} \pm t_\alpha \frac{\widehat{\sigma}}{\sqrt{n}}) = (18,50 - 21,34)$$

Table of Student's law

TABLE 4 **Loi de Student t_ν**

Valeur tabulée : argument en fonction de la probabilité et du nombre de degrés de liberté ν.
$P(t_\nu > c) = \alpha$
$\nu = 1(1)30, 40, 60, 120, \infty$

ν \ α	0,25	0,10	0.05	0,025	0,01	0,005	0,0025	0,001	0,0005
1	1,000	3,078	6,314	12,706	31,821	63,657	127,320	318,310	636,620
2	0,816	1,886	2,920	4,303	6,965	9,925	14,089	22,327	31,598
3	0,765	1,638	2,353	3,182	4,451	5,841	7,453	10,214	12,924
4	0,741	1,533	2,132	2,776	3,747	4,604	5,598	7,173	8,610
5	0,727	1,476	2,015	2,571	3,365	4,032	4,773	5,893	6,869
6	0,718	1,440	1,943	2,447	3,143	3,707	4,317	5,208	5,959
7	0,711	1,415	1,895	2,365	2,998	3,499	4,029	4,785	5,408
8	0,706	1,397	1,860	2,306	2,896	3,355	3,833	4,501	5,041
9	0,703	1,383	1,833	2,262	2,821	3,250	3,690	4,297	4,781
10	0,700	1,372	1,812	2,228	2,764	3,169	3,581	4,144	4,587
11	0,697	1,363	1,796	2,201	2,718	3,106	3,497	4,025	4,437
12	0,695	1,356	1,782	2,179	2,681	3,055	3,428	3,930	4,318
13	0,694	1,350	1,771	2,160	2,650	3,012	3,372	3,852	4,221
14	0,692	1,345	1,761	2,145	2,624	2,977	3,326	3,787	4,140
15	0,691	1,341	1,753	2,131	2,602	2,947	3,286	3,733	4,073
16	0,690	1,337	1,746	2,120	2,583	2,921	3,252	3,686	4,015
17	0,689	1,333	1,740	2,110	2,567	2,898	3,222	3,646	3,965
18	0,688	1,330	1,734	2,101	2,552	2,878	3,197	3,610	3,922
19	0,688	1,328	1,729	2,093	2,539	2,861	3,174	3,579	3,883
20	0,687	1,325	1,725	2,086	2,528	2,845	3,153	3,552	3,850
21	0,686	1,323	1,721	2,080	2,518	2,831	3,135	3,527	3,819
22	0,686	1,321	1,717	2,074	2,508	2,819	3,119	3,505	3,792
23	0,685	1,319	1,714	2,069	2,069	2,500	2,807	3,104	3,767
24	0,685	1,318	1,711	2,064	2,492	2,797	3,091	3,467	3,745
25	0,684	1,316	1,708	2,060	2,485	2,787	3,078	3,450	3,725
26	0,684	1,315	1,706	2,056	2,479	2,779	3,067	3,435	3,707
27	0,684	1,314	1,703	2,052	2,473	2,771	3,057	3,421	3,690
28	0,683	1,313	1,701	2,048	2,467	2,763	3,047	3,408	3,674
29	0,683	1,311	1,699	2,045	2,462	2,756	3,038	3,396	3,659
30	0,683	1,310	1,697	2,042	2,457	2,750	3,030	3,385	3,646
40	0,681	1,303	1,684	2,021	2,423	2,704	2,971	3,307	3,551
60	0,679	1,296	1,671	2,000	2,390	2,660	2,915	3,232	3,460
120	0,677	1,289	1,658	1,980	2,358	2,617	2,860	3,160	3,373

Estimation of any parameter

For any parameter θ for any law, we proceed in a similar way :

- we are looking for a suitable estimator $\widehat{\theta}$ whose variance can be estimated $\sigma^2_{\widehat{theta}}$ - this estimation is often done by replacing in the variance formula θ by $\widehat{\theta}$.
- for n large, $\widehat{\theta}$ will behave like a normal distribution and the general formula $(\widehat{\theta} \pm c_\alpha \sigma_{\widehat{\theta}})$ will give the confidence interval

Little flashback

- We apply the previous estimators without hesitation ... when the population is infinite (really infinite, with replacement, large in front of the sample size)
- When the population is finite, the variance estimators change!

The variance estimator is, for a population size N and a sample size n

$$\widehat{\sigma}_{\overline{X}} = \frac{\widehat{\sigma}}{\sqrt{n}} \sqrt{1 - \frac{n}{N}}$$

Confidence interval with a finite population

The level $1 - \alpha$ confidence interval for *mu* is then

$$\overline{X} \pm c_\alpha \frac{\widehat{\sigma}}{\sqrt{n}} \sqrt{1 - \frac{n}{N}}$$