Contexte

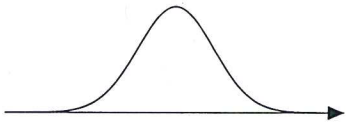
Les statistiques inférentielles ont pour but de déterminer les caractéristiques d'une population en utilisant les caractéristiques d'un échantillon provenant de cette population.

Il ne sera jamais possible de déterminer de manière exacte et certaine les caractéristiques de la population, mais il sera possible de les estimer plus ou moins précisément, en maîtrisant les risques de se tromper.

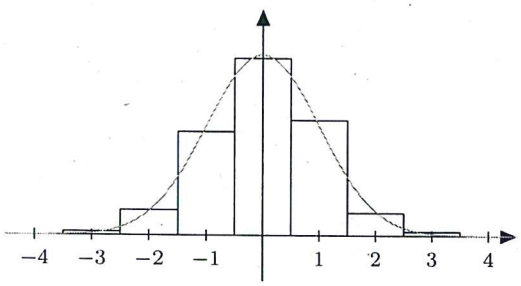
1 Loi normale

1.1 Définitions et notations

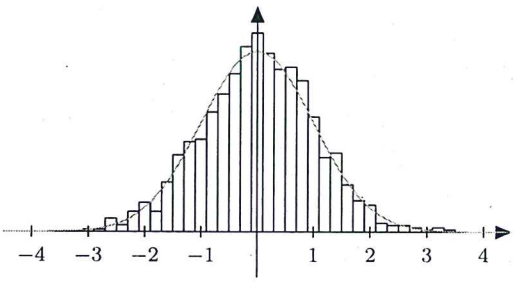
Dans de nombreux contextes, une variable statistique continue se distribue selon une courbe en cloche, appelée la courbe de Gauss.



Cette courbe en cloche correspond à la courbe de fréquence **théorique** de notre variable statistique. Si l'on disposait d'un échantillon extrêmement grand, et que l'on regroupait nos données en classes très petites, le polygone des fréquences que l'on obtiendrait ressemblerait à cette courbe.

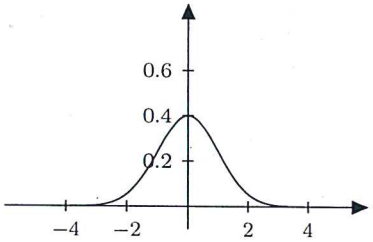


1000 données, 7 classes

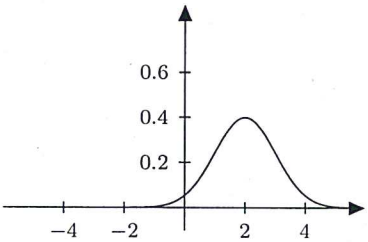


1000 données, 35 classes

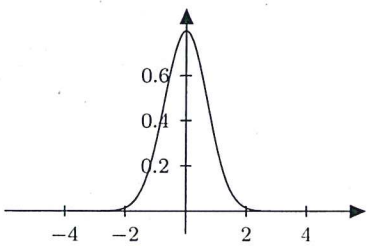
La distribution d'une variable suivant une loi normale ressemble toujours à une cloche, mais sa position et sa forme sont déterminées par la moyenne et la variance (ou l'écart-type) de la variable.



Moyenne $\mu = 0$, variance $\sigma^2 = 1$



Moyenne $\mu = 2$, variance $\sigma^2 = 1$



Moyenne $\mu = 0$, variance $\sigma^2 = 0.25$

$X \sim \mathcal{N}(2;1)$

Lorsqu'on dit qu'une variable suit une loi normale, on notera : $X \sim \mathcal{N}(\mu; \sigma^2)$

"suit une loi"

Propriétés

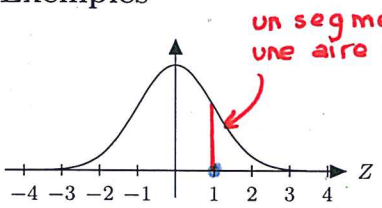
- La loi normale est symétrique autour de la moyenne μ
- L'aire sous la courbe vaut toujours 1

1.2 Représentation graphique

On considère une variable normale centrée réduite $Z \sim \mathcal{N}(0;1)$.
(centrée : moyenne $\mu = 0$, réduite : variance $\sigma^2 = 1$)

Sur la courbe de Gauss, on représente sur l'axe horizontal la valeur de la variable Z.
La probabilité d'obtenir certaines valeurs pour Z est donnée par l'aire sous la portion de courbe correspondante.

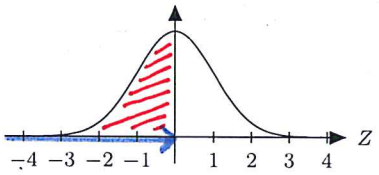
Exemples



$Z = 1$

la probabilité que Z soit égale à n'importe quelle valeur vaut toujours 0 car Z est continue

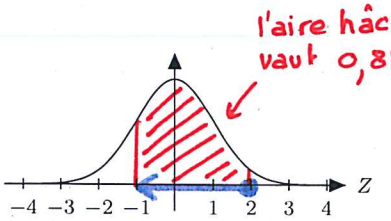
$P(Z = 1) = 0 = 0\%$



$Z < 0$

l'aire totale vaut 1 = 100% et la courbe est symétrique => la demi-aire vaut 50%

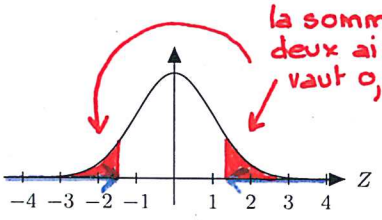
$P(Z < 0) = 0,5 = 50\%$



$-1 < Z \leq 2$ (noté aussi $Z \in]-1\,;\,2]$)

on apprendra à déterminer l'aire dans la partie 1.3.

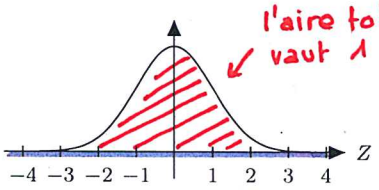
$P(-1 < Z \leq 2) = 0.8185 = 81.85\%$.



valeur absolue de Z : valeur de Z sans tenir compte du signe => soit $Z > 1,5$, soit $Z < -1,5$

$|Z| > 1.5$ (noté aussi $Z \in]-\infty\,;\,-1.5[\cup]1.5\,;\,+\infty[$)

$P(|Z| > 1.5) = 0.1336 = 13.36\%$.

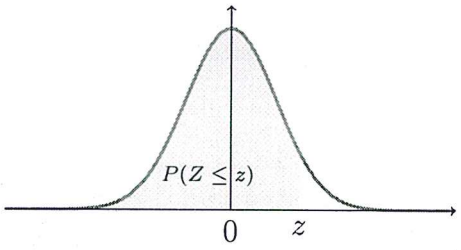


$Z \in \mathbb{R}$ (noté aussi $Z \in]-\infty\,;\,+\infty[$)

$P(Z \in \mathbb{R}) = 1 = 100\%$

1.3 Calculs de probabilités avec la table numérique

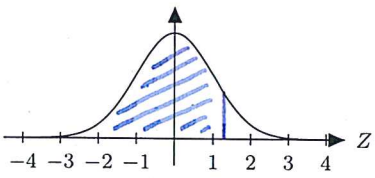
La table numérique de la loi normale centrée réduite (formulaire p.28) donne la probabilité d'obtenir pour la variable $Z \sim \mathcal{N}(0;1)$ une valeur inférieure ou égale à une valeur donnée $z \geq 0$:



Toutes les probabilités s'obtiennent en se ramenant à des calculs du type $P(Z \leq z)$, grâce aux deux propriétés de la loi normale :

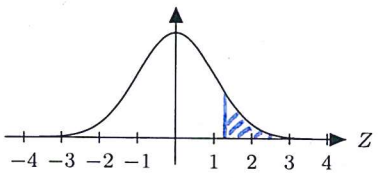
- symétrie autour de 0
- aire totale = 1 = 100%

Exemples



se trouve dans la table

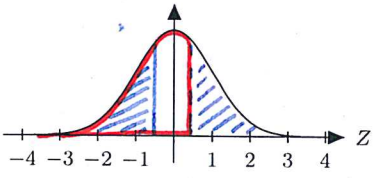
$P(Z \leq 1.25) =$ 0,8944 = 89,44%



aire totale — se trouve dans la table

$P(Z > 1.25) =$ 100% − P(Z ≤ 1,25)

= 100% − 89,44% = 10,56 %

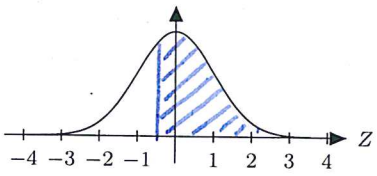


pas dans la table car −0,5 < 0 — symétrie — dans la table

$P(Z < -0.5) =$ P(Z > 0,5)

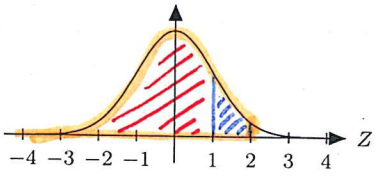
= 1 − P(Z ≤ 0,5)

= 1 − 0,6915 = 0,3085 = 30,85 %



$P(Z > -0.5) =$ $P(Z \leq 0{,}5) =$ (symétrie)

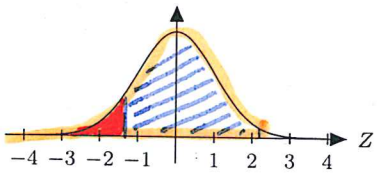
$= 0{,}6915 = 69{,}15\,\%$



$P(1 < Z < 2) =$ $P(Z < 2) - P(Z \leq 1)$ (dans la table)

$= 0{,}9772 - 0{,}8413 = 0{,}1359$

$= 13{,}59\,\%$



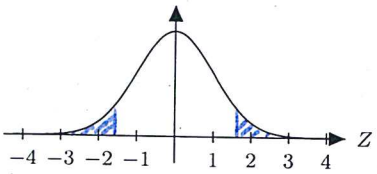
$P(Z - 1.34 < Z < 2.11) =$ $P(Z \leq 2{,}11) - P(Z \leq -1{,}34)$ (dans la table / pas dans la table)

$= 0{,}9826 - P(Z > 1{,}34)$

$= 0{,}9826 - (1 - P(Z \leq 1{,}34))$

$= 0{,}9826 - (1 - 0{,}9099) = 0{,}9826 - 0{,}0901$

$= 0{,}8925 = 89{,}25\,\%$



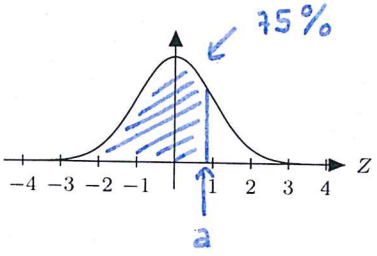
$P(|Z| > 1.65) =$ $2 \cdot P(Z > 1{,}65)$

$= 2 \cdot (1 - P(Z \leq 1{,}65)) = 2 \cdot (1 - 0{,}9505)$

$= 2 \cdot 0{,}0495 = 0{,}0990 = 9{,}90\,\%$

1.4 Recherche de quantiles avec la table numérique

Exemples

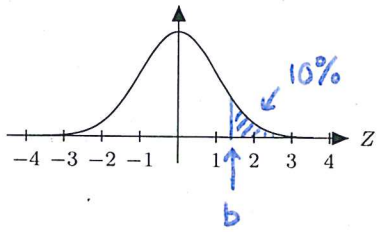


Déterminer a tel que $P(Z \leq a) = 75\%$ (chercher 0,7500 dans la table)

$\Rightarrow a \simeq 0{,}67$

le plus proche est $0{,}7486 = P(Z \leq 0{,}67)$

Remarque : $a = Q_3$ (troisième quartile).

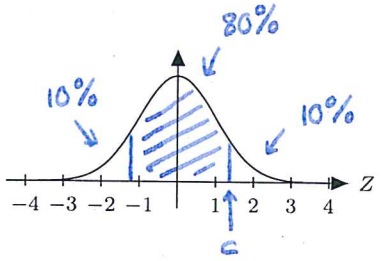


Déterminer b tel quel $P(Z \geq b) = 10\%$

$\Rightarrow P(Z \leq b) = 90\%$ (à chercher dans la table)

$P(Z \leq 1,28) = 0,8997$ (le plus proche de 0,90)

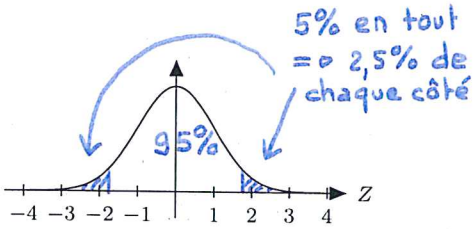
$\Rightarrow b \cong 1,28$



Déterminer c tel quel $P(|Z| < c) = 80\%$

$\Rightarrow P(Z \leq c) = 90\%$ (80% + 10%)

$\Rightarrow c \cong 1,28$



Déterminer d tel quel $P(|Z| \geq d) = 5\%$

$\Rightarrow P(Z \leq d) = 97,5\% = 0,975$

$\Rightarrow d = 1,96$

1.5 Normalisation et cote Z

Si une variable X suit une loi normale de moyenne μ et de variance σ^2, alors la cote Z de cette variable suit une loi normale centrée réduite :

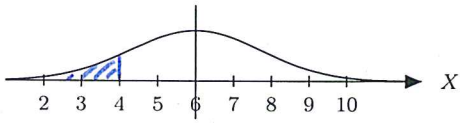
$$X \sim \mathcal{N}(\mu; \sigma^2) \quad \Longrightarrow \quad Z = \frac{X - \mu}{\sigma} \sim \mathcal{N}(0; 1)$$

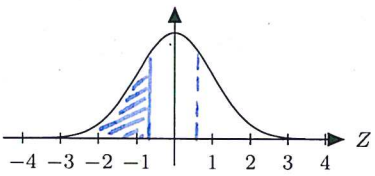
Lorsqu'on transforme une variable normale X en une variable normale centrée réduite Z, on dit qu'on **normalise** la variable X.

Comme on ne dispose pas d'une table numérique pour toutes les lois normales, on passera systématiquement par la cote Z lorsqu'on devra calculer des probabilités.

Exemples de calculs

a) Soit $X \sim \mathcal{N}(6;9)$. Calculer $P(X \leq 4)$.





cote z de 4

$\mathcal{N}(0;1)$

$P(X \leq 4) = P\left(Z \leq \frac{4-6}{3}\right)$

$= P(Z \leq -0{,}67)$

$= 1 - P(Z > -0{,}67)$

$= 1 - P(Z \leq 0{,}67) = 1 - 0{,}7486$

$= 0{,}2514 = 25{,}14\%$

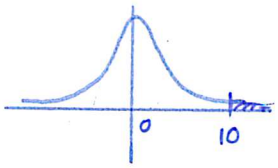
b) Soit $X \sim \mathcal{N}(-1\,;\,0.01)$. Calculer $P(X > 0)$.

$P(X > 0) = P\left(Z > \frac{0-(-1)}{\sqrt{0{,}01}}\right) = P(Z > 10) =$

$= 1 - P(Z \leq 10) = 1 - 1 \tilde{=} 0\%$

trop grand pour être dans la table

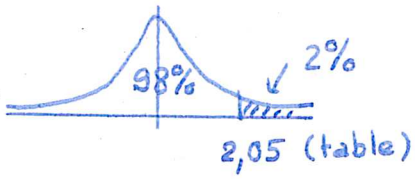
$\Rightarrow$ p ~ 100%



c) Soit $X \sim \mathcal{N}(3'000\,;\,40'000)$.
Déterminer a tel que $P(X > a) = 2\%$.

$P(X > a) = 2\%$

$P\left(Z > \frac{a - 3'000}{\sqrt{40'000}}\right) = 2\%$



D'après la table,

$\frac{a - 3'000}{200} = 2{,}05$ (2,05 est la cote z de a)

$\Rightarrow a = 3'000 + 2{,}05 \cdot 200 = 3410$

Exemple de problème

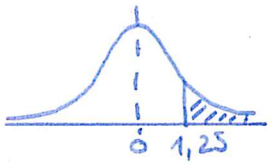
Un maraîcher observe que le poids de ses tomates suit une loi normale de moyenne 200 g et d'écart-type 40 g. $X \sim \mathcal{N}(200; 40^2)$

a) Quelle est la probabilité que le poids d'une tomate soit supérieur à 250 g ?

$P(X > 250) = P\left(Z > \frac{250-200}{40}\right) = P(Z > 1{,}25)$

$= 1 - P(Z \leq 1{,}25) = 1 - 0{,}8944$

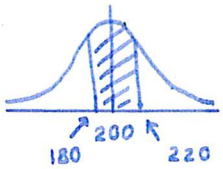
$= 0{,}1056 = 10{,}56\,\%$



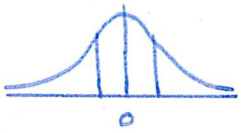
b) Quelle est la probabilité que le poids d'une tomate se situe à moins de 20 g du poids moyen ?

$P(180 < X < 220)$

$= P\left(\frac{180-200}{40} < Z < \frac{220-200}{40}\right)$



$= P(-0{,}5 < Z < 0{,}5)$



$= P(Z \leq 0{,}5) - P(Z \leq -0{,}5)$

$= P(Z \leq 0{,}5) - P(Z > 0{,}5) = P(Z \leq 0{,}5) - (1 - P(Z \leq 0{,}5))$

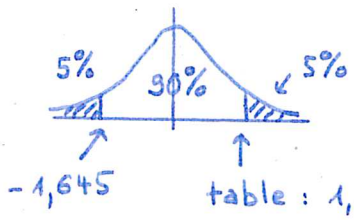
$= 0{,}6915 - (1 - 0{,}6915) = 38{,}3\,\%$

c) Le maraîcher garantit un poids minimal pour les tomates qu'il vend. Sachant qu'il n'y a que 5% de ses tomates qui ne peuvent pas être vendues à cause de leur poids, déterminer quel est le poids minimal garanti.

On cherche a tel que

$P(X \leq a) = 5\%$

$\Rightarrow P\left(Z \leq \frac{a-200}{40}\right) = 5\%$



(2 valeurs aussi proches de 0,95 : 1,64 et 1,65 $\Rightarrow$ on prend la moyenne)

$\Rightarrow \frac{a-200}{40} = -1{,}645$

$\Rightarrow a = 200 - 1{,}645 \cdot 40 = 134{,}2$

Le maraîcher garantit un poids minimal de 134,2 g

2 Théorème Central Limite (TCL)

Lorsqu'on étudie une certaine variable statistique, on va très souvent calculer la moyenne de cette variable sur un échantillon. Il sera alors utile de savoir de quelle manière se distribue cette moyenne par rapport aux caractéristiques réelles (et souvent inconnues) de la population (appelées les **paramètres** de la population).

2.1 Enoncé du TCL pour la moyenne d'un échantillon

On considère une variable X dont la moyenne vaut μ et la variance vaut σ^2.
On calcule la moyenne de cette variable sur un échantillon de taille n :

$$\bar{x} = \frac{x_1 + x_2 + ... + x_n}{n}$$

Alors, si X suit une loi normale ou si $n \geq 30$, cette moyenne suit également une loi normale.

$$\bar{x} \sim \mathcal{N}(\mu\,;\,\sigma_{\bar{x}}) \quad \text{avec} \quad \sigma_{\bar{x}} = \begin{cases} \frac{\sigma}{\sqrt{n}} & \text{si } n < \frac{N}{20} \text{ (petit échantillon)} \\ \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}} & \text{sinon (grand échantillon)} \end{cases}$$

s'appelle un facteur de correction

La moyenne $\bar{x}$ que l'on obtient se distribue donc selon une courbe en cloche, autour de la vraie moyenne de la population (paramètre μ). Plus l'échantillon est grand, plus l'écart-type de $\bar{x}$ est petit, donc plus on a de chances que la valeur obtenue pour $\bar{x}$ soit proche de la vraie moyenne de la population.

Remarque

Il ne faut pas confondre l'écart-type de la population, noté σ, et l'écart-type de $\bar{x}$, noté $\sigma_{\bar{x}}$!

2.2 Exemples

Mise en situation

Afin d'estimer le salaire moyen des anciens étudiants d'une université, on effectue un sondage auprès d'un échantillon de personnes en leur demandant leur salaire mensuel.

En interrogeant 10 personnes, on obtient un salaire mensuel moyen $\bar{x} = 7312$ francs.
En interrogeant 100 personnes, on obtient $\bar{x} = 7611$ francs.
Enfin, en interrogeant 1000 personnes, on obtient $\bar{x} = 7468$ francs.

Intuitivement, quelle estimation du salaire moyen est la plus précise ? 7468 francs

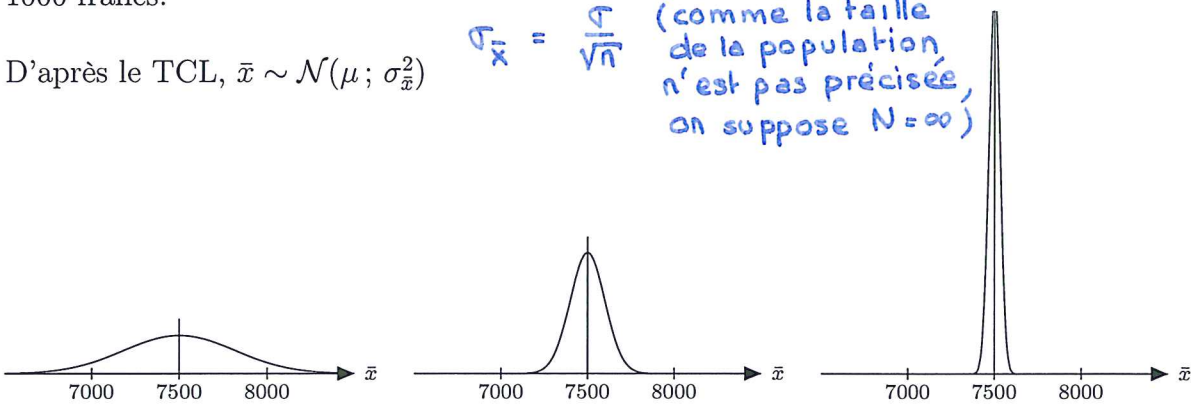
En interrogeant plus de personnes, on récolte plus d'informations sur la population ⇒ on devrait être plus précis

Cette intuition s'explique par le Théorème Central Limite :

Supposons que le salaire moyen en Suisse soit de 7500 francs par mois, avec un écart-type de 1000 francs.

D'après le TCL, $\bar{x} \sim \mathcal{N}(\mu\,;\,\sigma^2_{\bar{x}})$

$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ (comme la taille de la population n'est pas précisée, on suppose $N = \infty$)



n = 10	n = 100	n = 1000
$\mu =$ 7500	$\mu =$ 7500	$\mu =$ 7500
$\sigma_{\bar{x}} = \frac{1000}{\sqrt{10}} \cong 316$	$\sigma_{\bar{x}} = \frac{1000}{\sqrt{100}} = 100$	$\sigma_{\bar{x}} = \frac{1000}{\sqrt{1000}} \cong 32$

On voit bien que plus l'échantillon est grand, plus le risque d'obtenir une moyenne éloignée de 7500 est ..petit..... car la variance (et l'écart-type) de $\bar{x}$ devient plus .petit.....

Exemple de résolution d'exercice

Dans un supermarché, on a pu mesurer que les clients dépensaient en moyenne 250 francs le 24 décembre, avec un écart-type de 70 francs.
A la sortie du supermarché, on effectue un sondage auprès de 50 clients, en leur demandant le montant de leurs achats. On calcule ensuite la moyenne $\bar{x}$ des 50 réponses collectées.
On dénote par X le montant des achats des clients.

a) La variable X suit-elle une loi normale ?

Non, pas forcément ! (Rien n'est précisé dans l'énoncé)

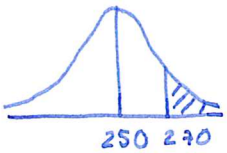
b) La variable $\bar{x}$ suit-elle une loi normale ?

pas de facteur de corr. car on suppose $N = \infty$

Oui, car $n \geq 30$

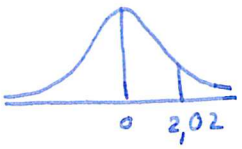
$\mu = 250$, $\sigma_{\bar{x}} = \frac{70}{\sqrt{50}} \cong 9{,}9$ francs $\Rightarrow \bar{x} \sim \mathcal{N}(250\,;\,9{,}9^2)$

c) Quelle est la probabilité que la moyenne obtenue par ce sondage soit supérieure à 270 francs ?



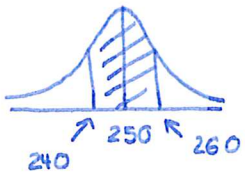
$$P(\bar{X} > 270) = P\left(Z > \frac{270 - 250}{9,9}\right)$$

$$= P(Z > 2,02) = 1 - P(Z \leq 2,02)$$



$$= 1 - 0,9783 = 0,0217 = 2,17\,\%$$

d) Quelle est la probabilité que la moyenne obtenue par ce sondage se situe à moins de 10 francs de la vraie moyenne ?



$$P(240 < \bar{X} < 260) = P\left(\frac{240 - 250}{9,9} < Z < \frac{260 - 250}{9,9}\right)$$

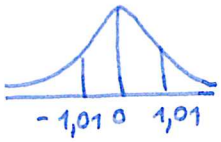
$$= P(-1,01 < Z < 1,01)$$

$$= P(Z \leq 1,01) - P(Z \leq -1,01)$$

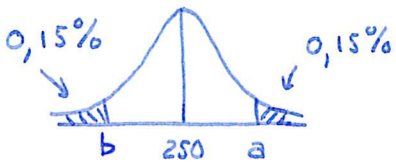
$$= P(Z \leq 1,01) - P(Z > 1,01)$$

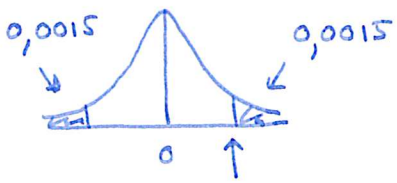
$$= P(Z \leq 1,01) - (1 - P(Z \leq 1,01))$$

$$= 0,8438 - (1 - 0,8438) = 0,6876 = 68,76\,\%$$



e) Si l'on néglige les valeurs ayant moins de 0.3% de chances de se produire, entre quelles valeurs devrait se situer $\bar{x}$?





on cherche 0,9985 dans la table

$\Rightarrow$ 2,967

On veut que

$P(-a < \bar{X} < a) = 99,7\,\%$

cote z de a : 2,967

$\Rightarrow a = 250 + 2,967 \cdot 9,9 \cong 279$

et $b = 250 - 2,967 \cdot 9,9 \cong 221$

$\Rightarrow$ entre 221 francs et 279 francs

3 Intervalles de confiance

3.1 Exemple d'introduction

Supposons que l'on cherche à estimer la taille moyenne μ des femmes suisses.
On demande donc à 1000 femmes suisses choisies au hasard leur taille, et on en calcule la moyenne : $\bar{x} = 164.2$ cm.

Peut-on alors affirmer que $\mu = 164.2$, donc que la taille moyenne de toutes les femmes suisse vaut 164.2 cm ? NON !!

Il est évidemment impossible de connaître la valeur de la vraie moyenne μ sans avoir interrogé toutes les femmes suisses. On peut par contre estimer que :

la taille moyenne des femmes suisses vaut environ 164.2 cm

Une telle estimation s'appelle une **estimation ponctuelle** (par une seule valeur).

Le problème avec une estimation ponctuelle est qu'on ne sait rien sur l'erreur que l'on risque de commettre. La vraie moyenne μ s'éloigne-t-elle de quelques millimètres, de quelques centimètres ou de quelques dizaines de centimètres de cette mesure ?

On va donc préférer estimer μ par un **intervalle de confiance** :

Après quelques calculs, on pourra par exemple affirmer que

il y a 95% de chances que la taille moyenne de toutes les femmes suisses se situe entre 163.4 cm et 165.0 cm

Autrement dit,

il y a 95% de chances que la taille moyenne de toutes les femmes suisses se situe au maximum à 0.8 cm de notre estimation ponctuelle (164.2 cm)

3.2 Calcul d'un intervalle de confiance pour estimer une moyenne

Exemple 1

Reprenons l'exemple des 1000 femmes mesurant en moyenne 164.2 cm, et supposons que l'on connaisse l'écart-type de la taille des femmes suisses : $\sigma = 12.3$ cm.

On décide de calculer un intervalle de confiance à un **niveau de confiance de 95%**.

Rappel du TCL

Si X suit une loi normale ou si $n \geq 30$:

$$\bar{x} \sim \mathcal{N}(\mu\,;\,\sigma_{\bar{x}}) \quad \text{avec} \quad \sigma_{\bar{x}} = \begin{cases} \frac{\sigma}{\sqrt{n}} & \text{si } n < \frac{N}{20} \text{ (petit échantillon)} \\ \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}} & \text{sinon (grand échantillon)} \end{cases}$$

Que représente la variable X dans ce cas ? La taille des femmes suisses

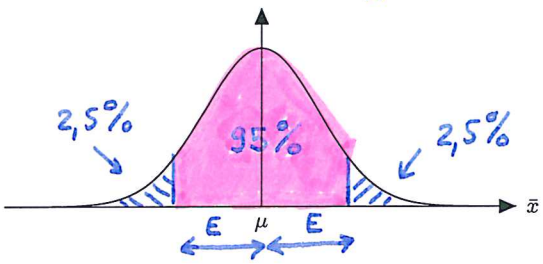
D'après le TCL, $\bar{x}$ suit bien une loi normale, car $n \geq 30$ (⚠ X ne suit peut-être pas une loi normale)

S'agit-il d'un petit échantillon ou d'un grand échantillon ?

Petit $\Rightarrow$ pas de facteur de correction

N n'est pas précisé $\Rightarrow$ on suppose que $N = \infty$

Calcul de $\sigma_{\bar{x}}$: $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{12,3}{\sqrt{1000}} \approx 0,39$



Calcul de la marge d'erreur pour un intervalle à 95% :

$$E = q_{0,975} \cdot \sigma_{\bar{x}} = 1,96 \cdot 0,39 = 0,76 \approx 0,8$$

Il y a 95% de chances que la zone colorée sur la courbe contienne la valeur 164.2 cm.

Autrement dit, il y a 95% de chances que la valeur obtenue pour $\bar{x}$ (164.2 cm), se situe à moins de 0,8 cm de la vraie moyenne μ.

Il y a donc aussi 95% de chances que la vraie moyenne μ (inconnue) se situe à moins de 0,8 cm de $\bar{x}$!

On peut donc construire un intervalle de confiance pour μ :

$$\mu \in [\bar{x} - E ; \bar{x} + E] = [164,2 - 0,8 ; 164,2 + 0,8]$$
$$= [163,4 ; 165,0]$$

Interprétation :

Il y a 95% de chances que la taille moyenne de toutes les femmes suisses se situe entre 163,4 et 165 cm

Quel est le risque d'erreur de cet intervalle ? 5%

Interprétation :

Il y a 5% de risques que la taille moyenne des femmes suisses soit plus petite que 163,4 ou plus grande que 165 cm

Ce risque d'erreur s'appelle le **seuil** de l'intervalle de confiance.

Exemple 2

Une ville de 17'000 habitants souhaite estimer le montant moyen dépensé par ses ménages chaque année au restaurant.

Pour ceci, elle interroge 1570 ménages. Le montant moyen dépensé par cet échantillon est de 1807 francs par année, avec un écart-type corrigé de 556 francs.

Construire un intervalle de confiance avec un niveau de confiance de 90% pour estimer le montant moyen dépensé par l'ensemble des ménages de cette ville.

Remarque importante

Ici, on ne connaît pas l'écart-type de la population. On va donc l'estimer en utilisant les données de l'échantillon.
L'écart-type de l'échantillon, noté s, n'est pas un bon estimateur pour σ. On va donc utiliser l'écart-type corrigé, noté $\hat{\sigma}$, qui se calcule comme suit :

$$\hat{\sigma}^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + (x_3 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n - 1}$$

$\sqrt{x_{n-1}}$

Calculatrice : touche $\boxed{\sqrt{x}}$ au lieu de $\boxed{\div}$ $\sqrt{x_n}$

— Résumé des données connues :

Population	Echantillon
N = 17'000	n = 1570
μ : à estimer	$\bar{x}$ = 1807
σ : inconnu	$\hat{\sigma}$ = 556

— Vérification des conditions d'application du TCL

$n \geq 30$

— Petit échantillon ou grand échantillon ?

Grand : $1570 > \frac{17'000}{20}$ ⇒ facteur de correction

— Calcul ou estimation de $\sigma_{\bar{x}}$

$$\sigma_{\bar{x}} \overset{\approx}{} \hat{\sigma}_{\bar{x}} = \frac{\hat{\sigma}}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}} = \frac{556}{\sqrt{1570}} \cdot \sqrt{\frac{15'430}{16'999}} \approx 13{,}4$$

— Calcul de la marge d'erreur

$$E = q_{0,95} \cdot \hat{\sigma}_{\bar{x}} = 1{,}645 \cdot 13{,}4 \approx 22$$

— Calcul de l'intervalle de confiance

$$I = [\bar{x} - E ; \bar{x} + E] = [1807 - 22 ; 1807 + 22] = [1785 ; 1829]$$

— Interprétation :

Il y a 90% de chances que le montant moyen dépensé au resto chaque année par les ménages de cette ville se situe entre 1785 francs et 1829 francs

3.3 Choix de la taille de l'échantillon

Supposons que l'on souhaite estimer la moyenne d'âge des étudiants d'une université par un intervalle de confiance.

L'intervalle se construit avec la formule suivante :

$$I = [\bar{x} - \underbrace{\sigma_{\bar{x}} \cdot q}_{E}\,;\, \bar{x} + \underbrace{\sigma_{\bar{x}} \cdot q}_{E}]$$

On observe que :

- Plus l'intervalle est ..petit....., plus l'estimation est précise.
- La taille de l'intervalle ne dépend pas de ....$\bar{x}$....., mais seulement de ...E.....
- Si l'on diminue la marge d'erreur E, l'intervalle devient plus ..petit.....
- Si l'on augmente le niveau de confiance, q devient plus ...grand....., et donc la marge d'erreur .augmente.....
- Si l'on augmente la taille n de l'échantillon, $\sigma_{\bar{x}}$ devient plus ..petit....., et donc la marge d'erreur .diminue.....

$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

Pour garantir une certaine précision dans notre estimation, on peut vouloir fixer une taille maximale pour l'intervalle de confiance.

Pour garantir que l'intervalle ne devienne pas trop grand, on peut jouer sur deux critères :

- **le niveau de confiance :**

  Pour diminuer la marge d'erreur, on doit ..diminuer..... le niveau de confiance.

  Inconvénient : .le.risque.d'erreur.augmente..=>..moins.fiable...

- **la taille de l'échantillon :**

  Pour diminuer la marge d'erreur, on doit ..augmenter..... la taille de l'échantillon.

  Inconvénient : .c'est.plus.cher..(plus.de.monde.à.interroger.).....

Revenons à l'estimation de l'âge moyen des étudiants de cette université.

Si l'on souhaite garder un niveau de confiance de 95%, et si l'on sait que des études antérieures ont donné un écart-type σ de 5.7 ans pour la population, quelle doit être la taille minimale de notre échantillon pour que la largeur de l'intervalle de confiance ne dépasse pas 3 ans ?

On veut que $E =$ 1,5 ans

Formule pour le calcul de la marge d'erreur : $E =$ $\sigma_{\bar{x}} \cdot q$

Niveau de confiance : 95 %

$\Rightarrow q = q_{0,975} = 1,96$ (table $\mathcal{N}(0;1)$)

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{5,7}{\sqrt{n}}$$

On veut donc que

$$1,5 = \frac{5,7}{\sqrt{n}} \cdot 1,96 \qquad | \cdot \sqrt{n}$$

$$1,5\sqrt{n} = 11,172 \qquad | : 1,5$$

$$\sqrt{n} = 7,448 \qquad | (\cdot)^2$$

$$n \cong 55,47$$

$\Rightarrow$ L'échantillon doit contenir au moins 56 personnes

4 Tests d'hypothèse pour une moyenne

4.1 Exemple 1

Un enseignant de sport observe depuis des années que le temps que mettent ses élèves pour courir 5 km suit une loi normale $\mathcal{N}(28; 2.25)$. Au début d'une année scolaire, il décide de faire courir ses élèves chaque semaine pour tenter d'améliorer leur endurance. A la fin de l'année, il mesure le temps que mettent ses 23 élèves pour courir 5 km, et obtient une moyenne de 27 minutes.

Peut-il affirmer que sa classe est meilleure que les élèves qu'il a eu au cours des dernières années, ou la différence est-elle trop petite pour être significative ?
Pourrait-elle simplement venir des variations normales des performances des élèves ? Autrement dit, est-il plausible d'obtenir une moyenne de 27 minutes dans un échantillon, alors que la vraie moyenne est de 28 minutes ?

Pour répondre à cette question, on va procéder à un **test d'hypothèse**.

1. **Formulation des hypothèses**

   On pose **l'hypothèse nulle**, qui est l'hypothèse sous laquelle on peut faire des calculs statistiques.

H_0 :	Le temps que mettent les élèves de cette classe suit une loi de moyenne ..$\mu = 28$..... minutes.

   On pose également **l'hypothèse alternative**

H_1 :	Le temps que mettent les élèves de cette classe suit une loi de moyenne ..$\mu < 28$..... minutes.

   Remarques

   — Le but de ce test est de ..rejeter..... H_0, afin de ..valider..... H_1.

   — Comme on teste $\boxed{\mu = 28}$ contre $\boxed{\mu < 28}$, il s'agit d'un test ..unilatéral..... (d'un seul côté). Si on voulait tester $\boxed{\mu = 28}$ contre $\boxed{\mu \neq 28}$, on procéderait à un test ..bilatéral..... (des deux côtés).

2. Seuil de signification du test

Comme dans le cas d'un intervalle de confiance, on doit choisir quel niveau de confiance on souhaite, et quel risque d'erreur on accepte.

Le **seuil** d'un test statistique représente le risque d'erreur.
C'est le risque de rejeter H_0 alors que cette hypothèse est vraie.

Posons comme seuil $\alpha =$ 1%

Remarque

Dans un test d'hypothèse, on peut fixer le risque de rejeter H_0 alors qu'elle est vraie, mais on ne maîtrise pas le risque de ne pas rejeter H_0 alors qu'elle est fausse.

3. Calcul du score du test

Rappel du TCL

Si X suit une loi normale ou si $n \geq 30$:

$$\bar{x} \sim \mathcal{N}(\mu\,;\,\sigma_{\bar{x}}) \quad \text{avec} \quad \sigma_{\bar{x}} = \begin{cases} \frac{\sigma}{\sqrt{n}} & \text{si } n < \frac{N}{20} \text{ (petit échantillon)} \\ \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}} & \text{sinon (grand échantillon)} \end{cases}$$

D'après le TCL, $\bar{x}$ suit bien une loi normale, car X suit une loi normale

S'agit-il d'un petit échantillon ou d'un grand échantillon ?

Petit, N supposé ∞ car pas précisé

Calcul de $\sigma_{\bar{x}}$: $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{\sqrt{2,25}}{\sqrt{23}} \approx 0,31$

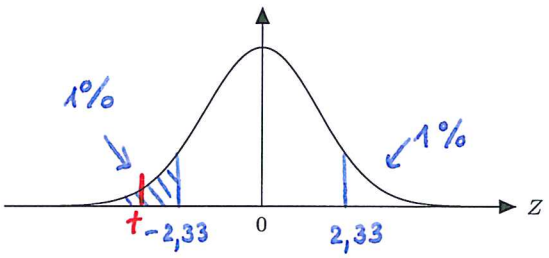
$\Rightarrow \bar{x} \sim \mathcal{N}(28\,;\,0,31^2)$ (sous l'hyp. nulle)

Score du test : on utilise la cote z :

$$t = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \frac{27 - 28}{0,31} \approx -3,20$$

4. **Décision et conclusion**

$H_1 : \mu < 28$ (unilatéral à gauche)
$\Rightarrow$ 1% à gauche



Règle : on rejette H_0 si $t < -2,33$

$t = -3,20 < -2,33 \Rightarrow$ on rejette H_0

Interprétation

Si H_0 est vraie, la valeur obtenue fait partie des 1% les plus rares $\Rightarrow$ on déduit plutôt que H_0 est fausse (rejet) $\Rightarrow$ oui, les élèves ont progressé

4.2 Exemple 2

Une machine remplit des bouteilles de sirop d'érable.
Le producteur veut contrôler le réglage de la machine pour s'assurer que le volume de sirop de chaque bouteille est bien de 540 ml en moyenne, comme l'indique l'étiquette. Il prélève donc 100 bouteilles au hasard, et obtient une moyenne de 541.5 ml, avec un écart-type corrigé de 5 ml.

Effectuer un test d'hypothèse au seuil de signification de 5% pour vérifier le réglage de la machine.

1. **Formulation des hypothèses**

   H_0 : La moyenne de la machine est $\mu = 540$ ml

   H_1 : $\mu \neq 540$ ml (test bilatéral)

2. **Seuil de signification du test :** $\alpha =$ 5 %

3. **Calcul du score du test**

   $\sigma_{\bar{x}} \overset{N}{=} \hat{\sigma}_{\bar{x}} = \frac{5}{\sqrt{100}} = 0,5$

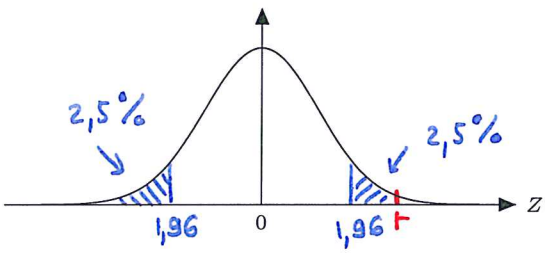
   $t = \frac{541,5 - 540}{0,5} = 3$

   Remarque

   Comme on utilise $\hat{\sigma}$ pour estimer σ, le score de test t suit en réalité une loi de Student, mais si $n \geq 30$, cette loi ressemble suffisamment à la loi normale pour que l'on puisse utiliser la table de la loi $\mathcal{N}(0;1)$.

4. **Décision et conclusion**

   $H_1 : \mu \neq 540$ (bilatéral)
   $\Rightarrow$ on répartit les 5% des deux côtés



Règle : on rejette H_0 si $t > 1,96$ ou $t < -1,96$ ($\Rightarrow$ si $|t| > 1,96$)
$t = 3 > 1,96$ $\Rightarrow$ on rejette H_0, la machine est mal réglée

Quels sont les risques de faire ajuster inutilement la machine ? 5 %

4.3 Exemple 3

On veut tester la durée, en kilomètres, d'un nouveau type de pneus de voiture.

La durée des pneus actuels suit une loi normale de moyenne 50'000 km. Sur 32 nouveaux pneus testés, on obtient une moyenne de 52'311 km, avec un écart-type corrigé de 9'780 km.

Au seuil de signification de 1%, peut-on affirmer que le nouveau type de pneus a une meilleure durée que l'ancien ?

⇒ unilatéral

1. **Formulation des hypothèses**

   H_0 : La durée des nouveaux pneus suit une loi de moyenne $\mu = 50'000$

   H_1 : $\mu > 50'000$ (test unilatéral)

2. **Seuil de signification du test :** $\alpha =$ 1%

3. **Calcul du score du test**

   $\hat{\sigma}$ ⇒ loi de Student, mais $n \geq 30$ ⇒ ok pour utiliser $\mathcal{N}(0;1)$

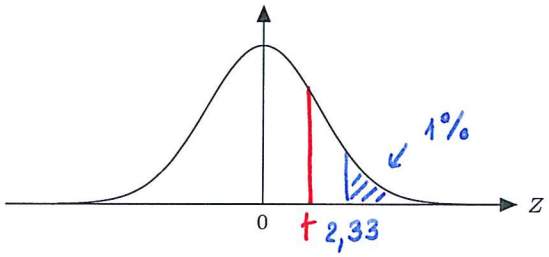
   $\sigma_{\bar{x}} \overset{\approx}{=} \hat{\sigma}_{\bar{x}} = \frac{9780}{\sqrt{32}} \overset{\approx}{=} 1728{,}9$ km

   $t = \frac{52'311 - 50'000}{1728{,}9} \overset{\approx}{=} 1{,}34$

4. **Décision et conclusion**

   $H_1 : \mu > 50'000$ (unilatéral)

   ⇒ 1% à droite



Règle : rejeter H_0 si $t > 2{,}33$

$t = 1{,}34$ ⇒ on ne peut pas rejeter H_0 : la différence obtenue avec les nouveaux pneus n'est pas significative

⚠ Ne pas rejeter H_0 ne signifie pas que H_0 est vraie !!