Chapitre C

Tests d'hypothèse et de comparaison

I) Test de comparaison

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.
A une première date t_1, le contrôle de 100 sacs a donné une moyenne de 58 kg et un écart type de 3 kg.
A 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 le poids des sacs a évolué entre les deux dates ?

Notons ... le poids moyen de tous les sacs (de la population entière), à la première date t_1, et ... à la seconde date t_2. Les deux poids moyens ... et ... ne sont pas connus et ne le seront pas.

1. **Variable aléatoire de décision.**
   Notons E_1 l'ensemble de tous les sacs produits par l'entreprise à la date t_1, et E_2 à la date t_2.

   - Soit $\overline{M}_1$ 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 :
     - $m_1 = 58$ est la moyenne des masses des sacs de l'échantillon n° 1 et $s_1 = 3$ son écart-type.
     - $\sigma_1 = 3 \times$ est une estimation ponctuelle de l'écart-type de la population à la date t_1.

     La taille des échantillons étant suffisamment grande, $\overline{M}_1$ suit la loi $\mathcal{N}\left(\mu_1; \dfrac{\sigma_1}{\sqrt{100}}\right) = \mathcal{N}(\mu_1\,;\ .....)$

     Attention !

     Ne pas confondre m_1 et μ : $m_1 = 58$ est une expression de la variable aléatoire $\overline{M}_1$ sur un échantillon de 100 sacs et μ est son espérance.

   - Soit $\overline{M}_2$ 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 :
     - $m_2 = 56$ est la moyenne des masses des sacs de l'échantillon n° 2 et $s_2 = 5$ son écart-type.
     - $\sigma_2 = 5 \times$ est une estimation ponctuelle de l'écart-type de la population à la date t_2.

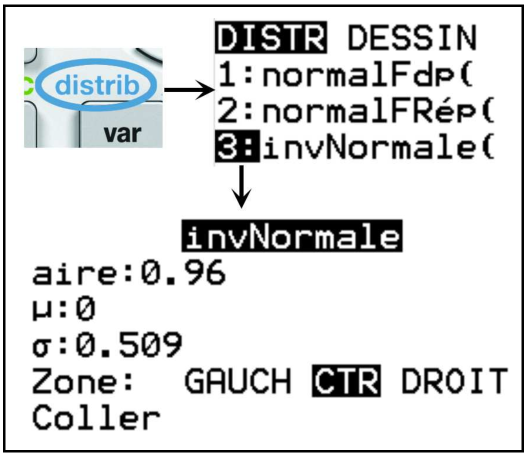
     La taille des échantillons étant suffisamment grande, $\overline{M}_2$ suit la loi $\mathcal{N}\left(\mu_2; \dfrac{\sigma_2}{\sqrt{150}}\right) = \mathcal{N}(\mu_2\,;\ .....)$.

- La variable aléatoire $D = \overline{M}_1 - \overline{M}_2$ suit également une loi normale de paramètres :
  - $E(D) = E(\overline{M}_1) - E(\overline{M}_2) =$ ....
  - $V(D) = V(\overline{M}_1) + V(\overline{M}_2) =$ .... D'où $\sigma_D =$...............................

  Donc, la **variable aléatoire de décision** D suit une loi $\mathcal{N}(\mu_1 - \mu_2\,;\ \ldots\ldots)$.

2. **Choix des hypothèses.**
   L'hypothèse nulle H_0 est .... (le poids n'a pas évolué).
   L'hypothèse alternative H_1 est .... (le poids a évolué).

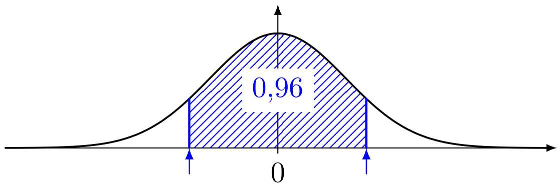
3. **Zone critique.**
   Supposons que l'hypothèse H_0 soit vraie, on a alors $\mu_1 - \mu_2 = \ldots$; alors D suit la loi normale $\mathcal{N}(\ \ldots\ ;\ 0{,}509)$.
   On cherche alors, le réel d tel que $P(-d \leq D \leq d) = \ldots$ car le risque imposé par l'énoncé est de 4%.



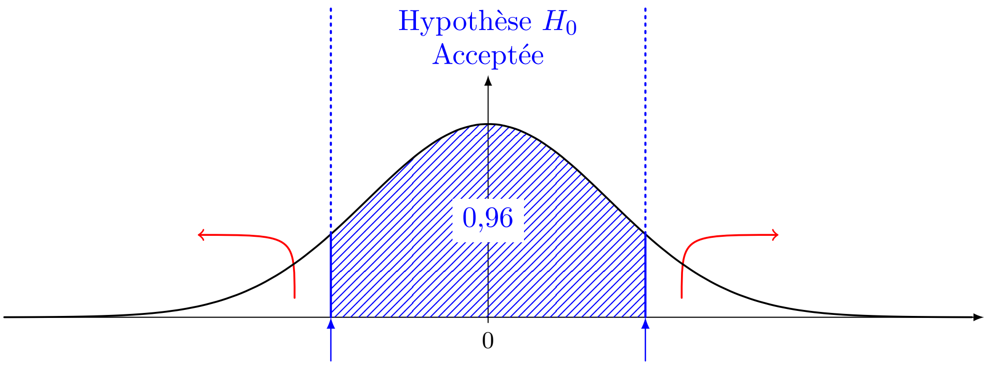
<u>La calculatrice affiche un intervalle :</u>

$[\ldots\ldots\ldots\ldots\ ;\ \ldots\ldots\ldots\ldots]$

Soit $d \simeq \ldots$



Pour un seuil de risque de 4%, la zone critique est : $]-\infty\,;\ \ldots\ldots\ [\ \cup\ ]\ \ldots\ldots\ ;+\infty[$.



4. **Règle de décision.**

☞ Si $m_1 - m_2 \in [\,.....\,;\,\,.....\,]$, alors l'hypothèse H_0 est validée ;

☞ sinon, l'hypothèse H_0 , n'est pas validée.

5. **Conclusion.**

La différence des moyennes des deux échantillons est $m_1 - m_2 = \ldots\ldots\ldots\ldots = \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$

D'après la règle de décision, on rejette H_0 et on décide que le poids des sacs a évolué entre les dates t_1 et t_2.

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. **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. Les proportions p_1 et p_2 ne sont pas connues.

Par contre, dans la première classe de $n_1 = 23 + 17 = 40$ élèves, le proportion d'élèves reçus est $f_1 = \ \ldots$

- Soit F_1 la variable qui, à chaque échantillon de $n_1 = 40$ élèves de la population C_1, associe sa proportion de reçus.

Les trois conditions : $\begin{cases} n_1 = \\ n_1 f_1 = \\ n_1(1-f_1) = \end{cases}$ sont vérifiées donc : $F_1 \hookrightarrow \mathcal{N}(p_1; \sigma_1)$

Une estimation ponctuelle de la fréquence et de l'écart-type pour la population C_1 est :

$f_1 =$, et $\sigma_1 =$ Donc, F_1 suit la loi $\mathcal{N}\,(p_1;\ .....\,)$.

- Soit F_2 la variable qui, à chaque échantillon de $n_2 = 15 + 25 = 40$ élèves de la population C_2, associe sa fréquence de succès.

Les trois conditions : $\begin{cases} n_2 = \\ n_2 f_2 = \\ n_2(1-f_2) = \end{cases}$ sont vérifiées donc : $F_2 \hookrightarrow \mathcal{N}(p_2; \sigma_2)$

Une estimation ponctuelle de la fréquence et de l'écart-type pour la population C_2 est :

$f_2 =$, et $\sigma_2 =$ Donc, F_2 suit la loi $\mathcal{N}\,(p_2;\ .....\,)$.

- La variable aléatoire $D = F_1 - F_2$ suit également une loi normale de paramètres :
  - $E(D) = E(F_1) - E(F_2) = \dots\dots\dots$
  - $V(D) = V(M_1) + V(M_2) = \dots\dots\dots\dots\dots\dots\dots\dots$ d'où $\sigma_D = \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots$

  Donc, **la variable aléatoire de décision** D suit une loi $\mathcal{N}(\dots\dots\dots ; \dots\dots)$

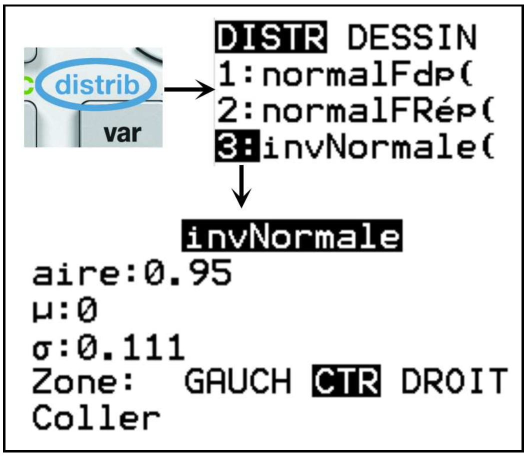
2. **Choix des hypothèses.**

   L'hypothèse nulle H_0 est $\dots\dots\dots$ (les deux populations ont le même niveau) ;

   l'hypothèse alternative H_1 est $\dots\dots\dots$ (les deux populations n'ont pas le même niveau).

3. **Zone critique.**

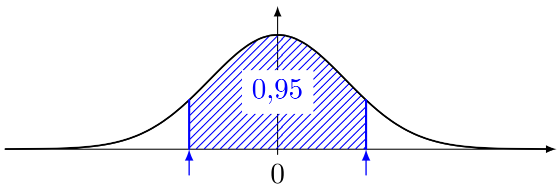
   Supposons que l'hypothèse H_0 soit vraie, on a alors $p_1 - p_2 = 0$ et D suit la loi normale $\mathcal{N}(0; 0{,}111)$.
   On cherche alors, le réel d tel que $P(-d \leq D \leq d) = \dots$ car le risque imposé par l'énoncé est de 5%.



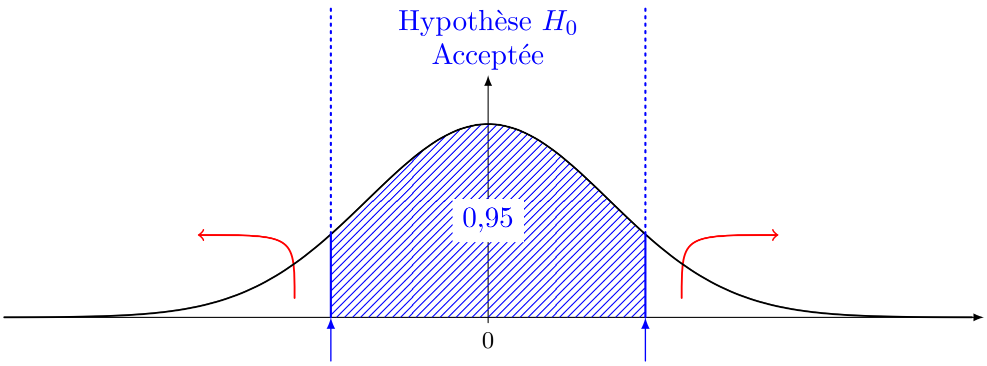
<u>**La calculatrice affiche un intervalle :**</u>

$[\dots\dots\dots\dots\dots ; \dots\dots\dots\dots\dots]$

Soit $d \simeq \dots$



Pour un seuil de risque de 5%, la zone critique est : $]-\infty ; \dots\dots [\cup] \dots\dots ; +\infty[$.



4. **Règle de décision.**

   ☞ Si $p_1 - p_2 \dots [-0{,}218 ; 0{,}218]$, alors l'hypothèse H_0 est validée ;

   ☞ sinon, l'hypothèse H_0, n'est pas validée.

5. **Conclusion.**

   La différence des moyennes des deux échantillons est $p_1 - p_2 =$

   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).

II) Exercices divers

Exercice C3 ***Délais de livraison***
Le service clientèle d'une entreprise de vente par correspondance sur internet cherche à améliorer les délais de livraison des commandes passées par les clients ; en 2016, les statistiques portant sur plusieurs dizaines de milliers de commandes ont montré que le délai moyen entre la commande passée par le client set la réception de la commande par le client était de 14,5 jours avec un écart-type de 8,5 jours.
Un échantillon de 200 commandes traitées au cours de l'année 2017 a permis de mesurer les délais suivants :

Délai en jours	*3*	*4*	*6*	*8*	*10*	*12*	*15*	*20*	*25*	*30*	*35*
Effectifs	*5*	*12*	*15*	*25*	*30*	*34*	*35*	*17*	*14*	*9*	*4*

On notera $\overline{X}$ la moyenne des délais de livraison observés dans des échantillons de taille 200 et on notera μ la moyenne inconnue des délais de livraison en 2017.

1. *Quelle est la loi de la variable d'échantillonnage $\overline{X}$?*
2. *A partir de l'échantillon observé, donner un intervalle de confiance à 95% pour la moyenne des délais de livraison au cours de l'année 2017.*
3. *On se propose de décider, au seuil de risque de 5%, entre l'hypothèse H_0 selon laquelle le délai de livraison est resté égal à 14,5 jours en moyenne, ou au contraire est devenu significativement différent de cette valeur au cours de l'année 2016. Dans toute la suite de l'énoncé, les calculs seront faits en supposant un écart-type connu $\sigma = 8{,}5$.*
   (a) *Préciser d'abord les deux hypothèses alternatives H_0 et H_1.*
   (b) *Par quelle loi est-il possible d'approcher la loi de $\overline{X}$ si H_0 est vraie ?*
   (c) *Déterminer l'intervalle d'acceptation et donner la conclusion du test.*
4. *Peut-on affirmer que le délai de livraison en 2017 est significativement inférieur à 14,5 jours, avec un risque inférieur à 5% de se tromper ? (Nouveau test à mettre en place)*
5. *Faire une figure*

Exercice C4 ***Distance moyenne parcourue en voiture***
La distance moyenne annuelle parcourue en voiture par les propriétaires de voitures à essence suit une loi d'espérance inconnue μ et d'écart-type inconnu σ. On pourra considérer qu'il y a plusieurs millions de propriétaires de véhicules à essence concernés.

Une enquête réalisée auprès de 16 propriétaires de véhicules à essence a permis d'obtenir les distances parcourues en 2017 ; ces distances sont reportées dans le tableau suivant :

17248	*15544*	*21202*	*10540*	*18082*	*19005*	*9906*	*10132*
18654	*16610*	*12250*	*17537*	*12945*	*15242*	*14799*	*10416*

On notera $\overline{X}$ la variable aléatoire qui à échantillon de taille 16 propriétaires associe la distance moyenne annuelle observées sur cet échantillon.

1. *A l'aide de la calculatrice, déterminer la moyenne $\overline{x}$ des distances annuelles parcourues en 2017, dans cet échantillon.*
2. *Préciser l'écart-type σ_e (à 10^{-2} près) des distances parcourues, dans cet échantillon.*
3. *Donner une estimation ponctuelle de la moyenne inconnue μ.*
4. *Donner une estimation ponctuelle σ^* de l'écart-type inconnu σ pour la population des distances annuelles parcourues par les propriétaires de véhicules à essence (arrondir au km le plus proche).*
5. *Détermination d'un intervalle de confiance de μ :*
   (a) *Les conditions sont-elles réunies pour pouvoir affirmer que la variable d'échantillonnage $\overline{X}$ suit une loi normale ? Pourquoi ?*
   (b) *On admettra que la distance annuelle parcourue en voiture par les propriétaires de voitures à essence suit une loi normale. Quelle est son espérance ? Quel est son écart-type ?*
   (c) *Donner un intervalle de confiance à 95% pour la moyenne inconnue μ des distances annuelles.*

Dans la suite de cet exercice, on continuera à admettre que la distance annuelle parcourue en voiture par les propriétaires de voitures à essence suit une loi normale, et on admettra aussi que l'écart-type σ des distances annuelles parcourues par l'ensemble de la population des propriétaires des véhicules à essence est de 3600 km.

6. *On se propose maintenant de tester si la distance moyenne parcourue par la population des propriétaires est ou non significativement différente de 13500 km.*
   (a) *Définir les hypothèses de ce test.*
   (b) *Par quelle loi est-il possible d'approcher la loi de $\overline{X}$ si H_0 est vraie ?*
   (c) *Déterminer l'intervalle d'acceptation et donner la conclusion du test.*
7. *On se propose maintenant de tester si la distance moyenne parcourue est ou non significativement supérieure à 13500 km.*
   (a) *Définir les hypothèses de ce test.*
   (b) *Calculer l'intervalle d'acceptation et donner la conclusion du test.*
8. *On se propose maintenant de tester si la distance moyenne parcourue est ou non significativement supérieure à 13500 km au risque de 2,5%.*
   (a) *Définir les hypothèses de ce test.*
   (b) *Calculer l'intervalle d'acceptation et donner la conclusion du test.*
   (c) *Comparer vos résultats à ceux de la question 6. Que concluez-vous ?*

Exercice C5 ***Les tâches ménagères, le retour***

L'enquête sur les taches ménagères réalisée par les étudiants d'un projet tuteuré a donné les résultats ci-contre à la question :« Pensez-vous qu'il existe des tâches ménagères spécifiques féminines ou masculines ? »

Oui	*49*	*38%*
Non	*80*	*62%*
Total	*129*	*100%*

On se demande si, au seuil de risque de 5%,

1. *La proportion de « Oui » dans l'ensemble de la population est significativement différente de 50% ?*

   Pour ce faire, on note $\overline{F}$ la variable aléatoire qui à un échantillon de 129 individus associe le proportion de « Oui », et p la proportion de « Oui » dans la population.

(a) *Définir les hypothèses de ce test.*

(b) *Si H_0 est vraie, quelle est la loi de $\overline{F}$?*

(c) *Calculer l'intervalle d'acceptation et donner la conclusion du test.*

2. *La proportion des « Non » est supérieure à 55% ?*

   Pour ce faire, on note $\overline{G}$ la variable aléatoire qui à un échantillon de 129 individus associe le proportion de « Non », et q la proportion de « Non » dans la population.

   (a) *Définir les hypothèses de ce test.*

   (b) *Si H_0 est vraie, quelle est la loi de $\overline{G}$?*

   (c) *Calculer l'intervalle d'acceptation et donner la conclusion du test.*

3. *Peut-on affirmer que la proportion des « Oui » est significativement inférieure à 40% avec un seuil de risque inférieur à 5% ?*

   (a) *Définir les hypothèses de ce test.*

   (b) *Si H_0 est vraie, quelle est la loi de variable aléatoire d'échantillonnage ?*

   (c) *Calculer l'intervalle d'acceptation et donner la conclusion du test.*

Exercice C6 ***Le test de l'association de consommateurs***

Pour l'un de ses numéros spéciaux, une association de consommateurs a testé la qualité des piles électriques de type LR6 utilisées dans un grand nombre de jouets ou d'appareils électroniques.
On s'intéresse ici à deux marques de piles, A et B. Les résultats des tests d'essais effectués sur deux échantillons sont donnés dans le tableau ci-dessous.

Marque des piles	*Effectif de l'échantillon*	*Nombre de piles d'une durée de vie supérieure à 100h*
A	*40*	*35*
B	*35*	*28*

On note p_A (respectivement p_B) la proportion inconnue des piles de la marque A (respectivement B) dont la durée de vie est supérieure à 100h.
Ces deux échantillons de piles de marques différentes peuvent être considérés comme indépendants l'un de l'autre.

Première partie : étude de la marque A

1. *On se propose de vérifier s'il est possible d'affirmer que la proportion p_A est significativement supérieure à 80%, au seuil de risque 5%, à partir de l'échantillon des 40 piles de la marque A.*

   (a) *Préciser les deux hypothèses alternatives H_0 et H_1.*

   (b) *Si H_0 est vraie, par quelle loi, est-il possible d'approcher la loi de la proportion F dans un échantillon ?*

   (c) *Calculer l'intervalle d'acceptation et donner la conclusion du test.*

2. *Faire la représentation graphique de ce test.*

Deuxième partie : étude des marques A et B

1. *On note f_A (resp. f_B) la proportion de piles d'une durée de vie supérieure à 100h dans l'échantillon de marque A (resp. B). Calculer f_A et f_B.*

2. *On note X la variable aléatoire qui, à un échantillon de 40 piles de marques A, associe la proportion de celles qui ont une durée de vie supérieure à 100h. Démontrer qu'on peut approcher la loi de X par la loi normale $\mathcal{N}(p_A\,;\,0{,}0530)$.*

3. *On note* Y *la variable aléatoire qui, à un échantillon de 35 piles de marques B, associe la proportion de celles qui ont une durée de vie supérieure à 100h. Démontrer qu'on peut approcher la loi de* Y *par la loi normale* $\mathcal{N}(p_B\,;\,0{,}0686)$.

4. *Pourquoi les variables aléatoires* X *et* Y *sont-elles indépendantes ?*

5. *Détermine la loi de la variable aléatoire* $D = A - B$.

6. *Interprète la variable aléatoire* D.

7. *Démontrer que si l'on fait l'hypothèse que les proportions* p_A *et* p_B *de piles des deux marques ayant une durée de vie supérieure à 100h sont les mêmes, alors* D *peut-être approcher la loi de* $\mathcal{N}(0\,;\,0{,}087)$.

8. *Les proportions* p_A *et* p_B *sont-elles significativement différentes, au seuil de risque 5%.*