# Chapitre 3 : Statistiques inférentielles

# Table des matières

# 1 Introduction

Les problèmes de l'échantillonnage et de l'estimation sont illustrés par l'étude de la situation suivante :

*Un industriel produit en très grand nombres des yaourts, pour lesquelles l'usinage doit respecter des normes sanitaires draconiennes. Á la suite de mauvais réglages de l'une des machines, l'industriel a produit* 1 *million de ces yaourts, dont beaucoup risquent ainsi de présenter des dangers pour le consommateur.*
*Il souhaite connaître la proportion de yaourts susceptibles de rendre malade un client, afin de savoir s'il doit détruire sa production, ce qui représentera un fort manque à gagner, ou s'il peut malgré courir le risque de quelques gênes isolées dans la population, sans craindre de campagne médiatique mettant en cause ces yaourts, ce qui lui causerait un préjudice encore plus grand.*
*Il est ainsi prêt à détruire son stock ainsi produit si la proportion de yaourts dangereux pour la santé dépasse les* $0,01\%$ *de sa production.*
*Il n'est bien entendu pas question d'analyser un par un tous les yaourts produits : cela lui reviendrait encore plus cher, et de toutes façons, il faudrait ouvrir les yaourts, ce qui les rendrait invendables. Il décide donc d'effectuer un sondage c'est à dire de prélever par exemple* 100 *yaourts, de les faire analyser, et de relever la proportion de yaourts contaminés dans cet échantillon.*
*Il obtient ainsi la résultat suivant : dans l'échantillon prélevé (au hasard) parmi les yaourts produits, on en a trouvé* 2% *qui contenaient des germes. Notre industriel est-il plus avancé après ces analyses pour résoudre son problème ?*

La réponse est bien sûr négative : en effet, il peut toujours se poser les questions suivantes :

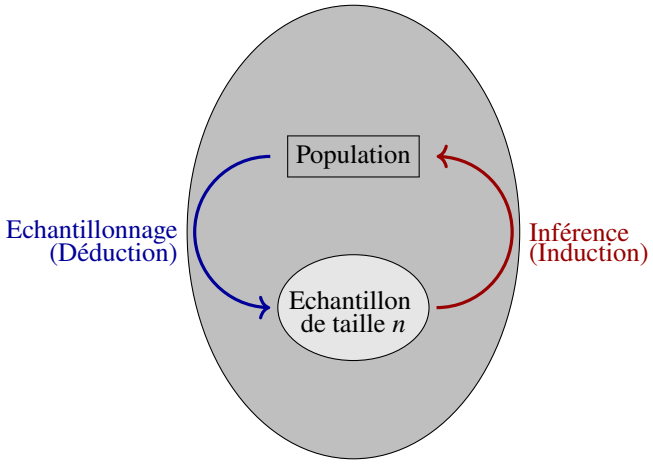
1. aurait-on obtenu le même pourcentage en prélevant un autre échantillon ? (autrement dit, la proportion inquiétante relevée dans le premier échantillon est-elle due à de la malchance ?)
2. l'analyse de 100 yaourts sur le million produit est-elle suffisante ?
3. quelle confiance peut-on accorder au fait que l'analyse d'un échantillon de 100 yaourts ait conduit à une proportion de 2% de produits contaminés ?
4. aurait-on gagné en fiabilité du pronostique si l'on en avait fait analyser 200, 1000, 10000 yaourts ?

La question 1., elle, relève du champ de l'échantillonnage. Cette théorie répond à la question : "comment varie la proportion relevée d'un échantillon à l'autre, sachant que tous sont de même taille donnée à l'avance ?". Ces questions ont des réponses fournies par le théorème de la limite centrée vue dans un précédent chapitre.
Les questions 2., 3. et 4., portant sur la taille de l'échantillon, et sur la confiance que l'on peut accorder au sondage sont du domaine de l'estimation : elles obtiennent une réponse avec les résultats sur la "loi des grands nombres".

L'échantillonnage est l'étude des liens existant entre les paramètres (moyenne ou fréquence) des échantillons issus d'une population et les paramètres de la population complète.

L'échantillonnage statistique consiste à prédire, à partir d'une population connue les caractéristiques des échantillons qui en seront prélevés.
On parle aussi de déduction des caractéristiques de l'échantillon.

Inversement, les statistiques inférentielles s'interessent à la détermination des paramètres de la population complète à partir de ceux d'un échantillon.
L'inférence statistique consiste à induire les caractéristiques inconnues d'une population à partir de celles d'un échantillon.
On parle aussi d'induction, ou encore d'extrapolation des caractéristiques à l'ensemble de la population.



# 2 Estimation

## 2.1 Estimation ponctuelle d'un paramètre

### 2.1.1 Moyenne

**Propriété :**
La valeur moyenne $m_e$ d'un paramètre observé sur un échantillon de population, dont la taille est fixée, fournit une estimation $\overline{x}$ de la moyenne réelle de ce paramètre sur la population considérée.

Exemple :
Une usine produit des vis cruciformes. On souhaite estimer la moyenne des longueurs des vis dans la production de la journée qui s'élève à 10000 pièces.
On choisit un échantillon de 150 vis et on obtient une moyenne de $m_e = 4,57$ cm.
On en déduit donc que la longueur moyenne des vis de la production journalière est $\overline{x} = 4,57$ cm.

### 2.1.2 Ecart type

Le problème est toujours le même, mais cette fois-ci, l'estimation de l'écart-type est moins intuitive . . .

**Propriété :**
L'écart-type $\sigma_e$ d'un paramètre observé sur un échantillon de population, dont la taille est fixée, fournit une estimation faussée de l'écart-type de ce paramètre dans toute la population considérée.
Une meilleure estimation $\sigma$ de l'écart-type réel est obtenue en considérant le nombre :

$$\sigma = \sigma_e \sqrt{\frac{n}{n-1}}$$

où $n$ est la taille de l'échantillon servant au calcul de $\sigma_e$.

Exemple :
La mesure de la longueur des vis produites dans l'échantillon précédent de 150 pièces conduit à relever un écart-type de 3 mm.
La meilleure estimation possible de l'écart-type de la production journalière n'est pas de 3 mm comme dans le cas précédent pour la moyenne, mais de $\sigma = 3\sqrt{\frac{150}{149}} \simeq 3,01$ mm.

**Remarque**

La correction devient assez rapidement minime lorsque la taille de l'échantillon augmente car :

$$\lim_{n\to\infty} \sqrt{\frac{n}{n-1}} = 1$$

La correction est ainsi :

- de l'ordre de $0,5\%$ pour des échantillons de taille 100
- de l'ordre de $0,05\%$ pour des échantillons de taille 1000.

### 2.1.3 Fréquence

**Propriété :**
La fréquence d'apparition $f_e$ d'un paramètre observé sur un échantillon de population, dont la taille est fixée, fournit une estimation $f$ de la fréquence réelle d'apparition de ce paramètre sur la population considérée.

Exemple :
Dans l'exemple précédent, On prélève un échantillon de 150 vis et on relève 3 pièces défectueuses.
On peut alors donner une estimation de la fréquence $f$ de vis défectueuses dans la production journalière :
On a $f_e = \dfrac{3}{150} = 0,02$ donc, $f = 0,02$.

**Remarque**

Notons qu'il revient exactement au même d'estimer un pourcentage : dans l'exemple précédent, on peut affirmer que 2% des vis ont une croix mal formée sur la tête.

## 2.2 Estimation par intervalle de confiance d'un paramètre

Les estimations ponctuelles proposées ci-dessus dépendent directement de l'échantillon prélevé au hasard.
Dans de très nombreux cas, l'importance attribuée au hasard est grande, cela conduit à s'interroger avant d'utiliser ces estimations pour prendre des décisions dont les conséquences peuvent être lourdes !
Aussi, sans rejeter les informations fournies par l'étude d'un échantillon, est-on amené à chercher un nouveau type d'estimation de la fréquence et de la moyenne d'une population, en utilisant le calcul de probabilités qui permet de "contrôler" l'influence d'un échantillon particulier.

### 2.2.1 Moyenne

On souhaite, à partir des observations faites sur un échantillon, déterminer un intervalle de confiance contenant la valeur moyenne avec un risque d'erreur décidé à l'avance.

On suppose que les conditions sont réunies pour faire l'approximation que la loi d'échantillonnage de la moyenne $\overline{X}$ est la loi normale $\mathcal{N}\left(m; \dfrac{\sigma}{\sqrt{n}}\right)$.
On pose $T = \dfrac{\overline{X} - m}{\frac{\sigma}{\sqrt{n}}}$, $T$ suit donc la loi normale centrée réduite $\mathcal{N}(0; 1)$.
Soit $\alpha$ la probabilité, fixée à l'avance, pour que $T$ n'appartienne pas à l'intervalle $[-t; t]$, on peut écrire :

$$\begin{aligned}
P(|T| > t) = \alpha &\iff 1 - P(|T| \leq t) = \alpha \\
&\iff P(|T| \leq t) = 1 - \alpha \\
&\iff P(-t \leq T \leq t) = 1 - \alpha \\
&\iff P\left(-t \leq \frac{\overline{X} - m}{\frac{\sigma}{\sqrt{n}}} \leq t\right) = 1 - \alpha \\
&\iff P\left(\overline{X} - t\frac{\sigma}{\sqrt{n}} \leq m \leq \overline{X} + t\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha
\end{aligned} \tag{1}$$

Autrement dit, $m$ appartient à l'intervalle $\left[\overline{X} - t\frac{\sigma}{\sqrt{n}}; \overline{X} + t\frac{\sigma}{\sqrt{n}}\right]$ pour $100(1-\alpha)\%$ des échantillons.

- Cet intervalle est appelé intervalle de confiance,
- $\alpha$ est le risque d'erreur ou le seuil de risque,
- $1-\alpha$ est le coefficient de confiance.

**Propriété :**
L'intervalle de confiance de la moyenne $m$ au seuil de confiance $1-\alpha$ (ou risque $\alpha$) est :

$$I_\alpha = \left[m_e - t\frac{\sigma}{\sqrt{n}}; m_e + t\frac{\sigma}{\sqrt{n}}\right]$$

**Remarque**

Les valeurs fréquentes du niveau de confiance sont $0,99$ et $0,95$.
Pour ces deux valeurs, on obtient successivement $t = 2,575$ et $t = 1,96$.

Exemple :
On suppose que la durée de vie, exprimée en heures, d'une ampoule électrique d'un certain type, suit la loi normale de moyenne $M$ inconnue et d'écart-type $\sigma = 20$.
Une étude sur un échantillon de 16 ampoules donne une moyenne de vie égale à 3000 heures.
On va déterminer un intervalle de confiance de $M$ au seuil de risque de 5%.

On a : $\alpha = 5\%$ d'où $t = 1,96$.
Un intervalle de confiance de $M$ est donc :

$$I_{0,05} = \left[3000 - 1,96\frac{20}{\sqrt{16}}; 3000 + 1,96\frac{20}{\sqrt{16}}\right] = [2990, 3009]$$

### 2.2.2 Fréquence

A l'aide d'un échantillon, nous allons définir, avec un coefficient de confiance choisi à l'avance, un intervalle de confiance de la fréquence $p$ des éléments de la population possédant une certaine propriété.

On se place dans le cas où on peut approximer la loi par la loi normale $\mathcal{N}\left(p; \sqrt{\frac{p(1-p)}{n}}\right)$.

**Propriété :**
L'intervalle de confiance de la moyenne $m$ au seuil de confiance $1-\alpha$ (ou risque $\alpha$) est :

$$I_\alpha = \left[f_e - t\sqrt{\frac{f_e(1-f_e)}{n}}; f_e + t\sqrt{\frac{f_e(1-f_e)}{n}}\right]$$

Exemple :
Un sondage dans une commune révèle que sur les 500 personnes interrogées, 42% sont mécontentes de l'organisation des transport. On veut déterminer, au seuil de risque 1%, un intervalle de confiance du pourcentage $p$ de personnes mécontentes dans la commune :
On a :

$$f = 0,42 \ ; \ n = 500 \ ; \ \alpha = 1\% \ \text{ donc } \ t = 2,575$$

Un intervalle de confiance du pourcentage $p$ est donc :

$$I_{0,01} = \left[0,42 - 2,575\sqrt{\frac{0,42 \times 0,58}{500}}; 0,42 + 2,575\sqrt{\frac{0,42 \times 0,58}{500}}\right] = [0,36; 0,48] = [36\%; 48\%]$$

### 2.3 Tableau récapitulatif

Le tableau ci-dessous regroupe toutes les situations dans lesquelles on doit savoir fournir une estimation ponctuelle ou par intervalle de confiance :

| Paramètre de la population totale à estimer | Valeur du paramètre dans l'échantillon de taille $n$ | Estimation ponctuelle pour la population totale | Estimation par intervalle de confiance au niveau de confiance $1-\alpha$ pour la population totale |
|---|---|---|---|
| Moyenne | $m_e$ | $m = m_e$ | $\left[m_e - t\dfrac{\sigma}{\sqrt{n}}; m_e + t\dfrac{\sigma}{\sqrt{n}}\right]$ |
| Écart-type | $\sigma_e$ | $\sigma = \sigma_e\sqrt{\dfrac{n}{n-1}}$ | |
| Fréquence | $f_e$ | $f = f_e$ | $\left[f_e - t\sqrt{\dfrac{f_e(1-f_e)}{n}}; f_e + t\sqrt{\dfrac{f_e(1-f_e)}{n}}\right]$ |

## 3 Tests d'hypothèse

Pour remplir des paquets de farine de 10 kg, on utilise une ensacheuse réglée avec précision, mais on ne peut espérer que tous les paquets sortant de la machine pèsent exactement 10 kg. On peut seulement exiger que l'espérance mathématique des masses de tous les paquets produits soit de 10 kg.
Ainsi, une palette de 50 paquets pèsera par exemple 497 kg. Doit-on en conclure que la machine est mal réglée ?
Si, après avoir réglé différemment la machine, une nouvelle palette de50 paquets pèse 502 kg, peut-on en conclure que la machine est mieux réglée ?
Ce sont les tests de validité d'hypothèse qui permettent de prendre une décision. Ces décisions seront prises avec un certain risque a priori.

Dans toute cette partie, les notions seront abordées grâce à des exemples.
Pour chaque test, on appliquera le cheminement suivant :

**Construction du test de validité d'hypothèse.**

- Étape 1 : choix des deux hypothèses : l'hypothèse nulle $H_0$ et l'hypothèse alternative $H_1$ ;
- Étape 2 : détermination de la variable aléatoire de décision et de ses paramètres (on utilisera en général la loi normale ;
- Étape 3 : l'hypothèse nulle étant considérée comme vraie et compte tenu de l'hypothèse alternative, détermination de la zone critique selon le niveau de risque $\alpha$ donné.
  On cherche $I_\alpha = [a, b]$ tel que $\mathbb{P}(a \leq X \leq b) = 1 - \alpha$ dans le cas bilatéral et $I_\alpha = [a; +\infty[$ ou $I_\alpha = ]-\infty; b]$ dans le cas unilatéral ;
- Étape 4 : rédaction d'une règle de décision.

**Utilisation du test d'hypothèse.**

- Étape 5 : calcul des caractéristiques d'un échantillon particulier puis application de la règle de décision.

## 3.1 Test bilatéral relatif à une moyenne

Exemple :
Une machine produit des rondelles dont l'épaisseur est une variable aléatoire $X$ d'écart type $0,3$ mm. La machine a été réglée pour obtenir des épaisseurs de 5 mm.
Un contrôle portant sur un échantillon de 100 rondelles a donné $5,07$ mm comme moyenne des épaisseurs de ces 100 rondelles.
Peut-on affirmer que la machine est bien réglée au seuil de risque de 5% ?

1. **Choix des hypothèses.**
   On estime que la machine est bien réglée, si la moyenne de toutes les rondelles produites par la machine est 5 mm. C'est donc l'hypothèse $m = 5$ que nous allons tester. On l'appelle l'hypothèse nulle $H_0$.
   Sinon, on choisit comme hypothèse alternative l'hypothèse $H_1$ : "$m \neq 5$".
   On a donc :

   - $H_0$ : "$m = 5$"
   - $H_1$ : "$m \neq 5$"

   Recherchons comment la moyenne $m_e$, d'un échantillon de 100 rondelles peut confirmer ou non l'hypothèse $H_0$.

2. **Variable aléatoire de décision.**
   Soit $m$ l'espérance mathématique de $X$, c'est-à-dire la moyenne des épaisseurs de toutes les rondelles produites par la machine ainsi réglée.
   Considérons la variable aléatoire $M$ qui, à chaque échantillon de taille 100, associe sa moyenne.
   La taille des échantillons étant suffisamment grande, on considère que $M$ suit la loi $\mathcal{N}\left(m; \dfrac{0,3}{\sqrt{100}}\right)$, c'est-à-dire $\mathcal{N}(m; 0,03)$.
   $M$ sera la variable aléatoire de décision.

3. **Zone critique.**
   Dons le cas où l'hypothèse $H_0$ est vraie, la variable aléatoire $M$ suit la loi $\mathcal{N}(5; 0,03)$.
   On cherche alors le réel $d$ tel que :

$$(E) : \quad \boxed{\mathbb{P}(5 - d \leq M \leq 5 + d) = 0,95}$$

   la variable aléatoire $T = \dfrac{M - 5}{0,03}$ suit la loi normale centrée réduite $\mathcal{N}(0, 1)$, on a alors :

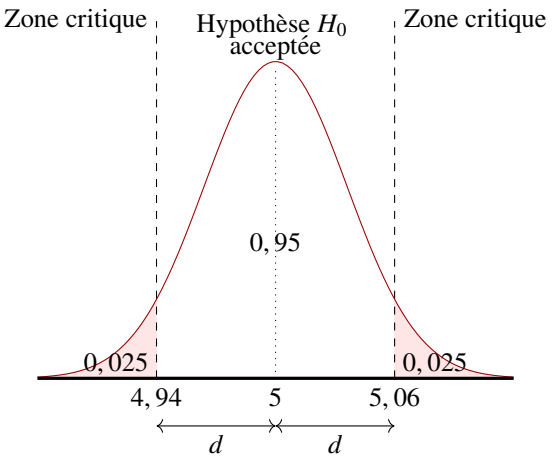
$$\begin{aligned}(E) &\iff \mathbb{P}(5 - d \leq 0,03T + 5 \leq 5 + d) = 0,95 \\ &\iff \mathbb{P}\left(-\frac{d}{0,03} \leq T \leq \frac{d}{0,03}\right) = 0,95 \\ &\iff 2\Pi\left(\frac{d}{0,03}\right) - 1 = 0,95 \\ &\iff \Pi\left(\frac{d}{0,03}\right) = 0,975\end{aligned} \tag{2}$$

   On trouve alors d'après la table que $\dfrac{d}{0,03} = 1,96$ soit $d = 0,0588 \approx 0,06$.
   L'intervalle de confiance est donc l'intervalle :

$$I_{0,05} = [5 - 0,06; 5 + 0,06] = [4,94; 5,06]$$

   On pourra également utiliser directement la calculatrice ou un logiciel de calcul formel pour déterminer la valeur du réel $d$ respectant $(E)$.

On obtient alors graphiquement la situation suivante :




La probabilité qu'un échantillon ait une moyenne située hors de cet intervalle étant $0,05$, on peut considérer que cet événement est rare. Ainsi, la moyenne de notre échantillon $m_e = 5,07$ nous amène à douter de la validité de l'hypothèse $H_0$.
Ne perdons pas de point de vue qu'il se peut, malgré tout, que la machine soit bien réglée et que notre échantillon fasse partie des 5% de ceux ayant une moyenne hors de l'intervalle trouvé. C'est pourquoi cette région est appelée zone critique.

4. **Règle de décision.**
   Si la moyenne de l'échantillon n'est pas située dans la zone critique, on accepte $H_0$, sinon, on refuse $H_0$ et on accepte $H_1$.

5. **Conclusion.**
   Puisque $5,07$ appartient à la zone critique, on décide de rejeter l'hypothèse $H_0$ et d'accepter l'hypothèse alternative $H_l$ : $m \neq 5$ (la machine n'est pas bien réglée).

Dans un test de validité d'hypothèse, le seuil de risque $\alpha$ est la probabilité de rejeter $H_0$ alors qu'elle est vraie.

### 3.2 Test unilatéral relatif à une moyenne

Exemple :
La durée de vie (en heures) des ampoules électriques produites par une usine est une variable aléatoire $X$ d'écart type 120. Le fabricant annonce qu'en moyenne, les ampoules ont une durée de vie de 1120 heures.
On demande de rédiger une règle de décision pour vérifier l'affirmation du fabriquant, au seuil de risque de de 5%, en testant un échantillon de 36 ampoules.

1. **Choix des hypothèses.**
   Soit l'hypothèse nulle $H_0 : m = 1120$ (l'affirmation du fabricant est vraie).
   Dans l'exemple précédent, les rondelles devaient avoir une épaisseur moyenne de 5 mm et cette mesure ne supportait ni excès, ni déficit. Ici, l'acheteur ne se plaindra que si la durée de vie des ampoules est inférieure à 1120 heures ; dans le cas où la moyenne $m_e$, de l'échantillon est supérieure à 1 120, l'hypothèse du fabricant se trouve immédiatement confirmée.
   L'hypothèse alternative $H_l$ est donc $m < 1120$ (l'affirmation du fabricant est fausse).
   On a donc :

   - $H_0$ : "$m = 1120$"
   - $H_1$ : "$m < 1120$"

2. **Variable aléatoire de décision.**
   Soit $m$ l'espérance mathématique de X, c'est-à-dire la moyenne des durée de vie de toutes les ampoules produites par l'usine. Considérons la variable aléatoire $M$ qui, à chaque échantillon de 36 ampoules associe la moyenne de durée de vie des 36 ampoules.
   La taille des échantillons étant suffisamment grande, on considère que $M$ suit la loi $\mathcal{N}\left(m; \frac{120}{\sqrt{36}}\right)$, c'est-à-dire $\mathcal{N}(m; 20)$.

3. **Zone critique.**
La zone critique se trouve donc d'un seul côté de la moyenne. On dit alors que le test est unilatéral par opposition au test bilatéral effectué au paragraphe précédent.
Dans le cas où l'hypothèse $H_0$ est vraie, la variable aléatoire $M$ suit la loi $\mathcal{N}(1120; 20)$.

On cherche alors le réel $d$ tel que :

$$(E) : \quad \boxed{\mathbb{P}(M < 1120 - d) = 0,05}$$

la variable aléatoire $T = \dfrac{M - 1120}{20}$ suit la loi normale centrée réduite $\mathcal{N}(0, 1)$, on a alors :

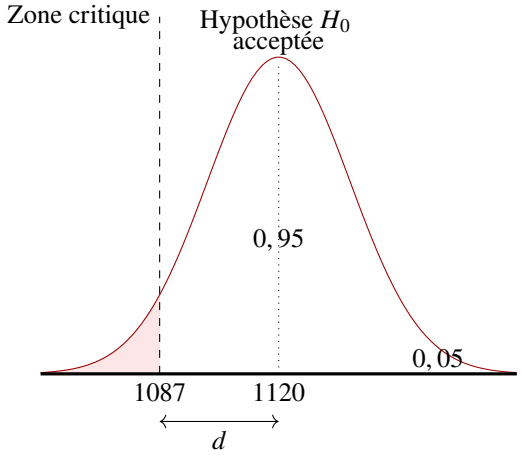
$$\begin{aligned}
(E) &\iff \mathbb{P}(20T + 1120 < 1120 - d) = 0,05 \\
&\iff \mathbb{P}\left(T < -\frac{d}{20}\right) = 0,05 \\
&\iff \mathbb{P}\left(T > \frac{d}{20}\right) = 0,05 \\
&\iff 1 - \mathbb{P}\left(T \leq \frac{d}{20}\right) = 0,05 \\
&\iff \Pi\left(\frac{d}{20}\right) = 0,95
\end{aligned} \tag{3}$$

On trouve alors d'après la table que $\dfrac{d}{20} = 1,645$ soit $d = 32,9 \approx 33$.
La zone critique est donc l'intervalle :

$$I_{0,05} = ]-\infty; 1120 - 33] = ]-\infty; 1087]$$

On pourra également utiliser directement la calculatrice ou un logiciel de calcul formel pour déterminer la valeur du réel $d$ respectant $(E)$.

On obtient alors graphiquement la situation suivante :




La zone critique est l'intervalle $]-\infty; 1087[$ : 5% seulement des échantillons de taille 36 ont en moyenne une durée de vie inférieure à 1087 heures.

4. **Règle de décision.**
Si la moyenne $m_e$ de l'échantillon observé est inférieure à 1087, on rejette l'hypothèse $H_0$ et on accepte l'hypothèse alternative $H_1$ (l'affirmation du fabricant est fausse).
Si la moyenne $m_e$ de l'échantillon observé est supérieure à 1087, on accepte l'hypothèse $H_0$.

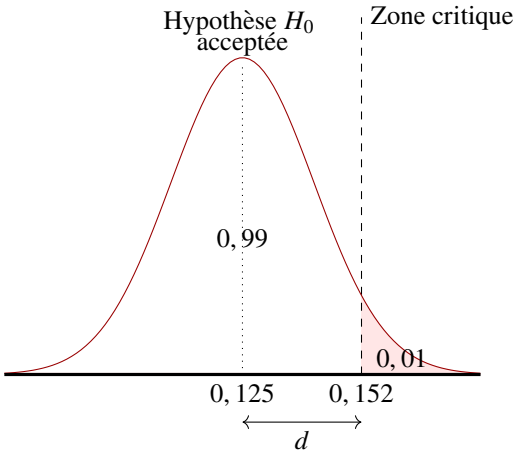
### 3.3 Test unilatéral relatif à une fréquence

On donne ici un exemple de test unilatéral relatif à une fréquence, mais d'autres cas peuvent amener à envisager des tests bilatéraux relatifs à une fréquence.

Exemple :
Un joueur qui doit choisir au hasard une carte dans un jeu de 32 cartes obtient certains avantages s'il découvre un roi. On constate qu'il a retourné 134 fois un roi sur 800 essais.
Peut-on présumer, au seuil de risque de 1%, que ce joueur est un tricheur ?

1. **Choix des hypothèses.**
   Si le joueur n'est pas un tricheur, la valeur de $p$ est $\frac{4}{32} = 0,125$.
   Donc, l'hypothèse nulle $H_0$ est $p = 0,125$ (le joueur n'est pas un tricheur).
   Si $p < 0,125$, on considérera que le joueur n'est pas un tricheur non plus, donc : l'hypothèse alternative $H_1$ est $p > 0,125$ (le joueur est un tricheur).
   On a donc :
   - $H_0$ : "$p = 0,125$"
   - $H_1$ : "$p > 0,125$"
2. **Variable aléatoire de décision.**
   Soit $p$ la fréquence de rois que le joueur découvrirait s'il jouait une infinité de fois.
   Soit $F$ la variable aléatoire qui, à chaque échantillon de 800 essais, associe la fréquence d'apparition du roi. La taille des échantillons étant suffisamment grande, on considère que $F$ suit la loi $\mathcal{N}\left(p; \sqrt{\frac{p(1-p)}{800}}\right)$. $F$ sera la variable aléatoire de décision.
3. **Zone critique.**
   Dans le cas où l'hypothèse $H_0$ est vraie, la variable aléatoire $F$ suit la loi $\mathcal{N}\left(0,125; \sqrt{\frac{0,125 \times 0,875}{800}}\right)$ soit $\mathcal{N}(0,125; 0,0117)$.
   On cherche alors le réel $d$ tel que :
   $$(E) : \quad \boxed{\mathbb{P}(F > 0,125 + d) = 0,01}$$
   la variable aléatoire $T = \frac{F - 0,125}{0,0117}$ suit la loi normale centrée réduite $\mathcal{N}(0,1)$, on a alors :
   $$\begin{aligned} (E) &\iff \mathbb{P}(0,0117T + 0,125 > 0,125 + d) = 0,01 \\ &\iff \mathbb{P}\left(T > \frac{d}{0,0117}\right) = 0,01 \\ &\iff 1 - \mathbb{P}\left(T \leq \frac{d}{0,0117}\right) = 0,01 \\ &\iff \Pi\left(\frac{d}{0,0117}\right) = 0,99 \end{aligned} \tag{4}$$
   On trouve alors d'après la table que $\frac{d}{0,0117} = 2,33$ soit $d = 0,027261 \approx 0,027$.
   La zone critique est donc l'intervalle :
   $$I_{0,01} = [0,125 + 0,027; +\infty[ = [0,152; +\infty[$$
   On pourra également utiliser directement la calculatrice ou un logiciel de calcul formel pour déterminer la valeur du réel $d$ respectant $(E)$.

On obtient alors graphiquement la situation suivante :




Donc la zone critique est $]0, 152; +\infty[$.

4. **Règle de décision.**
Si la fréquence de l'échantillon est supérieure à 0, 152, on rejette l'hypothèse $H_0$ et on accepte l'hypothèse $H_1$ : l'hypothèse $H_0$ n'est pas validée.
Si la fréquence de l'échantillon est inférieure à 0, 152, on accepte l'hypothèse $H_0$ : l'hypothèse $H_0$ est validée.

5. **Conclusion.**
L'échantillon observé a une fréquence égale à $\dfrac{134}{800} = 0, 1675$.
D'après la règle de décision, puisque $0, 1675 > 0, 152$, on accepte l'hypothèse $H_1$ : on décide que le joueur est un tricheur.

# 4 Test de comparaison

## 4.1 Comparaison de deux moyennes

Exemple :
Une entreprise fabrique des sacs en plastique pour déchets. Afin de surveiller la production, elle effectue des contrôles réguliers portant sur le poids maximum que les sacs peuvent supporter.
Á une première date $t_1$ , le contrôle de 100 sacs a donné une moyenne de 58 kg et un écart type de 3 kg.
À la seconde date $t_2$, le contrôle de 150 sacs a donné une moyenne de 56 kg et un écart type de 5 kg.
Peut-on considérer, au risque de 4%, que la qualité des sacs a évolué entre les deux dates ?

1. **Choix des hypothèses.**
L'hypothèse nulle $H_0$ est $m_1 = m_2$ (la qualité n'a pas évolué).
L'hypothèse alternative $H_1$ est $m_1 \neq m_2$ (la qualité a évolué).
On a donc :
   - $H_0$ : "$m_1 = m_2$"
   - $H_1$ : "$m_1 \neq m_2$"

2. **Variable aléatoire de décision.**
Appelons $E_1$ (resp. $E_2$) l'ensemble de tous les sacs produits par l'entreprise à la date $t_1$ (resp. $t_2$).
   - Soit $M_1$ la variable aléatoire qui, à chaque échantillon de 100 sacs issus de la population $E_1$, associe sa moyenne.
   Une estimation ponctuelle de la moyenne et de l'écart-type de à la date $t_1$ est : $m_1 = 58$, et $\sigma_1 = 3 \times \sqrt{\dfrac{100}{99}}$.
   La taille des échantillons étant suffisamment grande, $M_1$ suit la loi $\mathcal{N}\left(m_1; \dfrac{\sigma_1}{\sqrt{100}}\right) = \mathcal{N}\left(58; \dfrac{1}{\sqrt{11}}\right)$.

- Soit $M_2$ la variable aléatoire qui, à chaque échantillon de 150 sacs issus de la population $E_2$, associe sa moyenne. Une estimation ponctuelle de la moyenne et de l'écart-type à la date $t_2$ est : $m_2 = 56$, et $\sigma_2 = 5 \times \sqrt{\dfrac{150}{149}}$.
  La taille des échantillons étant suffisamment grande, $M_2$ suit la loi $\mathcal{N}\left(m_2; \dfrac{\sigma_2}{\sqrt{150}}\right) = \mathcal{N}\left(56; \dfrac{5}{\sqrt{149}}\right)$.
- La variable aléatoire $D = M_1 - M_2$ suit également une loi normale de paramètres :
  - $E(D) = E(M_1) - E(M_2) = m_1 - m_2$.
  - $V(D) = V(M_1) + V(M_2) = \dfrac{1}{11} + \dfrac{25}{149} = 0,2587$.

  D'où $\sigma_D = 0,51$.
  Donc $D$ suit la loi $\mathcal{N}(m_1 - m_2; 0,51)$. $D$ est la variable aléatoire de décision.

3. **Zone critique.**
Supposons que l'hypothèse $H_0$ soit vraie, on a alors $m_1 - m_2 = 0$ ; alors $D$ suit la loi normale $\mathcal{N}(0; 0,51)$.
On cherche alors le réel $d$ tel que :

$$(E) : \quad \boxed{\mathbb{P}(-d \leq D \leq d) = 0,95}$$

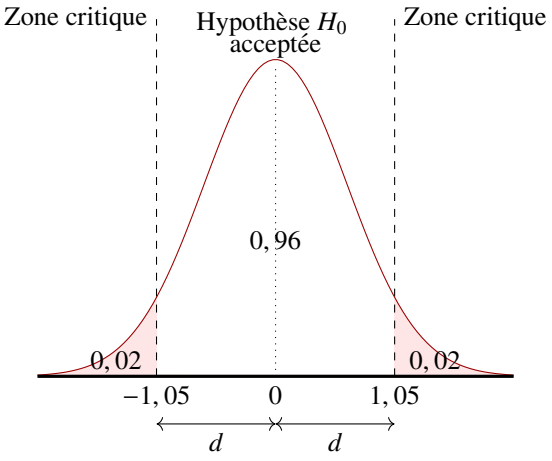
la variable aléatoire $T = \dfrac{D}{0,51}$ suit la loi normale centrée réduite $\mathcal{N}(0,1)$, on a alors :

$$\begin{aligned}(E) &\iff \mathbb{P}(-d < 0,51T < d) = 0,96 \\ &\iff \mathbb{P}\left(-\frac{d}{0,51} \leq T \leq \frac{d}{0,51}\right) = 0,96 \\ &\iff 2\Pi\left(\frac{d}{0,51}\right) - 1 = 0,96 \\ &\iff \Pi\left(\frac{d}{0,51}\right) = 0,98\end{aligned} \tag{5}$$

On trouve alors d'après la table que $\dfrac{d}{0,51} = 2,05$ soit $d = 1,0455 \approx 1,05$.
Pour un seuil de risque de 4%, la zone critique est :

$$I_{0,04} = ]-\infty; -1,05\,[ \cup ]\,1,05\,; +\infty[$$

On pourra également utiliser directement la calculatrice ou un logiciel de calcul formel pour déterminer la valeur du réel $d$ respectant $(E)$.
On obtient alors graphiquement la situation suivante :




4. **Règle de décision.**
Si la différence des moyennes des deux échantillons est inférieure à $-1,05$ ou supérieure à $1,05$, alors l'hypothèse $H_0$ , n'est pas validée.

Si la différence des moyennes des deux échantillons est comprise entre $-1,05$ et $1,05$,l'hypothèse $H_0$ est validée.

5. **Conclusion.**
   La différence des moyennes des deux échantillons est $58 - 56 = 2 > 1,05$.
   D'après la règle de décision, on rejette $H_0$ et on décide que la qualité des sacs a évolué entre les dates $t_1$ et $t_2$.

## 4.2 Comparaison de deux fréquences

Exemple :
A l'issue d'un examen, il y a 23 reçus et 17 ajournés dans une classe et 15 reçus et 25 ajournés dans une autre classe.
La différence observée entre les deux pourcentages de réussite est-elle significative d'une différence de niveau entre les deux classes, au seuil de 5% ?

1. **Choix des hypothèses.**
   L'hypothèse nulle $H_0$ est $p_1 = p_2$ (les deux populations ont le même niveau),
   l'hypothèse alternative $H_1$ est $p_1 \neq p_2$ (les deux populations n'ont pas le même niveau).
   On a donc :
   - $H_0$ : "$p_1 = p_2$"
   - $H_1$ : "$p_1 \neq p_2$"
2. **Variable aléatoire de décision.**
   On suppose que la première classe est issue d'une population $C_1$pour laquelle la fréquence de succès est $p_1$ , et que la deuxième classe est issue d'une population $C_2$ pour laquelle la fréquence de succès est $p_2$.

   - Soit $F_1$ la variable qui, à chaque échantillon de 40 élèves de la population $C_1$, associe sa fréquence de succès.
     La taille des échantillons étant suffisamment grande, on considère que $F_1$, suit la loi $\mathcal{N}\left(p_1; \sqrt{\dfrac{p_1(1-p_1)}{40}}\right)$.
     Une estimation ponctuelle de la fréquence et de l'écart-type pour la population $C_1$ est :
     $p_1 = \dfrac{23}{40} = 0,545$, et $\sigma_1 = \sqrt{\dfrac{40}{39}} \times \sqrt{\dfrac{0,545(1-0,545)}{40}} = 0,079$. Donc, $F_1$ suit la loi $\mathcal{N}(p_1; 0,079)$.
   - Soit $F_2$ la variable qui, à chaque échantillon de 40 élèves de la population $C_2$, associe sa fréquence de succès.
     La taille des échantillons étant suffisamment grande, on considère que $F_2$, suit la loi $\mathcal{N}\left(p_2; \sqrt{\dfrac{p_2(1-p_2)}{40}}\right)$.
     Une estimation ponctuelle de la fréquence et de l'écart-type pour la population $C_2$ est :
     $p_2 = \dfrac{15}{40} = 0,375$, et $\sigma_2 = \sqrt{\dfrac{40}{39}} \times \sqrt{\dfrac{0,375(1-0,375)}{40}} = 0,078$. Donc, $F_2$ suit la loi $\mathcal{N}(p_2; 0,078)$.
   - La variable aléatoire $D = F_1 - F_2$ suit également une loi normale de paramètres :
     - $E(D) = E(F_1) - E(F_2) = p_1 - p_2$.
     - $V(D) = V(F_1) + V(F_2) = 0,077^2 + 0,078^2$.

     D'où $\sigma_D = 0,11$.
     Donc $D$ suit la loi $\mathcal{N}(p_1 - p_2; 0,11)$. $D$ est la variable aléatoire de décision.

3. **Zone critique.**
   Supposons que l'hypothèse $H_0$ soit vraie, on a alors $p_1 - p_2 = 0$ ; alors $D$ suit la loi normale $\mathcal{N}(0; 0,11)$.
   On cherche alors le réel $d$ tel que :

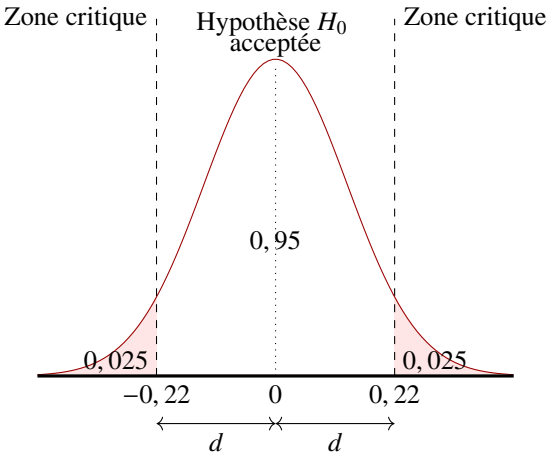
$$(E) : \quad \boxed{\mathbb{P}(-d \leq D \leq d) = 0,95}$$

la variable aléatoire $T = \dfrac{D}{0,11}$ suit la loi normale centrée réduite $\mathcal{N}(0,1)$, on a alors :

$$\begin{aligned}(E) &\iff \mathbb{P}(-d < 0,11T < d) = 0,95\\ &\iff \mathbb{P}\left(-\frac{d}{0,11} \leq T \leq \frac{d}{0,11}\right) = 0,95\\ &\iff 2\Pi\left(\frac{d}{0,11}\right) - 1 = 0,95\\ &\iff \Pi\left(\frac{d}{0,11}\right) = 0,975\end{aligned} \tag{6}$$

On trouve alors d'après la table que $\dfrac{d}{0,11} = 1,96$ soit $d = 0,2156 \approx 0,22$.
Pour un seuil de risque de 5%, la zone critique est :

$$I_{0,05} = ]-\infty; -0,22[ \cup ]0,22; +\infty[$$




4. **Règle de décision.**
   Si la différence des fréquences des deux échantillons est inférieure à $-0,22$ ou supérieure à $0,22$, alors l'hypothèse $H_0$ n'est pas validée. Sinon, l'hypothèse $H_0$ est validée.

5. **Conclusion.**
   La différence des fréquences de succès des deux échantillons est $\dfrac{23}{40} - \dfrac{15}{40} = 0,2 < 0,22$.
   D'après la règle de décision, on en conclut qu'au seuil de risque de 5%, la différence observée entre les deux échantillons n'est pas significative d'une différence de niveau entre les deux classes. (l'hypothèse $H_0$ est validée).