Estimation Theory
Théorie de l'estimation

Advanced Statistical Analysis

Feryal WINDAL

Punctual Estimation

Dans le chapitre précédent, nous nous sommes intéressés à deux statistiques :

- la moyenne d'échantillonnnage X pour estimer la moyenne de la population μ
- la proportion d'échantillonnage F pour estimer la proportion de la population p.

Chaque estimation que nous avons réalisée nous a donné une valeur unique, à partir des informations disponibles (taille de la population, taille de l'échantillon, valeurs observées dans l'échantillon).
Cette estimation est dite ponctuelle, au sens où elle donne une valeur échantillonnage unique pour le paramètre à estimer dans une population.

In the previous chapter, we looked at two statistics:

- the sampling mean X to estimate the population mean μ
- the sampling proportion F to estimate the proportion of the population p.

Each estimation we made gave us a unique value, based on the information available (population size, sample size, values observed in the sample).
This estimation is said to be point-in-time, in the sense that, it gives a single sampling value for the parameter to be estimated in a population.

Punctual Estimation

Cette statistique échantillonnale n'a que très peu de chances d'être égale à la vraie valeur du paramètre.

Il y a donc un écart entre l'estimation et la vraie valeur que l'on nomme erreur d'estimation.

Par ailleurs, nous avons vu que lorsque la taille de l'échantillon à partir duquel a été calculée l'estimation est suffisamment grande ($n \geq 30$), nous pouvons nous assurer de la proximité de la vraie valeur du paramètre dans le voisinage de la statistique échantillonnale estimée.

This sample statistic is very unlikely to be equal to the true value of the parameter.

Therefore, there is a difference between the estimate and the true value, which is called an estimate error.

Moreover, we have seen that when the size of the sample from which the estimation was calculated is enough larger ($n \geq 30$), we can ensure that the true value of the parameter is close to the neighborhood of the estimated sample statistic.

Punctual Estimation

L'intervalle de confiance obtenu permet alors d'établir qu'il existe une probabilité p que la vraie valeur du paramètre sur la population appartienne à un intervalle de valeurs centrées en la valeur échantillonnale (*estimateur ponctuel*) dont les bornes s'écartent à raison d'un nombre z d'écart-type de la distribution d'échantillonnage.

Dans ce chapitre, nous allons nous intéresser avec plus de détails à cette notion d'intervalle de confiance.

$$[\mu - \sigma; \mu + \sigma] \qquad [\mu - 2\sigma; \mu + 2\sigma] \qquad [\mu - 3\sigma; \mu + 3\sigma]$$

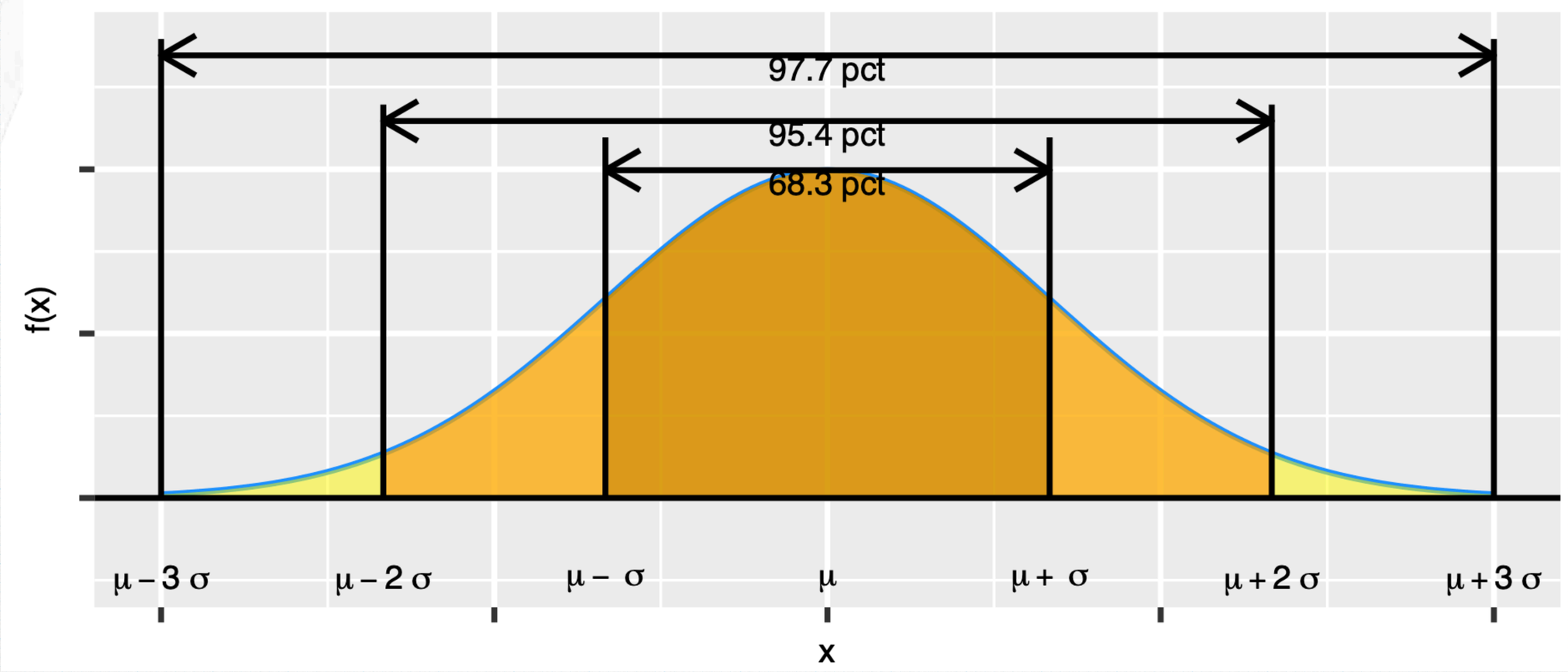
The confidence interval obtained then makes it possible to establish that there is a probability p that the true value of the parameter on the population belongs to an interval of values centered on the sample value (punctual estimator) whose limits diverge at a reasonable rate. a number z of the standard deviation of the sampling distribution.

In this chapter, we will take a closer look at this notion of confidence interval.

1.Mean (average) Estimation

1.1.Interval length, confidence level, and precision

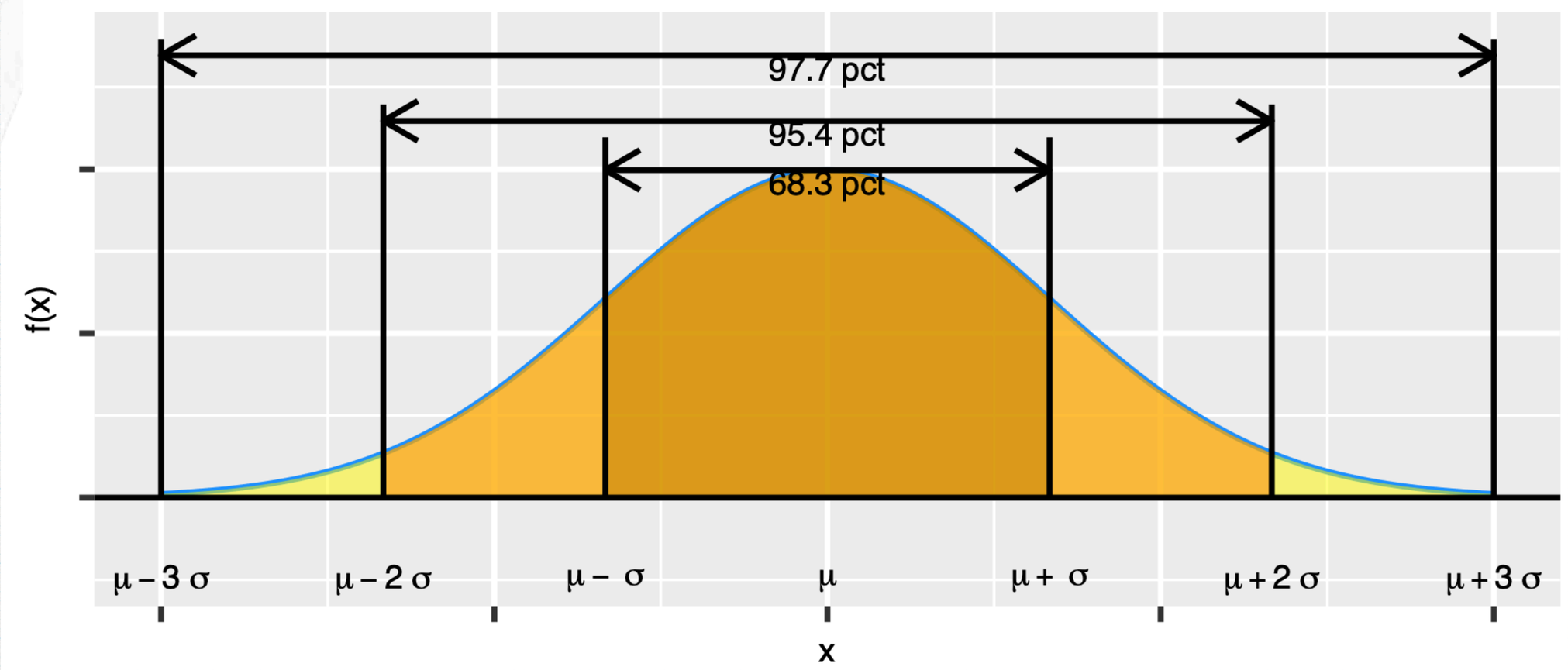
Interval length, confidence level, and precision



68,3% des valeurs d'une v.a normalement distribuée sont comprises dans l'intervalle $[\mu - \sigma; \mu + \sigma]$

68,3% values of r.v normally distributed are in the range $[\mu - \sigma; \mu + \sigma]$

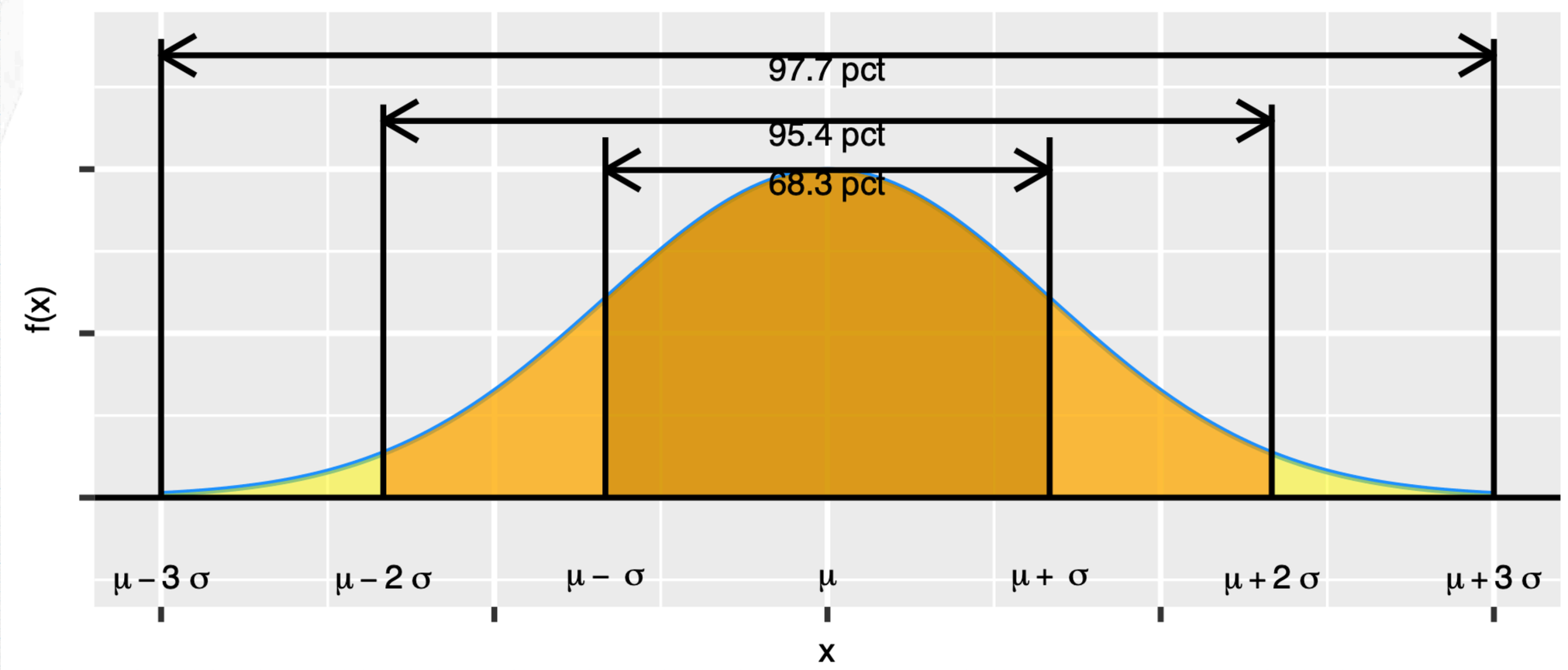
Interval length, confidence level, and precision



95,4% des valeurs d'une v.a normalement distribuée sont comprises dans l'intervalle $[\mu - 2\sigma; \mu + 2\sigma]$

95,4% values of r.v normally distributed are in the range $[\mu - 2\sigma; \mu + 2\sigma]$

Interval length, confidence level, and precision



99,7% des valeurs d'une v.a normalement distribuée sont comprises dans l'intervalle $[\mu - 3\sigma; \mu + 3\sigma]$

99,7% values of r.v normally distributed are in the range $[\mu - 3\sigma; \mu + 3\sigma]$

Interval length, confidence level, and precision

Rappelons-nous que, $\overline{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$, soit $Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$ (dans le cas d'une population de taille infinie).

Remember that, $\overline{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$, let $Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$ (in the case of a population of infinite size).

$$
\begin{aligned}
1 - \alpha = \mathbb{P}(-z_{\alpha/2} < Z < z_{\alpha/2}) &= \left(z_{\alpha/2} < \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}\right) \\
&= \mathbb{P}\left(-z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}} < \overline{X} - \mu < z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}}\right) \\
&= \mathbb{P}\left(-\overline{X} - z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}} < -\mu < -\overline{X} + z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}}\right) \\
&= \mathbb{P}\left(\overline{X} + z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}} > \mu > \overline{X} - z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}}\right) \\
&= \mathbb{P}\left(\overline{X} - z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}}\right)
\end{aligned}
$$

Interval length, confidence level, and precision

Il est donc envisageable d'encadrer la vraie valeur du paramètre à l'aide des valeurs suivantes :

It is therefore possible to frame the true value of the parameter using the following values:

$$\overline{X} - z_{\alpha/2} \times \sigma_{\overline{X}} < \mu < \overline{X} + z_{\alpha/2} \times \sigma_{\overline{X}}$$

où $z_{\alpha/2}$ est la valeur du quantile d'ordre $\alpha/2$, lue dans la table de la distribution de la normale centrée réduite.

where $z_{\alpha/2}$ is the value of the quantile of order $\alpha/2$, read from the table of reduced centered normal distribution.

z est égal au nombre d'écarts-types dont l'estimation s'écarte de l'estimateur ponctuel de la moyenne.

z is equal to the number of standard deviations whose estimate deviates from the estimator punctual from the average.

Confidence level

En général, on utilise de manière arbitraire un niveau de confiance de 90%, de 95% ou de 99%.

Le niveau de confiance représente le pourcentage d'intervalles qui incluraient le paramètre de la population si nous l'estimions à l'aide d'un très grand nombre d'échantillons différents de même taille, issus de la même population :

- admettons que l'on constitue 10 000 échantillons différents de taille n dans notre population de taille N
- sur chacun de ces échantillons, nous calculons l'estimation ponctuelle du paramètre (par exemple, la moyenne, estimée par la moyenne d'échantillonnage)
- sur chacun de ces échantillons, nous calculons également l'intervalle de confiance pour chaque estimation, en se donnant un niveau de confiance de 95% par exemple

Usually, a confidence level of 90%, 95% or 99% is used arbitrarily.

The confidence level represents the percentage of intervals that would include the population parameter if we estimated it using a very large number of different samples of the same size from the same population:

- let us assume that we constitute 10,000 different samples of size n in our population of size N
- on each of these samples, we calculate the punctual estimation of the parameter (for example, the mean, estimated by the sampling mean)
- on each of these samples, we also calculate the confidence interval for each estimate, giving ourselves a confidence level of 95% for example

Confidence level

Dans ce cas, nous nous attendons à ce que :

- dans 95% des 10 000 échantillons (soit 9 500 d'entre-eux), la vraie valeur du paramètre de la population soit incluse dans l'intervalle de confiance calculé
- dans les 5% restants (soit 500 d'entre-eux), la vraie valeur du paramètre de la population ne soit pas dans l'intervalle de confiance calculé.

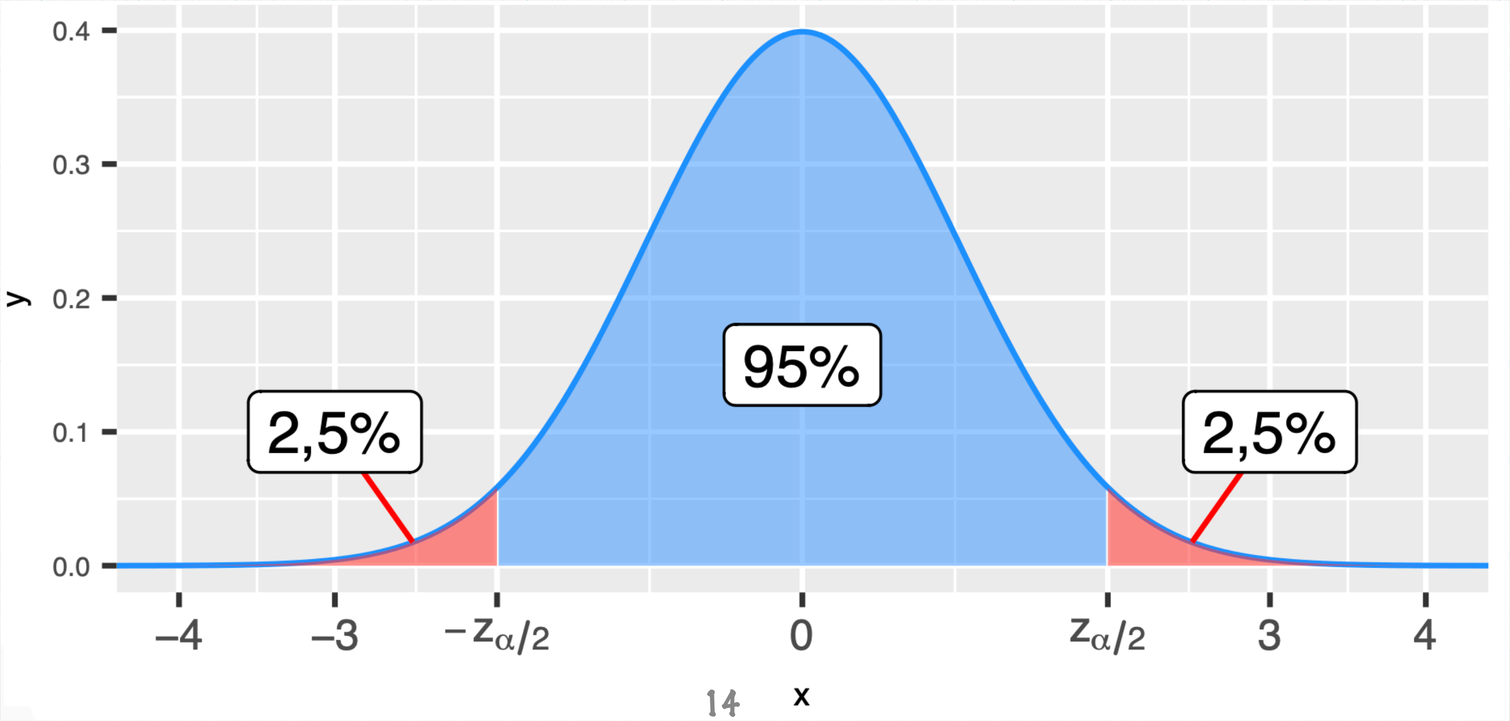
In this case, we expect that:

- in 95% of the 10,000 samples (or 9,500 of them), the true parameter value of the population is included in the calculated confidence interval
- in the remaining 5% (or 500 of them), the true parameter value of the population is not within the calculated confidence interval.

95% Confidence level

Admettons que l'on cherche à obtenir un intervalle de confiance à 95% (on définit $\alpha = 5\,\%$) pour le paramètre à estimer. Cherchons la valeur du quantile à identifier dans la table de la distribution normale centrée-réduite.

Let us assume that we are trying to obtain a 95% confidence interval (we define $\alpha = 5\,\%$) for the parameter to be estimated. Let us to find the value of the quantile to be identified in the table of the centered-reduced normal distribution.



95% Confidence level

Il s'agit donc de trouver, pour $Z \sim \mathcal{N}(0,1)$:

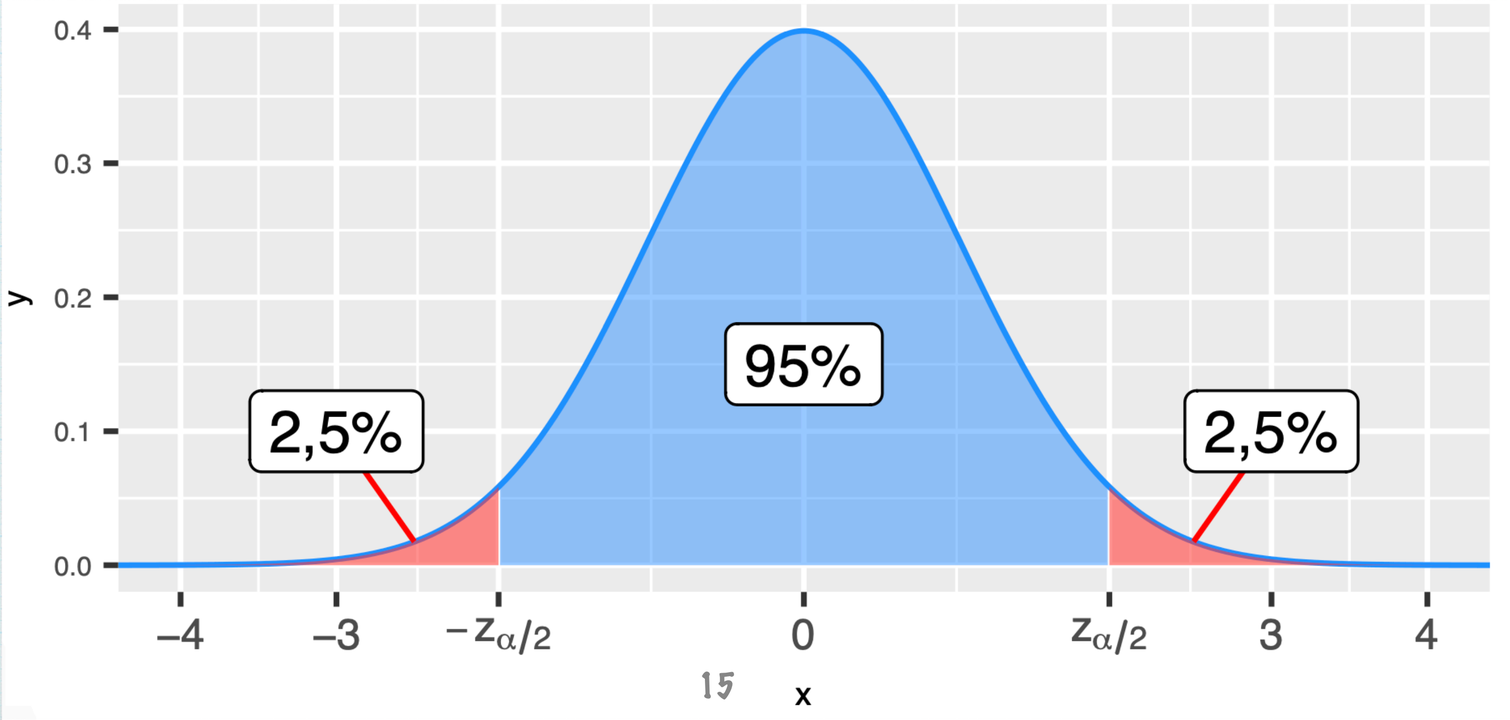
- Le quantile $-z_{\alpha/2}$ tel que $\mathbb{P}(Z < -z_{\alpha/2}) = 2{,}5\,\%$

➡ C'est-à-dire $-z_{\alpha/2}$ tel que $1 - \mathbb{P}(Z < z_{\alpha/2}) = 2{,}5\,\%$

It is therefore a question of finding, for $Z \sim \mathcal{N}(0,1)$

- The quantile $-z_{\alpha/2}$ as $\mathbb{P}(Z < -z_{\alpha/2}) = 2{,}5\,\%$

➡ that is to say $-z_{\alpha/2}$ as $1 - \mathbb{P}(Z < z_{\alpha/2}) = 2{,}5\,\%$



95% Confidence level

Il s'agit donc de trouver, pour $Z \sim \mathcal{N}(0,1)$:

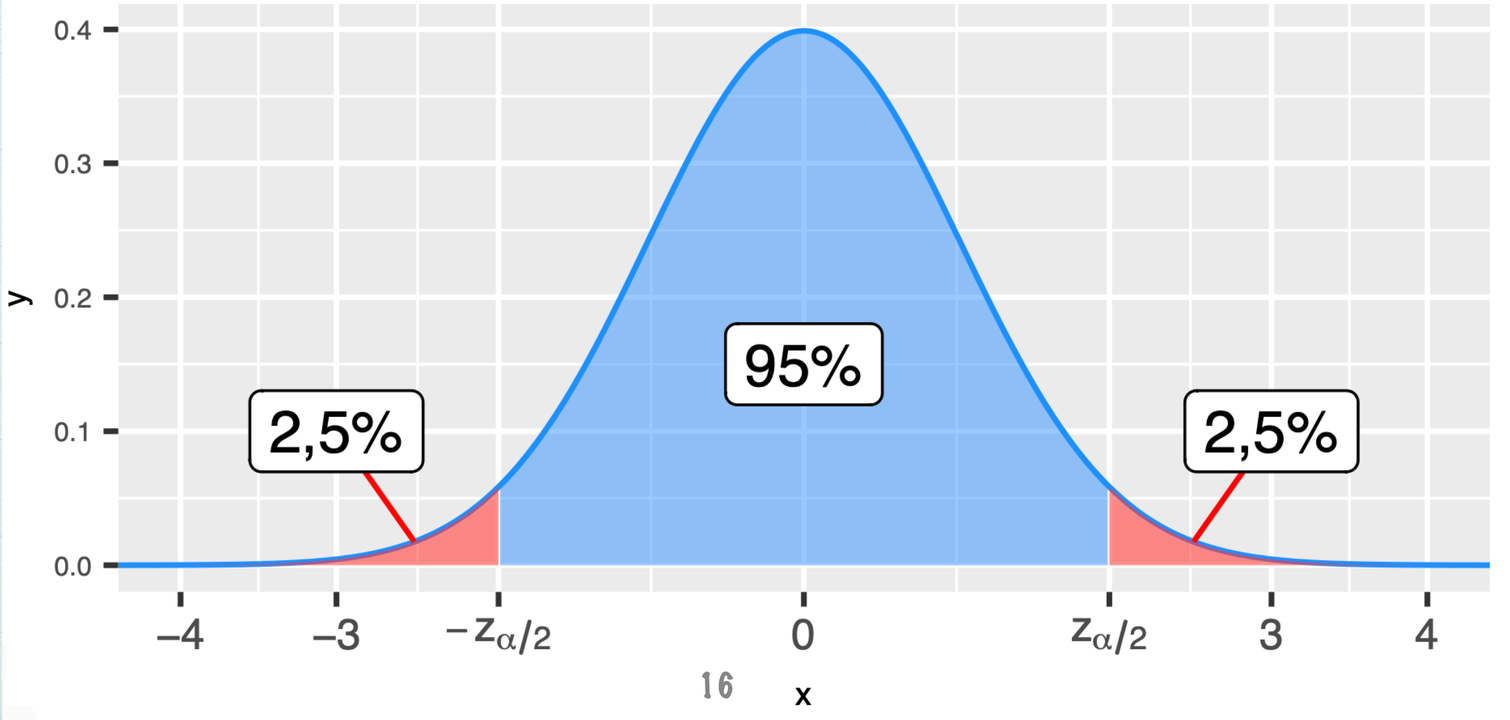
- Le quantile $z_{\alpha/2}$ tel que $\mathbb{P}(Z > z_{\alpha/2}) = 2{,}5\,\%$

➡ C'est-à-dire $z_{\alpha/2}$ tel que $1 - \mathbb{P}(Z < z_{\alpha/2}) = 2{,}5\,\%$

It is therefore a question of finding, for $Z \sim \mathcal{N}(0,1)$

- The quantile $z_{\alpha/2}$ as $\mathbb{P}(Z > z_{\alpha/2}) = 2{,}5\,\%$

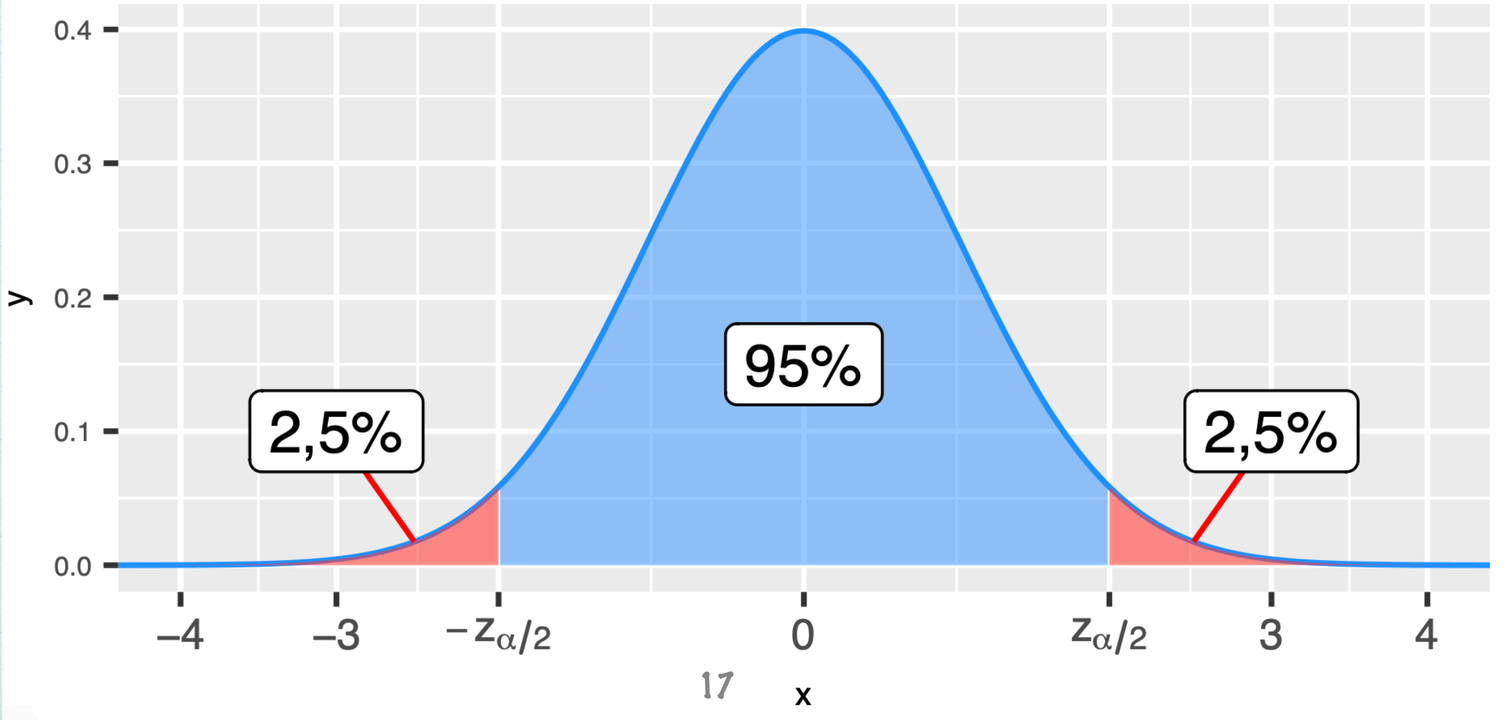
➡ that is to say $z_{\alpha/2}$ as $1 - \mathbb{P}(Z < z_{\alpha/2}) = 2{,}5\,\%$



95% Confidence level

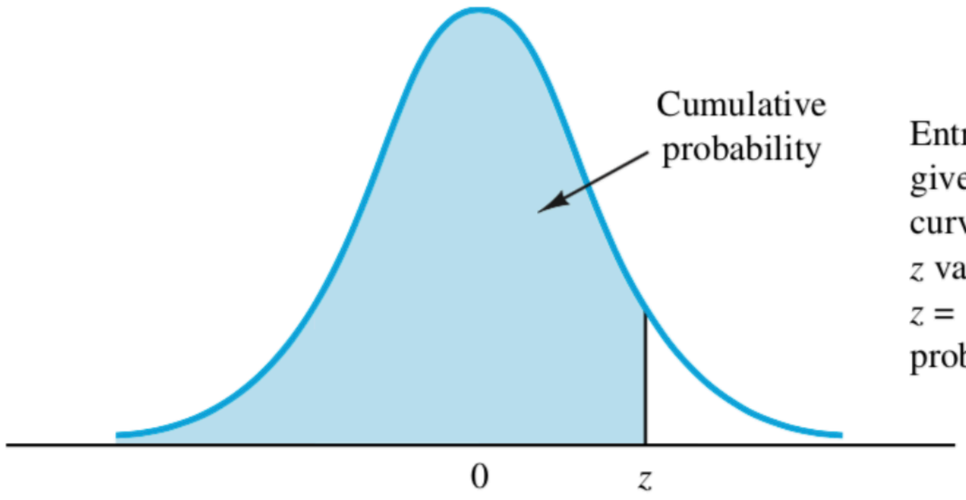
Il s'agit donc d'obtenir la valeur de $z_{\alpha/2}$ telle $\mathbb{P}(Z < z_{\alpha/2}) = 97{,}5$, dans la table de la $\mathcal{N}(0,1)$

It is therefore a question of obtaining the value of $z_{\alpha/2}$ such $\mathbb{P}(Z < z_{\alpha/2}) = 97{,}5$, in the table of $\mathcal{N}(0,1)$



95% Confidence level

CUMULATIVE PROBABILITIES FOR THE STANDARD NORMAL DISTRIBUTION



Entries in the table give the area under the curve to the left of the z value. For example, for $z = 1.25$, the cumulative probability is .8944.

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767

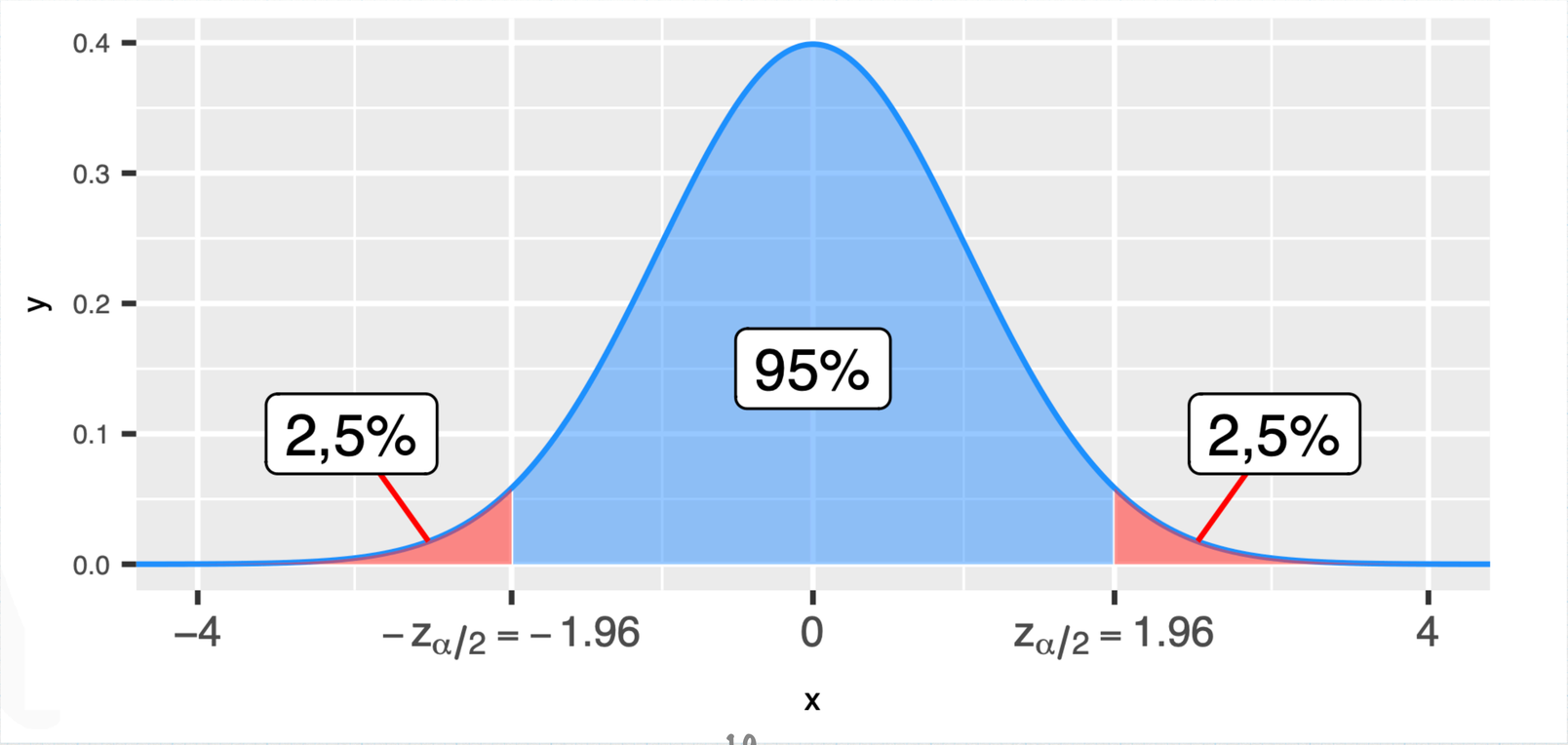
Figure 1: Source : Anderson D.R., Sweeney D.J., Williams T.A., *Statistics for Business and Economics (11th Edition)*, Cengage Learning, 2011

95% Confidence level

Nous avons donc lu dans la table que la valeur de $z_{\alpha/2}$ telle que $\mathbb{P}(Z < z_{\alpha/2})$ est $z_{\alpha/2}$ = 1,96.

We have therefore read in the table that the value of $z_{\alpha/2}$ such that $\mathbb{P}(Z < z_{\alpha/2})$ is $z_{\alpha/2}$= 1.96.

$$-z_{\alpha/2} = 1{,}96 \textbf{ and } z_{\alpha/2} = 1{,}96$$

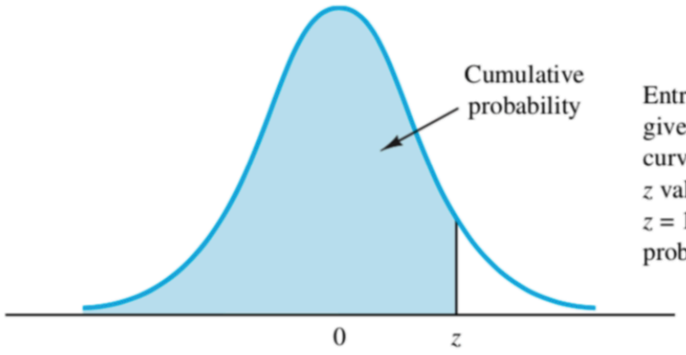


Confidence level

Même raisonnement pour des intervalles de confiance à 90% et 99%

Same reasoning for 90% and 99% confidence intervals:

CUMULATIVE PROBABILITIES FOR THE STANDARD NORMAL DISTRIBUTION



Entries in the table give the area under the curve to the left of the z value. For example, for $z = 1.25$, the cumulative probability is .8944.

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986

Figure 2: Source : Anderson D.R., Sweeney D.J., Williams T.A., *Statistics for Business and Economics (11th Edition)*, Cengage Learning, 2011

Confidence level

En résumant ces informations dans un tableau :

By summarizing this information in a table:

Niveau de confiance α	Valeur de $z_{\alpha/2}$	Intervalle de confiance
90%	$1,64$	$\overline{X} - 1,64\sigma_{\overline{X}} < \mu < \overline{X} + 1,64\sigma_{\overline{X}}$
95%	$1,96$	$\overline{X} - 1,96\sigma_{\overline{X}} < \mu < \overline{X} + 1,96\sigma_{\overline{X}}$
99%	$2,58$	$\overline{X} - 2,58\sigma_{\overline{X}} < \mu < \overline{X} + 2,58\sigma_{\overline{X}}$

Trade-off between confidence level and precision

Il existe ainsi, comme le montrent les résultats précédents, un arbitrage entre le niveau de confiance et la précision de l'estimation :

- lorsque le niveau de confiance α augmente, l'estimation perd en précision (la longueur de l'intervalle de confiance croît)
- lorsque le niveau de confiance α diminue, l'estimation gagne en précision (la longueur de l'intervalle de confiance diminue)

Ainsi, pour un niveau de confiance α donné, le seul moyen d'accroître la précision de l'estimation est d'augmenter la taille de l'échantillon : l'écart-type de l'estimateur dépend de la taille de l'échantillon.

As the previous results show, there is thus a trade-off between the level of confidence and the precision of the estimate:

- when the confidence level α increases, the estimation loses precision (the length of the confidence interval increases)
- when the level of confidence α decreases, the estimation becomes more precise (the length of the confidence interval decreases)

Thus, for a given confidence level α, the only way to increase the precision of the estimate is to increase the sample size: the standard deviation of the estimator depends on the sample size .

The confidence interval of the mean

Nous avons vu que pour calculer l'intervalle de confiance de la moyenne, le paramètre d'écart-type de X, c'est-à-dire $\sigma_{\overline{X}} = \dfrac{\sigma}{\sqrt{n}}$ fait intervenir l'écart-type σ du paramètre étudié dans la population.
Aussi, pour construire l'intervalle de confiance de la moyenne, il est nécessaire :

- soit de connaître cet écart-type
- soit de l'estimer.

We have seen that to calculate the confidence interval of the mean, the standard deviation parameter of X, i.e. $\sigma_{\overline{X}} = \dfrac{\sigma}{\sqrt{n}}$ involves the standard deviation σ of the parameter studied in the population.
Also, to construct the confidence interval of the mean, it is necessary:

- to know this standard deviation
- or to estimate it.

1.2. With known standard deviation

With known standard deviation

Considérons le cas où l'on connaît l'écart-type σ.

Rappelons que la définition de l'écart-type de la distribution d'échantillonnage des moyennes est donnée par :

- $\sigma_{\overline{X}} = \dfrac{\sigma}{\sqrt{n}}$ lorsque la taille de la population est infinie.
- $\sigma_{\overline{X}} = \dfrac{\sigma}{\sqrt{n}}\sqrt{\dfrac{N-n}{N-1}}$ lorsque la taille de la population est finie.

Consider the case where we know the standard deviation σ.

Recall that the definition of the standard deviation of the sampling distribution of the means is given by:

- $\sigma_{\overline{X}} = \dfrac{\sigma}{\sqrt{n}}$ when the size of the population is infinite.
- $\sigma_{\overline{X}} = \dfrac{\sigma}{\sqrt{n}}\sqrt{\dfrac{N-n}{N-1}}$ when the size of the population is finite.

With known standard deviation

Il ne reste plus alors qu'à déterminer le niveau de confiance avec lequel l'estimation doit être réalisée (et donc la valeur de $z_{\alpha/2}$ à préciser) pour effectuer l'estimation par intervalle de μ :

All that remains is to determine the level of confidence with which the estimate must be carried out (and therefore the value of $z_{\alpha/2}$ to be specified) to carry out the estimate by interval of μ:

$$\overline{X} - z_{\alpha/2} \times \sigma_{\overline{X}} < \mu < \overline{X} + z_{\alpha/2} \times \sigma_{\overline{X}}$$

Then

$$I.C._{\mu}(1-\alpha) = [\overline{X} - z_{\alpha/2} \times \sigma_{\overline{X}}; \overline{X} + z_{\alpha/2} \times \sigma_{\overline{X}}]$$

Example 1

Un responsable d'agence bancaire cherche à déterminer le temps que passe en moyenne un client à un DAB pour se faire délivrer de l'argent.
Pour un échantillon aléatoire de 36 retraits, le temps moyen est d'une minute et trente secondes.
En supposant que $\sigma = 0{,}3$ min, donner une estimation par intervalle du temps moyen de retrait à un DAB avec un niveau de confiance de 95%.

A bank branch manager seeks to determine the average time a customer spends at an ATM to receive cash.
For a random sample of 36 withdrawals, the average time is one minute and thirty seconds.
Assuming $\sigma = 0{,}3$ min, give an interval estimate of the mean ATM withdrawal time with a 95% confidence level.

Example 1

Aucune référence n'est faite quant à la taille de la population, que nous considérerons comme infinie.
À l'aide des informations disponibles, nous pouvons établir que :

No reference is made to the size of the population, which we will consider to be infinite.
Using the information available, we can establish that:

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}} = \frac{0{,}3}{\sqrt{36}} = 0{,}05$$

Le niveau de confiance doit-être de 95%. Ainsi, $\alpha = 1 - 95\,\% = 5\,\%$. Le quantile de la loi normale centrée réduite d'ordre $\alpha/2 = 0{,}025$ est $z_{0{,}025} = 1{,}96$

The confidence level should be 95%. Thus, $\alpha = 1 - 95\,\% = 5\,\%$. The quantile of the reduced centered normal distribution of order $\alpha/2 = 0{,}025$ is $z_{0{,}025} = 1{,}96$.

Example 1

Par conséquent, on a : Consequently, we have:

$$\overline{X} - 1{,}96 \times \sigma_{\overline{X}} < \mu < \overline{X} + 1{,}96 \times \sigma_{\overline{X}}$$

$$\mathbb{P}(\overline{X} - 1{,}96 \times \sigma_{\overline{X}} < \mu < \overline{X} + 1{,}96 \times \sigma_{\overline{X}}) = 0{,}95$$

On obtient : We obtain :

$$1{,}5 - 1{,}96 \times 0{,}05 < \mu < 1{,}5 + 1{,}96 \times 0{,}05$$

$$1{,}5 - 0{,}098 < \mu < 1{,}5 + 0{,}098$$

$$1{,}402 < \mu < 1{,}598$$

$$\mathbb{P}(1{,}402 < \mu < 1{,}598) = 0{,}95$$

Il y a ainsi 95% de chances que le temps nécessaire en moyenne pour retirer de l'argent au DAB soit compris entre 1,402 et 1,598 minutes (entre 1min 24 sec et 1min 36 sec).

There is a 95% chance that the average time it takes to withdraw money from an ATM will be between 1.402 and 1.598 minutes (between 1min 24sec and 1min 36sec).

Example 1

Nous venons de trouver

We just found

$$I.C_{\mu}(0{,}95) = [\overline{X} - 1{,}96 \times \sigma_{\overline{X}}; \overline{X} + 1{,}96 \times \sigma_{\overline{X}}] = [1{,}402; 1{,}598]$$

On remarque bien que la valeur $\mu = 1{,}5$ est au centre de cet intervalle :

Note that the value $\mu = 1{,}5$ is at the center of this interval:

$$\frac{1{,}402 + 1{,}598}{2} = 1{,}5$$

Example 2

Une GSB (grande surface de bricolage) vient de recevoir une livraison de 100 tuyaux d'arrosage.
Elle veut vérifier si le diamètre des tuyaux respecte la préconisation faite pour le diamètre.
Elle prélève 50 tuyaux, dont le diamètre moyen est de 2,55 cm.
Par le passé, le diamètre des tuyaux a admis comme écart-type 0,07cm.
Donner une estimation par intervalle du diamètre des tuyaux reçus, au niveau de confiance de 99%.

A GSB (DIY supermarket) has just received a delivery of 100 garden hoses.
It wants to check if the diameter of the pipes respects the recommendation made for the diameter.
It takes 50 pipes, the average diameter of which is 2.55 cm.
In the past, the diameter of the pipes allowed as standard deviation 0.07cm.
Give an interval estimation of the diameter of the pipes received, at the 99% confidence level.

Example 2

À l'aide des informations disponibles, on peut établir que :

Using the information available, it can be established that:

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} = \frac{0{,}07}{\sqrt{50}}\sqrt{\frac{100-50}{100-1}} = 0{,}007 \text{ cm}$$

Le niveau de confiance doit-être de 99%. Ainsi, $\alpha = 1 - 99\,\% = 1\,\% = 0{,}01$. Le quantile de la loi normale centrée réduite d'ordre $\alpha/2 = 0{,}005$ est $z_{0{,}005} = 2{,}58$

The confidence level should be 99%. Thus, $\alpha = 1 - 99\,\% = 1\,\% = 0{,}01$. The quantile of the reduced centered normal distribution of order $\alpha/2 = 0{,}005$ is $z_{0{,}005} = 2{,}58$.

Example 2

Par conséquent, on a : Consequently, we have:

$$\overline{X} - 2{,}58 \times \sigma_{\overline{X}} < \mu < \overline{X} + 2{,}58 \times \sigma_{\overline{X}}$$

$$\mathbb{P}(\overline{X} - 2{,}58 \times \sigma_{\overline{X}} < \mu < \overline{X} + 2{,}58 \times \sigma_{\overline{X}}) = 0{,}99$$

On obtient : We obtain :

$$2{,}55 - 2{,}58 \times 0{,}007 < \mu < 2{,}55 + 2{,}58 \times 0{,}007$$

$$2{,}55 - 0{,}018 < \mu < 2{,}55 + 0{,}018$$

$$2{,}532 < \mu < 2{,}568$$

$$\mathbb{P}(2{,}532 < \mu < 2{,}568) = 0{,}99$$

Le diamètre des tuyaux livrés a ainsi 99% de chances d'être compris entre 2,532 et 2,568 centimètres

The diameter of the pipes delivered thus has a 99% chance of being between 2,532 and 2,568 centimeters

Procedure

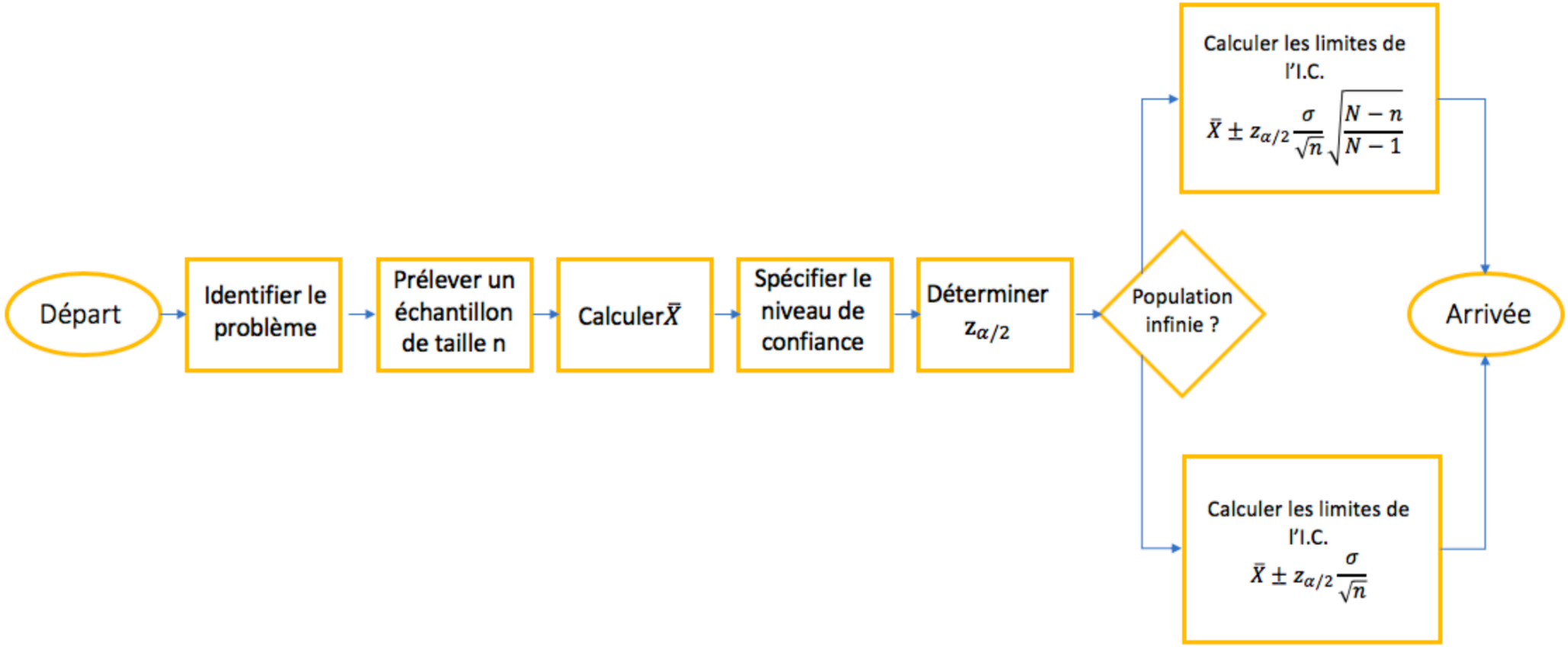


Figure 3: Algorithme de l'estimation d'une moyenne lorsque σ est connu.

1.3. With unknown standard deviation

With unknown standard deviation

Considérons le cas où l'on ne connaît pas l'écart-type σ dans la population.

Rappelons-nous que l'intervalle de confiance se calcule en faisant appel à la valeur de σ. Il est possible de connaître la valeur de σ si nous connaissons également la valeur de μ

- mais c'est justement μ que nous souhaitons estimer, donc nous ne connaissons pas sa valeur !

Aussi, l'hypothèse que nous avons effectuée précédemment, qui consiste à dire que l'on connaît σ ne peut durablement être tenue :

- dans des situations réelles, non seulement la moyenne μ mais aussi l'écart-type σ de la variable sur la population doivent être estimés.

Consider the case where we do not know standard deviation σ in the population.

Recall that the confidence interval is calculated using the value of σ. It is possible to know the value of σ if we also know the value of μ:

- but it is precisely μ that we want to estimate, so we do not know its value!

Also, the assumption that we made previously, which consists in saying that we know σ cannot be maintained in the long term:

- in real situations, not only the mean μ but also the standard deviation σ of the variable over the population must be estimated.

With unknown standard deviation

Nous avons vu dans le chapitre précédent que la moyenne d'échantillonnage $\overline{X}$ était un estimateur sans biais de la moyenne μ de la variable sur la population.

Nous pourrions ainsi être tentés de se dire que l'écart-type échantillonal doit dans le même esprit être un estimateur sans biais de l'écart-type σ de la variable sur la population :

We saw in the previous chapter that the sampling mean $\overline{X}$ was an unbiased estimator of the mean μ of the variable over the population.

We could thus be tempted to say that the sample standard deviation must in the same spirit be an unbiased estimator of the standard deviation σ of the variable over the population:

→ **It's wrong**

Standard deviation estimator

Notons S l'écart-type échantillonal, tel que :

Let S denote the sample standard deviation, such that:

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

Rappelons que l'écart-type de la variable sur la population est :

Recall that the standard deviation of the variable on the population is:

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{n}(x_i - \mu)^2}$$

Regardons pourquoi S n'est pas un estimateur sans biais de l'écart-type σ soit pourquoi S^2 n'est pas un estimateur sans biais de la variance σ^2

Let's see why S is not an unbiased estimator of the standard deviation σ, that is, why S^2 is not an unbiased estimator of the variance σ^2.

The expected value of S

Autrement dit, cherchons l'espérance de S , pour la comparer avec σ . Si ? on a $\mathbb{E}(S^2) \neq \sigma^2$,alors S^2 est un estimateur biaisé de σ^2 .

In other words, let us find the expectation of S, to compare it with σ. If ? on has $\mathbb{E}(S^2) \neq \sigma^2$, then S^2 is a biased estimator of σ^2.

$$\mathbb{E}(S^2) = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{X})^2\right)$$

$$= \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}\left[x_i^2 - 2\overline{X}X_i + \overline{X}^2\right]\right)$$

$$= \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}x_i^2 - \frac{2}{n}\sum_{i=1}^{n}\overline{X}X_i + \frac{1}{n}\sum_{i=1}^{n}\overline{X}^2\right)$$

$$= \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}X_i^2 - 2\overline{X}\underbrace{\frac{1}{n}\sum_{i=1}^{n}X_i}_{\overline{X}} + \frac{1}{n}\times n\overline{X}^2\right)$$

The expected value of S

$$\mathbb{E}(S^2) = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2 - 2\overline{X}\underbrace{\frac{1}{n}\sum_{i=1}^{n} X_i}_{\overline{X}} + \frac{1}{n} \times n\overline{X}^2\right)$$

$$= \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2 - 2\overline{X}^2 + \overline{X}^2\right)$$

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

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

The expected value of S

$$\mathbb{E}(S^2) = \frac{1}{n}\mathbb{E}\left(\sum_{i=1}^{n} X_i^2\right) - \mathbb{E}\left(\overline{X}^2\right)$$

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

$$= \frac{1}{n} \times n\mathbb{E}\left(X_i^2\right) - \mathbb{E}\left(\overline{X}^2\right)$$

$$= \mathbb{E}\left(X_i^2\right) - \mathbb{E}\left(\overline{X}^2\right) \qquad (1)$$

We know that : $\mathbb{V}ar(X) = \mathbb{E}\left[\left(X_i^2 - \mathbb{E}\left(\overline{X}^2\right)\right)\right] = \mathbb{E}(X^2) - \mathbb{E}(X)^2$

From where : $\mathbb{E}(X^2) = \mathbb{V}ar(X) + \mathbb{E}(X)^2$

The expected value of S

We know that : $\mathbb{V}ar(X) = \mathbb{E}\left[\left(X_i^2 - \mathbb{E}\left(\overline{X}^2\right)\right)\right] = \mathbb{E}(X^2) - \mathbb{E}(X)^2$

From where : $\mathbb{E}(X^2) = \mathbb{V}ar(X) + \mathbb{E}(X)^2$

Now, as : $\mathbb{V}ar(X) = \sigma^2$ **and** $\mathbb{E}(X) = \mu$

We deduce : $\mathbb{E}(X^2) = \sigma^2 + \mu^2$ $\longrightarrow$ (2)

The expected value of S

We also have : $\mathbb{V}ar(\overline{X}) = \mathbb{E}\left(\overline{X}^2\right) - \mathbb{E}\left(\overline{X}\right)^2$

From where : $\mathbb{E}(\overline{X}^2) = \mathbb{V}ar(\overline{X}) + \mathbb{E}(\overline{X})^2$

Now, as : $\mathbb{V}ar(\overline{X}) = \dfrac{\sigma^2}{n}$ **and** $\mathbb{E}(\overline{X}) = \mu$

We deduce : $\mathbb{E}(\overline{X}^2) = \dfrac{\sigma^2}{n} + \mu^2 \longrightarrow \quad (3)$

The expected value of S

$$\mathbb{E}(S^2) = \mathbb{E}\left(X_i^2\right) - \mathbb{E}\left(\overline{X}^2\right) \longrightarrow (1)$$

$$\mathbb{E}(X^2) = \sigma^2 + \mu^2 \longrightarrow (2)$$

$$\mathbb{E}(\overline{X}^2) = \frac{\sigma^2}{n} + \mu^2 \longrightarrow (3)$$

We deduce : $$\mathbb{E}(S^2) = \sigma^2 + \mu^2 - \left(\frac{\sigma^2}{n} + \mu^2\right)$$

$$= \sigma^2 \times \left(1 - \frac{1}{n}\right) = \frac{n-1}{n} \times \sigma^2$$

The expected value of S

Donc, $\mathbb{E}(S^2) \neq \sigma^2$: la variance échantillonnale S^2 est un estimateur biaisé de la variance σ^2 de la variable considérée sur la population .

Then, $\mathbb{E}(S^2) \neq \sigma^2$: the sample variance S^2 is a biased estimator of the variance σ^2 of the variable considered on the population.

the bias is equal to $\dfrac{n-1}{n}$

The consequences of bias

$$\mathbb{E}(S^2) = \frac{n-1}{n}\sigma^2$$

1. Le biais disparaît dès lors que la taille de l'échantillon devient très grande :

1. The bias disappears when the size of the sample becomes very large:

$$lim_{n\to+\infty}\mathbb{E}(S^2) = lim_{n\to\infty}\frac{n-1}{n}\sigma^2 = \sigma^2 - lim_{n\to\infty}\frac{1}{n}\sigma^2 = \sigma^2$$

The consequences of bias

$$\mathbb{E}(S^2) = \frac{n-1}{n}\sigma^2$$

2. Il faut donc corriger la variance échantillonnale S^2 par un facteur n pour obtenir un estimateur sans biais de σ^2 :

2. It is therefore necessary to correct the sample variance S^2 by a factor of n to obtain an unbiased estimator of σ^2:

$$\mathbb{E}\left(\frac{n}{n-1}S^2\right) = \frac{n}{n-1}\mathbb{E}(S^2) = \frac{n}{n-1}\frac{n-1}{n}\sigma^2 = \sigma^2$$

Cela signifie qu'à la variance échantillonnale:

This means that at the sample variance:

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

on préfèrera la variance échantillonnale corrigée:

we prefer the corrected sample variance:

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

The consequences of bias

$$\mathbb{E}(S^2) = \frac{n-1}{n}\sigma^2$$

3. La variance estimée de la distribution d'échantillonnage des moyennes $\hat{\sigma}^2_{\overline{X}}$ est un estimateur sans biais de la variance de la grandeur estimée sur la population :

3. The estimated variance of the sampling distribution of the means $\hat{\sigma}^2_{\overline{X}}$ is an unbiased estimator of the variance of the estimated quantity over the population:

$$\hat{\sigma}^2_{\overline{X}} = \frac{S_c^2}{n} = \frac{S^2}{n-1}$$

lorsque la taille de la population est infinie

when the population size is infinite

$$\hat{\sigma}^2_{\overline{X}} = \frac{S_c^2}{n} \times \frac{N-n}{N-1} = \frac{S^2}{n-1} \times \frac{N-n}{N-1}$$

lorsque la taille de la population est de taille finie

when the population size is finite size

The consequences of bias

$$\mathbb{E}(S^2) = \frac{n-1}{n}\sigma^2$$

4.L'écart-type estimé de la distribution d'échantillonnage des moyennes $\hat{\sigma}_{\overline{X}}$ s'écrit :

4.The estimated standard deviation of the sampling distribution of the means $\hat{\sigma}_{\overline{X}}$ is written

$$\hat{\sigma}_{\overline{X}} = \frac{S_c}{\sqrt{n}} = \frac{S}{\sqrt{n-1}}$$

lorsque la taille de la population est infinie

when the population size is infinite

$$\hat{\sigma}_{\overline{X}} = \frac{S_c}{n} \times \sqrt{\frac{N-n}{N-1}} = \frac{S^2}{n-1} \times \sqrt{\frac{N-n}{N-1}}$$

lorsque la taille de la population est de taille finie

when the population size is finite size

Confidence interval

L'intervalle de confiance pour μ devient alors :

The confidence interval for μ then becomes:

$$\overline{X} - z_{\alpha/2} \times \hat{\sigma}_{\overline{X}} < \mu < \overline{X} + z_{\alpha/2} \times \hat{\sigma}_{\overline{X}}$$

Ou, autrement écrit :

Or, otherwise written:

$$I.C_{\mu}(1-\alpha) = \left[\overline{X} \pm z_{\alpha/2} \times \hat{\sigma}_{\overline{X}}\right] = \left[\overline{X} - z_{\alpha/2} \times \hat{\sigma}_{\overline{X}}, \overline{X} + z_{\alpha/2} \times \hat{\sigma}_{\overline{X}}\right]$$

Cela n'est toutefois vrai que si la taille de l'échantillon est suffisamment grande (en pratique, si $n \geq 30$)

However, this is only true if the sample size is large enough (in practice, if $n \geq 30$)

Dans cette situation, on continue de repérer la valeur de $z_{\alpha/2}$ dans la table de la distribution de probabilités normale centrée réduite.

In this situation, we continue to locate the value of $z_{\alpha/2}$ in the table of the reduced centered normal probability distribution.

Example

Le patron d'un café veut estimer le "ticket" moyen payé par ses clients.
Sur la base d'un échantillon aléatoire de 101 clients, il obtient une moyenne échantillonnale de 3,50 euros et un écart-type échantillonnal de 0,75 euro.
Donner une estimation par intervalle du ticket moyen de sa clientèle, au niveau de confiance de 90%.

The cafe owner wants to estimate the average "ticket" paid by his customers.
Based on a random sample of 101 customers, he obtains a sample mean of 3.50 euros and a sample standard deviation of 0.75 euros.
Give an estimate by interval of the average ticket of its customers, at the 90% confidence level.

Example

Aucune référence n'est faite quant à la taille de la population que nous considérerons comme infinie.

No reference is made as to the size of the population which we consider to be infinite.

Nous sommes dans le cas où la variance de la variable observée (le coût des tickets) est inconnue.

We are in the case where the variance of the observed variable (the cost of tickets) is unknown.

À l'aide des informations disponibles, on peut établir que :

Using the information available, it can be established that:

$$\hat{\sigma}_{\overline{X}} = \frac{S}{\sqrt{n-1}} = \frac{0{,}75}{\sqrt{101-1}} = \frac{0{,}75}{10} = 0{,}075$$

Le niveau de confiance doit être de 90%. Ainsi, $\alpha = 1 - 90\,\% = 10\,\%$. Le quantile de la loi normale centrée réduite d'ordre $\alpha/2 = 0{,}05$ est $z_{0{,}05} = 1{,}648$

The confidence level should be 90%. Thus, $\alpha = 1 - 90\,\% = 10\,\%$. The quantile of the reduced centered normal distribution of order $\alpha/2 = 0{,}05$ is $z_{0{,}05} = 1{,}648$.

Example

Par conséquent, on a : Consequently, we have:

$$\overline{X} - 1{,}648 \times \hat{\sigma}_{\overline{X}} < \mu < \overline{X} + 1{,}6488 \times \hat{\sigma}_{\overline{X}}$$

$$\mathbb{P}(\overline{X} - 1{,}648 \times \hat{\sigma}_{\overline{X}} < \mu < \overline{X} + 1{,}648 \times \hat{\sigma}_{\overline{X}}) = 0{,}90$$

On obtient : We obtain :

$$3{,}50 - 1{,}648 \times 0{,}075 < \mu < 3{,}50 + 1{,}648 \times 0{,}075$$

$$3{,}50 - 0{,}1236 < \mu < 3{,}50 + 0{,}1236$$

$$3{,}3764 < \mu < 3{,}6236$$

$$\mathbb{P}(3{,}3764 < \mu < 3{,}6236) = 0{,}90$$

Le ticket moyen laissé par les clients du café a ainsi 90% de chances d'être compris entre 3, 38 euros et 3, 62 euros.

The average ticket left by cafe patrons has a 90% chance of being between 3.38 euros and 3.62 euros.

1.4.With unknown standard deviation and small sample size

With unknown standard deviation and small sample size

Lorsque l'échantillon est petit (en pratique, lorsque $n < 30$), il n'est plus possible d'utiliser la table de la distribution normale centrée réduite pour en tirer une valeur de $z_{\alpha/2}$

When the sample is small (in practice, when $n < 30$), it is no longer possible to use the reduced centered normal distribution table to derive a value of $z_{\alpha/2}$.

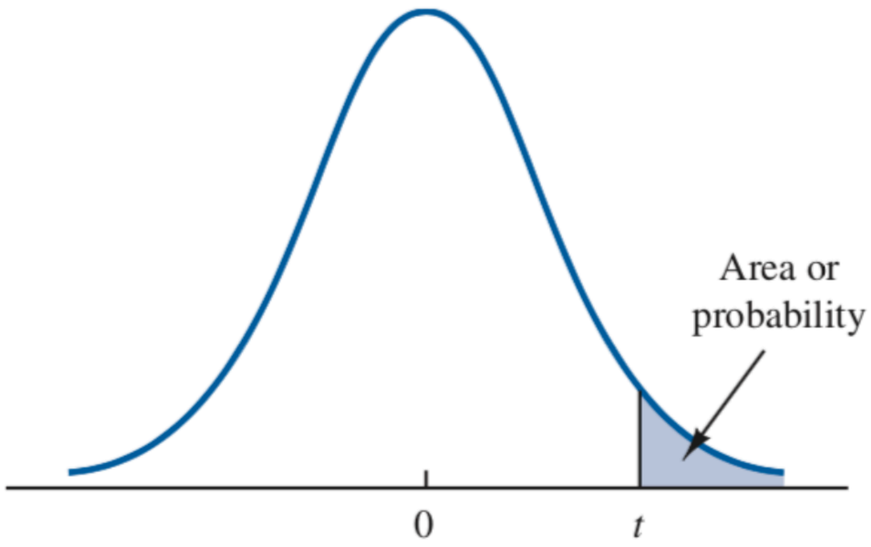
Dans ce cas, on utilise la table de la distribution de probabilité de Student

In this case, we use the table of the Student probability distribution

La distribution de Student a la même forme que la distribution normale (symétrie et concentration des valeurs autour de la moyenne), mais a des queues de distributions plus épaisses.

The Student distribution has the same shape as the normal distribution (symmetry and concentration of values around the mean), but has thicker tails of distributions.

With unknown standard deviation and small sample size



Entries in the table give t values for an area or probability in the upper tail of the t distribution. For example, with 10 degrees of freedom and a .05 area in the upper tail, $t_{.05} = 1.812$.

Degrees of Freedom	Area in Upper Tail					
	.20	.10	.05	.025	.01	.005
1	1.376	3.078	6.314	12.706	31.821	63.656
2	1.061	1.886	2.920	4.303	6.965	9.925
3	.978	1.638	2.353	3.182	4.541	5.841
4	.941	1.533	2.132	2.776	3.747	4.604

Figure 4: Source : Anderson D.R., Sweeney D.J., Williams T.A., *Statistics for Business and Economics (11th Edition)*, Cengage Learning, 2011

With unknown standard deviation and small sample size

En outre, on remarque que, pour un niveau de confiance donné (p dans la table), la valeur lue dans la table de Student tend vers la valeur lue dans la table de la $\mathcal{N}(0,1)$ lorsque n tend vers l'infini.

In addition, we notice that, for a given confidence level (p in the table), the value read in the Student table tends towards the value read in the table of the $\mathcal{N}(0,1)$ when n tends towards the infinite.

Par abus, on considère l'égalité des valeurs à partir de $n = 30$

By abuse, we consider the equality of values from $n = 30$.

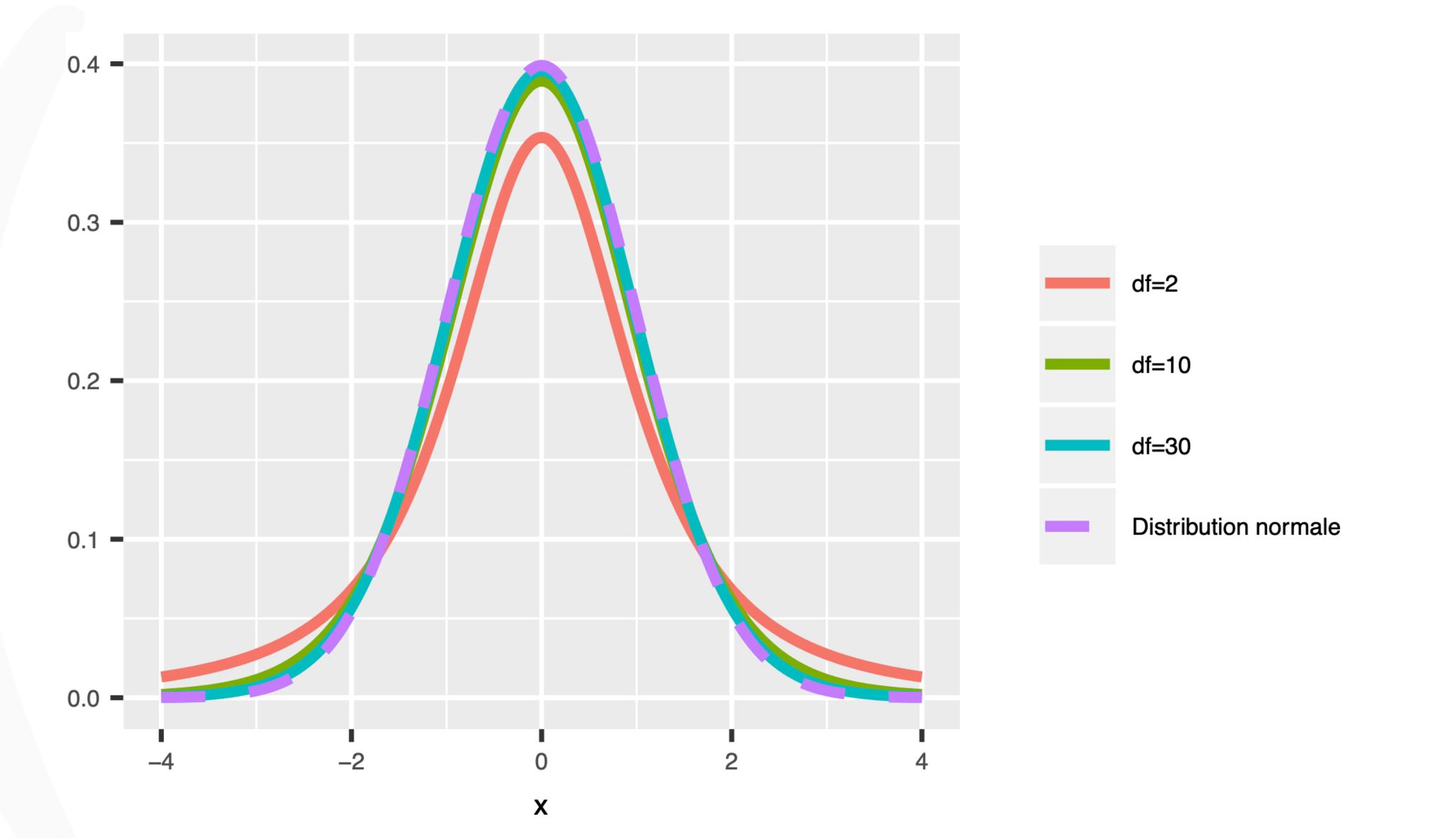
On remarque également dans la table de la distribution de Student qu'il est question, comparativement à la table de la $\mathcal{N}(0,1)$, du nombre de degrés de liberté :

We also notice in the table of the Student distribution that it is a question, compared to the table of the $\mathcal{N}(0,1)$, of the number of degrees of freedom

on utilise ici un nombre de degrés de liberté $v = n - 1$, soit la taille de l'échantillon moins un.

here we use a number of degrees of freedom $v = n - 1$, that is, the size of the sample minus one.

With unknown standard deviation and small sample size



With unknown standard deviation and small sample size

L'écriture de l'intervalle de confiance devient alors :

The writing of the confidence interval then becomes:

$$\overline{X} - t_{\alpha/2} \times \hat{\sigma}_{\overline{X}} < \mu < \overline{X} + t_{\alpha/2} \times \hat{\sigma}_{\overline{X}}$$

$$\mathbb{P}(\overline{X} - t_{\alpha/2} \times \hat{\sigma}_{\overline{X}} < \mu < \overline{X} + t_{\alpha/2} \times \hat{\sigma}_{\overline{X}}) = 1 - \alpha$$

Example

L'enseignante a utilisé un automate pour corriger les TPs du module Data Report. L'automate a corrigé donc 18 devoirs.

- Deux étudiants ont obtenu un 2
- Deux étudiants ont obtenu un 6
- Deux étudiants ont obtenu un 10
- Cinq étudiants on obtenu un 1
- Quatre étudiants ont obtenu un 16
- Deux étudiants ont obtenu un 18
- Un étudiant a obtenu un 20.

Estimer à l'aide d'un intervalle pour un niveau de confiance de 95% la note moyenne obtenue par les 180 étudiants de la promotion.

The teacher used an automaton to correct the practicals of the Data Report module. The automaton therefore corrected 18 assignments.

- Two students got a 2
- Two students got a 6
- Two students got a 10
- Five students got a 1
- Four students got a 16
- Two students got an 18
- One student got a 20.

Estimate, using an interval for a confidence level of 95%, the average mark obtained by the 180 students in the promotion.

Example

- Nous sommes dans le cas où nous devons estimer la moyenne d'une variable dans notre population (μ).
- Nous ne connaissons pas l'écart-type de cette variable σ.
- La taille de la population est finie : 180.
- L'échantillon est de petite taille : $n = 18 < 30$

À l'aide des informations disponibles, on peut calculer (à vérifier) : $\bar{x} = 12$ et $s = 5{,}16$

- We are in the case where we need to estimate the mean of a variable in our population (μ).
- We do not know the standard deviation of this variable σ.
- The size of the population is finite: 180.
- The sample is small: $n = 18 < 30$

Using the available information, we can calculate (to verify): $\bar{x} = 12$ and $s = 5{,}16$

We deduce that:

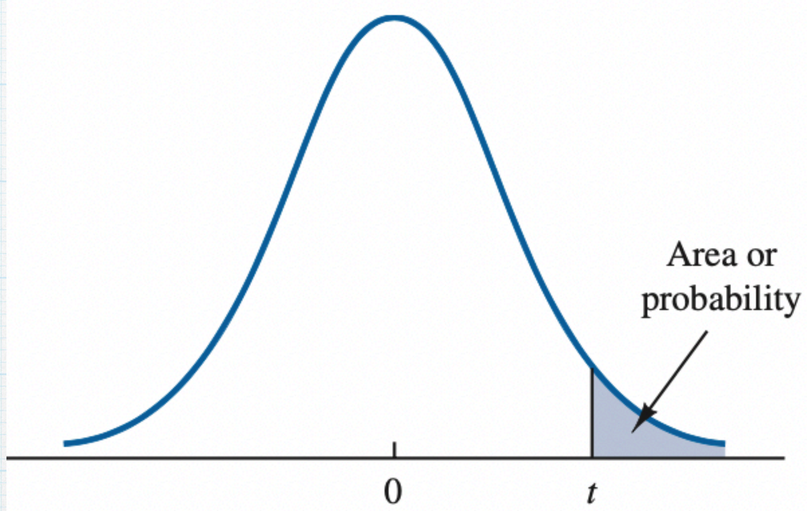
On en déduit que :

$$\hat{\sigma}_{\bar{x}} = \frac{S}{\sqrt{n-1}} \times \sqrt{\frac{N-n}{N-1}} = \frac{5{,}16}{\sqrt{18-1}} \times \sqrt{\frac{180-18}{180-1}} = \frac{5{,}16}{4{,}12} \times 0{,}95 = 1{,}19$$

Example

Le niveau de confiance doit être de 95%. Ainsi, $\alpha = 1 - 95\,\% = 5\,\%$. Le quantile de la loi de Student à $n - 1$ degrés de liberté d'ordre $\alpha/2 = 0{,}025$ est $t_{0{,}025} = 2{,}110$

The confidence level should be 95%. Thus, $\alpha = 1 - 95\,\% = 5\,\%$. The quantile of Student's law at $n - 1$degrees of freedom of order $\alpha/2 = 0{,}025$ is $t_{0{,}025} = 2{,}110$



Entries in the table give t values for an area or probability in the upper tail of the t distribution. For example, with 10 degrees of freedom and a .05 area in the upper tail, $t_{.05} = 1.812$.

Degrees of Freedom	Area in Upper Tail					
	.20	.10	.05	.025	.01	.005
15	.866	1.341	1.753	2.131	2.602	2.947
16	.865	1.337	1.746	2.120	2.583	2.921
17	.863	1.333	1.740	2.110	2.567	2.898
18	.862	1.330	1.734	2.101	2.552	2.878
19	.861	1.328	1.729	2.093	2.539	2.861

Example

Par conséquent, on a :

Consequently, we have:

$$\overline{X} - 2{,}110 \times \hat{\sigma}_{\overline{X}} < \mu < \overline{X} + 2{,}110 \times \hat{\sigma}_{\overline{X}}$$

$$\mathbb{P}(\overline{X} - 2{,}110 \times \hat{\sigma}_{\overline{X}} < \mu < \overline{X} + 2{,}110 \times \hat{\sigma}_{\overline{X}}) = 0{,}95$$

On obtient :

We obtain :

$$12 - 2{,}110 \times 1{,}19 < \mu < 12 + 2{,}110 \times 1{,}19$$

$$12 - 2{,}5109 < \mu < 12 + 2{,}5109$$

$$9{,}4891 < \mu < 14{,}5109$$

$$\mathbb{P}(9{,}492 < \mu < 14{,}51) = 0{,}95$$

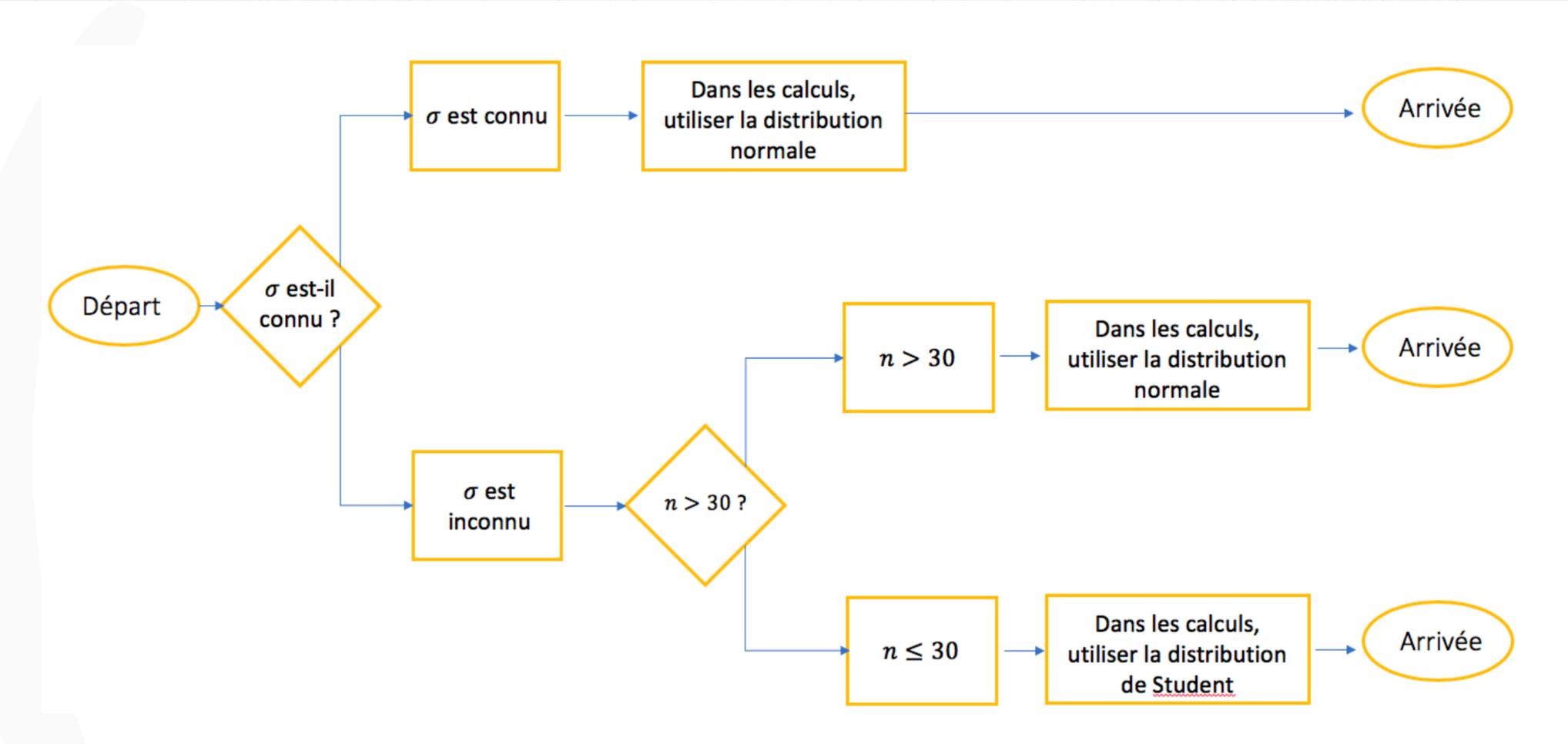
Il y a ainsi 95% de chances que la moyenne du test soit comprise entre 9,49 et 14,51.

There is a 95% chance that the test mean will be between 9.49 and 14.51.

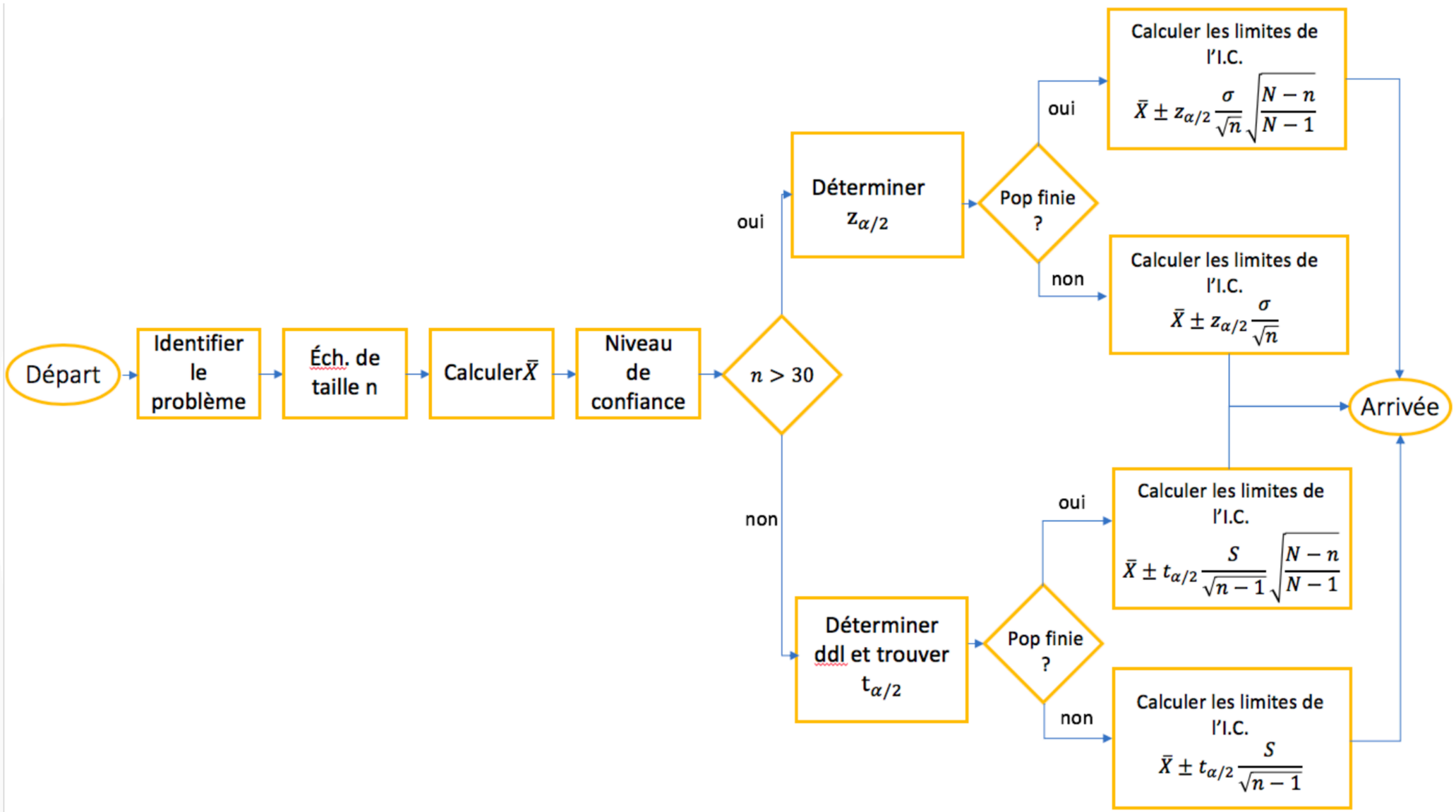
On remarque la faible précision de l'estimation, liée à la forte variabilité des notes relevées dans l'échantillon, ainsi qu'à la taille restreinte de l'échantillon.

Note the low precision of the estimation, linked to the high variability of the scores in the sample, as well as the small size of the sample.

Procedure


Départ
σ est-il connu ?
σ est connu
Dans les calculs, utiliser la distribution normale
Arrivée
σ est inconnu
$n > 30$?
$n > 30$
Dans les calculs, utiliser la distribution normale
Arrivée
$n \leq 30$
Dans les calculs, utiliser la distribution de Student
Arrivée

Procedure


Départ
Identifier le problème
Éch. de taille n
Calculer $\bar{X}$
Niveau de confiance
$n > 30$
oui
Déterminer $z_{\alpha/2}$
Pop finie ?
oui
Calculer les limites de l'I.C. $\bar{X} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$
non
Calculer les limites de l'I.C. $\bar{X} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$
non
Déterminer ddl et trouver $t_{\alpha/2}$
Pop finie ?
oui
Calculer les limites de l'I.C. $\bar{X} \pm t_{\alpha/2} \frac{S}{\sqrt{n-1}} \sqrt{\frac{N-n}{N-1}}$
non
Calculer les limites de l'I.C. $\bar{X} \pm t_{\alpha/2} \frac{S}{\sqrt{n-1}}$
Arrivée

Example 1

À l'aide des informations disponibles, on peut établir que :

Using the information available, it can be established that:

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} = \frac{0{,}07}{\sqrt{50}}\sqrt{\frac{100-50}{100-1}} = 0{,}007 \text{ cm}$$

Le niveau de confiance doit-être de 99%. Ainsi, $\alpha = 1 - 99\,\% = 1\,\% = 0{,}01$. Le quantile de la loi normale centrée réduite d'ordre $\alpha/2 = 0{,}005$ est $z_{0{,}005} = 2{,}58$

The confidence level should be 99%. Thus, $\alpha = 1 - 99\,\% = 1\,\% = 0{,}01$. The quantile of the reduced centered normal distribution of order $\alpha/2 = 0{,}005$ is $z_{0{,}005} = 2{,}58$.

Example 1

Par conséquent, on a : Consequently, we have:

$$\overline{X} - 2{,}58 \times \sigma_{\overline{X}} < \mu < \overline{X} + 2{,}58 \times \sigma_{\overline{X}}$$

$$\mathbb{P}(\overline{X} - 2{,}58 \times \sigma_{\overline{X}} < \mu < \overline{X} + 2{,}58 \times \sigma_{\overline{X}}) = 0{,}99$$

On obtient : We obtain :

$$2{,}55 - 2{,}58 \times 0{,}007 < \mu < 2{,}55 + 2{,}58 \times 0{,}007$$

$$2{,}55 - 0{,}018 < \mu < 2{,}55 + 0{,}018$$

$$2{,}532 < \mu < 2{,}568$$

$$\mathbb{P}(2{,}532 < \mu < 2{,}568) = 0{,}99$$

Le diamètre des tuyaux livrés a ainsi 99% de chances d'être compris entre 2,532 et 2,568 centimètres

The diameter of the pipes delivered thus has a 99% chance of being between 2,532 and 2,568 centimeters

Procedure

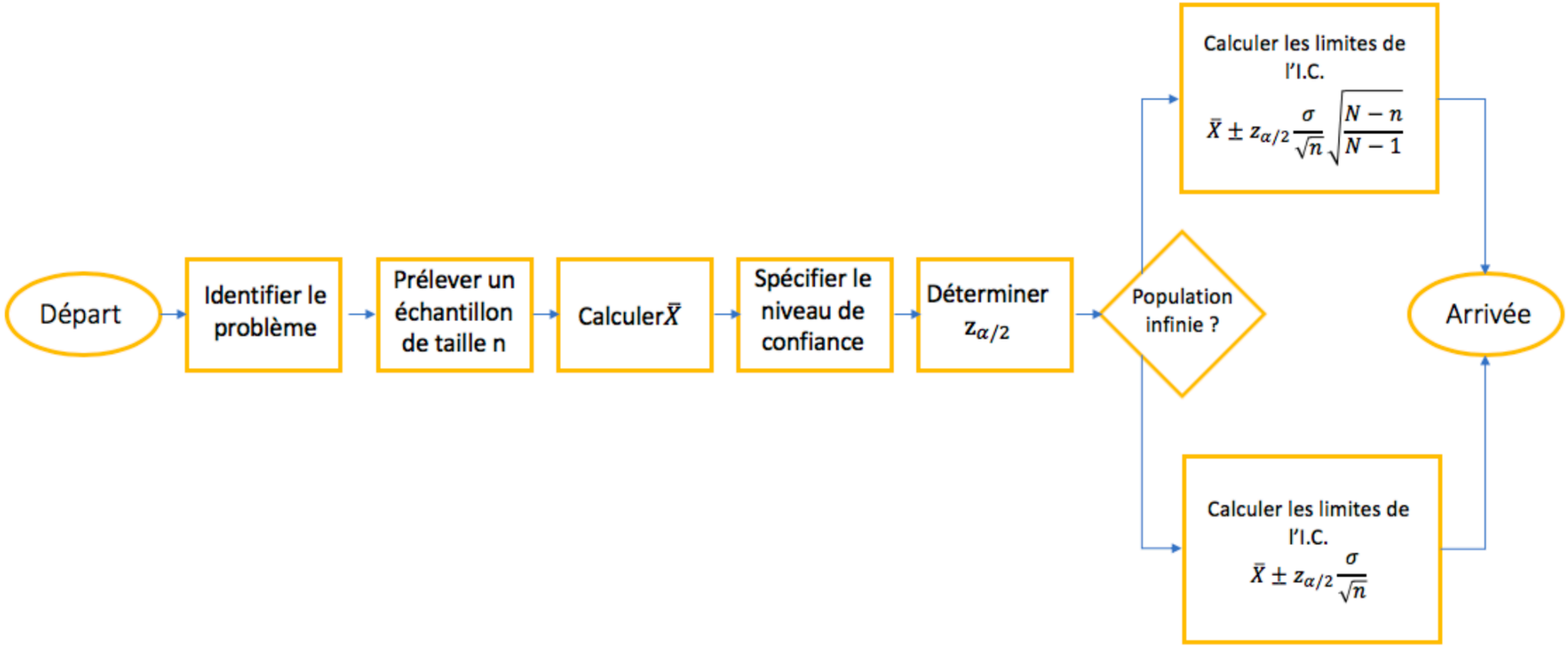


Figure 3: Algorithme de l'estimation d'une moyenne lorsque σ est connu.

2.Pourcentage Estimation

Pourcentage Estimation

À présent, concentrons-nous sur la construction d'un intervalle de confiance non plus pour la moyenne mais sur un pourcentage.

Par analogie avec ce qui a été écrit précédemment pour l'estimation par intervalle d'une moyenne, on obtient facilement :

Now, let's focus on building a confidence interval no longer for the mean but for a percentage.

By analogy with what was written previously for the interval estimation of an average, we easily obtain:

$$F - z_{\alpha/2} \times \sigma_F < p < F + z_{\alpha/2} \times \sigma_F$$

Pourcentage Estimation

Pour calculer l'intervalle de confiance, il suffirait de lire Z dans la table de la distribution de probabilité normale centrée-réduite

Rappelons que la définition de l'écart-type de la distribution d'échantillonnage des pourcentage est donnée par :

- $\sigma_F = \sqrt{\frac{p(1-p)}{n}}$ lorsque la taille de la population est infinie.
- $\sigma_F = \sqrt{\frac{p(1-p)}{n}}\sqrt{\frac{N-n}{N-1}}$ lorsque la taille de la population est finie.

To calculate the confidence interval, it would suffice to read Z in the table of the centered-reduced normal probability distribution

Recall that the definition of the standard deviation of the pourcentage sampling distribution is given by:

- $\sigma_F = \sqrt{\frac{p(1-p)}{n}}$ when the size of the population is infinite.
- $\sigma_F = \sqrt{\frac{p(1-p)}{n}}\sqrt{\frac{N-n}{N-1}}$ when the size of the population is finite.

Pourcentage Estimation

C'est exactement le même raisonnement que nous avons étudié pour la moyenne. pour encadrer la vra proportion p, il est nécessaire de connaître l'écart-type de la v.a. F, qui nécessite de connaître la vraie valeur de la proportion p. . .

Aussi, il est nécessaire d'obtenir une estimation de l'écart-type de la variable aléatoire F représentant la proportion d'échantillonnage :

- $\hat{\sigma}_F = \sqrt{\dfrac{f(1-f)}{n-1}}$ lorsque la taille de la population est infinie.
- $\hat{\sigma}_F = \sqrt{\dfrac{f(1-f)}{n-1}}\sqrt{\dfrac{N-n}{N-1}}$ lorsque la taille de la population est finie.

This is exactly the same reasoning we studied for the mean. to frame the true p-proportion, it is necessary to know the standard deviation of the v.a. F, which requires knowing the true value of the p-proportion. . .

Also, it is necessary to obtain an estimate of the standard deviation of the random variable F representing the sampling proportion

- $\hat{\sigma}_F = \sqrt{\dfrac{f(1-f)}{n-1}}$ when the size of the population is infinite.
- $\hat{\sigma}_F = \sqrt{\dfrac{f(1-f)}{n-1}}\sqrt{\dfrac{N-n}{N-1}}$ when the size of the population is finite.

Pourcentage Estimation

On peut montrer que $\hat{\sigma}_F$ la distribution d'échantillonnage des pourcentages est un estimateur sans biais de l'écart-type de la distribution d'échantillonnage des pourcentages

We can show that $\hat{\sigma}_F$ the sampling distribution of the percentages is an unbiased estimator of the standard deviation of the sampling distribution of percentages

The estimation interval then becomes:

L'intervalle d'estimation devient alors :

$$F - z_{\alpha/2} \times \hat{\sigma}_F < p < F + z_{\alpha/2} \times \hat{\sigma}_F$$

ou encore :

Or :

$$\mathbb{P}(F - z_{\alpha/2} \times \hat{\sigma}_{\overline{X}} < p < F + z_{\alpha/2} \times \hat{\sigma}_{\overline{X}}) = 1 - \alpha$$

où $1 - \alpha$ désigne le niveau de confiance

where $1 - \alpha$ denotes the confidence level.

NOTE :We will consider exclusively the situations where n ≥ 30.

Example 1

Un commerçant cherche à estimer la proportion de ses clients avec lesquels se produit un incident de paiement (moyen de paiement volé, compte non-approvisionné, etc.). Sur 1000 transactions étudiées, il parvient à identifier 50 incidents. Estimer la proportion d'incidents de paiement auxquels le commerçant s'expose avec sa clientèle pour un niveau de confiance de 95%.

A merchant seeks to estimate the proportion of his customers with whom a payment incident occurs (stolen means of payment, non-funded account, etc.).
Out of 1000 transactions studied, it manages to identify 50 incidents. Estimate the proportion of payment incidents to which the merchant is exposed with his customers for a confidence level of 95%.

Example 1

Aucune référence n'est faite quant à la taille de la population que nous considérerons comme infinie.

No reference is made as to the size of the population which we consider to be infinite.

Formalisons le problème. Soit X la variable aléatoire qui représente le nombre d'incidents dénombrés parmi 1000 transactions.

Let's formalize the problem. Let X be the random variable which represents the number of incidents counted among 1000 transactions.

Cette variable aléatoire décrit donc une répétition d'épreuves de Bernoulli indépendantes. On a donc :

This random variable therefore describes a repetition of independent Bernoulli tests. So we have :

$$X \sim \mathscr{B}in(n, p)$$

avec $n = 1000$ et p (la probabilité de succès, que l'on définit comme un défaut de paiement) inconnue.

with $n = 1000$ and p (the probability of success, which is defined as a default) unknown.

Example 1

Considérons la variable aléatoire F qui représente la proportion d'incidents parmi les 1000 transactions.

Consider the random variable F which represents the proportion of incidents among the 1000 transactions.

Cette variable aléatoire est ainsi définie comme :

This random variable is thus defined as:

$$F = \frac{X}{n} \sim \mathscr{B}in\left(p, \frac{p(1-p)}{n}\right)$$

On peut en effet vérifier que :

We can indeed verify that:

$$\mathbb{E}(F) = \mathbb{E}\left(\frac{X}{n}\right) = \frac{1}{n}\mathbb{E}(X) = \frac{np}{n} = p$$

$$\mathbb{V}ar(F) = \mathbb{V}ar\left(\frac{X}{n}\right) = \frac{1}{n^2}\mathbb{V}ar(X) = \frac{np(1-p)}{n^2} = \frac{p(1-p)}{n}$$

Example 1

Comme n est grand ($n \geq 30$), nous utilisons l'approximation de la loi Binomiale par une normale :

Since n is large ($n \geq 30$), we use the Binomial law approximation by a normal:

$$F \sim \mathcal{N}\left(p, \frac{p(1-p)}{n}\right)$$

Ou encore :

Or

$$Z = \frac{F - P}{\sqrt{\frac{p(1-p)}{n}}} \sim \mathcal{N}(0,1)$$

La table de la $\mathcal{N}(0,1)$ nous donne le quantile $z_{\alpha/2}$ tel que :

The table of $\mathcal{N}(0,1)$ gives us the quantile $z_{\alpha/2}$ such that:

$$\mathbb{P}(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha$$

$$\Leftrightarrow \mathbb{P}\left(-z_{\alpha/2} < \frac{F - p}{\sqrt{\frac{p(1-p)}{n}}} < z_{\alpha/2} \right) = 1 - \alpha$$

$$\Leftrightarrow \mathbb{P}\left(F - z_{\alpha/2} \times \sigma_F < p < F + z_{\alpha/2} \times \sigma_F\right) = 1 - \alpha$$

Example 1

L'intervalle de confiance d'ordre $1-\alpha$ pour la proportion est donc donnée par :

The confidence interval of order $1-\alpha$ for the proportion is therefore given by:

$$I.C._p(0{,}95) = \left[F - z_{\alpha/2} \times \sigma_F, F + z_{\alpha/2} \times \sigma_F\right]$$

Nous savons que $\sigma_F = \sqrt{\dfrac{p(1-p)}{n}}$, mais nous ne connaissons pas p (que nous cherchons à encadrer).

Nous allons utiliser une estimation $\hat{\sigma}_F$ de σ_F

we know that $\sigma_F = \sqrt{\dfrac{p(1-p)}{n}}$, but we do not know p (which we are trying to frame). We will use an estimate $\hat{\sigma}_F$ of σ_F

$$\hat{\sigma}_F = \sqrt{\frac{f(1-f)}{n-1}}$$

où f est la valeur observée de la proportion, qui, d'après l'énoncé, vaut $f = 50/1000 = 5\,\% = 0{,}05$

where f is the observed value of the proportion, which, according to the statement, is $f = 50/1000 = 5\,\% = 0{,}05$

Example 1

Ainsi, on a :

Thus, we have

$$\hat{\sigma}_F = \sqrt{\frac{f(1-f)}{n-1}} = \sqrt{\frac{0{,}05 \times 0{,}95}{1000-1}} = 0{,}0069 = 0{,}69\,\%$$

De fait, l'intervalle de confiance à 95% pour la proportion p vaut :

In fact, the 95% confidence interval for the proportion p is:

$$I.C._p(0{,}95) = \left[F - z_{\alpha/2} \times \sigma_F, F + z_{\alpha/2} \times \sigma_F\right]$$

$$\approx \left[F - z_{\alpha/2} \times \hat{\sigma}_F, F + z_{\alpha/2} \times \hat{\sigma}_F\right]$$

On lit dans la table de la loi normale centrée réduite que $z_{0{,}025} = 1{,}96$; comme par ailleurs la valeur observée de F vaut $0{,}05$, on a:

We read in the table of the reduced centered normal law that $z_{0{,}025} = 1{,}96$; as, moreover, the observed value of F is equal to $0{,}05$, we have:

$$I.C._p(0{,}95) \approx \left[0{,}05 - 1{,}96 \times 0{,}69\,\%, 0{,}05 + 1{,}96 \times 0{,}69\,\%\right] = [3{,}65\,\%\,;6{,}35\%]$$

Example 1

$$I.C._p(0{,}95) \approx \left[0{,}05 - 1{,}96 \times 0{,}69\,\%\,, 0{,}05 + 1{,}96 \times 0{,}69\,\%\right] = [3{,}65\,\%\,;\,6{,}35\%]$$

Il y a ainsi 95% de chances que le taux d'incident de paiement du commerçant soit compris entre 3, 65% et 6, 35%.

There is thus a 95% chance that the merchant's payment incident rate will be between 3.65% and 6.35%.

Example 2

Un poissonnier a fait l'objet d'une plainte de la part de clients mécontents de la fraîcheur de ses poissons. Lors du marché, il fait analyser 82 des 100 poissons qu'il vend ce jour. Sur les 82 poissons analysés, 27 dégagent une odeur insoutenable d'amoniaque, présentent des yeux foncés et s'avèrent de fait impropres à la consommation.
Estimer à l'aide d'un intervalle de confiance à 95% la part des poissons non consommables sur l'étal du vendeur.

A fish farmer was the subject of a complaint from customers unhappy with the freshness of his fish. At the market, he has 82 of the 100 fish he sells that day analyzed.
Of the 82 fish analyzed, 27 gave off an unbearable ammonia odor, had dark eyes and were in fact unfit for consumption.
Estimate using a 95% confidence interval the proportion of non-edible fish on the vendor's stall.

Example 2

À l'aide des informations disponibles, nous déduisons que nous sommes dans le cas d'une population finie de taille N = 100. Nous observons une proportion $f = 27/82 = 32\,\% = 0.3293$ pour un échantillon de taille $n = 82$

From the information available, we deduce that we are in the case of a finite population of size N = 100. We observe a proportion $f = 27/82 = 32\,\% = 0.3293$ for a sample size $n = 82$

Nous cherchons à encadrer l'estimation de la proportion dans la population des 100 poissons, à l'aide d'un intervalle de confiance à 95%. Nous avons besoin pour cela de connaître une estimation de la variance de la variable aléatoire F qui indique la proportion d'échantillonnage :

We seek to frame the estimate of the proportion in the population of 100 fish, using a 95% confidence interval. For this we need to know an estimate of the variance of the random variable F which indicates the sampling proportion:

$$\hat{\sigma}_F = \sqrt{\frac{f(1-f)}{n-1}}\sqrt{\frac{N-n}{N-1}}$$

$$\hat{\sigma}_F = \sqrt{\frac{0{,}3293 \times 0{,}6707}{82-1}}\sqrt{\frac{100-82}{100-1}} = 2.22\,\%$$

Example 2

De fait, l'intervalle de confiance à 95% pour la proportion p vaut :

In fact, the 95% confidence interval for the proportion p is:

$$I.C._p(0{,}95) = \left[F - z_{\alpha/2} \times \sigma_F, F + z_{\alpha/2} \times \sigma_F\right]$$
$$\approx \left[F - z_{\alpha/2} \times \hat{\sigma}_F, F + z_{\alpha/2} \times \hat{\sigma}_F\right]$$

On lit dans la table de la loi normale centrée réduite que $z_{0{,}025} = 1{,}96$; comme par ailleurs la valeur observée de F vaut $0{,}05$, on a:

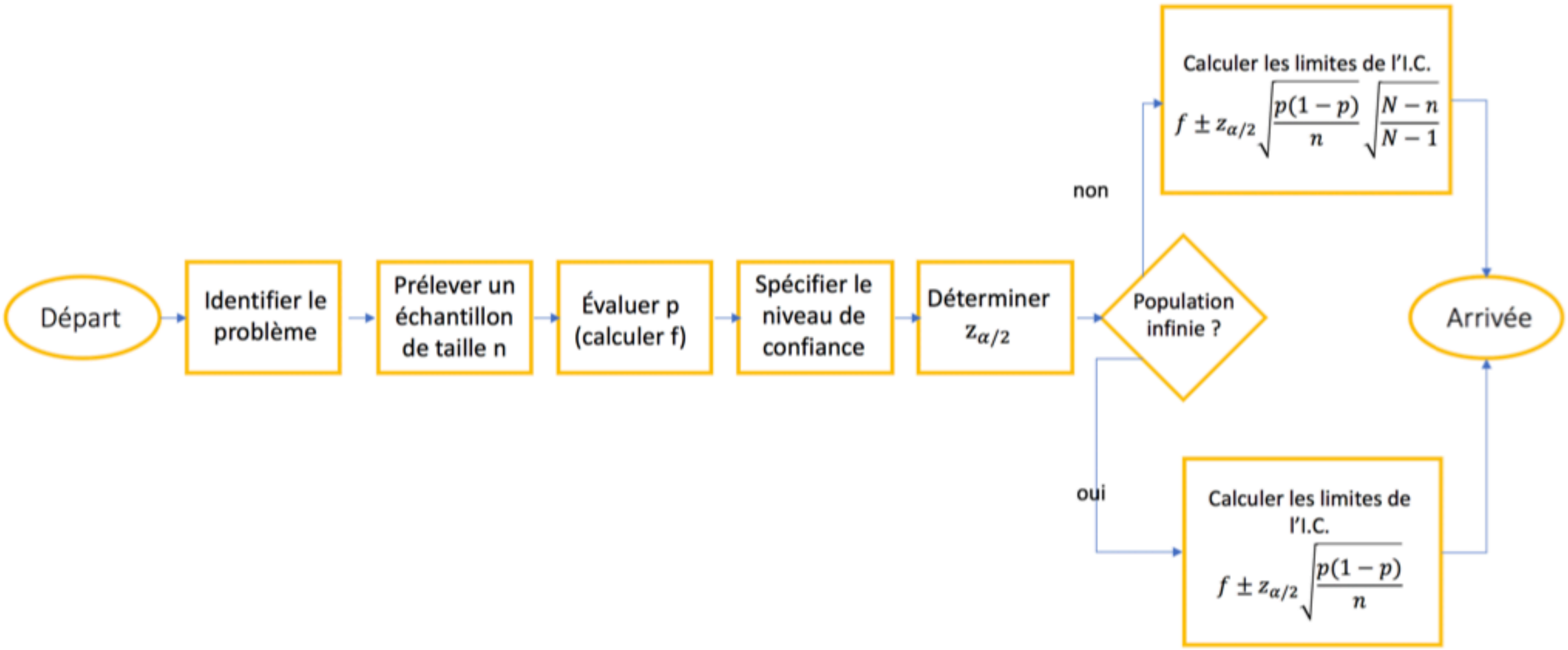
We read in the table of the reduced centered normal law that $z_{0{,}025} = 1{,}96$; as, moreover, the observed value of F is equal to $0{,}05$, we have:

$$I.C._p(0{,}95) \approx \left[0.329 - 1.96 \times 0.0222, 0.329 + 1.96 \times 0.0222\right] = [28.55\,\%, 37.25\%]$$

Il y a ainsi 95% de chances que la proportion de poissons impropres à la consommation soit comprise entre 28.55% et 37.25%.

Il y a ainsi 95% de chances que la proportion de poissons impropres à la consommation soit comprise entre 28, 55% et 37, 25%.

Procedure


Départ
Identifier le problème
Prélever un échantillon de taille n
Évaluer p (calculer f)
Spécifier le niveau de confiance
Déterminer $z_{\alpha/2}$
Population infinie ?
non
Calculer les limites de l'I.C.
$f \pm z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}\sqrt{\frac{N-n}{N-1}}$
oui
Calculer les limites de l'I.C.
$f \pm z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$
Arrivée