Cours de Tests paramétriques

F. Muri-Majoube et P. Cénac

2006-2007

Licence

Article 6

L'utilisateur doit respecter les droits moraux de l'auteur : aussi bien la paternité de l'auteur que l'intégrité de l'oeuvre. Ainsi, le nom de l'auteur doit toujours être clairement indiqué, même en cas de reproduction ou représentation partielle. En cas de reproduction ou représentation partielle, il devra être précisé qu'il ne s'agit que d'un extrait, et référence devra être faite à l'oeuvre intégrale. Aucune modification du document n'est autorisée, à l'exception des modifications apportées par l'auteur lui-même.

Article 7

Aucun apport ou actualisation n'est autorisé pour ce document. L'utilisateur qui désirerait apporter une précision ou actualiser le document, n'est pas autorisé à le faire. Il peut toutefois soumettre sa proposition à l'auteur, via l'association, afin d'obtenir une autorisation. Ainsi, cette licence n'ouvre aucun droit à contribution. De même, toute traduction est interdite, ou soumise à l'accord de l'auteur.

Article 8

Toute infraction à la présente licence pourrait constituer une atteinte aux droits de l'auteur sur son oeuvre. Dans tous les cas, le non-respect de la licence sera susceptible de fonder une action en justice.

Article 9

En cas de litige, la loi française sera la seule applicable.

Table des matières

Chapitre 1: Démarche

1.1 Objectifs du cours

Un test statistique est un procédé d'inférence : son but est d'énoncer des propriétés de la population en s'appuyant sur un échantillon d'observations. A l'aide d'un test, on construit aussi des intervalles de confiance qui expriment le degré d'incertitude associé à une simulation. L'objectif du test est de répondre à des problèmes décisionnels dans un environnement incertain. Par exemple, on peut se demander à partir d'un échantillon si la taille des hommes est différente de la taille des femmes, si un nouveau médicament est vraiment plus efficace ou encore si l'incidence du cancer du poumon est plus élevée actuellement qu'il y a vingt ans. On ne dispose en pratique que d'un échantillon soumis aux fluctuations pour répondre à ces questions. L'objet des tests statistiques est de distinguer ce qui est plausible de ce qui est trop peu vraisemblable.

Ce cours introduit les concepts nécessaires pour développer et appliquer les tests paramétriques.

1.2 Exemple introductif

Présentation du problème

On suppose que la teneur en cacao (en grammes par kg) des tablettes d'un certain fabricant est une v.a. X de loi $\mathcal{N}(600, 75^2)$. Le fabricant déclare avoir trouvé un nouveau procédé de conception qui assure une qualité supérieure du chocolat en garantissant une plus grande teneur en cacao, c'est-à-dire que la teneur en cacao de ses tablettes est une variable aléatoire de loi $\mathcal{N}(\mu, 75^2)$, avec $\mu > 600$, disons $\mu = 650$ pour fixer les idées.

On effectue un contrôle de qualité sur $n = 9$ tablettes fabriquées avec le nouveau procédé, et les teneurs en cacao (en grammes par kg) sont les suivantes :

$$580\ ;\ 614\ ;\ 700\ ;\ 590\ ;\ 660\ ;\ 630\ ;\ 640\ ;\ 645\ ;\ 620$$

Ces données, notées $x_1, \ldots, x_n$, sont les réalisations de n variables aléatoires indépendantes $(X_1, \ldots, X_n)$, de même loi $\mathcal{N}(\mu, 75^2)$, où X_i représente la teneur en cacao de la i^{eme} tablette testée $(1 \leq i \leq n)$. L'indépendance implique l'absence d'influence d'une variable sur les autres et le fait que les X_i aient la même loi implique en particulier que

$$\begin{aligned} \mathbb{E}(X_1) = \mathbb{E}(X_2) = \cdots = \mathbb{E}(X_n) &= \mu \\ \mathrm{Var}(X_1) = \mathrm{Var}(X_2) = \cdots = \mathrm{Var}(X_n) &= \sigma^2 = (75)^2 \end{aligned}$$

On veut savoir si le fabricant dit la vérité, c'est-à-dire déterminer laquelle des deux hypothèses est la bonne entre

▷ $\mu = 600$, le nouveau procédé n'est pas meilleur

▷ $\mu = 650$, le nouveau procédé est meilleur comme le prétend le fabricant.

Pour cela, on va effectuer un *test statistique*. : on va tester l'hypothèse

$$H_0 \ : \ \mu = 600 \quad \text{contre} \quad H_1 \ : \ \mu = 650.$$

Principe du test

On se place du *point de vue du consommateur* qui hésite à opter pour le nouveau procédé qui coûte plus cher. L'expérience doit être convaincante : *il n'abandonnera H_0 qu'en présence de faits expérimentaux contredisant nettement la validité de H_0*. On va établir une procédure qui va permettre de conclure quant à l'acceptation ou le rejet de H_0 avec un risque fixé à l'avance.

Comment décider ?

La question porte sur l'espérance μ de la variable. Pour prendre en compte les 9 expériences, il est logique de raisonner à partir d'un *estimateur* de μ. On pense de façon naturelle à l'estimateur empirique $\bar{X} = \dfrac{1}{n}\sum_{i=1}^{9} X_i$, dont on connaît la loi sous H_0 (et sous H_1). On rappelle que $\bar{X}$ est un estimateur sans biais de μ, $\boxed{\mathbb{E}(\bar{X}) = \mu}$ (il vise bien en moyenne), et que $\mathrm{Var}(\bar{X}) = \dfrac{\sigma^2}{n} = 25^2$. Ceci découle des propriétés suivantes.

Rappel 1 : Soient Y et Z deux v.a. et soient a et b deux réels, alors

$$\mathbb{E}\left(aY + bZ\right) = a\mathbb{E}(Y) + b\mathbb{E}(Z) \quad \text{et} \quad \mathrm{Var}\left(aY + b\right) = a^2\,\mathrm{Var}(Y).$$

$$\text{Si de plus, } Y \text{ et } Z \text{ sont indépendantes,} \quad \mathrm{Var}\left(Y + Z\right) = \mathrm{Var}(Y) + \mathrm{Var}(Z).$$

Vous avez vu (cf. cours d'estimation) que d'une part $\bar{X}$ converge en moyenne quadratique vers μ et d'autre part que $\bar{X}$ converge en probabilité vers μ. L'estimateur $\bar{X}$ est donc un estimateur consistant.

De plus, $\boxed{\bar{X} \sim \mathcal{N}(\mu, 25^2)}$ car les X_i sont indépendants et de même loi $\mathcal{N}(\mu, 75^2)$. En effet,

Rappel 2 : Soient Y et Z deux v.a. gaussiennes indépendantes telles que $Y \sim \mathcal{N}(\mu_1, \sigma_1^2)$ et $Z \sim \mathcal{N}(\mu_2, \sigma_2^2)$, alors

$$aY + bZ \sim \mathcal{N}\left(a\mu_1 + b\mu_2, a^2\sigma_1^2 + b^2\sigma_2^2\right).$$

Autrement dit, la somme de deux v.a. gaussiennes indépendantes est gaussienne, ce qui s'étend aux n v.a. $X_1, X_2, \ldots, X_n$, qui sont iid, avec $\mathbb{E}(X_i) = \mu$ et $\mathrm{Var}(X_i) = \sigma^2$.

Rappel 3 : Soit $X_1, \ldots, X_n$ un n-échantillon de la loi $\mathcal{N}(\mu, \sigma^2)$, alors

$$\sum_{i=1}^{n} X_i \sim \mathcal{N}\left(n\mu, n\sigma^2\right) \ ,$$

soit encore

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right).$$

Ainsi, sous H_0, $\bar{X} \sim \mathcal{N}\left(600, 25^2\right)$, et sous H_1, $\bar{X} \sim \mathcal{N}(650, 25^2)$. On notera

$$\bar{X} \sim_{H_0} \mathcal{N}(600, 25^2) \quad \text{et} \quad \bar{X} \sim_{H_1} \mathcal{N}(650, 25).$$

▷ Si $\bar{X}$ est *"petit"* , on choisira H_0.
▷ Si $\bar{X}$ est *"grand"*, on rejettera H_0, pour choisir H_1.

Le problème est de savoir à partir de quelle valeur on choisit H_1 plutôt que H_0, *i.e.* de déterminer *la zone de rejet et la zone d'acceptation de* H_0, *i.e.* de déterminer le réel c_α tel que
▷ si $\bar{X} \leq c_\alpha$, on accepte H_0,
▷ si $\bar{X} > c_\alpha$, on rejette H_0.

Erreurs commises

Lors de la prise de décision, on peut commettre deux erreurs :

Décision / Réalité	H_0	H_1
H_0	Décision correcte proba $1-\alpha$	Erreur de première espèce proba α
H_1	Erreur de seconde espèce proba β	Décision correcte proba $1-\beta$

- *L'erreur de première espèce* est l'erreur que l'on commet lorsqu'*on rejette H_0 à tort*, c'est-à-dire lorsqu'on choisit H_1 alors que H_0 est vraie.
  La probabilité de commettre cette erreur, que l'on appelle *le risque de première espèce*, est notée $\mathbb{P}(\text{rejeter } H_0 \text{ à tort}) = \mathbb{P}_{H_0}(\text{rejeter } H_0)$. Pour *un niveau de test* α, on impose que

  $$\mathbb{P}_{H_0}(\text{rejeter } H_0) \leq \alpha.$$

  Pour notre exemple, l'erreur de première espèce est

  $$\alpha = \mathbb{P}(\text{rejeter } H_0 \text{ à tort}) = \mathbb{P}_{H_0}(\text{rejeter } H_0) = \mathbb{P}_{H_0}(\bar{X} > c_\alpha).$$

- *L'erreur de deuxième espèce* est l'erreur que l'on commet lorsqu'*on accepte H_0 à tort*, c'est-à-dire lorsqu'on ne rejette pas H_0 alors qu'elle est fausse ; la probabilité de commettre cette erreur, que l'on appelle *le risque de deuxième espèce*, est notée erreur, que l'on appelle le risque de deuxième espèce, est notée β avec

  $$\beta = \mathbb{P}(\text{accepter } H_0 \text{ à tort}) = \mathbb{P}_{H_1}(\text{accepter } H_0).$$

  Dans notre exemple, l'erreur de deuxième espèce est

  $$\beta = \mathbb{P}_{H_1}(\bar{X} \leq c_\alpha).$$

On introduit une dissymétrie dans ces erreurs. On considère que l'erreur de première espèce est plus grave. On construit le test à partir de cette erreur. On fixe tout d'abord le risque de première espèce α, puis on cherche à minimiser celui de deuxième espèce, qui servira à juger de la qualité du test construit.

On commence par choisir c_α pour que la probabilité de rejeter H_0 à tort, *i.e.* de se tromper en acceptant H_1, soit petite, par exemple égale à $0,05$. Autrement dit, on fixe c_α pour qu'il y ait 5 chances sur 100 que, si H_0 est vraie, l'échantillon ne donne pas une valeur de $\bar{X}$ comprise dans la zone d'acceptation de H_0. H_0 est donc *privilégiée* : on veut avant tout contrôler le risque de rejeter H_0 à tort. On est donc prêt à rejeter H_0 si le résultat fait partie d'une éventualité improbable n'ayant que 5% de chances de se produire.

On cherche donc c_α tq

$$\mathbb{P}_{H_0}(\text{rejeter } H_0) = \mathbb{P}_{H_0}(\bar{X} > c_\alpha) = \alpha = 0,05$$

(ce seuil c_α n'a que 5% de chances d'être dépassé). Or pour notre exemple, on a vu que sous H_0, $\bar{X} \sim \mathcal{N}(600, 25^2)$ et $\dfrac{\bar{X} - 600}{25} \sim \mathcal{N}(0,1)$. Ainsi,

$$\mathbb{P}_{H_0}(\bar{X} > c_\alpha) = \mathbb{P}_{H_0}\left(\frac{\bar{X} - 600}{25} > \frac{c_\alpha - 600}{25}\right) = \mathbb{P}\left(Y > \frac{c_\alpha - 600}{25}\right),$$

où $Y \sim \mathcal{N}(0,1)$. Dans la table de la loi $\mathcal{N}(0,1)$, on lit la valeur $\dfrac{c_\alpha - 600}{25}$ telle que

$$\mathbb{P}_{H_0}\left(Y > \frac{c_\alpha - 600}{25}\right) = 0,05$$

On trouve $\dfrac{c_\alpha - 600}{25} = 1,645$, soit encore $c_\alpha = 600 + 25 * 1,645 = 641,125$.

On en déduit la règle de décision suivante :

▷ Si $\bar{X} \leq 641,125$, on accepte H_0 ;

▷ Si $\bar{X} > 641,125$, on rejette H_0 pour accepter H_1 (avec une proba de 5% de se tromper).

L'ensemble $\{\bar{X} > 641,125\}$ s'appelle *la zone ou région de rejet de H_0 au niveau* 5%, que l'on notera $ZR_{5\%}$, et $\{\bar{X} \leq 641,125\}$ la zone d'acceptation de H_0, notée $ZA_{5\%}$, au risque 5%.

Mise en oeuvre : on observe que $\bar{x} = 631 < 641,125$, donc on conserve H_0 : $\mu = 600$ (au risque 5%).

Attention : accepter H_0 *ne signifie pas que cette hypothèse est vraie* mais seulement que les données ne sont pas incompatibles avec H_0, et que l'on n'a pas de raison suffisante de lui préférer H_1 compte-tenu des résultats de l'échantillon. Ce n'est pas parce qu'une expérience ne conduit pas au rejet de H_0, que H_0 est vraie (d'autres expériences pourraient la rejeter).

On a choisi c_α pour que la probabilité de se tromper en rejetant H_0 soit $\leq 0,05$. Quelle est la probabilité de se tromper en acceptant H_0 ? Sous H_1, on a vu que $\bar{X} \sim \mathcal{N}(650, 25^2)$ et donc $\dfrac{\bar{X} - 650}{25} \sim \mathcal{N}(0,1)$. On en déduit

$$\begin{aligned}
\mathbb{P}_{H1}(\text{accepter } H_0) &= \mathbb{P}_{H1}(\bar{X} \leq 641,125) \\
&= \mathbb{P}_{H1}\left(\frac{\bar{X} - 650}{25} \leq -0,12\right) \\
&= \mathbb{P}_{H1}\left(\frac{\bar{X} - 650}{25} \geq 0,12\right) \\
&= 1 - \mathbb{P}_{H1}\left(\frac{\bar{X} - 650}{25} < 0,12\right) \\
&= 1 - 0,5478 = 0,4522
\end{aligned}$$

ce qui est considérable. La forme de ZR_α est indiquée par la nature de H_1 mais le seuil c_α ne dépend que de H_0 et de α, la deuxième erreur β n'intervient pas. H_0 est privilégiée : on veut avant tout contrôler le risque de rejeter H_0 à tort. On voit donc que les 2 hypothèses H_0 et H_1 ne jouent pas le même rôle.

- La probabilité de rejeter H_0 à tort est choisie au départ

  $$\mathbb{P}_{H_0}(\text{rejeter } H_0) = \alpha \text{ (parfois on a seulement une inégalité } \leq \alpha)$$

  où α est le niveau du test.
- La probabilité d'accepter H_0 à tort

  $$\beta = \mathbb{P}_{H_1}(\text{accepter } H_0)$$

  peut donc être relativement grande. On va chercher à minimiser cette erreur β, ce qui revient à maximiser la *puissance du test*

  $$\pi = P(\text{rejeter } H_0 \text{ quand elle est fausse}) = 1 - \beta = \mathbb{P}_{H_1}(\text{rejeter } H_0).$$

La puissance mesure l'aptitude du test à rejeter une hypothèse fausse. Dans notre exemple, la puissance est

$$1 - \beta = \mathbb{P}_{H_1}(\bar{X} > 641,125) = 1 - 0,4522 = 0,5478$$

La puissance du test est d'autant plus grande que la durée de vie des ampoules est importante (on s'éloigne de H_0). Le test de niveau 5% a 55 chances sur 100 de détecter une durée de vie de 650 (ou plus).

Que se passe-t'il quand le niveau α du test varie ?

- Si α grandit, $\mathbb{P}_{H_0}$(rejeter H_0) grandit, on tolère donc une plus grosse probabilité d'erreur en acceptant H_1 : on accepte plus facilement H_1 ; ainsi β diminue, et la puissance augmente.
- Si α diminue (décroît vers 0), $1 - \alpha$ augmente, et la règle de décision est plus stricte : on n'abandonne H_0 que dans des cas rarissimes, et on conserve H_0 bien souvent à tort. A force de ne pas vouloir abandonner H_0, "on la garde presque tout le temps", ainsi β augmente, et la puissance diminue.

- "L'idéal" serait d'avoir un niveau petit, et une grande puissance. Un moyen d'y parvenir est d'augmenter le nombre n d'observations ($n \to \infty$).

Nombre d'observations nécessaires pour obtenir, à un niveau fixé, une puissance donnée

Pour le niveau $\alpha = 5\%$ fixé, calculons le nombre d'observations nécessaires pour avoir une puissance de 80%. On a

$$\begin{aligned} \alpha &= 0,05 = \mathbb{P}_{H_0}(\bar{X} \leq c_\alpha) \\ 1-\beta &= 0,80 = \mathbb{P}_{H_1}(\bar{X} > c_\alpha) \end{aligned}$$

On a vu que sous H_0, $\bar{X} \sim \mathcal{N}\left(600, \frac{75^2}{n}\right)$ et sous H_1, $\bar{X} \sim \mathcal{N}\left(650, \frac{75^2}{n}\right)$. On en déduit,

$$\begin{aligned} \alpha &= 0,05 = \mathbb{P}_{H_0}\left(\frac{\sqrt{n}(\bar{X}-600)}{75} > \frac{\sqrt{n}(c_\alpha - 600)}{75}\right) \\ \alpha &= 0,05 = \mathbb{P}_{H_0}\left(Z > \sqrt{n}\frac{(c_\alpha - 600)}{75}\right) \end{aligned}$$

où $Z \sim \mathcal{N}(0,1)$. On lit dans la table de la loi normale $\mathcal{N}(0,1)$, $\frac{\sqrt{n}(c_\alpha - 40)}{75} = 1,645$, soit $\boxed{c_\alpha = 1,645\frac{75}{\sqrt{n}} + 600}$. De plus, on a aussi

$$\begin{aligned} 1-\beta &= 0,80 = \mathbb{P}_{H_1}\left(\frac{\sqrt{n}(\bar{X}-650)}{75} > \frac{\sqrt{n}(c_\alpha - 650)}{75}\right) \\ 1-\beta &= 0,80 = \mathbb{P}_{H_1}\left(Z > \sqrt{n}\frac{(c_\alpha - 650)}{75}\right). \end{aligned}$$

On lit dans la table de la loi normale $\mathcal{N}(0,1)$ que $\frac{\sqrt{n}(c_\alpha - 650)}{75} = -0,84$. En reportant la valeur de c_α du premier encadré, dans cette dernière équation, on trouve

$$\begin{aligned} -0,84 &= \sqrt{n}\frac{\left(1,645\frac{75}{\sqrt{n}} + 600 - 650\right)}{75} \\ &= 1,645 + \frac{\sqrt{n}(600-650)}{75} \\ \text{soit } \sqrt{n} &= (1,645 + 0,84)\frac{75}{(600-650)} \\ n &= (1,645+0,84)^2\frac{5625}{(600-650)^2} = 13,89 \end{aligned}$$

Il faudrait au moins $n = 14$ observations pour avoir, au niveau 5% une puissance de 80%.

Remarque (Lien avec les intervalles de confiance)**.** On a choisi c_α tel que $\mathbb{P}_{H_0}(\bar{X} > c_\alpha) = 0,05$, c'est-à-dire tel que $\mathbb{P}_{H_0}(\bar{X} \leq c_\alpha) = 0,95$, soit encore $\mathbb{P}_{H_0}\big(\bar{X} \in]-\infty, c_\alpha]\big) = 0,95$. L'intervalle $I_{5\%} =]-\infty, c_\alpha]$ est un intervalle de confiance non symétrique au risque 5% de l'estimateur $\bar{X}$,

sous l'hypothèse H_0 ($\mu \leq 600$). En effet, on a bien $\mathbb{P}_{H_0}(\bar{X} \in I_{5\%}) = 95\%$; autrement dit, sous H_0, on prend un risque de 5%, lorsqu'on parie, avant d'observer une réalisation de $\bar{X}$, que cette réalisation va tomber dans l'intervalle $I_{5\%}$. On a bien sûr $\mathbb{P}_{H_0}(\bar{X} \notin I_{5\%}) = 5\%$. Ainsi, on peut formuler la règle de décision suivante :

▷ Si $\bar{X} \notin I_{5\%}$, on rejette H_0,

▷ Si $\bar{X} \in I_{5\%}$, on accepte H_0.

Point de vue du fabricant

On se place cette fois du point de vue du fabricant qui hésite à opter pour H_1 car il ne gagnera de l'argent que si son procédé est utilisé. On reprend donc le même exemple, mais en testant cette fois

$$H_0 \ : \ \mu = 650 \quad \text{contre } H_1 \ : \ \mu = 600.$$

A nouveau le test est basé sur un *"bon"* estimateur de μ, l'estimateur empirique $\bar{X}$. Cette fois, sous H_0, $\bar{X} \sim \mathcal{N}(650, 25^2)$, et sous H_1 on a $\bar{X} \sim \mathcal{N}(600, 25^2)$. Contrairement au test précédent,

▷ Si $\bar{X}$ est "petit", on choisira H_1,

▷ Si $\bar{X}$ est "grand", on choisira H_0.

On en déduit la règle de décision,

▷ Si $\bar{X} < c_\alpha$, on rejette H_0,

▷ Si $\bar{X} \geq c_\alpha$, on accepte H_0,

où c_α est tel que $\mathbb{P}_{H_0}(\text{rejetter } H_0) = \mathbb{P}_{H_0}(\bar{X} < c_\alpha) = \alpha$ fixé.

Test au niveau 5%

- détermination de c_α : On cherche c_α tel que $\mathbb{P}_{H_0}(\bar{X} < c_\alpha) = 0,05$. Sous H_0, $\bar{X} \sim \mathcal{N}(650, 25^2)$, et donc $\frac{\bar{X}-650}{25} \sim \mathcal{N}(0,1)$. On en déduit

$$\mathbb{P}_{H_0}(\bar{X} < c_\alpha) = \mathbb{P}_{H_0}\left(\frac{\bar{X}-650}{25} < \underbrace{\frac{c_\alpha - 650}{25}}_{\leq 0}\right) = \mathbb{P}_{H_0}\left(\frac{\bar{X}-650}{25} > \frac{650-c_\alpha}{25}\right)$$

  (rappelons qu'une loi gaussienne est symétrique !) Ainsi on obtient

$$\begin{aligned}\mathbb{P}_{H_0}(\bar{X} < c_\alpha) = 0,05 &\iff 1 - \mathbb{P}_{H_0}\left(\frac{\bar{X}-650}{25} \leq \frac{650-c_\alpha}{25}\right) = 0,05 \\ &\iff \mathbb{P}_{H_0}\left(\frac{\bar{X}-650}{25} \leq \frac{650-c_\alpha}{25}\right) = 0,95\end{aligned}$$

  On lit sur la table de la loi $\mathcal{N}(0,1)$, $\dfrac{650-c_\alpha}{25} = 1,645$ et ainsi $c_\alpha = 608,875$.

- mise en oeuvre : $\bar{x} = 631 > 608,75$, donc on accepte H_0 ; cette fois, on conserve l'hypothèse $\mu = 650$.

Juste avec cet exemple, on voit que les hypothèses H_0 et H_1 ne jouent pas un rôle symétrique ; *il est donc très important de bien choisir les hypothèses H_0 et H_1.*

1.3 Cadre général

On s'intéresse à une population $\mathcal{P}$ au travers d'un caractère X (qualitatif ou quantitatif). Souvent la loi de X dépend d'un (ou plusieurs) paramètre inconnu(s). La statistique inférentielle permet de "transporter" les conclusions concernant les résultats d'un échantillon $(X_1, \cdots, X_n)$ de même loi que X à la population entière dont il est issu, *en quantifiant les erreurs que l'on risque de commettre*. Il existe essentiellement deux classes de méthodes :

- *la théorie de l'estimation* pour estimer un (ou plusieurs) paramètres par un nombre (estimation ponctuelle) ou par un intervalle (estimation par intervalle) ;
- *la théorie des tests* pour émettre et trancher entre deux hypothèses au vu des résultats de l'échantillon.

On dispose de n données $x_1, \cdots, x_n$ réalisations de n variables aléatoires $X_1, \cdots, X_n$. On va chercher à tester des hypothèses portant sur la loi de probabilité du n-uplet $(X_1, \cdots, X_n)$ modélisant les observations. On se placera essentiellement dans le cadre où les X_i sont indépendantes et identiquement distribuées (iid). On effectue un test de H_0 contre H_1, H_0 et H_1 étant deux hypothèses portant sur la loi de $X_1, \cdots X_n$.

La construction d'un test va consister à établir une règle de décision permettant de faire un choix entre les deux hypothèses H_0 et H_1 au vu d'un échantillon de même loi que X.

1.3.1 Description et nature des hypothèses

Définition 1. *L'hypothèse H_0 est appelée* hypothèse nulle, *et l'hypothèse H_1 est* hypothèse alternative.

Choix des hypothèses : H_0 est l'hypothèse privilégiée

L'hypothèse H_0 appelée *hypothèse nulle* est celle que l'on garde si le résultat de l'expérience n'est pas clair. On conserve H_0 sauf si les données conduisent à la rejeter. *Quand on accepte H_0, on ne prouve pas qu'elle est vraie* ; on accepte de conserver H_0 car on n'a pas pu accumulé suffisamment d'éléments matériels contre elle ; les données ne sont pas incompatibles avec H_0, et l'on n'a pas de raison suffisante de lui préférer H_1 compte-tenu des résultats de l'échantillon. Ne pas rejetter H_0, c'est "acquitter faute de preuve". On n'abandonne pas H_0 sans de solides raisons. L'hypothèse H_1 contre laquelle on teste H_0 est appelée *contre hypothèse* ou *hypothèse alternative*.

Analogie Test/Procès : tout suspect est présumé innocent et l'accusation doit apporter la preuve de sa culpabilité avant de le condamner. Cette preuve doit s'appuyer sur des éléments matériels comme le test qui utilise les données pour rejeter H_0.

Selon les cas, on choisit pour l'hypothèse nulle :

- *l'hypothèse qu'il est le plus grave de rejeter à tort.* On a vu que la probabilité de rejeter H_0 à tort est choisie aussi petite que l'on veut. On choisit donc H_0 et H_1 de telle sorte que le

scénario catastrophique soit d'accepter H_1 alors que H_0 est vraie : ce scénario "le pire" a ainsi une petite probabilité α (α est fixée) de se réaliser.

- *l'hypothèse dont on a admis jusqu'à présent la validité*, H_1 représentant la contre hypothèse suggérée par une nouvelle théorie ou une expérience récente.
- *l'hypothèse qui permet de faire le test* (seule hypothèse facile à formuler et sous laquelle on peut effectuer le calcul de la loi d'une variable aléatoire sur laquelle on peut fonder le test). Cette dernière raison, purement technique, est évidemment la moins bonne.

Exemples

1. On veut faire un test pour savoir si un pont va s'écrouler ou pas. Le scénario le pire est d'accepter que le pont ne s'écroulera pas alors qu'en fait il va s'écrouler. Ainsi, on testera H_0 : "le pont s'écroule" contre H_1 : "le pont ne s'écroule pas".
2. On veut faire un test pour savoir si un médicament a des effets secondaires dangereux ou pas. Le scénario le pire est d'accepter qu'il n'y a pas d'effets secondaires dangereux, alors qu'il y en a. On testera donc H_0 : "le médicament a des effets secondaires dangereux" contre H_1 : "le médicament n'a pas d'effets secondaires dangereux".

Remarque. *Le scénario catastrophique dépend souvent du point de vue considéré.*

Revenons à l'exemple introductif. Si le nouveau procédé coûte plus cher que l'ancien, le scénario catastrophique pour le consommateur est de décider $\mu = 650$ (le procédé le plus cher est utilisé) alors qu'en fait $\mu = 600$ (le plus cher n'est pas mieux que l'ancien). Donc, *du point de vue du consommateur*, on teste $H_0 : \mu = 600$ contre $H_1 : \mu = 650$. Mais si le fabricant qui a inventé le nouveau procédé de conception des tablettes gagne une grosse prime si son procédé est utilisé dans le commerce, le scénario catastrophique pour lui est que l'on accepte $\mu = 600$ (son procédé n'est pas utilisé) alors qu'en fait $\mu = 650$ (son procédé devrait être utilisé). Donc, *du point de vue du fabricant*, on teste $H_0 : \mu = 600$ contre $H_1 : \mu = 650$.

Comme en général, le test est réalisé par l'expérimentateur pour "prouver" que H_1 est vraie, le test est dit *significatif* lorsqu'il permet, une fois réalisé, de conclure au rejet de H_0. On "prouve" H_1 par le rejet de H_0.

Les différents types d'hypothèse et de test

Il existe plusieurs types d'hypothèses. Une hypothèse est dite *paramétrique* si elle porte sur la valeur d'un ou plusieurs paramètres, sinon elle est dite *non paramétrique*.

Une hypothèse est dite *simple* si elle spécifie complétement la loi de la variable aléatoire considérée, ou le paramètre considéré. Sinon elle est dite *multiple*.

1.3.2 Erreurs commises, qualité d'un test

Lors de la prise de la décision de rejetter ou non l'hypothèse H_0, on peut commettre deux erreurs. Les situations possibles lors de la prise de décision ont déjà été résumées dans le tableau de la section 1.2 de l'exemple introductif.

La stratégie consiste à fixer le niveau du test α et à imposer que le risque de première espèce soit inférieur à α. On est donc prêt à rejeter H_0 si le résultat fait partie d'une éventualité improbable. L'hypothèse H_0 est *privilégiée* : on veut avant tout contrôler le risque de rejeter H_0 à tort. Une fois le risque de première espèce controlé par α, on cherchera alors à minimiser le risque de deuxième espèce β, qui sera du coup une fonction de α.

Puissance

La qualité d'un test est donnée par sa capacité à séparer les deux hypothèses H_0 et H_1. Elle est mesurée par la puissance du test qui est la probabilité d'accepter H_0 quand H_1 est vraie. Cette puissance π est donc donnée par

$$\pi = \mathbb{P}(\text{rejeter } H_0 \text{ quand elle est fausse}) = 1 - \beta(\alpha) = \mathbb{P}_{H_1}(\text{rejeter } H_0).$$

La puissance mesure l'aptitude du test à rejeter une hypothèse fausse. C'est un mesure de *la qualité du test*. Pour un niveau α fixé, on va chercher à obtenir une grande puissance π.
Rien ne dit que conserver H_0 mette à l'abri de se tromper. La probabilité d'accepter H_0 à tort est β, et cette probabilité (qui ne se calcule que si l'on connaît la loi de T_n sous H_1) peut être très importante, contrairement à α qui est fixé aussi petit qu'on veut à l'avance. Les hypothèses H_0 et H_1 ne jouent donc pas un rôle symétrique.

1.4 Construction d'un test

A partir d'observations $x_1, \ldots, x_n$, qui sont les réalisations de n variables aléatoires $X_1, \ldots, X_n$, on va chercher à tester des hypothèses de modélisation portant sur la loi de $X_1, \ldots, X_n$. Les différentes étapes de la construction d'un test sont les suivantes.

- **Description de l'hypothèse nulle H_0 et de l'hypothèse alternative**. Souvent, la loi commune des X_i dépend d'un (ou plusieurs) paramètre θ, et H_0 est une assertion concernant θ. Dans l'exemple introductif, la loi des X_i est une loi normale dépendant de deux paramètres, espérance μ et variance σ^2. Nous avons fait un test portant sur le paramètre $\theta = \mu$.
- **Construction de la statistique de test**. Le test est basé sur l'utilisation d'une statistique de test, T_n, fonction des variables aléatoires $X_1, \cdots, X_n$, en général liée à un estimateur. La statistique de test T_n est une variable aléatoire dont *on connaît la loi sous H_0* et qui a *un comportement différent sous H_1*, que sous H_0 ($\bar{X}$ dans l'exemple précédent). La décision va porter sur la valeur prise par cette variable aléatoire .
- **Règle de décision et forme de la région de rejet**. La statistique de test T_n permet de construire la région de rejet de H_0. Intuitivement, on rejette H_0 quand T_n est *"loin"* de H_0, vers H_1. On détermine donc la règle de décision :

$$\begin{cases} \text{si } T \in ZR_\alpha & \text{alors on rejette } H_0, \\ \text{si } T \notin ZR_\alpha & \text{alors on accepte } H_0, \end{cases}$$

où ZR_α s'appelle la *région de rejet de H_0* du test de niveau α et doit être telle que

$$\mathbb{P}_{H_0}(T \in ZR_\alpha) = \mathbb{P}_{H_0}(\text{rejeter } H_0) = \mathbb{P}(\text{rejeter } H_0 \text{ à tort}) \leq \alpha,$$

où α est le niveau du test fixé à l'avance.
On peut également formuler la règle de décision en terme de zone d'acceptation :

$$\begin{cases} \text{si } T \in ZA_\alpha & \text{alors on accepte } H_0 \\ \text{si } T \notin ZA_\alpha & \text{alors on rejette } H_0 \end{cases}$$

ZA_α s'appelle la *région d'acceptation de H_0* du test de niveau α et doit être telle que

$$\mathbb{P}_{H_0}(T \in ZA_\alpha) = 1 - \alpha.$$

- **Mise en œuvre du test : calcul exact de la région de rejet et conclusions.** Soit t la réalisation de T,
  - ▷ Si $t \in ZR_\alpha$ alors on rejette H_0. En effet, si H_0 était vraie, il aurait été alors très peu probable d'obtenir les résultats que l'on a trouvés. *Les données sont en contradiction avec H_0*.
  - ▷ Si $t \notin ZR_\alpha$ alors on accepte H_0 : les donnés ne sont pas en contradiction avec H_0.

Remarque. Il est nécessaire de connaître la loi de la statistique de test T_n sous H_0 pour pouvoir calculer l'erreur de première espèce et ainsi construire la région de rejet. Il est également nécessaire que T_n ait un comportement différent sous H_0 et sous H_1 ; *mais* dans de nombreux cas, on ne connaît pas la loi de T_n sous H_1, ce qui n'empêche pas de construire le test.

Remarque (Lien entre tests et zones de confiance)**.** Construire une région de confiance I_α, au risque α, sous H_0, pour les réalisations de T_n, revient à construire un test de H_0 au niveau α. En effet, si I_α est une région de confiance de risque α pour les réalisations de T_n, on a

$$\mathbb{P}_{H_0}(T \in I_\alpha) = 1 - \alpha.$$

C'est-à-dire que l'on a

$$\mathbb{P}_{H_0}(T \notin I_\alpha) = 1 - \mathbb{P}_{H_0}(T \in I_\alpha) = \alpha.$$

On peut donc construire un test de règle de décision

$$\begin{cases} \text{si } T \notin I_\alpha & \text{alors on rejette } H_0, \\ \text{si } T \in I_\alpha & \text{alors on accepte } H_0. \end{cases}$$

Dans notre exemple, on a construit un test de

$$H_0 \ : \ \mu = 600 \quad \text{contre} \quad H_1 \ : \ \mu = 650.$$

On a montré que l'ensemble $\{\bar{X} > 641,125\}$ est une zone de rejet de H_0 au niveau 5%. L'intervalle $I_{5\%} =]-\infty; 641,125]$ est un intervalle de confiance non symétrique de $\bar{X}$ au risque 5%, sous l'hypothèse H_0. C'est-à-dire que sous H_0, on prend un risque de 5% lorsqu'on parie, avant d'observer une réalisation de $\bar{X}$, que cette réalisation va tomber dans $I_{5\%}$.

1.5 Degré de signification

Lorsque le test n'est pas significatif (c'est-à-dire qu'on ne rejette pas H_0, on peut se demander à quel niveau H_0 serait rejettée. Ce niveau est appelé *degré de signification* ou probabilité critique (*P-value*).

Soit t une réalisation de la statistique de test T. *Le degré de signification mesure la probabilité d'obtenir t sous l'hypothèse H_0.* C'est une mesure de l'accord entre l'hypothèse testée et le résultat obtenu. Plus il est proche de 0, plus forte est la contradiction entre H_0 et le résultat de l'échantillon, et plus on rejettera H_0 avec assurance.

Revenons à l'exemple introductif, où l'on testait $H_0 : \mu = 600$ contre $H_1 : \mu = 650$ au niveau $\alpha = 5\%$. La règle de décision était

▷ Si $\bar{X} > 641,125$, on rejette H_0 pour accepter H_1,

▷ Si $\bar{X} \leq 641,125$, on accepte H_0.

On a observé que $\bar{x} = 631 < 641,125$, donc on accepte $H_0 : \mu = 600$. Le test n'est donc *pas significatif à 5%*. On peut se demander s'il le serait par exemple à 10%. On peut aussi se demander dans quelle mesure la moyenne d'échantillon observée $\bar{x} = 631$ est compatible avec l'hypothèse $H_0 : \mu = 600$, c'est à dire, déterminer $\mathbb{P}_{H_0}(\bar{X} > 631)$ (probabilité que, sous H_0, la statistique de test $\bar{X}$ soit $>$ à la valeur réellement observée $\bar{x}$). C'est ce que l'on appelle le *degré de signification α_s*. Calculons, pour cet exemple, le degré de signification

$$\begin{aligned}\alpha_s &= \mathbb{P}_{H_0}(\bar{X} > 631) = \mathbb{P}_{H_0}\left(\frac{\bar{X}-600}{25} > \frac{631-600}{25}\right) = \mathbb{P}_{H_0}\left(\frac{\bar{X}-600}{25} > 1,24\right)\\ &= 1 - \mathbb{P}_{H_0}\left(\frac{\bar{X}-600}{25} \leq 1,24\right) = 1 - 0,8925 = 0,1075 \simeq 11\%\end{aligned}$$

Le degré de signification est assez élevé. Cela signifie que si la nouvelle conception de tablettes de chocolat n'était pas meilleure (*i.e.* sous H_0), les seules fluctuations d'échantillonage auraient 11 chances sur 100 de conduire à une valeur de $\bar{X}$ aussi grande (> 631). Ici, $\alpha_s > \alpha$ et donc $\bar{x} \notin ZR_\alpha$, on accepte H_0.

En fait, *le degré de signification est le plus petit niveau α_s pour lequel le test correspondant serait significatif.*

Remarque.

- Le degré de signification ne peut être calculé qu'une fois que les observations ont été faites : c'est le niveau du test obtenu en choisissant la zone de rejet de H_0 la plus petite possible qui contienne l'observation. Dans notre exemple, c'est le niveau du test de région de rejet $\{\bar{X} > 631\}$.
- Un test est d'autant plus significatif, c'est-à-dire *probant* en ce qui concerne l'alternative H_1, que son degré de signification est plus petit. *α_s mesure la conviction avec laquelle on affirme que le test est significatif.* Calculer le degré de signification évite le problème de se fixer un risque α à l'avance pour effectuer le test.
- Beaucoup de logiciels statistiques effectuent automatiquement les tests en donnant le degré de signification. En fonction du risque choisi, on décide si on accepte ou non H_0 : il suffit de comparer α_s à α : $\boxed{\text{si } \alpha_s < \alpha \text{ on rejette } H_0}$.

Généralement on admet que :

degré de signification	significativité du test
$0,01 < \alpha_s \leq 0,05$	significatif
$0,001 < \alpha_s \leq 0,01$	très significatif
$\alpha_s \leq 0,001$	hautement significatif

Remarque. Ne pas confondre *significativité* avec *d'importante distance* entre θ et H_0. Si le test est puissant, il peut être hautement significatif avec pourtant un écart faible entre θ et H_0. Il existe aussi des tests non significatifs avec pourtant une grande distance entre θ et H_0. Ce sont des tests peu puisants.

Chapitre 2: Tests sur les moyennes

(On parle aussi de tests sur l'espérance. On ne se limitera dans ce chapitre qu'aux tests sur les moyennes à un échantillon)

On va s'intéresser ici à un (ou plusieurs) caractère quantitatif X d'une population $\mathcal{P}$ dont on veut étudier la moyenne (l'espérance) $\mathbb{E}(X) \stackrel{\text{def}}{=} \mu$. On note $\sigma^2 \stackrel{\text{def}}{=} \text{Var}(X)$. On considèrera dans la suite, que l'on dispose d'une observation $(x_1, \cdots, x_n)$ constituée des mesures du caractère X faites sur un échantillon : on considère $(x_1, \cdots, x_n)$ comme la réalisation d'un $n-$uplet $(X_1, \cdots, X_n)$ de variables aléatoires indépendantes et de même loi que X. On dira que $(X_1, \cdots, X_n)$ est un n-échantillon de (la loi de) X.

2.1 Comparaison d'une moyenne observée à une moyenne théorique

On considère donc un n-échantillon $(X_1, \cdots, X_n)$ de X d'espérance $\mathbb{E}(X_i) = \mathbb{E}(X) = \mu$ et de variance $\text{Var}(X_i) = \text{Var}(X) = \sigma^2$.

Faire un test de comparaison d'une moyenne observée avec une moyenne théorique revient à répondre à la question suivante. La variable quantitative X a-t-elle une moyenne μ égale à une valeur μ_0 donnée à l'avance ? Il s'agit donc de juger de la pertinence de l'hypothèse

$$H_0 : \mu = \mu_0$$

où μ_0 est la *moyenne théorique*. La quantité dont on dispose pour pouvoir juger est la réalisation $\bar{x}$ de la moyenne empirique du n-échantillon de X

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

$\bar{X}$ est un estimateur de μ. Il est naturel de rejeter H_0 si l'écart $\bar{X} - \mu_0$ entre μ_0 et cet estimateur est "trop grand". La difficulté est de *quantifier* l'écart *acceptable* dû à l'aléa de l'estimateur. Il va s'agir de reconnaître un *faible* écart dû à l'aléa de l'estimateur, d'un écart *significatif* non expliquable par les simples fluctuations de l'estimateur.

▷ Quelle est l'alternative qui nous intéresse quand H_0 est inacceptable ?

1. Une alternative unilatérale $H_1' : \mu > \mu_0$,
2. $H_1'' : \mu < \mu_0$ (alternative unilatérale) ?
3. ou une alternative bilatérale : $H_1 : \mu \neq \mu_0$?

Si H_0 est rejetée au profit de

1. H_1', on rejetera H_0 pour les valeurs "trop grandes" de $\bar{X} - \mu_0$
2. H_1'', on rejetera H_0 pour les valeurs "trop petites" de $\bar{X} - \mu_0$, ou encore pour les valeurs "trop grandes" de $\mu_0 - \bar{X}$,
3. H_1, on rejetera H_0 pour les valeurs "trop grandes" de l'écart absolu $|\bar{X} - \mu_0|$.

▷ Quelle est la loi de $\bar{X} - \mu_0$ sous H_0 ?

En pratique, on va distinguer plusieurs cas :

- la variance σ^2 de X est-elle connue ou inconnue ?
- a-t'on un grand échantillon (de loi quelconque) pour utiliser le TCL, ou un petit échantillon de loi normale ?

On rappelle que

Rappel 4 : (Théorème Central Limite) Soit $X_1, X_2, \ldots, X_n$ un n-échantillon de loi X, d'espérance μ et de variance σ^2. On note la moyenne empirique

$$\overline{X_n} = \frac{X_1 + X_2 + \cdots + X_n}{n}.$$

La loi de $\overline{X_n}$, lorsque n est suffisamment grand, est approximativement une $\mathcal{N}\left(\mu, n^{-1}\sigma^2\right)$. Plus précisément, si l'on note

$$Z_n = \frac{\sqrt{n}\left(\overline{X_n} - \mu\right)}{\sigma}$$

et $F_n(x)$ la fonction de répartition de Z_n, alors $\lim_{n\to\infty} F_n(x) = F(x)$, où

$$F(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{z^2}{2}} dz \quad \text{est la fonction de répartition de la loi } \mathcal{N}(0,1).$$

C'est-à-dire que l'on a

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{\sqrt{n}\left(\overline{X_n} - m\right)}{\sigma} < x \right) = \mathbb{P}\left(Z < x\right) \ , \text{ où } Z \sim \mathcal{N}(0,1).$$

On dit alors que Z_n converge en loi vers Z, et on note $Z_n \xrightarrow[\mathrm{n}\to\infty]{\mathcal{L}} Z$.

2.1.1 Grands échantillons de loi quelconque

On considère un n-échantillon $(X_1, \ldots, X_n)$ de X avec $\mathbb{E}(X_i) = \mu$, et $\mathrm{Var}(X_i) = \sigma^2$. On rappelle (cf. cours d'estimation) que l'estimateur empirique $\bar{X}_n$ est un estimateur consistant de μ. De plus, pour un "grand échantillon" (c'est-à-dire pour n grand), on peut appliquer le TCL et ainsi

$$\frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \approx \mathcal{N}(0,1).$$

Sous l'hypothèse $H_0 : \mu = \mu_0$ on a alors

$$T_n \overset{\text{def}}{=} \frac{\sqrt{n}(\bar{X} - \mu_0)}{\sigma} \approx \mathcal{N}(0,1). \tag{2.1.1}$$

La statistique T_n ne dépend que des observations et des paramètres connus que si σ est connue. Sinon, il faudra utiliser une estimation de cet écart-type afin de ne conserver dans la statistique que des variables connues ou observées.

Cas où la variance σ^2 est connue

- **Test de $H_0 : \mu = \mu_0$ contre $H_1 : \mu \neq \mu_0$**
  On considère comme statistique de test la variable T_n définie en (2.1.1), de sorte que, sous H_0,
  $$T_n \approx \mathcal{N}(0,1).$$
  Sous H_1, T_n prend des valeurs plus petites ou plus grandes que sous H_0. En effet, sous H_1, $\overline{X_n}$ se comporte en moyenne comme μ, et $\mu \neq \mu_0$. Ainsi T aura tendance à prendre des "grandes" valeurs positives ou négatives.

  Pour un test de niveau α (fixé par exemple à 5%), on en déduit la règle de décision suivante.
  ▷ Si $T_n > c_\alpha$ ou $T_n < -c_\alpha$, c'est-à-dire si $|T_n| > c_\alpha$, on rejette H_0.
  ▷ Si $-c_\alpha \leq T_n \leq c_\alpha$, c'est-à-dire si $|T_n| \leq c_\alpha$, on accepte H_0.
  Le seuil c_α est tel que $\mathbb{P}_{H_0}(\text{rejeter } H_0) = \mathbb{P}_{H_0}(|T_n| > c_\alpha) = \alpha$. On cherche c_α tel que $\mathbb{P}_{H_0}(T_n \leq -c_\alpha) = \mathbb{P}_{H_0}(T_n > c_\alpha) \leq \alpha/2$. Comme sous H_0, $T_n \approx \mathcal{N}(0,1)$, on choisit le quantile c_α tel que $\mathbb{P}(\mathcal{N}(0,1) \leq c_\alpha) = 1 - \alpha/2$.
  Mise en oeuvre : on note $t_n = \sqrt{n}(\overline{x_n} - \mu_0)/\sigma$ la réalisation de T_n, on rejetera H_0 si $|t_n| > c_\alpha$, on acceptera H_0 si $|t_n| \leq c_\alpha$.

- **Test de $H_0 : \mu = \mu_0$ contre $H_1 : \mu > \mu_0$**
  Sous H_0, $T_n \approx \mathcal{N}(0,1)$, et sous H_1, T_n prend des valeurs plus grandes que sous H_0.
  On en déduit la règle de décision suivante.
  ▷ Si $T_n > c_\alpha$, on rejette H_0.
  ▷ Si $T_n \leq c_\alpha$, on accepte H_0.
  c_α est tel que $\mathbb{P}_{H_0}(\text{rejeter } H_0) = \mathbb{P}_{H_0}(T > c_\alpha) = \alpha$, pour un niveau de test α fixé.
  Mise en oeuvre : on note $t_n = \sqrt{n}(\bar{x} - \mu_0)/\sigma$ la réalisation de T_n, on rejetera H_0 si $t_n > c_\alpha$.

- **Test de $H_0 : \mu = \mu_0$ contre $H_1 : \mu < \mu_0$**
  Sous H_0, $T_n \approx \mathcal{N}(0,1)$, et sous H_1, T_n prend des valeurs plus petites que sous H_0.

  On en déduit la règle de décision suivante.
  ▷ Si $T_n < c_\alpha (< 0)$, on rejette H_0.
  ▷ Si $T_n \geq c_\alpha$, on accepte H_0.
  c_α est tel que $\mathbb{P}_{H_0}(\text{rejeter } H_0) = \mathbb{P}_{H_0}(T < c_\alpha) = \alpha$. *Notons que c_α est négatif.*
  Mise en oeuvre : soit $t_n = \sqrt{n}(\bar{x} - \mu_0)/\sigma$ la réalisation de T_n. On rejetera H_0 si $t_n < c_\alpha$.

  Remarque. On peut aussi effectuer les tests
  $$H_0 : \mu \leq \mu_0 \quad \text{contre} \quad H_1 : \mu > \mu_0$$

ou encore

$$H_0 : \mu \geq \mu_0 \quad \text{contre} \quad H_1 : \mu < \mu_0.$$

Pour le premier test, la construction est basée sur le choix de c_α tel que $\mathbb{P}_{H_0}(T_n > c_\alpha) = \alpha_\mu$; le risque dépend de μ, et sous H_0, μ peut prendre toutes les valeurs réelles telles que $\mu \leq \mu_0$. Mais en fait, α correspond au risque maximum que l'on accepte de prendre en rejetant H_0 à tort, c'est-à-dire $\alpha = \sup_{\mu \in H_0} \alpha_\mu$. On construira donc finalement la zone de rejet en utilisant la valeur μ_0 *seule* qui correspond *au risque le plus élevé*. Le niveau du test α est la valeur la plus élevée du risque atteinte en $\mu = \mu_0$. On peut donc confondre le niveau α du test (risque maximum) et le risque α_μ qui dépend de μ, car lorsque l'on teste des hypothèses simples du type $H_0 : \mu = \mu_0$, le risque maximum est atteint en $\mu = \mu_0$.

Exemple : Une petite usine fabrique des portes dont la hauteur est une variable aléatoire d'espérance $\mu = 2.5$ mètres et de variance $\sigma^2 = 10^{-4}$. Pour vérifier si ses machines sont bien réglées, le nouveau directeur de l'usine fait mesurer les hauteurs de cents portes tirées au hasard et observe $\bar{x}_{100} = 2.4$ mètres.
Les machines sont-elles bien réglées ? (on prendra comme niveau de test $\alpha = 5\%$)

Soit X la hauteur d'une porte d'espérance $\mathbb{E}(X) = \mu$ (taille moyenne d'une porte) et de variance $\mathrm{Var}(X) = \sigma^2 = 10^{-4}$. On note X_i la hauteur de la $i^{ème}$ porte, pour $i = 1, \ldots, 100$. On a donc $\mathbb{E}(X_i) = \mu$ et $\mathrm{Var}(X_i) = 10^{-4}$.
On souhaite donc tester $H_0 : \mu = 2,5$ contre $H_1 : \mu \neq 2,5$.
On considère la statistique de test

$$T_n = \frac{\sqrt{100}(\bar{X}_{100} - 2,5)}{\sqrt{10^{-4}}}$$

On a vu que sous H_0, $T_n \approx \mathcal{N}(0,1)$ d'après le TLC.
Règle de décision :
▷ si $|T_n| > c_\alpha$ on rejette H_0,
▷ si $|T_n| \leq c_\alpha$ on accepte H_0,
où c_α est tel que $\mathbb{P}_{H_0}(|T_n| > c_\alpha) = 0,05$. Pour lire la valeur de c_α dans la table de la loi gaussienne centrée réduite, on peut utiliser la symétrie de cette loi :

$$\begin{aligned} \mathbb{P}_{H_0}(|T_n| > c_\alpha) &= \mathbb{P}_{H_0}(T_n > c_\alpha) + \mathbb{P}_{H_0}(T_n < -c_\alpha) \\ &= \mathbb{P}_{H_0}(T_n > c_\alpha) + \mathbb{P}_{H_0}(T_n > c_\alpha) \\ &= 2\big((1 - \mathbb{P}_{H_0}(T_n \leq c_\alpha)\big) \\ \mathbb{P}_{H_0}(|T_n| > c_\alpha) = \alpha \iff & \mathbb{P}_{H_0}(T_n \leq c_\alpha) = 1 - \frac{\alpha}{2} \end{aligned}$$

Dans la table de la $\mathcal{N}(0,1)$, on lit la valeur de c_α telle que $\mathbb{P}_{H_0}(T_n \leq c_\alpha) = 1 - \frac{0,05}{2} = 0,975$. On trouve $c_\alpha = 1,96$. Donc si $|T_n| > 1,96$ on rejetera H_0.

Mise en oeuvre : $\overline{x_n} = 2,4$ et ainsi $t_n = 10(2,4 - 2,5)/10^{-2} = -100$. On rejette donc H_0 ;

la différence est significative au niveau 5% (les machines sont significativement mal réglées).

Degré de signification

$$\begin{aligned} \alpha_s &= \mathbb{P}_{H_0}(|T_n| > 100) = 2\big(1 - \mathbb{P}_{H_0}(T_n \leq 100)\big) \\ &\simeq 2(1-1) \simeq 0 \end{aligned}$$

Donc $\alpha_s \simeq 0\%$. L'écart entre la moyenne observée et la moyenne théorique $2,5$ est significatif à 0%. Ce qui signifie que, si les machines étaient bien réglées (*i.e.* sous H_0), les seules fluctuations d'échantillonnage n'auraient que 0 chances sur 100 de conduire au résultat observé, c'est-à-dire une moyenne de 2.4 au lieu de 2.5.

Cas où la variance σ^2 est inconnue

Lorsque la variance $\sigma^2 = \mathrm{Var}(X)$ est inconnue, il faut, comme pour les intervalles de confiance, remplacer σ^2 par un estimateur (consistant) de σ^2.

♦ **Test de $H_0 : \mu = \mu_0$ contre $H_1 : \mu \neq \mu_0$**
Lorsque σ^2 est connue, on utilise comme statistique de test T_n. On peut prendre comme estimateur de σ^2,

$$S_e^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 \quad \text{ou} \quad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2,$$

ou tout autre estimateur consistant de σ^2. On choisit le plus souvent S^2 qui est un *estimateur sans biais de σ^2*.

Résultat (admis) **:** Sous H_0, $U_n = \dfrac{\sqrt{n}(\bar{X} - \mu_0)}{\sqrt{S^2}}$ suit approximativement une $\mathcal{N}(0,1)$.

On procède alors de la même manière que dans le cas où σ^2 est connue.
Règle de décision :
▷ si $|U_n| > c_\alpha$, on rejette H_0,
▷ si $|U_n| \leq c_\alpha$, on accepte H_0,
où le seuil c_α est tel que $\mathbb{P}_{H_0}(\text{rejeter } H_0) = \mathbb{P}_{H_0}(|U_n| > c_\alpha) = \alpha$, pour un niveau de test α fixé (par exemple à 5%). Or sous H_0, $U_n \approx \mathcal{N}(0,1)$, on cherche donc c_α tel que $\mathbb{P}(\mathcal{N}(0,1) \leq c_\alpha) = 1 - \alpha/2$.

Mise en oeuvre : si $u_n = \sqrt{n}(\overline{x_n} - \mu_0)/S$ est la réalisation de U_n, on rejetera H_0 si $|u_n| > c_\alpha$ et on acceptera H_0 si $|u_n| \leq c_\alpha$.

Exemple : Le bénéfice mensuel moyen d'une succursale d'une chaîne de magasins est égal à 300 000€. Afin d'augmenter ses marges, cette chaîne de magasins a décidé d'adopter une nouvelle politique de gestion des stocks pour l'ensemble de ses succursales. On note X_i la

variable aléatoire représentant le bénéfice de la $i^{ème}$ succursale. Une étude menée sur 50 succursales a donné les valeurs observées suivantes :

$$\frac{1}{50}\sum_{i=1}^{50} x_i = 314\ 600\ € \qquad \sqrt{\frac{1}{49}\sum_{i=1}^{50}(x_i - \overline{x_n})^2} = 25\ 000\ €$$

La nouvelle politique est-elle plus efficace que l'ancienne ?
La loi de X_i est inconnue ; $\mathbb{E}(X_i) = \mu$ bénéfice mensuel moyen, $\mathrm{Var}(X_i) = \sigma^2$ inconnue.
On teste $H_0 : \mu = 300\ 000$ contre $H_1 : \mu > 300\ 000$
Sous H_0, $U_{50} = \dfrac{\sqrt{50}(\overline{X_n} - 300\ 000)}{\sqrt{S^2}} \approx \mathcal{N}(0,1)$, et sous H_1, U_{50} prend des valeurs plus grandes que sous H_0. La règle de décision est si $u_{50} > c_\alpha$ on rejette H_0, sinon on accepte H_0, où c_α est tel que $\mathbb{P}_{H_0}(U_n > c_\alpha) = 0,05$. On lit dans la table de la loi gaussienne, $c_\alpha = 1,64$.

Mise en oeuvre : $u_{50} = \sqrt{50}(314\ 600 - 300\ 000)/25\ 000 = 4,13 > 1,64$, donc le test est significatif au niveau 5% : la nouvelle politique est significativement plus efficace.

Degré de signification :

$$\alpha_s = \mathbb{P}_{H_0}(U_n > 4,13) = 1 - \mathbb{P}_{H_0}(U_n \leq 4,13) \approx 0$$

Remarque. Pour les tests de H_0 contre toutes les autres alternatives ($\mu \neq \mu_0$, $\mu < \mu_0$, $\mu > \mu_0$), on procède aussi comme dans le cas où la variance est connue, en remplaçant cette variance par un estimateur consistant.

2.1.2 Petits échantillons gaussiens

Lorsque l'on a affaire à des petits échantillons, on ne peut plus appliquer le TLC. Il faut donc une hypothèse supplémentaire sur la loi commune des X_i, pour connaître la loi de T_n sous H_0.
On suppose maintenant que l'on a un n-échantillon $(X_1, \ldots, X_n)$ gaussien, $X_i \sim \mathcal{N}(\mu, \sigma^2)$.
Sous H_0, on connaît alors la *loi exacte* (et non *plus une approximation de la loi*) de $\overline{X_n}$ et par conséquent de T_n aussi.

$$\text{Sous } H_0, \quad T_n = \frac{\sqrt{n}(\overline{X_n} - \mu_0)}{\sigma} \sim \mathcal{N}(0,1)$$

A nouveau, T_n n'est utilisable directement comme statistique de tests (c'est-à-dire comme fonction des observations) que si σ^2 est connue.

Cas où la variance σ^2 est connue

On utilise la statistique de test T_n dont la loi exacte sous H_0 est une $\mathcal{N}(0,1)$. La procédure pour effectuer chacun des tests est alors la même que précedemment.

Exemple : Le PDG d'une chaîne de magasins à grande surface voudrait savoir s'il doit entreprendre une campagne publicitaire à l'échelon national pour augmenter la vente de ses produits. Pour prendre une décision, il fait une campagne publicitaire dans 5 villes et observe les accroissements X_i de bénéfices correspondants. On suppose que $X_i \sim \mathcal{N}(\mu, 2)$ où $\mu = 0$ si la campagne est inefficace, et $\mu > 0$ si la campagne est efficace. Les accroissements de bénéfice observés dans les 5 villes sont : $1\ ;\ 1,5\ ;\ 0\ ;\ 1\ ;\ 0,5$.
La campagne est-elle significativement efficace ? (on fera le test à 5%).
On teste donc $H_0 : \mu = 0$ contre $H_1 : \mu > 0$.
Sous H_0, $T_n = \sqrt{5}(\bar{X} - 0)/2 \sim \mathcal{N}(0, 1)$, et sous H_1, T_n a tendance à prendre des valeurs plus grandes.

Règle de décision :
▷ Si $T_n > c_\alpha$, on rejette H_0,
▷ si $T_n \leq c_\alpha$ on accepte H_0,
où c_α est tel que $\mathbb{P}_{H_0}(T_n > c_\alpha) \leq 0,05$. On lit dans la table de la loi gaussienne $c_\alpha = 1,64$.

Mise en oeuvre : $t_n = \sqrt{5}(0,8 - 0)/2 = 0,89 < 1,64$, donc on accepte H_0, la campagne n'est pas significativement efficace.

Degré de signification :

$$\alpha_s = \mathbb{P}_{H_0}(T_n > 0,89) = 1 - \mathbb{P}_{H_0}(T_n \leq 0,89) = 1 - 0,8133 = 0,1867.$$

Donc $\alpha_s = 18,67\% > \alpha = 5\%$.

Cas où la variance σ^2 est inconnue

On effectue encore le test de $H_0 : \mu = \mu_0$ contre $H_1 : \mu \neq \mu_0$.

On ne peut plus utiliser la statistique de test précédente puisque σ^2 est inconnue. On estime alors σ^2 par S^2. Dans le cas d'un échantillon gaussien, on sait alors que

$$U_n = \frac{\sqrt{n}(\bar{X} - \mu_0)}{\sqrt{S^2}} \sim_{H_0} St(n-1),$$

où $St(d)$ est une loi de Student à d degrés de liberté.

Rappel 5 : Si Y et Z sont deux variables aléatoires indépendantes telles que $Y \sim \mathcal{N}(0, 1)$ et $Z \sim \chi^2(n)$, alors

$$\frac{Y}{\sqrt{\frac{Z}{n}}} \sim St(n) \quad \text{(loi de Student à } n \text{ degrés de liberté).}$$

Rappel 6 : Si $Y_1, \ldots, Y_n$ sont n variables aléatoires indépendantes de même loi $\mathcal{N}(0, 1)$, alors $Y = \sum_{i=1}^{n} Y_i^2 \ \sim\ \chi^2(n)$ (Loi du chi deux à n degrés de liberté).

Rappel 7 : Soit $(X_1, \ldots, X_n)$ un n-échantillon gaussien de loi $\mathcal{N}(\mu, \sigma)$. Alors

$$\frac{\overline{X_n} - \mu}{\sqrt{\frac{S^2}{n}}} = \frac{\sqrt{n}(\overline{X_n} - \mu)}{S} \sim St(n-1) \quad \text{et} \quad \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1).$$

En effet,

$$\frac{\sqrt{n}(\overline{X_n} - \mu)}{\sqrt{S^2}} = \sqrt{n-1}\frac{\frac{\sqrt{n}(\overline{X_n} - \mu)}{\sigma}}{\sqrt{(n-1)\frac{S^2}{\sigma^2}}} \; '' = '' \; \frac{\mathcal{N}(0,1)}{\sqrt{\frac{\chi^2(n-1)}{n-1}}}.$$

Nous avons donc $(n-1)\frac{S^2}{\sigma^2} \sim \chi^2(n-1)$ et sous H_0, $\frac{\sqrt{n}(\overline{X_n} - \mu_0)}{\sigma} \sim \mathcal{N}(0,1)$. Donc sous H_0, $U_n = \frac{\sqrt{n}(\overline{X_n} - \mu_0)}{\sqrt{S^2}} \sim St(n-1)$.

Règle de décision :
▷ si $|U_n| > c_\alpha$, on rejette H_0,
▷ si $|U_n| \leq c_\alpha$, on accepte H_0.
c_α est tel que $\mathbb{P}_{H_0}(|U_n| > c_\alpha) = \alpha$. La valeur du seuil c_α se lit cette fois dans la table de la $St(n-1)$.

On procède de même pour les autres hypothèses alternatives ($\mu < \mu_0$, $\mu > \mu_0$, $\mu = \mu_1$, ...).

Remarque. Ces résultats sont valables pour n'importe quel échantillon gaussien, qu'il soit grand ou petit.

Exemple : On suppose que le taux de cholestérol d'un individu est une variable aléatoire X de loi normale d'espérance μ et de variance σ^2. Une étude épidémiologique a mis en évidence que les individus atteints d'une maladie C présentent un fort taux de cholestérol de 4 g / litre. Des médecins décident de s'assurer de l'efficacité d'un nouveau médicament M supposé faire baisser le taux de cholestérol. Pour cela, ils mesurent le taux de cholestérol sur un échantillon de $n = 10$ individus atteints de la maladie C, traités par le nouveau médicament M. On observe les résultats suivants :

$$\sum_{i=1}^{10} x_i = 37.2 \quad \text{et} \quad \sum_{i=1}^{10} x_i^2 = 139.58,$$

où x_i désigne la réalisation de la variable aléatoire X_i représentant le taux de cholestérol obtenu après traitement par le médicament M, du $i^{ème}$ individu testé. Le nouveau médicament M est-il significativement efficace ?

On souhaite donc tester $H_0 : \mu = 4$ contre $H_1 : \mu < 4$.

Sous H_0, $U_{10} = \sqrt{10}(\overline{X_n} - 4)/\sqrt{S^2} \sim St(9)$. On vérifie aisément que U_{10} s'écrit en fonction des deux estimateurs $10\overline{X_{10}}$ et $\sum X_i^2$ dont on connaît les observations,

$$U_{10} = \frac{\sqrt{10}(\overline{X_n} - 4)}{\sqrt{\frac{1}{9}\sum_{i=1}^n X_i^2 - \frac{1}{90}(10\overline{X_{10}})^2}}.$$

Règle de décision :
▷ si $U_{10} < c_\alpha$, on rejette H_0,
▷ si $U_{10} \leq c_\alpha$, on accepte H_0.
c_α est tel que $\mathbb{P}_{H_0}(U_{10} < c_\alpha) = \alpha$. La valeur du seuil c_α se lit cette fois dans la table de la $St(9)$. Pour un niveau α de 5%, on trouve $c_\alpha = -1,833$.

Mise en oeuvre : $u_n = \sqrt{10}(3,72 - 4)/0.36 = -2,43 < -1.833$, donc on rejette H_0, le médicament est significativement efficace.

Degré de signification :

$$\alpha_s = \mathbb{P}_{H_0}(U_n < -2,43) \approx 1 - 0,975 = 0,025.$$

Donc $\alpha_s = 2,5\% < \alpha = 5\%$.

2.2 Comparaison d'une proportion observée à une proportion théorique

On s'intéresse cette fois à une proportion p inconnue d'individus possèdant un caractère X dans une population $\mathcal{P}$. Par exemple, la proportion d'individus d'intentions de votes favorables à un candidat. La variable X est définie par $\{X = 1\}$ si l'individu possède la propriété en question (dans l'exemple, si il est favorable au candidat) et $\{X = 0\}$ sinon (défavorable). On a $\mathbb{P}(X = 1) = p$, $\mathbb{P}(X = 0) = 1 - p$, et $X \sim \mathcal{B}(p)$. On rappelle que pour une loi de Bernoulli de paramètre p on a $\mathbb{E}(X) = p$ et $\mathrm{Var}(X) = p(1 - p)$. Ce cadre est donc un cas particulier de test de moyennes avec cette fois un n-échantillon $(X_1, \ldots, X_n)$ de X de loi $\mathcal{B}(p)$, d'espérance $\mu \overset{\text{def}}{=} \mathbb{E}(X_i) = \mathbb{E}(X) = p$ et de variance $\sigma^2 \overset{\text{def}}{=} \mathrm{Var}(X_i) = \mathrm{Var}(X) = p(1 - p)$.
On rappelle que la proportion aléatoire d'individus de l'échantillon (possèdant le caractère) notée $P_n = \overline{X_n} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} X_i$ est un bon estimateur de la proportion p inconnue dans la population.

♦ **Test de $H_0 : p = p_0$ contre $H_1 : p \neq p_0$**
Le test va être construit autour de l'estimateur P_n de p, plus exactement à partir de la différence $P_n - p_0$. Si on dispose d'un échantillon suffisamment grand ($n > 30$), on peut appliquer le théorème central limite pour approcher la loi de P par une loi normale :

$$P_n \approx \mathcal{N}\left(p, \frac{p(1-p)}{n}\right),$$

soit encore

$$\frac{\sqrt{n}(P_n - p)}{\sqrt{p(1-p)}} \approx \mathcal{N}(0,1).$$

On pourrait estimer la variance $p(1 - p)$ par un estimateur consistant ($P_n(1 - P_n)$ par exemple) mais c'est inutile puisque sous H_0, $p = p_0$, et donc sous H_0, $p(1-p) = p_0(1-p_0)$. On construit donc le test à partir de la statistique de test

$$T_n = \frac{\sqrt{n}(P_n - p_0)}{\sqrt{p_0(1-p_0)}}$$

Et sous H_0,

$$T_n \approx_{H_0} \mathcal{N}(0,1)$$

Règle de décision :
▷ si $|T_n| > c_\alpha$, on rejette H_0,
▷ si $|T_n| \leq c_\alpha$, on accepte H_0.
Le seuil c_α est tel que $\mathbb{P}_{H_0}(|T_n| > c_\alpha) = \alpha$. La valeur du seuil c_α se lit dans la table de la $\mathcal{N}(0,1)$.
Si on note p_n la réalisation de P_n sur l'échantillon, et $t_n = \sqrt{n}(\hat{p} - p_0)/\sqrt{p_0(1-p_0)}$ celle de T_n, on rejetera H_0 si $|t_n| > c_\alpha$.
On procède de la même manière pour les autres formes d'hypothèse alternative H_1.

Exemple 1 Une machine doit être mise à la casse si elle produit strictement plus de 10% de pièces défectueuses. Pour savoir si l'on doit remplacer une certaine machine, on prélève au hasard 70 pièces fabriquées par cette machine, et on constate que 8 d'entre elles sont défectueuses. On notera p la proportion de pièces défectueuses fabriquées par cette macine. Doit-on acheter une nouvelle machine ? (test à 5%)
Soit p la proportion de pièces défectueuses produites par une machine. On veut tester l'hypothèse $H_0 : p = 0,1$ contre $H_1 : p > 0,1$ ($p_0 = 0,1$).
Soit X_i la v.a. représentant l'état de la ième pièce fabriquée par la machine ($X_i = 1$ si la ième pièce est défectueuse et 0 sinon), $X_i \sim \mathcal{B}(p)$.
p est estimée par $P_n = \sum_{i=1}^{n} X_i$ la proportion aléatoire de pièces défectueuses de l'échantillon.
La taille $n = 70$ de l'échantillon est suffisamment grande pour appliquer le TCL, et la statistique de test est

$$T_n = \frac{\sqrt{n}(P_n - 0.1)}{\sqrt{0.1(1-0.1)}} \approx_{H_0} \mathcal{N}(0,1)\,.$$

Règle de décision :
▷ si $T_n > c_\alpha$ on rejette H_0,
▷ si $T_n \leq c_\alpha$ on accepte H_0.
Le seuil c_α est tel que $\mathbb{P}_{H_0}(T_n > c_\alpha) = \alpha = 5\%$. Comme $T_n \approx_{H_0} \mathcal{N}(0,1)$, on lit c_α dans la table de la $\mathcal{N}(0,1)$ tel que $\mathbb{P}(\mathcal{N}(0,1) \leq c_\alpha) = 1 - \alpha = 0.95$. On trouve $c_\alpha = 1.645$.
Mise en oeuvre : Sur l'échantillon, on trouve $p_n = 8/70 = 0,11$ (0.11 est une estimation ponctuelle de p), et $t_n = 0.398$. $t_n < c_\alpha$ donc on accepte H_0 à 5%. Le test est non significatif.

Exemple 2 En 2004, 42% des franciliens travaillaient à plus de 30 kms de leur domicile. Pour estimer la proportion de franciliens qui, en 2006, travaillent à plus de 30 kms de leur domicile, on a effectué un sondage sur 2000 franciliens ; parmi les 2000 personnes interrogées, 860 d'entre elles ont déclaré travailler à plus de 30 kms de leur domicile. La proportion de franciliens travaillant à plus de 30 kms de leur domicile a-t-elle évolué de manière significative entre 2004 et 2006 ?

Soit p la proportion de franciliens travaillant à plus de 30 kms de leur domicile en 2006. On veut tester $H_0 : p = 0,42$ contre $H_1 : p \neq 0,42$.

Exemple 3 En 1990, 38% des parisiens allaient régulièrement (au moins une fois par semaine) au cinéma. Une enquête réalisée en 1997 indique que sur 1000 personnes interrogées, 350 d'entre elles vont régulièrement au cinéma. Y a t-il une baisse significative de la fréquentation des salles de cinéma parisiennes entre 1990 et 1997 ?
Soit p la proportion de parisiens allant régulièrement au cinéma en 1997. On veut tester $H_0 : p = 0,35$ contre $H_1 : p < 0,35$.

2.3 Test d'égalité d'une variance à une valeur fixe

On considère donc un n-échantillon $(X_1, \ldots, X_n)$ de X d'espérance $\mathbb{E}(X_i) = \mathbb{E}(X) = \mu$ et de variance $\mathrm{Var}(X_i) = \mathrm{Var}(X) = \sigma^2$.

On se propose d'étudier la question : la variable (quantitative) X a-t-elle une variance $\mathrm{Var}(X) = \sigma^2$ égale à une valeur σ_0^2 donnée à l'avance ? Pour répondre à cette question, nous allons effectuer le test de l'hypothèse nulle

$$H_0 \ : \ \sigma^2 = \sigma_0^2$$

où σ_0^2 est une valeur fixée, contre les alternatives $\sigma^2 > \sigma_0^2$, $\sigma^2 < \sigma_0^2$, ou $\sigma^2 \neq \sigma_0^2$. Ce test est basé sur l'estimation de la variance σ^2 par l'estimateur $\widehat{\Sigma^2} = n^{-1} \sum_{i=1}^n (X_i - \mu)^2$, quand μ est connue et par l'estimateur $S^2 = (n-1)^{-1} \sum_{i=1}^n (X_i - \overline{X_n})^2$, quand μ est inconnue. On suppose dans la suite que $\mathbb{E}(X_i^4)$ existe et est finie. Comme précédemment, nous allons distinguer le cas des grands échantillons de lois quelconques des petits échantillons gaussiens.

2.3.1 Cas des grands échantillons de loi quelconque

Cas où μ est connue

La variance σ^2 peut être estimée par $\widehat{\Sigma_n^2} = \frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2$. Notons $Z_i \stackrel{\text{def}}{=} (X_i - \mu)^2$. Ainsi, $\widehat{\Sigma_n^2}$ s'écrit en fonction des variables Z_i puisque $\widehat{\Sigma_n^2} = \overline{Z_n}$. De plus, les variables aléatoires Z_i sont *i.i.d.*, d'espérance σ^2 et de variance $\sigma^4 + \mathbb{E}(X_i - \mu)^4$. On peut donc appliquer le Théorème Central Limite aux variables Z_i et obtenir que

$$\frac{\sqrt{n}(\widehat{\Sigma_n^2} - \sigma^2)}{\sqrt{\frac{1}{n-1} \sum_{i=1}^n (Z_i - \overline{Z_n})^2}} \xrightarrow[\mathrm{n}\to\infty]{\mathcal{L}} \mathcal{N}(0,1).$$

Tester une égalité sur la variance de X_i revient à tester une égalité sur la moyenne de Z_i. On construit donc le test de la même façon, à partir de la statistique de test

$$\frac{\sqrt{n}(\widehat{\Sigma_n^2} - \sigma^2)}{\sqrt{\frac{1}{n-1} \sum_{i=1}^n (Z_i - \overline{Z_n})^2}}.$$

Cas où μ est inconnue

2.3.2 Cas des petits échantillons gaussiens

Cas où μ est connue

La variance σ^2 peut être estimée par $\widehat{\Sigma_n^2}$. Notons à nouveau $Z_i \stackrel{\text{def}}{=} (X_i - \mu)^2$. Les variables aléatoires Z_i sont *i.i.d.* de loi gaussienne $\mathcal{N}\big(\sigma^2, \sigma^4 + \mathbb{E}(X_i - \mu)^4\big)$. Ainsi, on a

$$\frac{\sqrt{n}(\widehat{\Sigma_n^2} - \sigma^2)}{\sqrt{\frac{1}{n-1}\sum_{i=1}^n (Z_i - \overline{Z_n})^2}} \sim St(n-1).$$

Cas où μ est inconnue

On utilise la statistique de test $(n-1)S^2/\sigma_0^2$. En effet, sous H_0 cette statistique suit une loi $\chi^2(n-1)$.

Chapitre 3: Tests de comparaison

Nous avons vu jusqu'à présent des test de moyennes (et proportions) à partir d'un échantillon. On comparait la moyenne d'un caractère d'une population à une valeur de référence. Nous allons maintenant effectuer des tests de comparaison de moyennes d'un même caractère à partir de deux échantillons extraits de deux populations. La généralisation à plus de deux populations sera faite dans le cas gaussien dans le cadre du cours de l'analyse de la variance (modèle linéaire).

3.1 Comparaison de deux moyennes observées

On dispose de mesures d'une même grandeur sur deux échantillons extraits, indépendamment, de deux populations différentes :

- $(X_1, \ldots, X_{n_1})$, d'effectif n_1, extrait d'une population de moyenne $\mathbb{E}(X_i) = \mu_1$ et de variance $\mathrm{Var}(X_i) = \sigma_1^2$; on note $\overline{X_n}$ la moyenne, et S_1^2 la variance de cet échantillon ;
- $(Y_1, \cdots, Y_{n_2})$, d'effectif n_2, extrait d'une population de moyenne $\mathbb{E}(Y_i) = \mu_2$, et de variance $\mathrm{Var}(Y_i) = \sigma_2^2$; on note $\overline{Y_n}$ la moyenne, et S_2^2 la variance de cet échantillon ;

Problème posé : On se demande si la moyenne de cette grandeur est la même dans les deux populations. On veut donc tester

$$H_0 \ : \ \mu_1 = \mu_2$$

- ♦ contre l'alternative bilatérale $H_1 \ : \ \mu_1 \neq \mu_2$,
- ♦ ou contre l'alternative unilatérale $H_1 \ : \ \mu_1 < \mu_2$,
- ♦ ou contre l'autre l'alternative unilatérale $H_1 \ : \ \mu_1 > \mu_2$.

Par exemple, on s'intéresse à la taille des hommes et des femmes d'un pays ; on note μ_1 la taille moyenne des femmes et μ_2 la taille moyenne des hommes. On peut chercher à tester si la taille des hommes est significativement supérieure à celle des femmes, c'est-à-dire $H_0 \ : \ \mu_1 = \mu_2$ contre $H_1 \ : \ \mu_1 < \mu_2$.

Il est naturel de fonder les tests de H_0 sur l'écart $\overline{X_n} - \overline{Y_n}$ entre les moyennes observées des deux échantillons. Sous l'hypothèse H_0, la différence observée $\overline{X_n} - \overline{Y_n}$ doit avoir une espérance nulle puisque $\mathbb{E}(\overline{X_n}) - \mathbb{E}(\overline{Y_n}) = \mu_1 - \mu_2 = 0$. Il faut donc connaître la loi de $\overline{X_n} - \overline{Y_n}$ sous H_0. On va une nouvelle fois distinguer deux cas : les grands échantillons de loi quelconque, et les petits échantillons gaussiens.

3.1.1 Grands échantillons de loi quelconque

Si les tailles d'échantillons n_1 et n_2 sont grandes, on sait d'après le TLC que, approximativement, on a

$$\overline{X_n} \approx \mathcal{N}\left(\mu_1, \frac{\sigma_1^2}{n_1}\right) \quad \text{et} \quad \overline{Y_n} \approx \mathcal{N}\left(\mu_2, \frac{\sigma_2^2}{n_2}\right).$$

Comme les deux échantillons sont indépendants, $\overline{X_n}$ et $\overline{Y_n}$ sont indépendants et qu'approximativement sous H_0, on a

$$\overline{X_n} - \overline{Y_n} \approx \mathcal{N}\left(0, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right),$$

soit encore,

$$\frac{\overline{X_n} - \overline{Y_n}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \approx \mathcal{N}(0,1).$$

On ne peut utiliser directement cette variable aléatoire comme statistique de test que lorsque les variances des deux populations sont connues.

Cas où les variances σ_1^2 et σ_2^2 sont connues

- **Test de** $H_0 : \mu_1 = \mu_2$ **contre** $H_1 : \mu_1 \neq \mu_2$
  La statistique de test est

  $$T_n = \frac{\overline{X_n} - \overline{Y_n}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

  Sous H_0, $T_n \approx \mathcal{N}(0,1)$. Sous H_1, T_n a tendance à prendre des valeurs plus petites ou plus grandes.
  Règle de décision :
  ▷ si $|T_n| > c_\alpha$ on rejette H_0,
  ▷ si $|T_n| \leq c_\alpha$, on accepte H_0.
  Le seuil c_α est tel que $\mathbb{P}_{H_0}(|T_n| > c_\alpha) = \alpha$, *i.e.* tel que $\mathbb{P}_{H_0}(T_n > c_\alpha) = \alpha/2$. La valeur de c_α est lue dans la table de la $\mathcal{N}(0,1)$.

- **Test de** $H_0 : \mu_1 = \mu_2$ **contre** $H_1 : \mu_1 > \mu_2$
  Sous H_0, $T_n = \frac{\overline{X_n} - \overline{Y_n}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \approx \mathcal{N}(0,1)$. Sous H_1, T_n a tendance a prendre des valeurs plus grandes.
  Règle de décision :
  ▷ si $T_n > c_\alpha$ on rejette H_0,
  ▷ si $T_n \leq c_\alpha$, on accepte H_0,
  où c_α est tel que $\mathbb{P}_{H_0}(T_n > c_\alpha) = \alpha$.

- **Test de** $H_0 : \mu_1 = \mu_2$ **contre** $H_1 : \mu_1 < \mu_2$
  Sous H_0, $T_n = \frac{\overline{X_n} - \overline{Y_n}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \approx \mathcal{N}(0,1)$. Sous H_1, T_n a tendance a prendre des valeurs plus petites.
  Règle de décision :
  ▷ si $T_n < -c_\alpha$ $(c_\alpha > 0)$ on rejette H_0,
  ▷ si $T_n \geq -c_\alpha$, on accepte H_0,
  où c_α est tel que $\mathbb{P}_{H_0}(T_n < -c_\alpha) = \alpha$.

Cas où les variances σ_1^2 et σ_2^2 sont inconnues

On remplace respectivement σ_1^2 et σ_2^2 par leurs estimateurs consistants et sans biais

$$S_X^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(X_i - \overline{X_n})^2 \quad \text{et} \quad S_Y^2 = \frac{1}{n_2 - 1}\sum_{i=1}^{n_2}(Y_i - \overline{Y_n})^2.$$

On utilise alors la statistique de test

$$T_n = \frac{\overline{X_n} - \overline{Y_n}}{\sqrt{\frac{S_X^2}{n_1} + \frac{S_Y^2}{n_2}}}$$

Sous H_0, $T_n \approx \mathcal{N}(0,1)$ approximativement, et on procède de même que précédemment.

Cas où les variances σ_1^2 et σ_2^2 sont inconnues mais égales

Si les variances sont inconnues mais égales $\sigma_1^2 = \sigma_2^2 = \sigma^2$, on estime la valeur commune σ^2 par

$$S^2 = \frac{(n_1 - 1)S_X^2 + (n_2 - 1)S_Y^2}{n_1 + n_2 - 2} = \frac{1}{n_1 + n_2 - 2}\Big(\sum_{i=1}^{n_1}(X_i - \overline{X_n})^2 + \sum_{i=1}^{n_2}(Y_i - \overline{Y_n})^2\Big).$$

On montre facilement que l'estimateur obtenu S^2 est sans biais. En effet, S_X^2 est un estimateur sans biais de σ_1^2 donc $\mathbb{E}(S_X^2) = \sigma_1^2 = \sigma^2$, et S_Y^2 est sans biais de σ_2^2 donc $\mathbb{E}(S_Y^2) = \sigma_2^2 = \sigma^2$. On en déduit que

$$\mathbb{E}(S^2) = \frac{(n_1 - 1)\mathbb{E}(S_X^2) + (n_2 - 1)\mathbb{E}(S_Y^2)}{n_1 + n_2 - 2} = \frac{(n_1 - 1)\sigma^2 + (n_2 - 1)\sigma^2}{n_1 + n_2 - 2} = \sigma^2$$

On estime alors $\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2} = \sigma^2\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)$ par $S^2\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)$ et on utilise la statistique de test

$$T_n = \frac{\overline{X_n} - \overline{Y_n}}{\sqrt{S^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}.$$

Sous H_0, $T_n \approx \mathcal{N}(0,1)$ approximativement, et on procède de même que précédemment.

Exemple

Un fabriquant de câbles en acier étudie un nouveau traitement de câbles pour améliorer leur résistance. Il choisit au hasard 200 câbles traités et 100 câbles non traités. On suppose que la charge de rupture est une variable aléatoire. On note X_i la charge de rupture du ième câble *traité* et Y_i la charge de rupture du ième câble *non traité*. On observe $\overline{x_n} = 30,82$, $\overline{y_n} = 29,63$,

$\frac{1}{199}\sum_{i=1}^{200}(x_i - \overline{x_n})^2 = 27,25$ et $\frac{1}{99}\sum_{i=1}^{100}(y_i - \overline{y_n})^2 = 23,99$. Peut-on conclure à l'efficacité du traitement ?

On note μ_1 la charge de rupture moyenne (dans la population) des câbles traités, σ_1^2 la variance, et μ_2 la charge de rupture moyenne (dans la population) des câbles non traités, σ_2^2 la variance. On suppose que les deux échantillons $X_1, \ldots, X_{n_1}$, $n_1 = 200$, et $Y_1, \ldots, Y_{n_2}$, $n_2 = 100$ sont indépendants. On a $\mathbb{E}(X_i) = \mu_1$, $\mathrm{Var}(X_i) = \sigma_1^2$, $\mathbb{E}(Y_i) = \mu_2$, $\mathrm{Var}(Y_i) = \sigma_2^2$. μ_1 et μ_2 sont estimées par $\overline{X_n}$ et $\overline{Y_n}$ les charges moyennes aléatoires des câbles traités et non traités des échantillons. Une estimation ponctuelle de μ_1 (resp. μ_2) est la réalisation $\overline{x_n} = 30.82$ (resp. $\overline{y_n} = 29.63$) de l'estimateur $\overline{X_n}$ (resp. $\overline{Y_n}$) sur l'échantillon des câbles traités (resp. non traités). Les variances σ_1^2 et σ_2^2 sont inconnues et estimées par les variances aléatoires S_X^2 et S_Y^2. Une estimation ponctuelle de σ_1^2 (resp. σ_2^2) est la réalisation $s_X^2 = 27.25$ (resp. $s_X^2 = 23.99$) de l'estimateur S_X^2 (resp. S_Y^2) sur l'échantillon.
On souhaite tester $H_0 \ : \ \mu_1 = \mu_2$ contre $H_1 \ : \ \mu_1 > \mu_2$.
Le TCL s'applique (les échantillons sont suffisamment importants), et la statistique de test est en l'absence d'information sur les variances

$$T_n = \frac{\overline{X_n} - \overline{Y_n}}{\sqrt{\frac{S_X^2}{n_1} + \frac{S_Y^2}{n_2}}} \approx \mathcal{N}(0,1) \quad sous\ H_0$$

Si les variances étaient supposées inconnues mais égales $\sigma_1^2 = \sigma_2^2 = \sigma^2$, on estimerait la valeur commune σ^2 par la variance empirique aléatoire calculée sur les deux échantillons traités/non traités réunis (normalisée pour avoir un estimateur sans biais)

$$S^2 = \frac{(n_1 - 1)S_X^2 + (n_2 - 1)S_Y^2}{n_1 + n_2 - 2}.$$

La statistique de test serait

$$T_n = \frac{\overline{X_n} - \overline{Y_n}}{\sqrt{S^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \approx \mathcal{N}(0,1) \quad sous\ H_0$$

Règle de décision :
▷ si $T_n > c_\alpha$ on rejette H_0,
▷ si $T_n \leq c_\alpha$, on accepte H_0,
où le seuil c_α est tel que $\mathbb{P}_{H_0}(T_n > c_\alpha) = \alpha = 5\%$. Dans la table de la loi $\mathcal{N}(0,1)$, on lit $c_\alpha = 1,645$. Si on note t_n la réalisation de T_n sur les deux échantillons, si $t_n > 1.645$ on rejettera H_0 (avec un risque de 5% de se tromper), si $t_n \leq 1.645$ on acceptera H_0.

Mise en oeuvre : Sur l'échantillon, on trouve $t_n = 1,94 > 1,645$. On rejette H_0. Le traitement est donc efficace.

3.1.2 Petits échantillons gaussiens

On considère à nouveau deux échantillons indépendants, le premier $(X_1, \ldots, X_{n_1})$, de moyenne $\mathbb{E}(X_i) = \mu_1$, $\mathrm{Var}(X_i) = \sigma_1^2$, et le second $(Y_1, \ldots, Y_{n_2})$, $\mathbb{E}(Y_i) = \mu_2$, $\mathrm{Var}(Y_i) = \sigma_2^2$. Si les effectifs respectifs n_1 et n_2 des deux échantillons ne sont pas assez importants pour appliquer le TCL, on fait alors à nouveau l'hypothèse supplémentaire que ces échantillons sont gaussiens : $(X_1, \ldots, X_{n_1})$, $X_i \sim \mathcal{N}(\mu_1, \sigma_1^2)$ et $(Y_1, \cdots, Y_{n_2})$, $Y_i \sim \mathcal{N}(\mu_2, \sigma_2^2)$. Bien sûr les résultats suivants seront valables pour de grands échantillons gaussiens !

♦ **Test de $H_0 : \mu_1 = \mu_2$ contre $\mu_1 \neq \mu_2$**
Le test est encore fondé sur la différence $\overline{X_{n_1}} - \overline{Y_{n_2}}$. Dans le cas gaussien, on connaît la loi exacte de $\overline{X_{n_1}}$ et $\overline{Y_{n_2}}$. On a

$$\overline{X_{n_1}} \sim \mathcal{N}\left(\mu_1, \frac{\sigma_1^2}{n_1}\right) \quad \text{et} \quad \overline{Y_{n_2}} \sim \mathcal{N}\left(\mu_2, \frac{\sigma_2^2}{n_2}\right).$$

Par indépendance des deux échantillons,

$$\overline{X_{n_1}} - \overline{Y_{n_2}} \sim \mathcal{N}\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right).$$

Sous H_0,

$$\overline{X_{n_1}} - \overline{Y_{n_2}} \sim \mathcal{N}\left(0, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right) \quad \text{soit encore} \quad \frac{\overline{X_{n_1}} - \overline{Y_{n_2}}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim \mathcal{N}(0,1).$$

On ne peut utiliser directement cette variable aléatoire comme statistique de test que lorsque les variances des deux populations sont connues.

Cas où les variances σ_1^2 et σ_2^2 sont connues

La statistique de test est

$$T_n = \frac{\overline{X_{n_1}} - \overline{Y_{n_2}}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}.$$

Sous H_0, $T_n \sim \mathcal{N}(0,1)$ et on procède de la même manière que précédemment.

Cas où les variances σ_1^2 et σ_2^2 sont inconnues mais égales

Les solutions ne sont pas satisfaisantes pour des petits échantillons lorsque les variances sont différentes et inconnues. On supposera donc l'égalité des variances $\sigma_1^2 = \sigma_2^2 = \sigma^2$. Nous verrons une procédure permettant de tester l'égalité des variances dans le cas gaussien, *le test de Fisher*, que vous utiliserez également dans le modèle linéaire. On va donc estimer la valeur commune σ^2

par la moyenne pondérée par les effectifs $(n_1 - 1)$ et $(n_2 - 1)$ des deux variances d'échantillons S_1^2 et S_2^2 :

$$S^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

de sorte que $\sigma_1^2/n_1 + \sigma_2^2/n_2 = \sigma^2\,(1/n_1 + 1/n_2)$ est estimée par $S^2\,(1/n_1 + 1/n_2)$. On utilise alors la statistique de test

$$T_n = \frac{\overline{X_{n_1}} - \overline{Y_{n_2}}}{\sqrt{S^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}.$$

On montre que dans le cas gaussien, sous H_0, $T_n \sim St(n_1 + n_2 - 2)$. En effet, on a

$$T_n = \frac{\overline{X_{n_1}} - \overline{Y_{n_2}}}{\sqrt{S^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{\dfrac{\overline{X_{n_1}} - \overline{Y_{n_2}}}{\sqrt{\sigma^2\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}}{\dfrac{\sqrt{\dfrac{(n_1 + n_2 - 2)S^2}{\sigma^2}}}{\sqrt{n_1 + n_2 - 2}}}\text{"} = \text{"}\frac{\mathcal{N}(0,1)}{\sqrt{\dfrac{\chi^2(n_1 + n_2 - 2)}{n_1 + n_2 - 2}}}\text{"} = \text{"}St(n_1 + n_2 - 2).$$

On a besoin de l'égalité des variances pour avoir un χ^2 au dénominateur (on aurait sinon une somme de χ^2). La statistique de test est donc T_n, mais cette fois, sous H_0, $T_n \sim St(n_1 + n_2 - 2)$. Les règles de décision se construisent de la même manière mais on recherche cette fois les valeurs seuils "c_α" dans la table de la student $St(n_1 + n_2 - 2)$ et non dans la table de la $\mathcal{N}(0,1)$.

Exemple 1

On cherche à savoir si le rythme cardiaque d'un individu augmente lorsque l'individu est soumis à un stress. Pour cela, on mesure le rythme cardiaque de 5 individus avant une séance de cinéma passant un film d'horreur, et le rythme cardiaque de 5 autres individus après la séance de cinéma. On suppose que le rythme cardiaque d'un individu est une variable aléatoire de loi normale. Les rythmes cardiaques observés sont les suivants :

avant la séance	90	76	80	87	83
après la séance	98	77	88	90	89

1. Le ryhtme cardiaque augmente t'il de manière significative avec le stress ?
2. Calculer le degré de signification du test.

Exemple 2

On admet que la production de lait d'une vache normande (respectivement hollandaise) est une variable aléatoire de loi $\mathcal{N}(\mu_1, \sigma_1^2)$ (respectivement $\mathcal{N}(\mu_2, \sigma_2^2)$). Un producteur de lait souhaite comparer le rendement des vaches normandes et hollandaises de son unité de production.

Les relevés de production de lait (exprimée en litres) de 10 vaches normandes et hollandaises ont donné les résultats suivants :

vaches normandes	552	464	483	506	497	544	486	531	496	501
vaches hollandaises	487	489	470	482	494	500	504	537	482	526

Les deux races de vaches laitières ont-elles le même rendement ? (on supposera l'égalité des variances)

3.2 Comparaison de deux proportions

On souhaite comparer deux proportions p_1 et p_2 d'individus possèdant un même caractère dans deux populations différentes. Par exemple, p_1 représente la proportion de favorables à un candidat dans une ville V1, et p_2 la proportion de favorables à ce candidat dans une autre ville V2 : on peut se demander si la proportion de favorables au candidat est la même dans les deux villes, auquel cas on testera $H_0 : p_1 = p_2$ contre $H_1 : p_1 \neq p_2$. On va donc considérer deux échantillons indépendants $(X_1, \cdots, X_{n_1})$ de loi $\mathcal{B}(p_1)$, et $(Y_1, \cdots, Y_{n_2})$ de loi $\mathcal{B}(p_2)$. Dans le cas de l'exemple, X_i représente l'opinion du ième électeur dans V1, Y_i dans V2. La proportion théorique p_1 est estimée par la proportion aléatoire du premier échantillon $n_1^{-1}\sum_{i=1}^{n_1} X_i = \overline{X_{n_1}}$ et p_2 par $n_2^{-1}\sum_{i=1}^{n_2} Y_i = \overline{Y_{n_2}}$, proportion aléatoire du 2ème échantillon. On rappelle que $\mathbb{E}(X_i) = p_1$, $\mathrm{Var}(X_i) = p_1(1-p_1)$, $\mathbb{E}(Y_i) = p_2$, $\mathrm{Var}(Y_i) = p_2(1-p_2)$, $\mathbb{E}(\overline{X_{n_1}}) = p_1$, $\mathrm{Var}(\overline{X_{n_1}}) = p_1(1-p_1)/n_1$, $\mathbb{E}(\overline{Y_{n_2}}) = p_2$ et $\mathrm{Var}(\overline{Y_{n_2}}) = p_2(1-p_2)/n_2$.

♦ **Test de $H_0 : p_1 = p_2$ contre $H_1 : p_1 \neq p_2$**
Le test est fondé sur l'écart $\overline{X_{n_1}} - \overline{Y_{n_2}}$ entre les deux proportions aléatoires observées dans les échantillons, variables aléatoires dont il faut connaître la loi sous H_0. Si les tailles d'échantillons n_1 et n_2 sont suffisamment importantes, le TCL s'applique et on a

$$\overline{X_{n_1}} \approx \mathcal{N}\left(p_1, \frac{p_1(1-p_1)}{n_1}\right) \quad \text{et} \quad \overline{Y_{n_2}} \approx \mathcal{N}\left(p_2, \frac{p_2(1-p_2)}{n_2}\right)$$

Par indépendance,

$$\overline{X_{n_1}} - \overline{Y_{n_2}} \approx \mathcal{N}\left(p_1 - p_2, \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}\right).$$

Sous H_0,

$$\overline{X_{n_1}} - \overline{Y_{n_2}} \approx \mathcal{N}\left(0, \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}\right)$$

soit encore,

$$\frac{\overline{X_{n_1}} - \overline{Y_{n_2}}}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \approx \mathcal{N}(0,1)$$

On connaît la loi de cette variable aléatoire sous H_0, mais on ne peut l'utiliser comme statistique de test car on ne connaît pas la valeur de $p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2$. Il faut

donc estimer cette quantité sous H_0. Sous H_0, $p_1 = p_2 = p$, et on estime la valeur commune p par

$$P_n = \frac{\sum_{i=1}^{n_1} X_i + \sum_{i=1}^{n_2} Y_i}{n_1 + n_2} = \frac{n_1 \overline{X_{n_1}} + n_2 \overline{Y_{n_2}}}{n_1 + n_2}.$$

P_n est la proportion aléatoire observée sur les deux échantillons. On estime alors (sous H_0) $p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2 = p(1-p)(1/n_1 + 1/n_2)$ par $P_n(1-P_n)(1/n_1 + 1/n_2)$. On construit le test à partir de la statistique

$$T_n = \frac{\overline{X_{n_1}} - \overline{Y_{n_2}}}{\sqrt{P_n(1-P_n)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Sous H_0, $T_n \approx \mathcal{N}(0,1)$.

Règle de décision :

▷ si $|T_n| > c_\alpha$ on rejette H_0,

▷ si $|T_n| \leq c_\alpha$, on accepte H_0,

où le seuil c_α est choisi tel que $\mathbb{P}_{H_0}(|T_n| > c_\alpha) = \alpha$; la valeur de c_α est lue dans la table de la $\mathcal{N}(0,1)$.

Exemple

A la sortie de deux salles de cinéma donnant le même film, on a interrogé des spectateurs quant à leur opinion sur le film. Les résultats de ce sondage d'opinion sont les suivants :

Opinion	Mauvais film	Bon film	Total
Salle 1	30	70	100
Salle 2	48	52	100
Total	78	122	200

L'opinion est-elle significativement liée à la salle ?

Soit p_1, resp. p_2 la proportion de gens de la salle 1, resp. salle 2, ayant une mauvaise opinion du film. On veut tester $H_0 : p_1 = p_2$ contre $H_1 : p_1 \neq p_2$.

Soit X_i, resp. Y_i, la variable aléatoire représentant l'opinion du ième spectateur interrogé dans la salle 1, resp. salle 2. $X_i = 1$ si le ième spectateur interrogé dans la salle 1 a une mauvaise opinion, $X_i = 0$ sinon. $X_i \sim \mathcal{B}(p_1)$. De même, $Y_i \sim \mathcal{B}(p_2)$. On note $\overline{X_{n_1}}$, resp. $\overline{Y_{n_2}}$, la proportion aléatoire de spectateurs interrogés dans la salle 1, resp. salle 2, ayant une mauvaise opinion. Les effectifs $n = n_1 = n_2 = 100$ sont suffisamment grands pour appliquer le TCL. La statistique de test est

$$T_n = \frac{\overline{X_n} - \overline{Y_n}}{\sqrt{P_n(1-P_n)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

où sous H_0 $P_n = (n_1\overline{X_{n_1}} + n_2\overline{Y_{n_2}})/(n_1 + n_2)$ est un estimateur de la proportion commune p de spectateurs interrogés ayant une mauvaise opinion salle 1 et 2 confondues. Sous H_0, $T_n \approx \mathcal{N}(0,1)$.

Règle de décision :

▷ si $|T_n| > c_\alpha$ on rejette H_0,
▷ si $|T_n| \leq c_\alpha$, on accepte H_0,
où le seuil c_α est choisi tel que $\mathbb{P}_{H_0}(|T_n| > c_\alpha) = \alpha = 5\%$; la valeur de $c_\alpha = 1,96$ est lue dans la table de la $\mathcal{N}(0,1)$.

Mise en oeuvre : soient $\overline{x_{n_1}}$, $\overline{y_{n_2}}$, p_n les réalisations de $\overline{X_{n_1}}$, $\overline{Y_{n_2}}$, P_n (ce sont des estimations ponctuelles de p_1, p_2, p). $\overline{x_{n_1}} = 30/100$, $\overline{y_{n_2}} = 48/100$, $p_n = (30+48)/200$.

Ainsi, $t_n = (\overline{x_{n_1}} - \overline{y_{n_2}})/\sqrt{p_n(1-p_n)(1/100+1/100)} = -2.61$. $|t_n| = 2.61 > 1.96 = c_\alpha$ donc on rejette H_0 à 5% : il y a une différence significative entre les deux salles (il y a une influence significative de la salle sur l'opinion).

Degré de signification :

$$\alpha_s = \mathbb{P}_{H_0}(|T_n| > 2.61) = 2(1 - \mathbb{P}_{H_0}(T \leq 2.61)) = 2(1 - 0.995) = 0.01.$$

3.3 Comparaison de deux variances (cas gaussien)

On souhaite comparer les variances σ_1^2 et σ_2^2 d'un même caractère dans deux populations différentes. Nous ne traiterons que le cas où le caractère est distribué dans les deux populations suivant une loi normale. On considère deux échantillons gaussiens $(X_1, \ldots, X_{n_1})$, $X_i \sim \mathcal{N}(\mu_1, \sigma_1)$ et $(Y_1, \ldots, Y_{n_2})$, $Y_i \sim \mathcal{N}(\mu_2, \sigma_2)$. Soient $\overline{X_{n_1}} = \sum_{i=1}^{n_1} X_i/n_1$ estimateur de μ_1, $\overline{Y_{n_2}} = \sum_{i=1}^{n_2} Y_i/n_2$ estimateur de μ_2, $S_1^2 = \sum_{i=1}^{n_1}(X_i - \overline{X_{n_1}})^2/(n_1-1)$ estimateur de σ_1^2, et $S_2^2 = \sum_{i=1}^{n_2}(Y_i - \overline{Y_{n_2}})^2/(n_2-1)$ estimateur de σ_2^2.

On veut comparer σ_1^2 et σ_2^2 : à partir des variances observées S_1^2 et S_2^2, peut-on dire si $\sigma_1^2 = \sigma_2^2$? On va donc construire le test de $H_0 : \sigma_1^2 = \sigma_2^2$ contre $H_1 : \sigma_1^2 \neq \sigma_2^2$.

Sous H_0, $\sigma_1^2 = \sigma_2^2$, *i.e.* $\sigma_1^2/\sigma_2^2 = 1$. Le test va alors être fondé sur le rapport S_1^2/S_2^2 que l'on comparera à la valeur de référence 1. Si ce rapport est significativement différent de 1, on rejettera H_0, sinon on acceptera H_0.

Quelle est la loi de la statistique de test $T = S_1^2/S_2^2$ sous H_0 ? Puisque les échantillons sont gaussiens, on sait que

$$\frac{(n_1-1)S_1^2}{\sigma_1^2} \sim \chi^2(n_1-1) \quad \text{et} \quad \frac{(n_2-1)S_2^2}{\sigma_2^2} \sim \chi^2(n_2-1).$$

De plus, ces deux variables aléatoires sont indépendantes.

Rappel 8 : Soient X, Y deux variables aléatoires indépendantes tq $X \sim \chi^2(d_1)$ et $Y \sim \chi^2(d_2)$. Alors la variable aléatoire $(X/d_1)/(Y/d_2) \sim \mathcal{F}(d_1, d_2)$ (loi de Fisher à d_1 et d_2 degrés de liberté). La loi de Fisher est une loi non symétrique et une variable aléatoire de Fisher ne prend que des valeurs positives.

Donc on a

$$\frac{\frac{(n_1-1)S_1^2}{\sigma_1^2}/(n_1-1)}{\frac{(n_2-1)S_2^2}{\sigma_2^2}/(n_2-1)}\text{"} = \text{"}\frac{\chi^2(n_1-1)/(n_1-1)}{\chi^2(n_2-1)/(n_2-1)}\text{"} = \text{"}\mathcal{F}(n_1-1, n_2-1).$$

Or sous H_0, $\sigma_1^2 = \sigma_2^2$, donc $T = S_1^2/S_2^2 \sim \mathcal{F}(n_1 - 1, n_2 - 1)$.

Règle de décision :
- ▷ si $T < c_\alpha$ ($c_\alpha < 1$) ou si $T > d_\alpha$ ($d_\alpha > 1$) on rejette H_0,
- ▷ si $c_\alpha \leq T \leq d_\alpha$, on accepte H_0.

Les seuils c_α et d_α sont choisis tels que

$$\mathbb{P}_{H_0}(\text{rejetter } H_0) = \mathbb{P}_{H_0}(T < c_\alpha \text{ ou } T > d_\alpha) = \mathbb{P}_{H_0}(T < c_\alpha) + \mathbb{P}_{H_0}(T > d_\alpha) = \alpha$$

fixé à l'avance. Or sous H_0, $T \sim \mathcal{F}(n_1 - 1, n_2 - 1)$, donc les valeurs de c_α et d_α sont lues dans la table de la Fisher $(n_1 - 1, n_2 - 1)$ tels que $\mathbb{P}(\mathcal{F}(n_1 - 1, n_2 - 1) < c_\alpha) = \alpha/2$ et $\mathbb{P}(\mathcal{F}(n_1 - 1, n_2 - 1) > d_\alpha) = \alpha/2$, *i.e.* $\mathbb{P}(\mathcal{F}(n_1 - 1, n_2 - 1) \leq d_\alpha) = 1 - \alpha/2$.

Remarque. Il y a donc deux bornes à lire dans la table de la Fisher. La valeur de d_α (> 1) se lit directement dans la table de $\mathcal{F}(n_1 - 1, n_2 - 1)$. En revanche, la lecture de c_α pose plus de problème car $c_\alpha < 1$. Mais on rappelle que si $F \sim \mathcal{F}(d_1, d_2)$, alors $\frac{1}{F} \sim \mathcal{F}(d_2, d_1)$, et donc pour lire une valeur plus petite que 1, il suffit de prendre l'inverse de la valeur lue dans la table $\mathcal{F}(d_2, d_1)$ où l'on a inversé les degrés de liberté. Ainsi, si $F \sim \mathcal{F}(d_1, d_2)$ et si $a < 1$, $\mathbb{P}(F < a) = P(1/F > 1/a) = \alpha/2$ et $\mathbb{P}(1/F \leq 1/a) = 1 - \alpha/2$.

Remarque. Il est bien entendu possible de lire les valeurs des deux seuils c_α et d_α. Mais on peut ne lire qu'une seule valeur, celle située du même côté que 1 que le rapport observé $t = s_1^2/s_2^2$. En effet si $t > 1$, il suffit de lire d_α puisque l'on sait que $c_\alpha < 1$: si $t > d_\alpha$ on rejettera H_0, et si $t < d_\alpha$, on sait aussi que $c_\alpha < 1 < t < d_\alpha$, et on acceptera H_0. De même si $t < 1$, il suffit de lire la valeur c_α puisque l'on sait que $d_\alpha > 1$. Attention, même si on ne peut lire la valeur que d'une seule borne pour conclure, la règle de décision est bien :
- ▷ si $T < c_\alpha$ ou si $T > d_\alpha$ on rejette H_0,
- ▷ si $c_\alpha < T < d_\alpha$, on accepte H_0.

En pratique, on calculera au départ les réalisations s_1^2 et s_2^2 des variances : si $s_1^2 > s_2^2$ on utilisera le rapport $t = s_1^2/s_2^2$ qui sera plus grand que 1, il suffira alors de lire la valeur de d_α dans la table $\mathcal{F}(n_1 - 1, n_2 - 1)$. Si $s_1^2 < s_2^2$ on utilisera le rapport $t = s_2^2/s_1^2$ qui sera plus grand que 1, il suffira alors de lire la valeur de d_α mais cette fois dans la table $\mathcal{F}(n_2 - 1, n_1 - 1)$. Ainsi pour ne calculer que la borne supérieure d_α (plus simple à lire dans la table), on choisit au prélable *le bon rapport à calculer*, *i.e.* celui où la variance la plus grande est au numérateur (en prenant garde d'utiliser les bons degrés de liberté).

Exemple 1

Tester l'égalité des variances dans l'exemple du test de comparaison de moyennes du cas gaussien sur "les vaches hollandaises et normandes".

Exemple 2

Dans deux Unités d'Enseignement et de Recherche (UER), $U1$ et $U2$, de psychologie, on suppose

que les notes de statistiques des étudiants suivent des lois normales. On observe les résultats suivants sur un échantillon de 25 étudiants de l'UER $U1$ et de 10 étudiants de l'UER $U2$:

$$\sum_{i=1}^{25} x_i = 310 \quad \sum_{i=1}^{25} x_i^2 = 3916 \quad \sum_{i=1}^{10} y_i = 129 \quad \text{et} \quad \sum_{i=1}^{10} y_i^2 = 1709.1$$

où x_i (respectivement y_i) désigne la note obtenue par le $i^{ème}$ étudiant de l'échantillon de $U1$ (respectivement $U2$).
Peut-on dire que les variances des notes de statistiques dans les deux UER sont significativement différentes ? (on prendra un risque de 5%).