Statistiques inférentielles 3C

Contexte

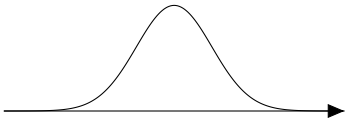
Les statistiques inférentielles ont pour but de déterminer les caractéristiques d'une population en utilisant les caractéristiques d'un échantillon provenant de cette population.

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.

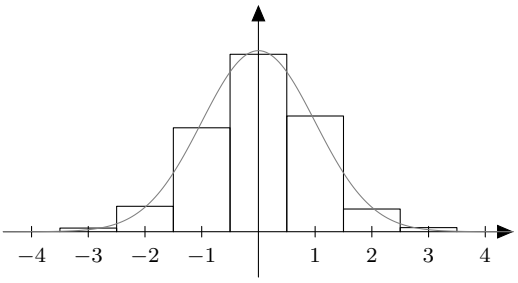
1 Loi normale

1.1 Définitions et notations

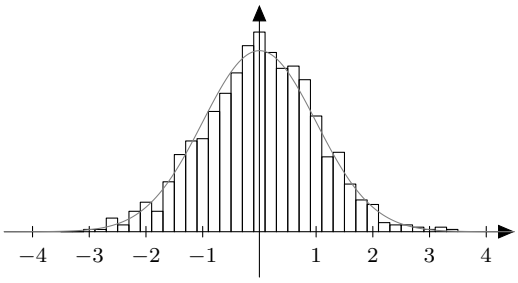
Dans de nombreux contextes, une variable statistique continue se distribue selon une courbe en cloche, appelée la courbe de Gauss.



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.

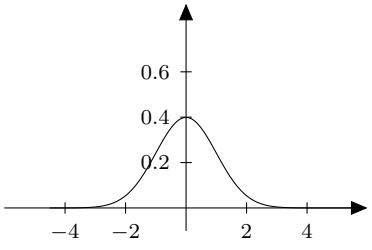


1000 données, 7 classes

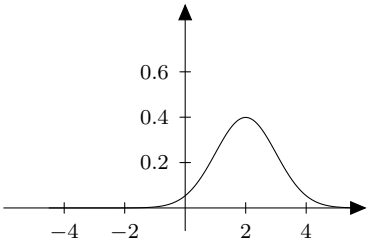


1000 données, 35 classes

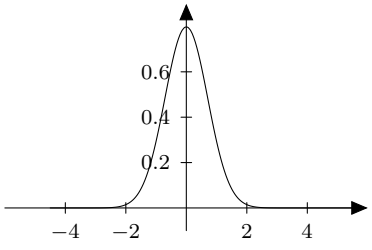
La distribution d'une variable suivant une loi normale ressemble toujours à une cloche, mais sa position et sa forme sont déterminées par la moyenne et la variance (ou l'écart-type) de la variable.



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



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



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

Lorsqu'on dit qu'une variable suit une loi normale, on note : .....

Propriétés

— La loi normale est symétrique autour de la moyenne μ

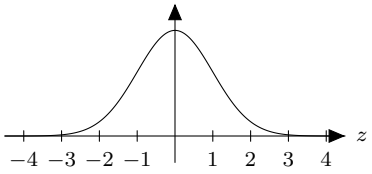
— L'aire sous la courbe vaut toujours 1

1.2 Représentation graphique

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

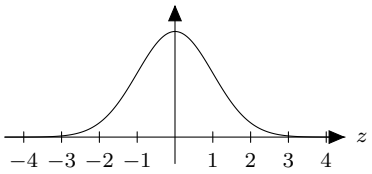
Sur la courbe de Gauss, on représente sur l'axe horizontal la valeur de la variable Z.
La probabilité d'obtenir certaines valeurs pour Z est donnée par l'aire sous la portion de courbe correspondante.

Exemples



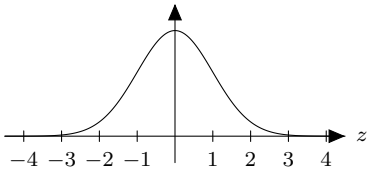
$Z = 1$

$P(Z = 1) =$ . . . . . .



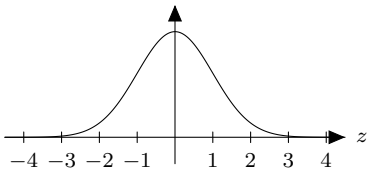
$Z < 0$

$P(Z < 0) =$ . . . . . .



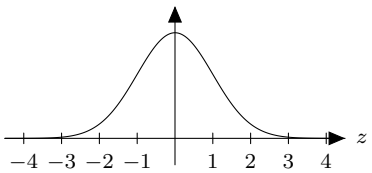
$-1 < Z \leq 2$ (noté aussi $Z \in]-1\,;\,2]$)

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



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

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

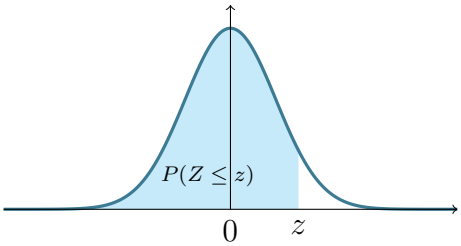


$Z \in \mathbb{R}$ (noté aussi $Z \in]-\infty\,;\,+\infty\,[$)

$P(Z \in \mathbb{R}) =$ . . . . . .

1.3 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 $z \geq 0$:

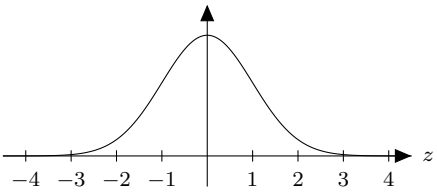


Toutes les probabilités s'obtiennent en se ramenant à des calculs du type $P(Z \leq z)$, grâce aux deux propriétés de la loi normale :

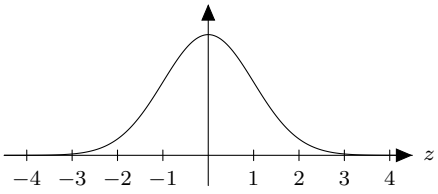
— symétrie autour de 0

— aire totale = 1 = 100%

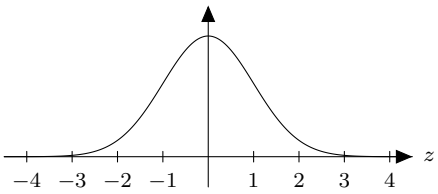
Exemples



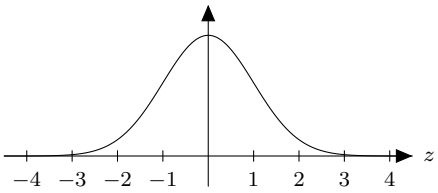
$P(Z \leq 1.25) =$ . . . . . .



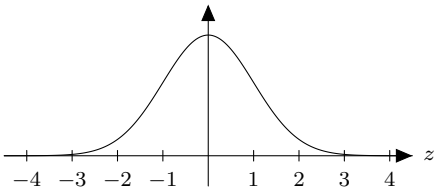
$P(Z > 1.25) =$ . . . . . .



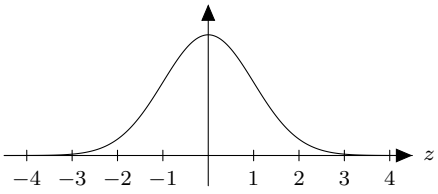
$P(Z < -0.5) =$ . . . . . .



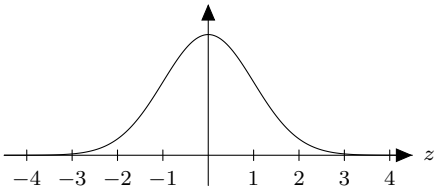
$P(Z > -0.5) =$ . . . . . .



$P(1 < Z < 2) =$ . . . . . .



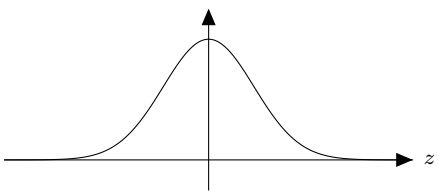
$P(-1.34 < Z < 2.11) =$ . . . . . .



$P(|Z| > 1.65) =$ . . . . . .

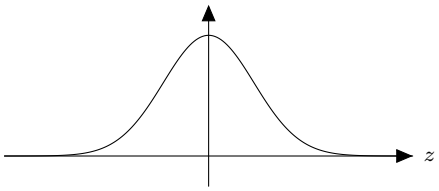
1.4 Recherche de quantiles avec la table numérique

Exemples

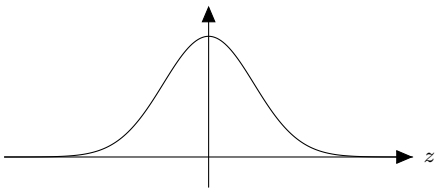


Déterminer a tel quel $P(Z \leq a) = 75\%$

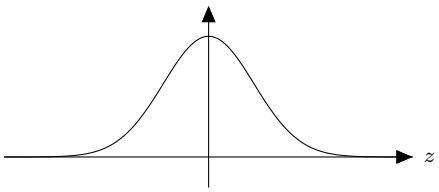
Remarque : $a = Q_3$ (troisième quartile).



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



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



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

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 :

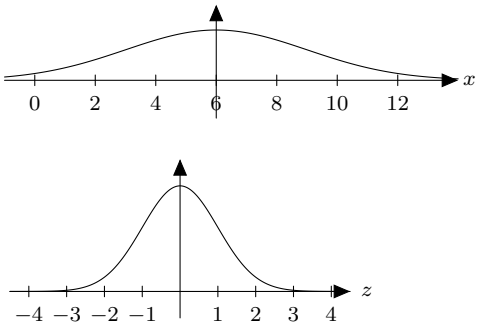
$$X \sim \mathcal{N}(\mu\,;\,\sigma^2) \qquad \Longrightarrow \qquad Z = \frac{X-\mu}{\sigma} \sim \mathcal{N}(0;1)$$

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 $P(X \leq 4)$.



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

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

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.

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

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

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.

2 Théorème Central Limite (TCL)

Lorsqu'on étudie une certaine variable statistique sur une population de taille N, on va très souvent calculer la moyenne de cette variable sur un échantillon de taille n (avec $n < N$). 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 $n \geq 30$, la moyenne $\bar{X}$ suit également une loi normale.

$$\bar{X} \sim \mathcal{N}(\mu\,;\,\sigma^2_{\bar{X}}) \quad \text{avec} \quad \sigma_{\bar{X}} = \begin{cases} \frac{\sigma}{\sqrt{n}} & \text{si } n < \frac{N}{20} \text{ (petit échantillon)} \\ \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}} & \text{sinon (grand échantillon)} \end{cases}$$

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 30 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 ? .....

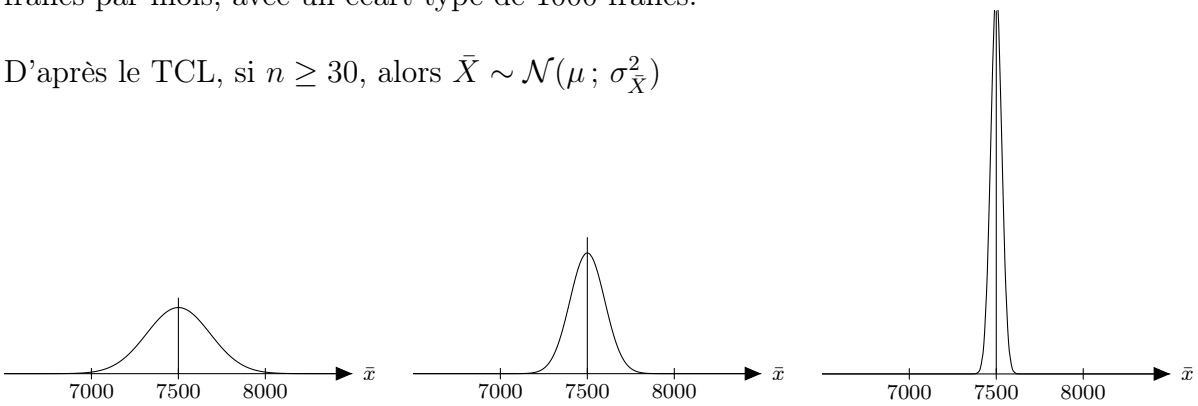
Pourquoi ? .....

.....

Cette intuition s'explique par le Théorème Central Limite :

Supposons que le salaire moyen de tous les anciens étudiants de cette université soit de 7500 francs par mois, avec un écart-type de 1000 francs.

D'après le TCL, si $n \geq 30$, alors $\bar{X} \sim \mathcal{N}(\mu\,;\,\sigma^2_{\bar{X}})$



$n = 30$	$n = 100$	$n = 1000$
$\mu =$	$\mu =$	$\mu =$ .....
$\sigma_{\bar{X}} =$	$\sigma_{\bar{X}} =$	$\sigma_{\bar{X}} =$ .....

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

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 ?

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

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

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

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

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 suisses vaut 164.2 cm ?

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 $n \geq 30$:

$$\bar{X} \sim \mathcal{N}(\mu\,;\,\sigma^2_{\bar{X}}) \quad \text{avec} \quad \sigma_{\bar{X}} = \begin{cases} \frac{\sigma}{\sqrt{n}} & \text{si } n < \frac{N}{20} \text{ (petit échantillon)} \\ \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}} & \text{sinon (grand échantillon)} \end{cases}$$

Que représente la variable X dans ce cas ? .....

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

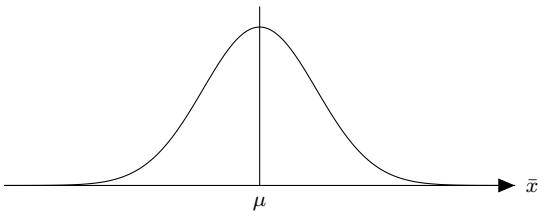
S'agit-il d'un petit échantillon ou d'un grand échantillon ?

.....

.....

Calcul de $\sigma_{\bar{X}}$: .....

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



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

Autrement dit, il y a 95% de chances que la valeur que prend $\bar{X}$ (ici $\bar{x} = 164.2$ cm), se situe à moins de ..... de la vraie moyenne μ.

Il y a donc aussi ..... de chances que la vraie moyenne μ (inconnue) se situe à moins de ..... de $\bar{x}$!

On peut donc construire un intervalle de confiance pour μ :

Interprétation :

. . . . . .

. . . . . .

Quel est le risque d'erreur de cet intervalle ? . . . . . .

Interprétation :

. . . . . .

. . . . . .

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

Exemple 2

Une ville comptant 17'000 ménages souhaite estimer le montant moyen qu'ils dépensent 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 :

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

— Résumé des données connues :

Population	
N	
μ	
σ	

Echantillon	
n	
$\bar{x}$	
$\hat{\sigma}$	

— Vérification des conditions d'application du TCL

. . . . . .

— Petit échantillon ou grand échantillon ?

. . . . . .

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

— Calcul de la marge d'erreur

— Calcul de l'intervalle de confiance

— Interprétation :

. . . . . .

. . . . . .

. . . . . .

. . . . . .

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 ....., plus l'estimation est précise.

— La taille de l'intervalle ne dépend pas de ....., mais seulement de .....

— Si l'on diminue la marge d'erreur E, l'intervalle devient plus .....

— Si l'on augmente le niveau de confiance, q devient plus ....., et donc la marge d'erreur .....

— Si l'on augmente la taille n de l'échantillon, $\sigma_{\bar{X}}$ devient plus ....., et donc la marge d'erreur .....

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 ..... le niveau de confiance.

Inconvénient : .....

— **la taille de l'échantillon :**

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

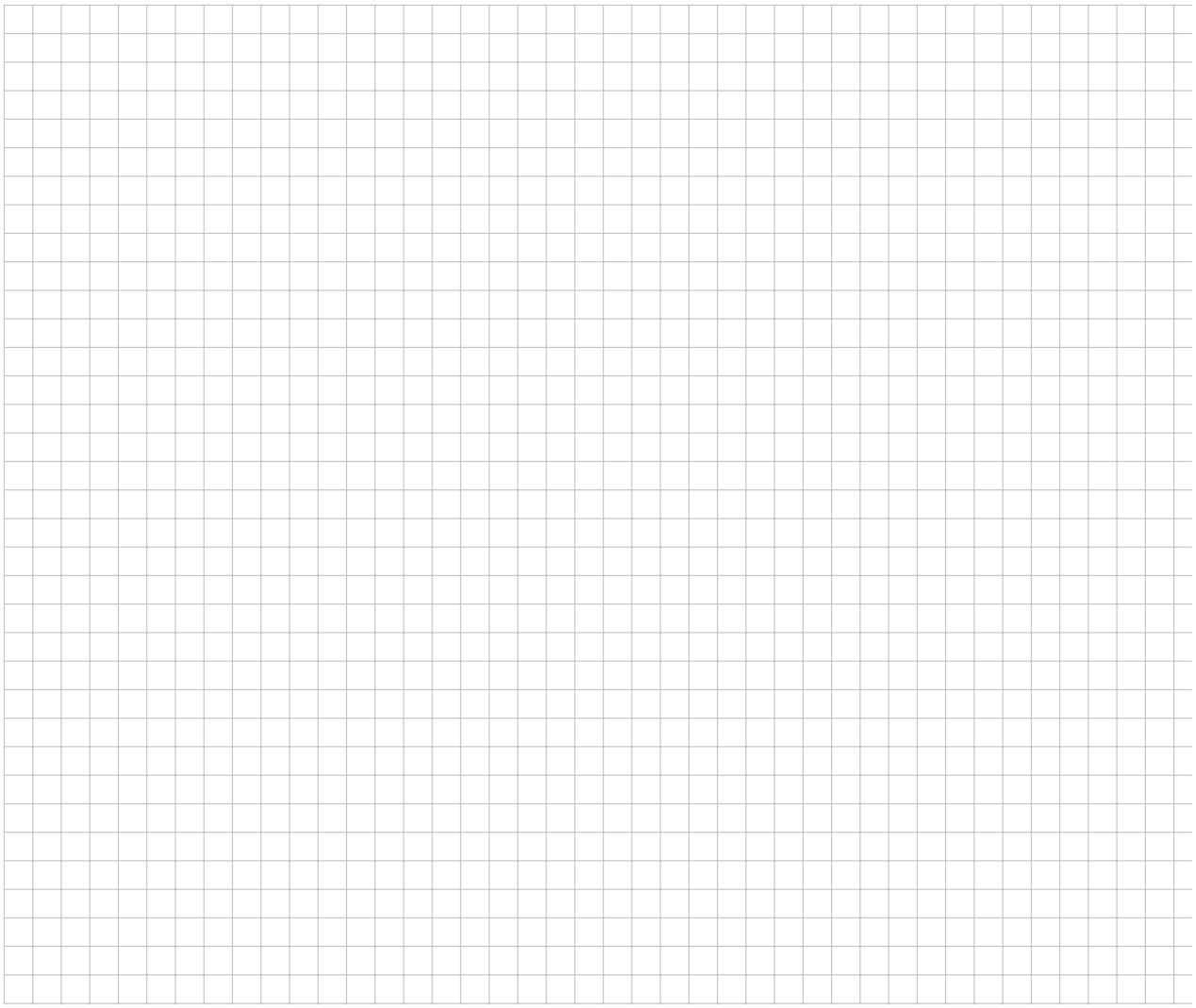
Inconvénient : .....

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 $E =$ .....

Formule pour le calcul de la marge d'erreur : $E =$ .....



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 ..... minutes.

   On pose également **l'hypothèse alternative**

H_1 :	Le temps que mettent les élèves de cette classe suit une loi de moyenne ..... minutes.

   Remarques

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

   — Comme on teste $\boxed{\mu = 28}$ contre $\boxed{\mu < 28}$,
   il s'agit d'un test ..... (d'un seul côté).

   Si on voulait tester $\boxed{\mu = 28}$ contre $\boxed{\mu \neq 28}$,
   on procéderait à un test ..... (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 ..... H_0 alors que cette hypothèse est .....

Posons comme seuil $\alpha = 1\%$.

Remarque

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

3. **Règle de décision**

Rappel du TCL

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

$$\bar{X} \sim \mathcal{N}(\mu\,;\,\sigma_{\bar{X}}^2) \quad \text{avec} \quad \sigma_{\bar{X}} = \begin{cases} \frac{\sigma}{\sqrt{n}} & \text{si } n < \frac{N}{20} \text{ (petit échantillon)} \\ \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}} & \text{sinon (grand échantillon)} \end{cases}$$

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

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

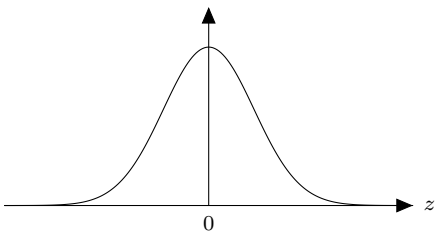
.....

Calcul de $\sigma_{\bar{X}}$:

.....

Ainsi, **sous l'hypothèse nulle,** $\bar{X} \sim$ .....

et $Z = \dfrac{\bar{X} - \text{.....}}{\text{.....}} \sim$ .....



On peut donc affirmer que **si l'hypothèse nulle est vraie**, Z a ..... % de chances .....

4. **Calcul du score de test et conclusion**

.....

.....

.....

.....

.....

4.2 Exemple 2

Une entreprise produit et vend du sirop d'érable. La machine qui remplit les bouteille est précise à 2 ml, valeur considérée comme l'écart-type des volumes de l'ensemble de la production. Le producteur veut contrôler le réglage de la machine pour s'assurer que le contenu moyen des bouteilles de la production est bien de 500 ml, comme spécifié sur l'étiquette.

Il prélève 100 bouteilles de la production et obtient un volume moyen de 499.2 ml.
Sur la base de ce résultat, le producteur doit-il douter du réglage de sa machine ?

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 volume de remplissage des bouteilles suit une loi d'espérance ..... ml.

   On pose également **l'hypothèse alternative**

H_1 :	Le volume de remplissage des bouteilles suit une loi d'espérance ..... ml.

   Remarque

   Comme on teste $\boxed{\mu = 500}$ contre $\boxed{\mu \neq 500}$,
   il s'agit d'un test ..... (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 un risque d'erreur.
C'est le risque de ..... H_0 alors que cette hypothèse est .....
Posons comme seuil $\alpha = 1\%$.

Remarque

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

3. **Règle de décision**

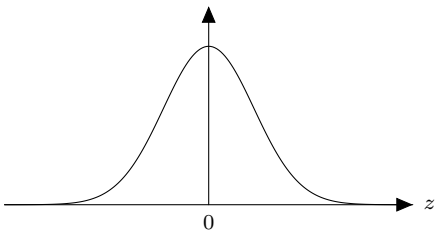
Rappel du TCL : Si X suit une loi normale ou si $n \geq 30$:

$$\bar{X} \sim \mathcal{N}(\mu\,;\,\sigma^2_{\bar{X}}) \quad \text{avec} \quad \sigma_{\bar{X}} = \begin{cases} \frac{\sigma}{\sqrt{n}} & \text{si } n < \frac{N}{20} \text{ (petit échantillon)} \\ \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}} & \text{sinon (grand échantillon)} \end{cases}$$

D'après le TCL, $\bar{X}$ suit bien une loi normale, car .....
S'agit-il d'un petit échantillon ou d'un grand échantillon ? .....
Calcul de $\sigma_{\bar{X}}$:
.....

Ainsi, **sous l'hypothèse nulle**, $\bar{X} \sim$ .....

et $Z = \dfrac{\bar{X} - \ldots\ldots}{\ldots\ldots} \sim$ .....



On peut donc affirmer que **si l'hypothèse nulle est vraie**, Z a ..... % de chances de se trouver entre ..... et ..... .

La règle de décision du test d'hypothèse sera donc :

Rejeter H_0 si la cote z de la valeur $\bar{x}$ de l'échantillon est ..... ou .....

4. **Calcul du score de test et conclusion**

Quels sont les risques de faire ajuster inutilement la machine ? .....

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 d'espérance 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 5%, peut-on affirmer que le nouveau type de pneus a une meilleure durée que l'ancien ?

1. **Formulation des hypothèses**

   H_0 : .....

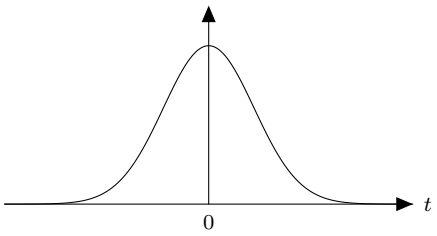
   H_1 : .....

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

3. **Règle de décision**

   $\hat{\sigma}_{\bar{X}} =$ .....

   Comme on utilise $\hat{\sigma}$ pour estimer σ, le score de test $T \sim$ .....



⇒ La règle de décision est :

.....

4. **Calcul du score de test et conclusion**

. . . .

. . . .

. . . .

. . . .

. . . .

. . . .