Statistiques inférentielles : tests d'hypothèse

Table des matières

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 tout ce chapitre, 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 : détermination de la variable aléatoire de décision et de ses paramètres,
- Étape 2 : choix des deux hypothèses : l'hypothèse nulle H_o et l'hypothèse alternative H_l,
- 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 α donné,
- 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.

I Tests d'hypothèse

I.1 Test bilatéral relatif à une moyenne

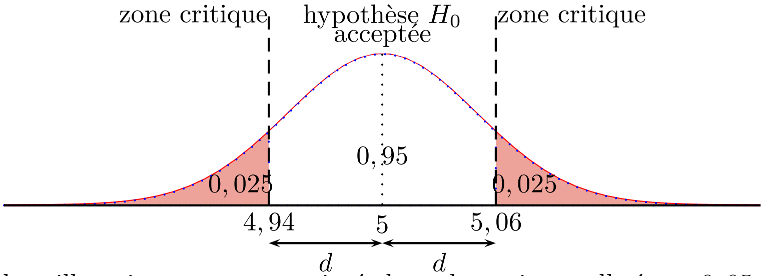
Exemple 1

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. **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.

2. **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 : « $m \neq 5$ ».
   Recherchons comment la moyenne m_e, d'un échantillon de 100 rondelles peut confirmer ou non l'hypothèse H_0.

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 $P(5 - d \leq M \leq 5 + d) = 0,95$. (E)
   la variable aléatoire $T = \dfrac{M-5}{0,03}$ suit la loi normale centrée réduite $\mathcal{N}(0,1)$, on a alors :
   (E) $\iff P(5 - d \leq 0,03T + 5 \leq 5 + d) = 0,95 \iff P\left(-\dfrac{d}{0,03} \leq T \leq \dfrac{d}{0,03}\right) = 0,95$
   (E) $\iff 2\Pi\left(\dfrac{d}{0,03}\right) - 1 = 0,95 \iff \Pi\left(\dfrac{d}{0,03}\right) = 0,975$
   On trouve alors $\dfrac{d}{0,03} = 1,96$ soit $d = 0,0588 \approx 0,06$.
   L'intervalle de confiance est donc l'intervalle : $[5 - 0,06; 5 + 0,06] = [4,94; 5,06]$.



   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 α est la probabilité de rejeter H_0 alors qu'elle est vraie.

I.2 Test unilatéral relatif à une moyenne

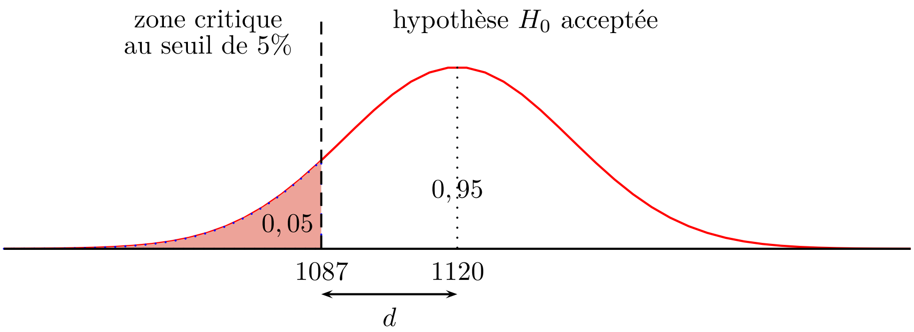
Exemple 2

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. **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 $\mathcal{N}\left(m; \dfrac{120}{\sqrt{36}}\right)$, c'est-à-dire $\mathcal{N}(m; 20)$.

2. **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).

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 hypothèse H_0 est vraie, la variable aléatoire M suit la loi $\mathcal{N}(1120; 20)$
   On cherche alors le réel d tel que $P(M < 1120 - d) = 0,05$. (E)
   la variable aléatoire $T = \dfrac{M - 1120}{20}$ suit la loi normale centrée réduite $\mathcal{N}(0, 1)$, on a alors :
   (E) $\iff P(20T + 1120 < 1120 - d) = 0,05 \iff P\left(T < -\dfrac{d}{20}\right) = 0,05$
   (E) $\iff P\left(T > \dfrac{d}{20}\right) = 0,05 \iff 1 - P\left(T \leq \dfrac{d}{20}\right) = 0,05 \iff \Pi\left(\dfrac{d}{20}\right) = 0,95$
   On trouve alors $\dfrac{d}{20} = 1,645$ soit $d = 32,9 \approx 33$.
   La zone critique est donc l'intervalle $]-\infty; 1120 - 33] =]-\infty; 1087]$.



   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.

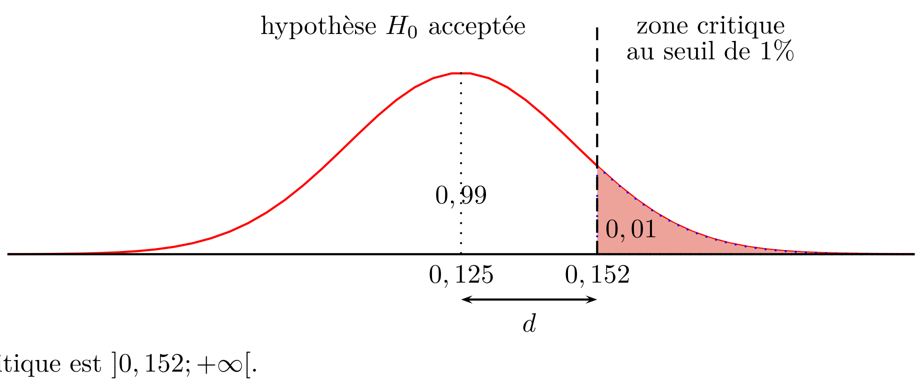
I.3 Test unilatéral relatifs à une fréquence

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 3
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. **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 $\mathcal{N}\left(p; \sqrt{\dfrac{p(1-p)}{800}}\right)$. F sera la variable aléatoire de décision.

2. **Choix des hypothèses.**
   Si le joueur n'est pas un tricheur, la valeur de p est $\dfrac{4}{32} = 0,125$.
   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).
3. **Zone critique.**
   Dans le cas où l'hypothèse H_0 est vraie, la variable aléatoire F suit la loi $\mathcal{N}\left(0,125; \sqrt{\dfrac{0,125 \times 0,875}{800}}\right)$ soit $\mathcal{N}(0,125; 0,0117)$.
   On cherche alors le réel d tel que $P(F > 0,125 + d) = 0,01$. (E)
   la variable aléatoire $T = \dfrac{F - 0,125}{0,0117}$ suit la loi normale centrée réduite $\mathcal{N}(0,1)$, on a alors :
   (E) $\iff P(0,0117T + 0,125 > 0,125 + d) = 0,01 \iff P\left(T > \dfrac{d}{0,0117}\right) = 0,01$
   (E) $\iff 1 - P\left(T \leq \dfrac{d}{0,0117}\right) = 0,01 \iff \Pi\left(\dfrac{d}{0,0117}\right) = 0,99$
   On trouve alors $\dfrac{d}{0,0117} = 2,33$ soit $d = 0,027261 \approx 0,027$.
   La zone critique est donc l'intervalle $[0,125 + 0,027; +\infty[= [0,152; +\infty[$.



   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.

II Test de comparaison

II.1 Comparaison de deux moyennes

Exemple 4

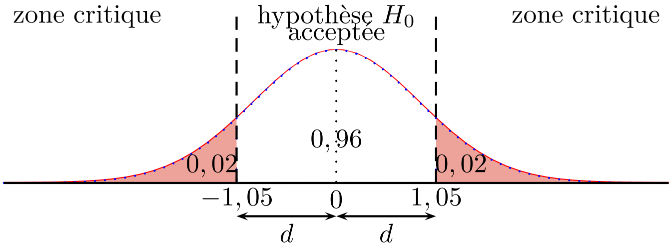
Une entreprise fabrique des sacs en plastique pour déchets. Afin de surveiller la production, elle effectue des contrôles réguliers portant sur le poids maximum que les sacs peuvent supporter.
Á une première date t_1 , le contrôle de 100 sacs a donné une moyenne de 58 kg et un écart type de 3 kg.
À la 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. **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 $\sigma_1 = 3 \times \sqrt{\frac{100}{99}}$.
     La taille des échantillons étant suffisamment grande, M_1 suit la loi $\mathcal{N}\left(m_1; \frac{\sigma 1}{\sqrt{100}}\right) = \mathcal{N}\left(58; \frac{1}{\sqrt{11}}\right)$.
   - 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 $\sigma_2 = 5 \times \sqrt{\frac{150}{149}}$.
     La taille des échantillons étant suffisamment grande, M_2 suit la loi $\mathcal{N}\left(m_2; \frac{\sigma 2}{\sqrt{150}}\right) = \mathcal{N}\left(56; \frac{5}{\sqrt{149}}\right)$.
   - La variable aléatoire $D = M_1 - M_2$ suit également une loi normale de paramètres :
     $E(D) = E(M_1) - E(M_2) = m_1 - m_2$.
     $V(D) = V(M_1) + V(M_2) = \frac{1}{11} + \frac{25}{149} = 0,2587$. 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.

2. **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).

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 $P(-d \leq D \leq d) = 0,95$. (E)
   la variable aléatoire $T = \frac{D}{0,51}$ suit la loi normale centrée réduite $\mathcal{N}(0,1)$, on a alors :
   (E) $\iff P(-d < 0,51T < d) = 0,96 \iff P\left(-\frac{d}{0,51} \leq T \leq \frac{d}{0,51}\right) = 0,96$
   (E) $\iff 2\Pi\left(\frac{d}{0,51}\right) - 1 = 0,96 \iff \Pi\left(\frac{d}{0,51}\right) = 0,98$. On trouve alors $\frac{d}{0,51} = 2,05$ soit $d = 1,0455 \approx 1,05$.
   Pour un seuil de risque de 4%, la zone critique est : $]-\infty; -1,05[\cup]1,05 ; +\infty[$.



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 $58 - 56 = 2$.
   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.

II.2 Comparaison de deux fréquences

Exemple 5

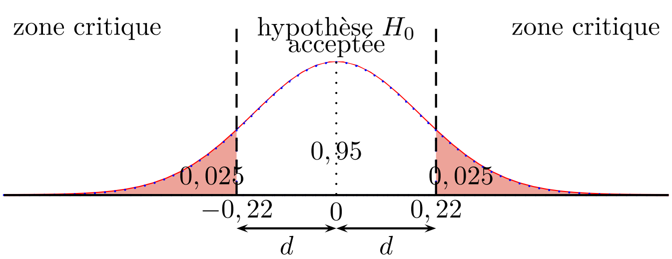
A l'issue d'un examen, il y a 23 reçus et 17 ajournés dans une classe et 15 reçus et 25 ajournés dans une autre classe.
La différence observée entre les deux pourcentages de réussite est-elle significative d'une différence de niveau entre les deux classes, au seuil de 5%

1. **Variable aléatoire de décision.**
   On suppose que la première classe est issue d'une population C_1pour 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 $\mathcal{N}\left(p_1; 0,079\right)$.
   - 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 :
     $p_1 2 = \dfrac{15}{40} = 0,375$, et $\sigma_2 = \sqrt{\dfrac{40}{39}} \times \sqrt{\dfrac{0,375(1-0,375)}{40}} = 0,078$. Donc, F_2 suit la loi $\mathcal{N}\left(p_2; 0,078\right)$.
   - 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.

2. **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).

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 $P(-d \leq D \leq d) = 0,95$. (E)
   la variable aléatoire $T = \dfrac{D}{0,11}$ suit la loi normale centrée réduite $\mathcal{N}(0,1)$, on a alors :
   (E) $\iff P(-d < 0,11T < d) = 0,95 \iff P\left(-\dfrac{d}{0,11} \leq T \leq \dfrac{d}{0,11}\right) = 0,95$
   (E) $\iff 2\Pi\left(\dfrac{d}{0,11}\right) - 1 = 0,95 \iff \Pi\left(\dfrac{d}{0,11}\right) = 0,975$. On trouve alors $\dfrac{d}{0,11} = 1,96$ soit $d = 0,2156 \approx 0,22$.
   Pour un seuil de risque de 5%, la zone critique est : $]-\infty; -0,22[\cup]0,22; +\infty[$.



4. **Règle de décision.**
   Si la différence des moyennes 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 $\dfrac{23}{40} - \dfrac{15}{40} = 0,2$.
   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).