CHAPITRE

9 Échantillonage, estimation et test d'hypothèse

L'inférence statistique est l'ensemble de techniques permettant d'induire les caractéristiques d'un groupe général (la population) à partir de celles d'un groupe particulier (l'échantillon), en fournissant une mesure de la certitude de la prédiction : la probabilité d'erreur.

Strictement, l'inférence s'applique à l'ensemble des membres (pris comme un tout) de la population représentée par l'échantillon, et non pas à tel ou tel membre particulier de cette population. Par exemple, les intentions de vote indiquées par l'échantillon ne peuvent révéler l'intention de vote qu'à tel ou tel membre particulier de la population des électeurs de la circonscription électorale.

L'inférence statistique est donc un ensemble de méthodes permettant de tirer des conclusions fiables à partir de données d'échantillons statistiques. L'interprétation de données statistiques est, pour une large part, le point clé de l'inférence statistique. Elle est guidée par plusieurs principes et axiomes.

L'union entre les méthodes statistiques rudimentaires de Pierre-Simon de Laplace et de Carl Friedrich Gauss, confinées à l'astronomie, et la science de l'État, circonscrite à la démographie et aux sciences sociales naissantes, a lieu à la charnière des XIX^e siècle et XX^e siècle, dans le domaine intermédiaire de la biologie, lorsque l'évolution fut reformulée en tant que problème statistique grâce à l'influence de l'eugénisme et de la biométrie1.

Les méthodes d'inférence statistiques ont connu deux grandes phases de développement. La première commence à la fin du XIX^e siècle, avec les travaux de K. Pearson, R. Fisher, Jerzy Neyman, Egon Pearson et Abraham Wald qui dégagent les notions fondamentales de vraisemblance, de puissance des tests d'hypothèse et d'intervalle de confiance.

La seconde période, qui perdure aujourd'hui, a été rendue possible grâce à la puissance de calcul des ordinateurs et à la banalisation de l'outil informatique à partir de la fin des années 1940. Ces calculateurs ont permis de dépasser les hypothèses traditionnelles d'indépendance et de normalité, commodes du point de vue mathématique mais souvent simplistes, pour donner toute leur fécondité à des concepts même anciens comme l'hypothèse bayésienne. L'informatique a permis aussi l'explosion des techniques de simulation par application des techniques de rééchantillonnage : méthode de Monte Carlo, bootstrap, jackknife etc. imaginées par John von Neumann, Stanislas Ulam, Bradley Efron, Richard von Mises.

https ://fr.wikipedia.org/wiki/Inférence_statistique

Table des matières

1 Échantillonage **4**

1.1 Lois limites 4

1.1.1 Rappels sur les lois usuelles 4

1.1.2 Loi faible des grands nombres 5

1.1.3 Théorème de la limite centrée 5

1.2 Applications 6

1.2.1 Approximations de la loi Binomiale par la loi normale 6

1.2.2 Application : lois d'échantillonnage 7

2 Estimation **9**

2.1 Estimation ponctuelle d'un paramètre 9

2.1.1 Moyenne 9

2.1.2 Écart-type 9

2.1.3 Fréquence 10

2.2 Estimation par intervalle de confiance d'un paramètre 10

2.2.1 Moyenne 11

2.2.2 Fréquence 12

2.3 Tableau récapitulatif 13

3 Test d'hypothèse **14**

3.1 Test bilatéral relatif à une moyenne 15

3.2 Test unilatéral relatif à une moyenne 16

3.3 Test unilatéral relatifs à une fréquence 17

3.4 Test de comparaison 19

3.4.1 Comparaison de deux moyennes 19

3.4.2 Comparaison de deux fréquences 20

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

Exemple

Un industriel produit en très grand nombres des yaourts, pour lesquelles l'usinage doit respecter des normes sanitaires draconiennes. À 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 tout 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 avait fait analyser 200, 1000, 10000 yaourts ?

La question 1, 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**.

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 »**.

Chapitre 1

Échantillonage

1.1 Lois limites

1.1.1 Rappels sur les lois usuelles

Voici un tableau récapitulatif représentant les principales formules des lois usuelles vues :

(Dans toutes les formules, on a $p+q=1$ c'est-à-dire $q=1-p$).

Loi	Notation	Formule	Espérance	Variance
Loi de Bernoulli	$\mathcal{B}(p)$	$P(X=1)=p$; $P(X=0)=q$	$E(X)=p$	$V(X)=pq)$
Loi Binomiale	$\mathcal{B}(n;p)$	$P(X=k)=C_n^k \times p^k \times q^{n-k}$	$E(X)=np$	$V(X)=npq)$
Loi de Poisson	$\mathcal{P}(\lambda)$	$P(X=k)=e^{-\lambda}\dfrac{\lambda^k}{k!}$	$E(X)=\lambda$	$V(X)=\lambda$
Loi Normale	$\mathcal{N}(m;\sigma)$	$f(x)=\dfrac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-m}{\sigma}\right)^2}$	$E(X)=m$	$V(X)=\sigma^2$
Centrée réduite	$\mathcal{N}(0;1)$	$f(x)=\dfrac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2}$	$E(X)=0$	$V(X)=1$

1.1.2 Loi faible des grands nombres

Proposition 1.

Soit X_1, X_2, ..., X_n n variables aléatoires indépendantes de même loi, d'espérance m et d'écart-type σ et soit $\overline{X_n} = \dfrac{X_1 + X_2 + \cdots + X_n}{n}$, alors :

$$\text{Pour tout } \epsilon > 0, \lim_{n \to +\infty} P(|\overline{X_n} - m| < \epsilon) = 1.$$

Concrètement, ce théorème signifie que plus n est grand plus la variable aléatoire $\overline{X_n}$ se rapproche de l'espérance mathématique m.

Exemple

On lance un dé. Si on obtient 6, c'est gagné et on marque 1 point. Sinon, c'est perdu et on marque 0 point.
Soit X_i la variable aléatoire correspondant au nombre de point obtenu lors du i^e lancer.
On a donc : $P(X = 0) = \dfrac{5}{6}$, $P(X = 1) = \dfrac{1}{6}$ et $E(X) = \dfrac{1}{6}$.
On répète n fois cette même expérience, les n variables aléatoires X_1, X_2, ..., X_n ont la même loi de probabilité.
Pour connaitre le nombre de succès, on étudie la variable aléatoire $\overline{X_n}$: « Fréquence des succès »
avec $\overline{X_n} = \dfrac{\text{Nombre de succès}}{\text{Nombre d'expériences aléatoires}} = \dfrac{X_1 + X_2 + \cdots + X_n}{n}$.
Y-a-t-il une forte probabilité pour que $\overline{X_n}$ soit proche de $\dfrac{1}{6}$?

- Pour $n = 3$ par exemple, il y a peu de chance pour que l'on trouve $\overline{X_3} = \dfrac{1}{6}$.
- Pour $n = 30$, la probabilité de trouver $\overline{X_{30}} = \dfrac{1}{6}$ augmente sans être très forte.
- Pour $n = 1000$, on se rapproche de cette valeur de $\dfrac{1}{6}$.

Le théorème dit que plus n est grand, plus $\overline{X_n}$ se rapproche de la valeur théorique $\dfrac{1}{6}$.

1.1.3 Théorème de la limite centrée

Proposition 2.

Soit X_1, X_2, ..., X_n n variables aléatoires indépendantes de même loi, d'espérance m et d'écart-type σ et soit $\overline{X_n} = \dfrac{X_1 + X_2 + \cdots + X_n}{n}$, alors :
Pour n suffisamment grand, $\overline{X_n}$ suit approximativement la loi normale $\mathcal{N}\left(m, \dfrac{\sigma}{\sqrt{n}}\right)$.

Remarque

Dans la plupart des cas, on considère que n est « suffisamment grand » lorsque n atteint quelques dizaines, par exemple lorsque $n \geq 30$, mais cela dépend de la nature, de la population et du contexte de l'étude

1.2 Applications

1.2.1 Approximations de la loi Binomiale par la loi normale

Proposition 3.

Pour n « assez grand » et pour p « ni proche de 0 ni de 1 » tels que $np(1-p)$ ne soit « pas trop grand », on peut approcher la loi binomiale $\mathcal{B}(n;p)$ par la loi normale $\mathcal{N}(m;\sigma)$ où $m = np$ et $\sigma = \sqrt{np(1-p)}$.

On convient de faire cette approximation pour $n \geq 50$, $p \leq 50$ et $np(1-p) > 10$.

Exemple

On lance 300 fois une pièce de monnaie truquée ce qui constitue une partie. La probabilité d'obtenir « face » est $\frac{2}{3}$.

On désigne par X la variable aléatoire qui à chaque partie associe le nombre de « face » obtenus.

1. Justifier que X suit une loi binomiale, en préciser les paramètres.
2. Peut-on calculer simplement $P(X > 210)$?
3. Montrer qu'une approximation de la loi binomiale par une loi normale se justifie.
4. Calculer $P(X > 210)$ à l'aide de cette approximation.

Solution

1. Pour chaque jet, on a **deux résultats possibles** : ou bien on obtient « face » avec une probabilité de $p = \frac{2}{3}$, ou bien on obtient « pile » avec une probabilité de $q = 1 - p = \frac{1}{3}$.
   On lance 300fois la pièce de manière **indépendante**.
   On peut donc conclure que X suit la loi Binomiale $\mathcal{B}\left(300; \frac{2}{3}\right)$.
2. $P(X > 210) = \sum_{i=211}^{300} C_{300}^{i} \times \frac{2}{3}^{i} \times \frac{1}{3}^{300-i}$, la calculatrice ne peut pas toujours effectuer un tel calcul.
3. On a $n \geq 50$, $p = \frac{2}{3}$ et $np(1-p) = 66,66 > 10$.
   On peut donc faire une approximation par la loi normale $\mathcal{P}\left(300 \times \frac{2}{3}; \sqrt{300 \times \frac{2}{3} \times \frac{1}{3}}\right) = \mathcal{P}(200; 8,16)$.
4. On utilise le changement de variable $T = \dfrac{X - 200}{8,16}$. T suit la loi normale $\mathcal{N}(0,1)$.

$$\begin{aligned} P(X > 210) &= P(8,16T + 200 > 210) \\ &= P(T > 1,22) \\ &= 1 - P(T \leq 1,22) \\ &= 1 - 0,8888 \\ &= 0,1112. \end{aligned}$$

1.2.2 Application : lois d'échantillonnage

En statistiques, il est en général impossible d'étudier un caractère sur toute une population de taille N élevée. La théorie de l'échantillonnage se pose la question suivante :

En supposant connus les paramètres statistiques de la population, que peut-on en déduire sur les échantillons prélevés dans la population ?

On suppose que ces échantillons sont prélevés au hasard et que le tirage de ces échantillons est effectué avec remise. L'ensemble de ces échantillons de taille n est appelé échantillonnage de taille n.

On peut étudier dans ces conditions :

— la loi d'échantillonnage des moyennes,

— la loi d'échantillonnage des fréquences,

Loi d'échantillonnage des moyennes

Étant donné une population de taille N et X une variable aléatoire telle que $E(X) = m$ et $\sigma(X) = \sigma$.

Pour prélever les échantillons de taille n, on a procédé à n épreuves indépendantes de variables aléatoires X_1, X_2, …, X_n de même loi que X.

La variable aléatoire $\overline{X_n} = \dfrac{X_1 + X_2 + \cdots + X_n}{n}$ associe à tout échantillon de taille n sa moyenne.

D'après le théorème de la limite centrée, pour n assez grand, on a :

Proposition 4.

La loi d'échantillonnage de taille n de la moyenne $\overline{X_n}$ quand $n \geq 30$, peut être approchée par la loi normale $\mathcal{N}\left(m, \dfrac{\sigma}{\sqrt{n}}\right)$.

Exemple

Une machine fabrique des pièces en grande série. À chaque pièce tirée au hasard, on associe sa longueur exprimée en millimètres ; on définit ainsi une variable aléatoire X.

On suppose que X suit la loi normal $\mathcal{N}(28,20; 0,027)$.

Soit M_n la variable aléatoire qui à tout échantillon aléatoire de taille n associe la moyenne des longueurs des n pièces de l'échantillon.

La propriété nous dit alors que pour n assez grand, M_n suit la loi normale $\mathcal{N}\left(m, \dfrac{\sigma}{\sqrt{n}}\right)$ soit $\mathcal{N}\left(28,20; \dfrac{0,027}{\sqrt{n}}\right)$.

Supposons que les échantillons soient de taille 100, alors M_{100} suit la loi $\mathcal{N}(28,20; 0,0027)$.

Loi d'échantillonnage des fréquences

On étudie, dans une population de taille N, un caractère X suivant une loi de bernoulli $\mathcal{B}(p)$, c'est-à-dire que les éléments possèdent une certaine propriété de fréquence p.

Dans un échantillon de taille n, on répète n fois la même épreuve de façon indépendante. On obtient n variables aléatoires X_1, X_2, ..., X_n de même loi que X.

La variable aléatoire $f_n = \dfrac{X_1 + X_2 + \cdots + X_n}{n}$ associe à tout échantillon de taille n la fréquence de succès sur cet échantillon.

Proposition 5.

La loi d'échantillonnage de taille n de la fréquence f_n pour n « assez grand » peut être approchée par la loi normale $\mathcal{N}\left(p; \sqrt{\dfrac{p(1-p)}{n}}\right)$.

On convient de dire que n est « assez grand » lorsque $n \geq 50$.

Remarque

Ce résultat est un cas particulier du précédent en l'appliquant à $m = p$ et $\sigma = \sqrt{p(1-p)}$

Exemple

Une urne contient 100 boules numérotées de 1 à 100, indiscernables au toucher. Lors d'un tirage aléatoire d'une boule, la probabilité d'obtenir un nombre inférieur ou égal à 37 est $p = 0,37$. On appelle succès l'événement qui consiste à tirer une des boules numérotées de 1 à 37.

Un échantillon de taille 50 est obtenu par un tirage aléatoire, avec remise, de 50 boules. On s'intéresse à la fréquence d'apparition d'un succès lors du tirage de ces 50 boules.

Soit f_{50} la variable aléatoire qui à chaque échantillon de taille 50 associe sa fréquence de succès.

X_i est la variable aléatoire qui à chaque échantillon associe 1 si la $i-ième$ boule apporte un succès, O sinon.

Les X_i sont des variables aléatoires indépendantes et suivent le même loi de Bernoulli de paramètre $p = 0,37$ d'espérance $E(X_i) = 0,37$ et d'écart-type $\sigma(X) = \sqrt{p(1-p)} = 0,48$.

On a $f_{50} = \dfrac{X_1 + X_2 + \cdots + X_{50}}{50}$ qui a pour espérance mathématique $p = 0,37$ et pour écart-type $\sqrt{\dfrac{0,37 \times 0,63}{50}} = 0,068$.

Chapitre 2

Estimation

2.1 Estimation ponctuelle d'un paramètre

2.1.1 Moyenne

Proposition 6.

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.

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 Écart-type

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

Proposition 7.

L'écart-type σ_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.

Une meilleure estimation σ de l'écart-type réel est obtenue en considérant le nombre $\sigma = \sigma_e\sqrt{\dfrac{n}{n-1}}$, où n est la taille de l'échantillon servant au calcul de σ_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 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*, et de l'ordre de* $0,05\%$ *pour des échantillons de taille* 1000.

2.1.3 Fréquence

Proposition 8.

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 $f_e = \frac{3}{150} = 0,02$ 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.
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 de la loi d'échantillonnage de la moyenne $\overline{X}$ par la loi normale $\mathcal{N}\left(m; \frac{\sigma}{\sqrt{n}}\right)$.

On pose $T = \dfrac{\overline{X} - m}{\frac{\sigma}{\sqrt{n}}}$, T suit donc la loi normale centrée réduite $\mathcal{N}(0;1)$.

Soit α la probabilité, fixée à l'avance, pour que T n'appartienne pas à l'intervalle $[-t;t]$, on peut écrire :

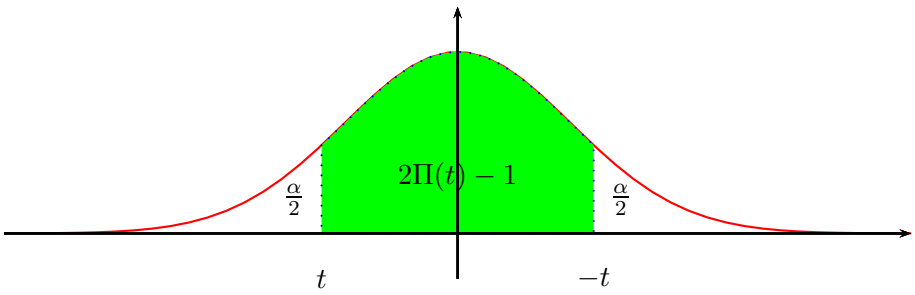
$$\begin{aligned}
P(|T| > t) = \alpha &\iff 1 - P(|T| \leq t) = \alpha \\
&\iff P(|T| \leq t) = 1 - \alpha \\
&\iff P(-t \leq T \leq t) = 1 - \alpha \\
&\iff P\left(-t \leq \frac{\overline{X} - m}{\frac{\sigma}{\sqrt{n}}} \leq t\right) = 1 - \alpha \\
&\iff P\left(\overline{X} - t\frac{\sigma}{\sqrt{n}}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

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

Proposition 9.

L'intervalle $\left[\overline{X} - t\frac{\sigma}{\sqrt{n}}; \overline{X} + t\frac{\sigma}{\sqrt{n}}\right]$ est l'intervalle de confiance de la moyenne m de la population avec le coefficient de confiance $2\Pi(t) - 1 = 1 - \alpha$.



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 10%.

On a : $\alpha = 10\%$ d'où $2\Pi(t) - 1 = 0,90 \iff \Pi(t) = 0,95 \iff t = 1,645$.

Un intervalle de confiance de M est donc : $\left[3000 - 1,645\frac{20}{\sqrt{16}}; 3000 + 1,645\frac{20}{\sqrt{16}}\right] = [2992, 3008]$.

2.2.2 Fréquence

À 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 (théorique) des éléments de la population possédant une certaine propriété.

Soit F la variable aléatoire qui à chaque échantillon de taille n associe la fréquence du nombre d'élément qui appartiennent à la catégorie choisie.

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

On sait que l'écart type associé à la fréquence f (pratique) de la loi d'échantillonage des fréquences de l'échantillon de taille n est $\sigma' = \sqrt{\frac{f(1-f)}{n}}$ et on se sert de l'estimation ponctuelle de σ puisque p est inconnue :

$$\sigma = \sigma'\sqrt{\frac{n}{n-1}} = \sqrt{\frac{f(1-f)}{n}} \times \sqrt{\frac{n}{n-1}} = \sqrt{\frac{f(1-f)}{n-1}}.$$

Donc la variable aléatoire $T = \dfrac{F-p}{\sigma}$ suit approximativement une loi normale centrée réduite.

On cherche un intervalle de confiance de la fréquence p, c'est à dire un intervalle tel que la probabilité que la fréquence p appartienne à cet intervalle soit égale à α où $\alpha \in [0; 1]$.

C'est un intervalle de confiance avec le coefficient de confiance α ou avec le risque $1 - \alpha$.

Le risque que l'on prend à dire que p appartient à cet intervalle est donc de $1 - \alpha$.

Déterminons cet intervalle de confiance : Soit t le nombre réel positif tel que $P(-t \leq T \leq t) = \alpha$ on a donc $2\Pi(t) - 1 = \alpha$ d'où t est tel que $\Pi(t) = \frac{1+\alpha}{2}$ et :

$$\begin{aligned}
P(-t \leq T \leq t) = \alpha &\iff P\left(-t \leq \frac{F-p}{\sigma} \leq t\right) = \alpha \\
&\iff P\left(-t\sigma \leq F - p \leq t\sigma\right) = \alpha \\
&\iff P\left(F - t\sigma \leq p \leq F + t\sigma\right) = \alpha \\
&\iff P\left(F - t\sqrt{\frac{f(1-f)}{n-1}} \leq p \leq F + t\sqrt{\frac{f(1-f)}{n-1}}\right) = \alpha
\end{aligned}$$

L'intervalle de confiance de la fréquence p avec un coefficient de confiance de α est :

$$\left[f - t\sqrt{\frac{f(1-f)}{n-1}}; f + t\sqrt{\frac{f(1-f)}{n-1}}\right]$$

Proposition 10.

L'intervalle $\left[f - t\sqrt{\frac{f(1-f)}{n-1}}; f + t\sqrt{\frac{f(1-f)}{n-1}}\right]$ est l'intervalle de confiance d'une fréquence p de la population avec le coefficient de confiance $2\Pi(t)-1 = \alpha$ ayant pour centre la fréquence f de l'échantillon considéré.

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 : $f = 0,42$; $n = 500$; $\alpha = 1\%$ donc $t = 2,575$.

Un intervalle de confiance du pourcentage p est donc :

$$\left[0,42 - 2,58\sqrt{\frac{0,42 \times 0,58}{499}}; 0,42 + 2,575\sqrt{\frac{0,42 \times 0,58}{499}}\right] = [0,36; 0,48] = [36\%; 47\%].$$

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 $2\Pi(t) - 1$ pour la population totale
Moyenne	m_e	$m = m_e$	$\left[m_e - t\frac{\sigma}{\sqrt{n}}; m_e + t\frac{\sigma}{\sqrt{n}}\right]$
Écart-type	σ_e	$\sigma = \sigma_e\sqrt{\frac{n}{n-1}}$	
Fréquence	f_e	$f = f_e$	$\left[f_e - t\sqrt{\frac{f_e(1-f_e)}{n-1}}; f_e + t\sqrt{\frac{f_e(1-f_e)}{n-1}}\right]$

Chapitre 3

Test 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 de 50 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_1,
- **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.

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. **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. Nous allons donc tester l'hypothèse $m = 5$. C'est 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)

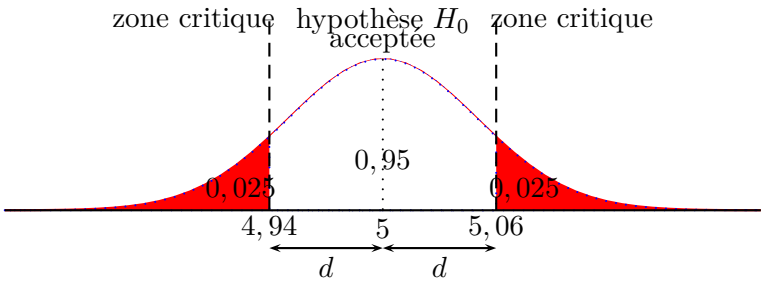
   la variable aléatoire $T = \dfrac{M-5}{0,03}$ suit la loi normale centrée réduite $\mathcal{N}(0,1)$, on a alors :

   $$(\text{E}) \iff P(5-d \leq 0,03T+5 \leq 5+d) = 0,95 \iff P\left(-\frac{d}{0,03} \leq T \leq \frac{d}{0,03}\right) = 0,95$$

   $$(\text{E}) \iff 2\Pi\left(\frac{d}{0,03}\right) - 1 = 0,95 \iff \Pi\left(\frac{d}{0,03}\right) = 0,975$$

   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.

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. **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)

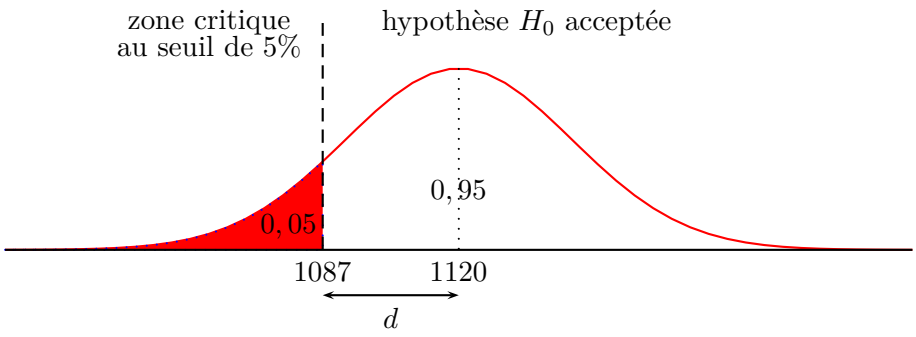
la variable aléatoire $T = \dfrac{M - 1120}{20}$ suit la loi normale centrée réduite $\mathcal{N}(0,1)$, on a alors :

$$(\text{E}) \iff P(20T + 1120 < 1120 - d) = 0,05 \iff P\left(T < -\frac{d}{20}\right) = 0,05$$

$$(\text{E}) \iff P\left(T > \frac{d}{20}\right) = 0,05 \iff 1 - P\left(T \leq \frac{d}{20}\right) = 0,05 \iff \Pi\left(\frac{d}{20}\right) = 0,95$$

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.

3.3 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

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.**

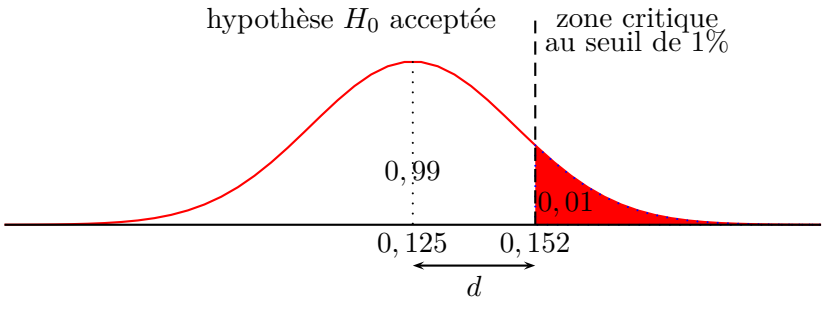
   Dans le cas où l'hypothèse H_0 est vraie, la variable aléatoire F suit la loi $\mathcal{N}\left(0,125; \sqrt{\dfrac{0,125 \times 0,875}{800}}\right)$ soit $\mathcal{N}(0,125; 0,0117)$. On cherche alors le réel d tel que $P(F > 0,125 + d) = 0,01$. (E)

   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 $\frac{134}{800} = 0,1675$.

   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.

3.4 Test de comparaison

3.4.1 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.

À 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 des 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 est suffisamment grande, $M_1 \sim \mathcal{N}\left(m_1; \frac{\sigma_1}{\sqrt{100}}\right) = \mathcal{N}\left(58; \frac{1}{\sqrt{11}}\right)$.

   — 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 est suffisamment grande, $M_2 \sim \mathcal{N}\left(m_2; \frac{\sigma_2}{\sqrt{150}}\right) = \mathcal{N}\left(56; \frac{5}{\sqrt{149}}\right)$.

   — 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 $D \sim \mathcal{N}(0; 0,51)$.

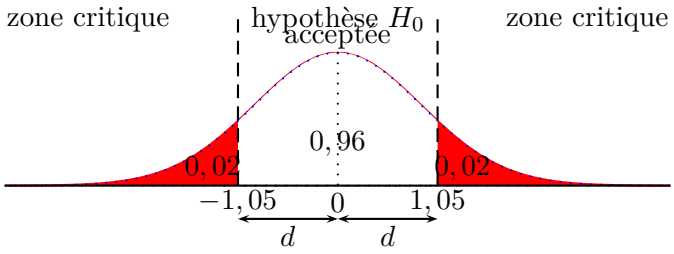
   On cherche alors le réel d tel que $P(-d \leq D \leq d) = 0,96$. (E)

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

   $(E) \iff P(-d < 0,51T < d) = 0,96 \iff P\left(-\dfrac{d}{0,51} \leq T \leq \dfrac{d}{0,51}\right) = 0,96$

   $(E) \iff 2\Pi\left(\dfrac{d}{0,51}\right) - 1 = 0,96 \iff \Pi\left(\dfrac{d}{0,51}\right) = 0,98$. On trouve alors $\dfrac{d}{0,51} = 2,05$ 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 $1,05$ alors 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.

3.4.2 Comparaison de deux fréquences

Exemple

À 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. 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_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 $F_1 \sim \mathcal{N}\left(p_1; \sqrt{\dfrac{p_1(1-p_1)}{40}}\right)$.

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 $F_2 \sim \mathcal{N}\left(p_2; \sqrt{\dfrac{p_2(1-p_2)}{40}}\right)$.

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.

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 $D \sim \mathcal{N}(0; 0,11)$.

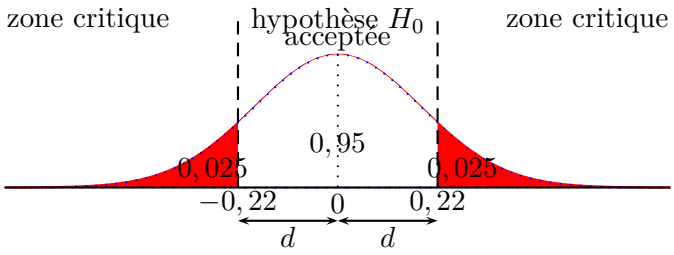
On cherche alors le réel d tel que $P(-d \leq D \leq d) = 0,95$. (E)

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).