

Chapitre 3: Les tests

Christelle Dumas

Contents

1	Concepts	3
1.1	Définitions	3
1.2	Deux types d'erreurs	4
1.3	Tests unilatéraux et bilatéraux	7
2	Test sur la moyenne μ d'une population	9
2.1	Test sur la moyenne μ lorsque la variance est connue	10
	2.1.1 Méthode	10
	2.1.2 Application	12
2.2	Mémento-Démarche pour exécuter un test	13
2.3	Test sur la moyenne μ lorsque la variance est inconnue	14
2.4	Test d'égalité de 2 moyennes	15
	2.4.1 Distribution d'échantillonnage de la différence de deux moyennes	16
	2.4.2 Test d'égalité de deux moyennes	18
	2.4.3 Application	19

3	Test sur une proportion	20
3.1	Test d'égalité d'une proportion à une valeur	21
3.1.1	Méthode	21
3.1.2	Application	22
3.2	Test d'égalité de deux proportions	22
3.2.1	Distribution d'échantillonnage de la différence de deux proportions	23
3.2.2	Test d'égalité de deux proportions	23
3.2.3	Application	25
4	Test sur une corrélation	25
4.1	Estimation de la corrélation entre deux variables	26
4.1.1	Rappels et propriétés sur le coefficient de corrélation	26
4.1.2	Application: estimation d'une corrélation	27
4.2	Test d'hypothèse sur une corrélation	28

Introduction

Les tests d'hypothèse constituent un autre aspect important de l'inférence statique.

Le principe général est le suivant : considérons une population P dont les éléments possèdent un caractère statistique mesurable ou dénombrable et dont la valeur du paramètre θ , relative au caractère étudié, est inconnue. **Une hypothèse est formulée sur la valeur de ce paramètre θ** (sur la base de considérations théoriques ou pratiques) et **l'on cherche alors à porter un jugement sur cette hypothèse sur la base d'un échantillon prélevé de cette population.**

Si par exemple, on se demande si la consommation de tabac a diminué après une hausse du prix de la cartouche, on compare la proportion de fumeurs avant et après. On trouve une diminution de 4%. Est-ce que ces 4% peuvent être imputés à l'échantillon particulier que nous avons interrogé ou est-ce

qu'ils correspondent à un réel changement? (dans le dernier cas, on parlera de changement statistiquement significatif).

Ainsi, en raison des fluctuations d'échantillonnage, la statistique d'échantillon servant d'estimation au paramètre de la population ne prendra pas une valeur rigoureusement égale à la valeur théorique proposée dans l'hypothèse. Ainsi, **pour décider si l'hypothèse formulée est supportée ou non empiriquement, il faut une méthode qui permettra de conclure si l'écart observé entre la valeur de la statistique obtenue de l'échantillon et celle du paramètre spécifié dans l'hypothèse est trop important pour être uniquement imputable au hasard de l'échantillonnage.**

Précisément, la construction d'un test d'hypothèse consiste à déterminer entre quelles valeurs peut varier la statistique (comme par exemple l'estimation du paramètre) en supposant l'hypothèse vraie et sur la base d'un échantillon particulier. Si l'estimation effectivement obtenue n'est pas comprise entre ces valeurs alors on considèrera que l'hypothèse n'est pas valide.

1 Concepts

1.1 Définitions

Définissons pour commencer les concepts **d'hypothèse statistique** et **de test d'hypothèse**.

Définition 1 *Une hypothèse statistique est un énoncé (une affirmation) concernant les caractéristiques (valeurs des paramètres, forme de la distribution) d'une population.*

Définition 2 *Un test d'hypothèse (ou test statistique) est une démarche qui a pour but de fournir une règle de décision permettant, sur la base de résultats d'échantillon, de faire un choix entre deux hypothèses statistiques.*

Bien qu'il soit possible de tester l'adéquation d'une variable aléatoire à une loi (normale, par exemple), nous nous contenterons dans ce cours de traiter des tests dits paramétriques et qui portent uniquement sur la valeur de certains paramètres.

Les hypothèses statistiques qui sont envisagées a priori s'appellent **l'hypothèse nulle**, notée H_0 , et **l'hypothèse alternative** (ou contraire, rivale), notée H_1 ou H_a .

L'hypothèse nulle est une affirmation sur un paramètre de la population. Par exemple, est-ce que la valeur du paramètre θ de la population peut-être égale à θ_0 ? Dans ce cas, l'hypothèse nulle sera formulée comme suit :

$$H_0 : \theta = \theta_0$$

C'est l'hypothèse sur laquelle le statisticien se base en l'absence de preuve du contraire. Si on reprend l'exemple de l'introduction, θ représentera la différence de proportions de fumeurs dans la population entre les deux dates et θ_0 sera égal à 0. L'hypothèse nulle est qu'il n'y a pas de changement et on regarde si l'on trouve des indications statistiques d'un changement.

La définition de l'hypothèse nulle est toujours accompagnée de **la définition d'une hypothèse contraire**, H_1 , qui peut être par exemple:

$$H_1 : \theta \neq \theta_0 \quad \text{ou} \quad H_1 : \theta > \theta_0 \quad \text{ou} \quad H_1 : \theta < \theta_0$$

Cela permet de définir quelles sont les autres possibilités que l'on envisage pour la valeur de θ . De nouveau, dans l'exemple, définir l'hypothèse alternative revient à définir ce que l'on envisage comme possibilités pour la proportion de fumeurs dans la population: qu'elle ait changé (dans n'importe quel sens: $\theta \neq 0$)? qu'elle ait augmenté ($\theta > 0$)? qu'elle ait diminué ($\theta < 0$)?

1.2 Deux types d'erreurs

Par définition, **un test statistique est une procédure pour rejeter ou non l'hypothèse nulle, H_0 , à partir d'un échantillon particulier de taille n .**

Il convient cependant de remarquer que la décision de retenir l'hypothèse nulle (ou l'hypothèse alternative) est basée sur une information partielle, à savoir les résultats sur un échantillon. Il est donc statistiquement impossible de toujours prendre la bonne décision (à moins d'observer toute la population, ce qui n'est généralement pas le cas).

Un test d'hypothèse peut se ramener à un problème de décision concernant les deux états de l'hypothèse H_0 :

Table 1: États de H_0

H_0 est vraie	H_0 est fausse
-----------------	------------------

et les deux décisions possibles:

Table 2: Décisions

Ne pas rejeter H_0	Rejeter H_0
----------------------	---------------

Puisque l'information est partielle, on ne prendra pas toujours la bonne décision¹, et ce sera notamment le cas lorsque:

- on rejette H_0 alors qu'elle est vraie
- on ne rejette pas H_0 alors qu'elle est fausse

A contrario, on prend la bonne décision lorsque:

- on ne rejette pas H_0 et elle est vraie
- on rejette H_0 alors qu'elle est fausse

La procédure de test consiste à prendre le plus rarement possible une mauvaise décision et donc à réduire autant que possible les risques de rejeter à tort l'hypothèse nulle et de ne pas rejeter l'hypothèse nulle alors qu'elle est fausse. Ceci conduit à la notion de risque ou de probabilité de se tromper. Au tableau suivant, nous présentons les décisions et les risques correspondants:

Il est en fait assez clair que les 2 risques de prendre une des mauvaises décisions sont reliés. Prenons un exemple non statistique pour illustrer cela.

¹Si l'hypothèse porte sur la vraie valeur de la moyenne d'une population, l'estimation de cette moyenne sur un échantillon ne vaut a priori pas exactement le vrai paramètre; or on est obligé de baser le test de l'hypothèse sur cette information partielle puisqu'on en n'a pas d'autre.

Table 3: Décisions et risques associés

	Ne pas rejeter H_0	Rejeter H_0
H_0 est vraie	Bonne décision	Mauvaise décision; erreur de type I; $\alpha = P(\text{rejeter } H_0 H_0 \text{ vraie})$.
H_0 est fausse	Mauvaise décision; erreur de type II; $\beta = P(\text{ne pas rejeter } H_0 H_0 \text{ fausse})$.	Bonne décision

La question est: OJ Simpson est-il coupable de l'assassinat de sa femme? je définis H_0 comme l'hypothèse selon laquelle il est innocent. L'hypothèse alternative (H_1) est le cas où il est coupable. Les risques d'erreur dans une décision de justice sont les suivants:

- risque de première espèce: condamner OJ Simpson alors qu'il est innocent
- risque de seconde espèce: relâcher OJ Simpson alors qu'il est criminel

Il apparaît que si l'on veut réduire le risque de première espèce (risque d'incarcérer OJ Simpson alors qu'il est innocent), on est amené à libérer des accusés contre lesquels on n'a pas suffisamment de preuves et on augmente le risque de laisser impunis des criminels. Inversement, si l'on voulait se protéger contre le risque de laisser courir des criminels, on serait amené à emprisonner des innocents. Diminuer le risque de 1ère espèce conduit donc à augmenter celui de 2ème espèce.

Dans la mesure où on ne peut pas à la fois diminuer le risque de seconde espèce et celui de première espèce, on fait le choix en statistique de privilégier le contrôle du risque de 1ère espèce (c'est-à-dire le risque de rejeter H_0 à tort). Ceci induit une dissymétrie de traitement entre les deux types d'erreurs et donc entre l'hypothèse nulle et l'hypothèse alternative. L'hypothèse nulle est l'hypothèse conservatrice: on fait apparaître une préférence pour l'hypothèse nulle qui ne sera rejetée que si l'on accumule suffisamment de preuves contre

elle. En pratique, cela ne veut pas dire qu'on réduit à 0 le risque de première espèce (ce n'est pas possible puisqu'il y a toujours des incertitudes dues à l'échantillonnage), mais qu'on le contrôle. Pour reprendre notre parallèle avec la justice, le fait d'être présumé innocent indique que l'hypothèse nulle est l'innocence, qui ne sera rejetée que si l'on accumule suffisamment de preuves de culpabilité. Ce n'est pas du tout équivalent à être présumé coupable et blanchi uniquement s'il y a des preuves d'innocence.

Remarque sémantique: on ne dit pas que l'on "accepte l'hypothèse nulle": en pratique, tout ce que l'on sera en mesure de faire, c'est de ne pas la rejeter: on n'a pas suffisamment de preuves empiriques pour dire qu'elle est fautive et donc pour la rejeter.

Définition 3 *Le risque consenti à l'avance et que nous notons α de rejeter à tort l'hypothèse nulle H_0 alors qu'elle est vraie s'appelle le seuil de significativité du test et s'énonce en probabilité comme suit :*

$$\alpha = \Pr(\text{rejeter } H_0 \mid H_0 \text{ vraie}) = \Pr(\text{choisir } H_1 \mid H_0 \text{ vraie})$$

*On parle également de **risque de première espèce**.*

À ce seuil de significativité, on fait correspondre sur la distribution d'échantillonnage de la statistique **une région de rejet** de l'hypothèse nulle (ou **région critique**). L'aire de cette région correspond à la probabilité α .

Par exemple, si l'on prend $\alpha = 0,05$, cela signifie que l'on admet par avance que la statistique (la variable d'échantillonnage) peut prendre dans 5% des cas une valeur se situant dans la région de rejet de H_0 bien que l'hypothèse H_0 soit vraie.

Sur la distribution d'échantillonnage correspondra aussi une région complémentaire, dite **région de non rejet** (ou **région d'acceptation**). L'aire de cette région correspond à la probabilité $1 - \alpha$.

1.3 Tests unilatéraux et bilatéraux

La définition de l'hypothèse alternative indique si le test effectué est unilatéral ou bilatéral.

Test bilatéral Lorsque l'on s'intéresse au changement du paramètre θ dans l'une ou l'autre des directions (soit $\theta > \theta_0$ ou $\theta < \theta_0$), il convient d'opter pour un test bilatéral dont la formalisation est la suivante:

$$\begin{aligned}H_0 & : \theta = \theta_0 \\H_1 & : \theta \neq \theta_0\end{aligned}$$

C'est par exemple le cas si l'on observe l'évolution d'un candidat à la présidentielle dans les sondages. On observe le changement entre deux dates et on se demande s'il est nul ou s'il a varié.

Il est alors possible de schématiser les régions de rejet et de non-rejet de l'hypothèse H_0 comme suit: inclure graphe.

Si suite à la réalisation de l'échantillon, la valeur de la statistique $T_n(\theta)$ se situe dans l'intervalle $\theta_{c1} \leq T_n(\theta) \leq \theta_{c2}$, il ne sera pas possible de rejeter H_0 au seuil de significativité choisi. En revanche, si $T_n(\theta) > \theta_{c2}$ ou $T_n(\theta) < \theta_{c1}$, l'hypothèse H_0 est rejetée et l'on favorise H_1 .

Test unilatéral Lorsque l'on s'intéresse au changement du paramètre θ dans une seule direction, il convient d'opter pour un test unilatéral. Les hypothèses sont les suivantes si l'on s'intéresse à un changement du côté gauche:

$$\begin{aligned}H_0 & : \theta = \theta_0 \\H_1 & : \theta < \theta_0\end{aligned}$$

Il est alors possible de schématiser ce test unilatéral à gauche de la façon suivante: inclure graphe.

L'hypothèse H_0 sera alors rejetée lorsque $T_n(\theta) < \theta_c$ (il convient de favoriser H_1 dans ce cas).

Les hypothèses sont les suivantes si l'on s'intéresse à un changement du côté droit:

$$\begin{aligned}H_0 & : \theta = \theta_0 \\H_1 & : \theta > \theta_0\end{aligned}$$

Il est alors possible de schématiser ce test unilatéral à droite de la façon suivante: inclure graphe.

L'hypothèse H_0 sera alors rejetée lorsque $T_n(\theta) > \theta_c$ (il convient de favoriser H_1 dans ce cas).

À quelle intuition correspondent les tests unilatéraux? Dans le cas où l'on se demandait si la proportion de fumeurs avait diminué au cours du temps, l'hypothèse nulle est qu'il n'y a pas de changement alors que l'hypothèse alternative est que $\theta < 0$. Il s'agit donc d'un test unilatéral à gauche. Lorsqu'on se demande si le taux de CO_2 dans l'air est supérieur à un niveau limite de pollution, on fait un test unilatéral à droite.

2 Test sur la moyenne μ d'une population

Considérons une population P , de dimension N , de moyenne μ et de variance σ^2 . Supposons que nous disposions d'un échantillon de grande taille (dans ce cas, il importe peu que la population soit distribuée normalement ou selon une loi inconnue). À partir d'un échantillon aléatoire de taille n , un estimateur de la moyenne est donné par la moyenne d'échantillonnage \bar{X}_n définie comme suit:

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

La fonction de distribution de \bar{X} dépend de la connaissance, ou non, de la variance σ^2 . En conséquence, pour ce test d'hypothèse, deux cas doivent être distingués :

- variance de la population σ^2 connue;
- variance de la population σ^2 inconnue;

Nous commençons donc par traiter ces deux cas dans le cadre du test de l'égalité d'une moyenne à une valeur; puis nous passons au test d'égalité de deux moyennes.

2.1 Test sur la moyenne μ lorsque la variance est connue

2.1.1 Méthode

Supposons que nous sommes dans la situation où **nous voulons soumettre au test l'hypothèse nulle selon laquelle la moyenne μ est égale à une valeur particulière m_0 contre l'hypothèse alternative qu'elle diffère de m_0** , il vient:

$$\begin{aligned}H_0 &: \mu = m_0 \\H_1 &: \mu \neq m_0\end{aligned}$$

Considérons également que la population est distribuée normalement de variance connue σ^2 . Nous prélevons de cette population, un échantillon de taille n . La statistique adéquate est donc la moyenne d'échantillonnage \bar{X}_n . Nous savons d'après les chapitres précédents que :

$$\bar{X}_n \rightsquigarrow \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right) \Leftrightarrow Z = \frac{\bar{X}_n - \mu}{\sqrt{\frac{\sigma^2}{n}}} \rightsquigarrow \mathcal{N}(0, 1)$$

En supposant l'hypothèse H_0 vraie alors nous avons :

$$Z = \frac{\bar{X}_n - m_0}{\sqrt{\frac{\sigma^2}{n}}} \rightsquigarrow \mathcal{N}(0, 1)$$

Fixons maintenant le seuil de significativité du test à la valeur α , i.e. fixons la probabilité de rejeter H_0 alors que H_0 est vraie ou plus simplement :

$$\Pr[\text{rejeter } H_0 | H_0 \text{ vraie}] = \alpha$$

Ayant fixé, le seuil de significativité, **il est alors possible à l'aide de la distribution d'échantillonnage de la statistique \bar{X}_n de déterminer les valeurs critiques \bar{x}_1 et \bar{x}_2** telle que l'intervalle $\bar{x}_1 \leq \bar{X}_n \leq \bar{x}_2$ constitue la région de non rejet de H_0 et les deux extrémités de la distribution $\bar{X}_n < \bar{x}_1$ et $\bar{X}_n > \bar{x}_2$ constituent la région de rejet de H_0 . Graphiquement, nous obtenons quelque chose de la forme: inclure graphe.

La région de non-rejet est telle que:

$$\Pr [\text{non-rejet de } H_0 | H_0 \text{ vraie}] = \Pr [\bar{x}_1 \leq \bar{X}_n \leq \bar{x}_2 | H_0 \text{ vraie}] = 1 - \alpha$$

et

$$\begin{aligned} \Pr [\text{rejeter } H_0 | H_0 \text{ vraie}] &= \Pr [\bar{X}_n < \bar{x}_1 | H_0 \text{ vraie}] + \Pr [\bar{X}_n > \bar{x}_2 | H_0 \text{ vraie}] \\ &= \frac{\alpha}{2} + \frac{\alpha}{2} = \alpha \end{aligned}$$

Il est évidemment possible (voire usuel) de réécrire ses relations sous forme centrée-réduite. Il vient alors:

$$\begin{aligned} \Pr [\bar{x}_1 \leq \bar{X}_n \leq \bar{x}_2 | H_0 \text{ vraie}] &= \Pr \left[\frac{\bar{x}_1 - m_0}{\sqrt{\frac{\sigma^2}{n}}} \leq \frac{\bar{X}_n - m_0}{\sqrt{\frac{\sigma^2}{n}}} \leq \frac{\bar{x}_2 - m_0}{\sqrt{\frac{\sigma^2}{n}}} \middle| H_0 \text{ vraie} \right] \\ &= 1 - \alpha \end{aligned}$$

Puisque l'aire sous la distribution d'échantillonnage de \bar{X}_n est fixée à $\alpha/2$ à chaque extrémité de la distribution, il en sera de même aux extrémités de la distribution de l'écart-réduit. Il s'agit donc de lire la table de la loi normale centrée réduite $z_{\alpha/2}$ de telle sorte que :

$$\Pr \left[-z_{\alpha/2} \leq \frac{\bar{X}_n - m_0}{\sqrt{\frac{\sigma^2}{n}}} \leq z_{\alpha/2} \right] = 1 - \alpha$$

où $-z_{\alpha/2}$ et $z_{\alpha/2}$ sont les valeurs critiques de l'écart-réduit.

Puisque $-z_{\alpha/2} = \frac{\bar{x}_1 - m_0}{\sqrt{\frac{\sigma^2}{n}}}$ et $z_{\alpha/2} = \frac{\bar{x}_2 - m_0}{\sqrt{\frac{\sigma^2}{n}}}$, il est aisé de déduire les valeurs critiques de \bar{X}_n , nous aurons alors :

$$\begin{aligned} \bar{x}_1 &= m_0 - z_{\alpha/2} \sqrt{\frac{\sigma^2}{n}} \\ \bar{x}_2 &= m_0 + z_{\alpha/2} \sqrt{\frac{\sigma^2}{n}} \end{aligned}$$

où m_0, σ^2, n et $z_{\alpha/2}$ sont connus.

Il nous reste donc **à déterminer les règles de décision du test**. Ces règles de décisions s'énoncent comme suit pour les hypothèse $H_0 : \mu = m_0$ et $H_1 : \mu \neq m_0$:

- Si l'on utilise les valeurs critiques de \bar{X}_n , nous adoptons alors la règle de décision suivante :

$$\text{Rejeter } H_0 \text{ si } \bar{X}_n > \bar{x}_1 = m_0 - z_{\alpha/2} \sqrt{\frac{\sigma^2}{n}}$$

$$\text{Rejeter } H_0 \text{ si } \bar{X}_n < \bar{x}_2 = m_0 + z_{\alpha/2} \sqrt{\frac{\sigma^2}{n}}$$

Accepter H_0 sinon

Dans le premier cas, on rejette parce que la moyenne estimée (\bar{X}_n) prend une valeur trop grande par rapport à m_0 et dans le second trop petite.

- Si l'on utilise les valeurs critiques de Z , nous adoptons alors la règle de décision suivante :

$$\text{Rejeter } H_0 \text{ si } Z > z_{\alpha/2}$$

$$\text{Rejeter } H_0 \text{ si } Z < -z_{\alpha/2}$$

Accepter H_0 sinon

La règle de décision consiste donc à préciser de combien la moyenne d'échantillon peut s'écarter de m_0 , pour un seuil α et une taille d'échantillon n , pour que H_0 ou H_1 soient respectivement considérées comme dépourvues de soutien expérimental.

Ainsi, si l'écart observé $(\bar{x} - m_0)$ est plus grand, en valeur absolue que $(\bar{x}_1 - m_0)$ ou $(\bar{x}_2 - m_0)$, nous dirons que la différence $(\bar{x} - m_0)$ est statistiquement significative au seuil α . Cette hypothèse est alors anormalement élevée et permet de rejeter l'hypothèse H_0 .

A contrario, $(\bar{x} - m_0)$ est plus petit, en valeur absolue que $(\bar{x}_1 - m_0)$ ou $(\bar{x}_2 - m_0)$, nous dirons que la différence $(\bar{x} - m_0)$ est statistiquement non significative au seuil α . La différence est alors imputable aux seules fluctuations d'échantillonnages et l'on ne peut pas rejeter l'hypothèse nulle.

2.1.2 Application

Considérons un exemple simple. Nous nous intéressons à la moyenne du nombre de vêtements achetés par les individus de 18 à 25 ans en une année. Nous

voulons tester au seuil de significativité $\alpha = 0.05$, les hypothèses statistiques suivantes:

$$H_0 : \mu = 12$$

$$H_1 : \mu \neq 12$$

en prélevant au hasard un échantillon de taille $n = 9$ d'une population normale de variance $\sigma^2 = 4$. Entre quelles valeurs doit se situer la moyenne d'échantillonnage pour considérer au seuil $\alpha = 0.05$ l'hypothèse H_0 comme vraisemblable?

La statistique pertinente est \bar{X}_n ou encore en se ramenant à la loi normale centrée réduite et en supposant l'hypothèse H_0 vraie $Z = \frac{\bar{X}_n - m_0}{\sqrt{\frac{\sigma^2}{n}}}$ avec $m_0 =$

12. Nous savons que $\Pr \left[-z_{\alpha/2} \leq \frac{\bar{X}_n - m_0}{\sqrt{\frac{\sigma^2}{n}}} \leq z_{\alpha/2} \right] = 0.95$. A partir de la table de la loi normale centrée réduite, nous obtenons $z_{\alpha/2} = 1.96$. Il est alors très facile de déterminer les valeurs critiques pour \bar{X}_n , il vient :

$$\bar{x}_1 = m_0 - z_{\alpha/2} \sqrt{\frac{\sigma^2}{n}} = 12 - 1,96 \cdot \frac{2}{3} = 10,69$$

$$\bar{x}_2 = m_0 + z_{\alpha/2} \sqrt{\frac{\sigma^2}{n}} = 12 + 1,96 \cdot \frac{2}{3} = 13,31$$

La règle de décision suivante est alors adoptée pour vérifier la crédibilité de l'hypothèse H_0 sur la base d'un échantillon de taille $n = 9$ et au seuil de significativité $\alpha = 0.05$, nous avons :

$$\begin{aligned} & \text{rejeter } H_0 \text{ si } \bar{X}_n < 10.69 \text{ ou si } \bar{X}_n > 13.31 \\ & \text{ne pas rejeter } H_0 \text{ si } 10.69 \leq \bar{X}_n \leq 13.31 \end{aligned}$$

2.2 Mémento-Démarche pour exécuter un test

Un test d'hypothèse est une procédure en plusieurs étapes. Par convention, nous choisirons de travailler systématiquement avec l'écart-réduit. Dans l'exécution d'un test d'hypothèse, nous proposons la démarche suivante en 7 étapes :

- (i) Formuler l'hypothèse nulle H_0 et l'hypothèse alternative H_1 ;

- (ii) Fixer d'avance (avant la réalisation du sondage) le seuil de significativité α , i.e. spécifier le risque de rejeter à tort une hypothèse H_0 vraie;
- (iii) Préciser les conditions d'application du test. Spécification ou non de la forme de la population échantillonnée, indication si nous sommes en présence d'un grand échantillon, si la variance de la population est connue ou inconnue etc...;
- (iv) Spécifier la statistique qui convient pour le test et définir l'écart-réduit. En déduire sa distribution d'après les conditions d'application;
- (v) Adopter une règle de décision qui conduira au rejet ou au non-rejet de H_0 au seuil α ;
- (vi) calculer la valeur numérique de l'écart réduit, valeur déduite des résultats de l'échantillon;
- (vii) Décision et conclusion. Comparer la valeur numérique obtenue pour l'écart-réduit avec la règle de décision adoptée en (v). Décider entre les deux hypothèses formulées en (i) et conclure;

2.3 Test sur la moyenne μ lorsque la variance est inconnue

Le principe est le même lorsque la variance est inconnue. Il est seulement nécessaire dans ce cas de remplacer la variance σ^2 par son estimateur. La variable centrée réduite sera de la forme :

$$\frac{\bar{X}_n - m_0}{\sqrt{\frac{S_n^2}{n}}}$$

où S_n^2 est l'estimateur de la variance lorsque la moyenne est inconnue. Lorsque la variance σ^2 est inconnue, il convient de la remplacer pour son estimateur. Nous avons vu qu'un estimateur centré de la variance lorsque la moyenne de la population est inconnue est défini par :

$$S_n^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$$

et nous avons montré que :

$$\frac{(n-1)S_n^2}{\sigma^2} \rightsquigarrow \chi_{n-1}^2$$

À ce stade, il est nécessaire de distinguer deux cas selon la taille de l'échantillon.

- Si l'échantillon est de petite taille ($n < 30$) alors la statistique vérifie :

$$T = \frac{\bar{X}_n - m_0}{\sqrt{\frac{S_n^2}{n}}} \rightsquigarrow T_{n-1}$$

et est distribuée selon une loi de Student avec $\nu = n - 1$ degrés de liberté.

- Si l'échantillon est de grande taille ($n > 30$) alors la statistique vérifie :

$$Z = \frac{\bar{X}_n - m_0}{\sqrt{\frac{S_n^2}{n}}} \rightsquigarrow \mathcal{N}\left(0, \frac{n-1}{n-3}\right)$$

Par approximation, il est alors possible de se ramener à $n - 1 \simeq n - 3$ et nous avons alors :

$$Z = \frac{\bar{X}_n - m_0}{\sqrt{\frac{S_n^2}{n}}} \rightsquigarrow N(0, 1)$$

2.4 Test d'égalité de 2 moyennes

Il est très fréquent de vouloir comparer deux moyennes; c'est notamment le cas si l'on veut savoir si les garçons achètent plus de vêtements que les filles. On va estimer la moyenne du nombre de vêtements achetés pour la population des filles, puis pareil pour les garçons et on se demande si les deux moyennes sont égales. Comme précédemment, on se demande si la différence observée est due à l'aléa de l'échantillonnage ou à une réelle différence.

2.4.1 Distribution d'échantillonnage de la différence de deux moyennes

On peut assez facilement se ramener aux cas précédents en considérant que le paramètre sous-jacent est la différence entre la moyenne pour les filles et la moyenne pour les garçons. Commençons donc par décrire la distribution de l'estimateur de la différence entre deux moyennes. Si l'on estime la moyenne d'un caractère dans une population 1 par la moyenne sur un échantillon de cette population (\bar{X}_1) et de même pour la moyenne d'une population 2 (\bar{X}_2) alors la différence entre les deux moyennes peut être estimée par $\bar{X}_1 - \bar{X}_2$. On veut connaître la distribution de cette statistique pour pouvoir faire des tests dessus.

Trois cas peuvent se présenter:

- populations normales et variances σ_1^2 et σ_2^2 connues;
- grands échantillons ($n_1 \geq 30$ et $n_2 \geq 30$), variances σ_1^2 et σ_2^2 inconnues;
- populations normales, un ou les deux échantillons petits, variances des populations inconnues mais supposées égales ($\sigma_1^2 = \sigma_2^2 = \sigma^2$).

Cas 1. Populations normales et variances σ_1^2 et σ_2^2 connues On prélève au hasard et indépendamment deux échantillons de tailles n_1 et n_2 , respectivement de deux populations normales dont les éléments possèdent un caractère mesurable de paramètres μ_1 et σ_1^2 sur la population 1, μ_2 et σ_2^2 sur la population 2 où μ_1 et μ_2 sont inconnues.

La différence de moyennes échantillonnales $\bar{X}_1 - \bar{X}_2$:

- est distribuée normalement;
- a pour espérance $E(\bar{X}_1 - \bar{X}_2) = \mu_1 - \mu_2$;
- a pour variance $V(\bar{X}_1 - \bar{X}_2) = V(\bar{X}_1) + V(\bar{X}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$ (indépendance des échantillons).

Par conséquent, l'écart-réduit:

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \rightsquigarrow \mathcal{N}(0, 1)$$

Cas 2. Grands échantillons ($n_1 \geq 30$ et $n_2 \geq 30$), **variances σ_1^2 et σ_2^2 inconnues** On prélève au hasard et indépendamment deux échantillons de grandes tailles $n_1 \geq 30$ et $n_2 \geq 30$ respectivement de deux populations dont les éléments possèdent un caractère mesurable de paramètres μ_1 et σ_1^2 sur la population 1, μ_2 et σ_2^2 sur la population 2 qui sont tous inconnus.

La différence de moyennes échantillonnales $\bar{X}_1 - \bar{X}_2$:

- est distribuée de façon approximativement normale;
- a pour espérance $E(\bar{X}_1 - \bar{X}_2) = \mu_1 - \mu_2$;
- a pour variance estimée $V(\bar{X}_1 - \bar{X}_2) = V(\bar{X}_1) + V(\bar{X}_2) = \frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}$.

Par conséquent, l'écart-réduit:

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \rightsquigarrow \mathcal{N}(0, 1)$$

(Il s'agit d'une loi de Student qui peut s'approximer par une loi normale du fait de la grande taille des échantillons).

Populations normales, un ou les deux échantillons petits, variances des populations inconnues mais supposées égales ($\sigma_1^2 = \sigma_2^2 = \sigma^2$) Il arrive fréquemment qu'on ne dispose pas de grands échantillons. On veut toujours comparer deux moyennes mais l'un ou les deux échantillons est petit. Si les échantillons proviennent de populations normales, on est dans le cadre d'applicabilité du cas 1 mais les variances sont inconnues. Une modification importante doit être apportée: il faut faire l'hypothèse que les variances des deux populations, bien qu'inconnues, soient égales. Dans le cas de petits échantillons, on ne peut remplacer σ_1^2 et σ_2^2 par leurs estimations calculées sur chacun des échantillons car elles seront peu précises; il vaut mieux combiner les deux échantillons, puisqu'on suppose les deux variances égales à une valeur commune, pour obtenir une estimation unique s_C^2 de la variance commune σ^2 .

On obtient cette estimation en combinant la variabilité observée dans chaque échantillon comme suit:

$$\begin{aligned} s_C^2 &= \frac{\sum(x_{i1} - \bar{x}_1)^2 + \sum(x_{i2} - \bar{x}_2)^2}{(n_1 - 1) + (n_2 - 1)} \\ &= \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \end{aligned}$$

La variance de $\bar{X}_1 - \bar{X}_2$ s'écrit alors:

$$V(\bar{X}_1 - \bar{X}_2) = s_C^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right).$$

Pour résumer, la différence de moyennes échantillonales $\bar{X}_1 - \bar{X}_2$:

- est distribué approximativement normalement;
- a pour espérance $E(\bar{X}_1 - \bar{X}_2) = \mu_1 - \mu_2$;
- a pour variance estimée $V(\bar{X}_1 - \bar{X}_2) = s_C^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)$.

Par conséquent, l'écart-réduit:

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \rightsquigarrow \mathcal{T}_{n_1+n_2-2}$$

Note: tout comme nous sommes en train de développer une méthode de test d'égalités de moyenne, nous pourrions tester l'égalité des variances dans les deux échantillons au lieu de simplement en faire l'hypothèse. Cependant, pour des soucis de simplicité, nous ne le détaillerons pas dans ce cours, bien que ce soit la procédure standard.

2.4.2 Test d'égalité de deux moyennes

Lorsque l'on veut tester l'égalité de deux moyennes, l'hypothèse nulle est qu'elle sont égales:

$$\begin{aligned} H_0 &: \mu_1 = \mu_2 \\ \Leftrightarrow H_0 &: \mu_1 - \mu_2 = 0 \end{aligned}$$

Les hypothèses alternatives peuvent être de différences ou d'inégalités dans un sens particulier. On voit que l'on peut aisément se ramener au cas du test d'un paramètre égal à 0, en utilisant les résultats énoncés précédemment pour estimer ce paramètre (différence des moyennes entre les deux populations).

Donnons rapidement les règles de décision pour chacun des trois cas énoncés précédemment si l'hypothèse alternative est $H_1 : \mu_1 \neq \mu_2$, on fera en application un cas différent.

Table 4: Tests bilatéraux d'égalité de moyennes

Cas	Statistique	Rejet de H_0 ssi
1: Population normale Variances connues	$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$	ou $Z > z_{\alpha/2}$ $Z < -z_{\alpha/2}$
2: Grands échantillons	$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$	ou $Z > z_{\alpha/2}$ $Z < -z_{\alpha/2}$
3: Population normale Variances supposées égales Petit échantillon	$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2(n_1-1) + s_2^2(n_2-1)}{n_1+n_2-2}} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$	ou $T > t_{\alpha/2; n_1+n_2-2}$ $T < -t_{\alpha/2; n_1+n_2-2}$

2.4.3 Application

Nous proposons ici une application de l'exemple au cas où on veut tester que les garçons achètent autant de vêtements que les filles. L'hypothèse alternative étant qu'ils en achètent moins (au hasard). On veut fournir un test au seuil de 5%. On ne connaît pas les variances mais on les estime et la distribution du nombre de vêtements est supposée normale pour les 2 groupes. La taille de l'échantillon est de 15 couples (donc 15 observations pour les filles et autant pour les garçons).

Définition des hypothèses:

$$H_0 : \mu_1 - \mu_2 = 0$$

$$H_1 : \mu_1 - \mu_2 > 0$$

où 1 représente le groupe des filles et 2 celui des garçons. On devra donc faire un test unilatéral à droite.

Comme la variance des populations est inconnue mais que la loi est normale, la statistique de test va suivre un Student à $n_1 + n_2 - 2 = 30 - 2 = 28$ degrés de liberté. Comme la taille de chaque échantillon est petite, on ne peut pas utiliser l'approximation par loi normale. Par conséquent, sous l'hypothèse H_0 :

$$T = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{s_1^2(n_1-1) + s_2^2(n_2-1)}{n_1+n_2-2}} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \rightsquigarrow T_{n_1+n_2-2}$$

Comme on veut un test unilatéral, on cherche à partir de quelle différence entre la moyenne estimée pour les filles et pour les garçons, on ne peut plus accepter que la moyenne est la même et on est obligé de considérer que la moyenne pour les filles est supérieure à celle pour les garçons. Cela correspond à:

$$\begin{aligned} \alpha &= P(T > t_{n_1+n_2-2;\alpha} | H_0 \text{ vraie}) \\ &= P\left(\overline{X}_1 - \overline{X}_2 > t_{n_1+n_2-2;\alpha} \cdot \sqrt{\frac{s_1^2(n_1-1) + s_2^2(n_2-1)}{n_1+n_2-2}} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right) \end{aligned}$$

Par conséquent, la règle de décision est la suivante:

- si $\overline{X}_1 - \overline{X}_2 > t_{n_1+n_2-2;\alpha} \cdot \sqrt{\frac{s_1^2(n_1-1) + s_2^2(n_2-1)}{n_1+n_2-2}} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$, on rejette l'égalité des moyennes avec une probabilité α de se tromper;
- si $\overline{X}_1 - \overline{X}_2 < t_{n_1+n_2-2;\alpha} \cdot \sqrt{\frac{s_1^2(n_1-1) + s_2^2(n_2-1)}{n_1+n_2-2}} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$, on ne rejette pas l'hypothèse selon laquelle les moyennes sont égales.

Application numérique.

3 Test sur une proportion

On peut appliquer une méthode tout à fait comparable pour tester l'égalité d'une proportion à une valeur ou pour tester l'égalité de deux proportions.

3.1 Test d'égalité d'une proportion à une valeur

C'est ce que l'on veut mettre en œuvre lorsque l'on cherche à répondre à une question du type: "la proportion de fumeurs dans la population est-elle de 15%?".

3.1.1 Méthode

Dans ce cas, l'hypothèse nulle est de la forme:

$$H_0 : p = p_0$$

où p représente la proportion de fumeurs. L'hypothèse alternative considérée peut soit être $H_1 : p \neq p_0$, qui donnera lieu à un test bilatéral, soit $H_1 : p > p_0$ ou $H_1 : p < p_0$, qui donneront lieu à un test unilatéral.

Plaçons-nous dans le cadre où l'on peut appliquer l'approximation par loi normale à l'estimateur de la proportion \hat{p} ; pour cela, il faut que le nombre d'observations soit suffisamment grand pour que $np > 5$ et $n(1 - p) > 5$. Ainsi,

$$\hat{p}_n \rightsquigarrow \mathcal{N}\left(p, \frac{p(1-p)}{n}\right).$$

Sous l'hypothèse H_0 , $p = p_0$ et par conséquent l'écart-réduit s'exprime comme suit et distribué selon une normale centrée réduite:

$$Z = \frac{\hat{p}_n - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \rightsquigarrow \mathcal{N}(0, 1).$$

Cette statistique va nous permettre de tester l'hypothèse nulle. En effet, si l'on fixe le seuil de significativité à α et que l'on considère un test bilatéral, alors il faut rejeter l'hypothèse nulle dès que l'on est dans un des deux cas suivants:

$$Z > z_{\alpha/2} \quad \text{ou} \quad Z < -z_{\alpha/2}$$

où z_α représente le quantile α de la normale centrée réduite. A contrario, si:

$$-z_{\alpha/2} < Z < z_{\alpha/2}$$

alors on ne peut pas rejeter l'hypothèse nulle que $p = p_0$.

La région de non-rejet de l'hypothèse nulle peut aussi s'écrire directement en fonction de la valeur p_0 . En effet, on ne rejette pas si et seulement si:

$$p_0 - z_{\alpha/2} \sqrt{\frac{p_0(1-p_0)}{n}} < \hat{p}_n < p_0 + z_{\alpha/2} \sqrt{\frac{p_0(1-p_0)}{n}}$$

Si l'estimateur de la proportion excède la borne de droite ou est inférieur à celle de gauche alors on rejette l'hypothèse nulle (trop de différence entre l'estimation et la valeur p_0 pour que l'hypothèse nulle soit crédible).

Attention, si l'on veut faire un test unilatéral, il faut adapter les seuils pour le prendre en compte.

3.1.2 Application

Aux dernière élections, un parti politique a obtenu 42% des suffrages. Un récent sondage a révélé que sur 1041 personnes interrogées entre le 26 et le 29 février, 458 accorderaient leur appui à ce parti. Le chef du parti déclara que la popularité de son parti était à la hausse. Que penser de cette affirmation au seuil de signification $\alpha = 0.05$?

- Hypothèses statistiques
- Conditions d'application
- Statistique de test
- Règle de décision
- Application numérique et conclusion

3.2 Test d'égalité de deux proportions

On met en œuvre un test d'égalité de deux proportions lorsque l'on compare les proportions estimées sur deux échantillons et que l'on se demande si elles sont égales. Ceci permet notamment de répondre à des questions du type: "la proportion d'hommes qui mangent des pizzas deux fois par semaine est-elle la même que la proportion de femmes qui mangent des pizzas deux

fois par semaine?” ou “la proportion de fumeurs dans la population a-t-elle diminué entre l’année dernière et cette année?”. Il s’agit de déterminer si l’écart observé entre deux proportions estimées est significatif ou s’il est plutôt attribuable au hasard de l’échantillonnage.

3.2.1 Distribution d’échantillonnage de la différence de deux proportions

Pour juger de la comparabilité de deux proportions, on construit la statistique qui est la différence entre ces deux proportions. Comme dans le cas de la moyenne, il nous faut décrire la distribution de cette statistique pour construire un test à partir de celle-ci.

Nous ne traitons que le cas où nous disposons de grands échantillons. On prélève au hasard et indépendamment deux échantillons de grande taille respectivement de deux populations donc les éléments possèdent, dans une proportion p_1 un certain caractère dans la population 1 et dans une proportion p_2 le même caractère dans la population 2, où p_1 et p_2 sont bien sûr inconnues. L’estimation de ces proportions est donnée par les proportions dans les échantillons, \hat{p}_1 et \hat{p}_2 , d’éléments présentant le caractère.

La différence de proportion $\hat{P}_1 - \hat{P}_2$ est un estimateur de $p_1 - p_2$ et donc une variable aléatoire dont la distribution possède les propriétés suivantes:

- la distribution de $\hat{P}_1 - \hat{P}_2$ est approximativement normale (somme de deux normales);
- l’espérance de $\hat{P}_1 - \hat{P}_2$ est

$$E(\hat{P}_1 - \hat{P}_2) = p_1 - p_2$$

- la variance de la distribution de $\hat{P}_1 - \hat{P}_2$ est

$$V(\hat{P}_1 - \hat{P}_2) = V(\hat{P}_1) + V(\hat{P}_2) = \frac{p_1(1 - p_1)}{n} + \frac{p_2(1 - p_2)}{n}.$$

3.2.2 Test d’égalité de deux proportions

L’hypothèse nulle que l’on veut tester est que les proportions sont les mêmes dans les deux populations:

$$H_0 : p_1 = p_2 \quad \text{ou encore} \quad p_1 - p_2 = 0.$$

Les valeurs p_1 et p_2 sont donc inconnues mais supposées égales à une valeur commune: $p_1 = p_2 = p$. On obtient une estimation de cette valeur commune en combinant les proportions observées dans chaque échantillon comme suit:

$$\hat{p} = \frac{n_1 \hat{p}_1 + n_2 \hat{p}_2}{n_1 + n_2}$$

Sous l'hypothèse nulle, la variance estimée de $\hat{P}_1 - \hat{P}_2$ s'écrit donc:

$$V(\hat{P}_1 - \hat{P}_2) = \hat{p}(1 - \hat{p}) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)$$

où n_1 est la taille de l'échantillon 1 et n_2 la taille de l'échantillon 2.

Par conséquent, sous l'hypothèse nulle, $p_1 - p_2 = 0$ et

$$\hat{P}_1 - \hat{P}_2 \rightsquigarrow \mathcal{N} \left(0, \hat{p}(1 - \hat{p}) \left(\frac{1}{n_1} + \frac{1}{n_2} \right) \right)$$

ainsi l'écart-réduit:

$$Z = \frac{\hat{P}_1 - \hat{P}_2}{\sqrt{\hat{p}(1 - \hat{p}) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

est distribué selon une normale centrée réduite.

Le test de comparaison se conduit ensuite exactement comme d'habitude, avec des régions de rejet qui dépendent de l'hypothèse alternative considérée (voir tableau 5).

Table 5: Test de l'hypothèse $H_0 : p_1 = p_2$

Hypothèse alternative	Rejet de H_0 si
$H_1 : p_1 \neq p_2$	$Z > z_{\alpha/2}$ ou $Z < -z_{\alpha/2}$
$H_1 : p_1 > p_2$	$Z > z_{\alpha}$
$H_1 : p_1 < p_2$	$Z < -z_{\alpha}$

3.2.3 Application

Deux types de publicité sont envisagés par une entreprise pour lancer un nouveau déodorant. Après avoir visionné les deux types de publicité mis au point par des spécialistes en communication, la direction émet l'hypothèse que la publicité de type A sera plus efficace que celle de type B. Deux régions, considérées comme marché-test et possédant sensiblement les mêmes caractéristiques de consommation sont choisies pour évaluer l'efficacité des deux types de publicité: la publicité A sera utilisée dans une région et la B dans l'autre.

Un sondage auprès de 125 individus ayant vu la publicité de type A indique que 44 se sont procurés le nouveau déodorant alors que sur 100 ayant vu la publicité de type B, 32 se sont procurés le nouveau déodorant. À l'aide d'une procédure de test, indiquez si vous souhaitez confirmer ou infirmer l'hypothèse émise par la direction au seuil de signification $\alpha = 0.05$.

4 Test sur une corrélation

Dans le premier chapitre, nous avons défini la corrélation entre deux variables aléatoires qui permettait de mesurer la force d'une dépendance linéaire entre deux variables. Bien que nous ne nous soyons que peu servi de cet outil, la corrélation, en tant que fonction variables aléatoires calculées sur un échantillon est une statistique et donc une variable aléatoire, sur laquelle il est possible de mettre en oeuvre des tests.

Or ceci est particulièrement important dans la mesure où s'il l'on soupçonne un lien statistique entre deux variables, disons, la proportion de fumeurs et le prix du paquet de cigarettes, il ne suffira pas de calculer une corrélation temporelle entre ces variables pour s'assurer que le lien est significatif. Pour adapter l'application précédente par exemple, imaginons que l'on observe une corrélation positive entre l'intensité du matraquage publicitaire et le chiffre de ventes de l'entreprise mais que la corrélation calculée n'est que de 13% sur l'échantillon observé. Dans la mesure où la corrélation est soumise à l'aléa de l'échantillonnage, peut-on considérer que cette corrélation est significativement différente de 0? ou même supérieure à un certain niveau? c'est ce dont il faut s'assurer avant de dépenser des millions en publicité et nous allons voir comment faire dans ce chapitre.

4.1 Estimation de la corrélation entre deux variables

4.1.1 Rappels et propriétés sur le coefficient de corrélation

Pour rappel, nous avons vu que le coefficient de corrélation était défini par:

$$\rho = \frac{Cov(X, Y)}{\sqrt{V(X)V(Y)}} = \frac{E(XY) - E(X)E(Y)}{\sqrt{V(X)V(Y)}} = \frac{E[(X - E(X))(Y - E(Y))]}{\sqrt{V(X)V(Y)}}$$

La contrepartie empirique de ρ sur un échantillon de taille n s'exprime:

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}.$$

Le calcul du coefficient de corrélation sur l'échantillon permet donc d'obtenir une estimation du degré de corrélation linéaire entre deux variables aléatoires X et Y d'une même population. En raison de la symétrie de la définition, il mesure aussi bien l'intensité de la liaison linéaire entre Y et X qu'entre X et Y .

On se souvient aussi du fait que le coefficient de corrélation est compris entre -1 et 1 et que:

- s'il vaut 0, alors il n'y a pas de corrélation linéaire entre les deux variables;
- s'il vaut 1, il y a une dépendance linéaire parfaite et positive entre les deux variables (elles sont proportionnelles et le coefficient de proportionnalité est positif);
- s'il vaut -1, il y a une dépendance linéaire parfaite et négative entre les deux variables.

On ne le démontrera pas dans le cadre de ce cours mais l'estimateur

$$\hat{\rho} = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2}}$$

(qui a pour réalisation r une fois l'échantillon tiré) est une variable aléatoire qui a les propriétés suivantes:

- $E(\hat{\rho}) = \rho$;
- et la variance estimée de $\hat{\rho}$ est $V(\hat{\rho}) = \frac{1-r^2}{n-2}$.
- l'écart réduit

$$T = \frac{\hat{\rho} - \rho}{\sqrt{\frac{1-\hat{\rho}^2}{n-2}}}$$

suit une Student à $n-2$ degrés de liberté.

Note: sans vouloir entrer dans la démonstration, il est légitime de se demander d'où proviennent les $n - 2$ degrés de liberté. Nous disposons de n observations mais tout comme utiliser une moyenne estimée dans le cas de la variance introduisait une relation linéaire et donc faisait perdre un degré de liberté, nous ajoutons ici une autre relation linéaire qui provient du remplacement de la moyenne de Y par son estimateur. Une seconde relation linéaire est donc ajoutée et il ne reste que $n - 2$ degrés de liberté.

4.1.2 Application: estimation d'une corrélation

On s'intéresse à la corrélation entre le temps moyen passé à étudier ses statistiques par semaine et les notes au partiel. On obtient les 20 observations consignées dans le tableau 6.

Table 6: Temps passé à étudier et notes en statistiques

Observation (i)	1	2	3	4	5	6	7	8	9	10
Temps (X_i)	50	60	25	0	120	80	75	30	20	10
Note (Y_i)	10	8	4	6	15	13	14	9	5	6
Observation (i)	11	12	13	14	15	16	17	18	19	20
Temps (X_i)	60	70	45	35	85	150	140	95	65	40
Note (Y_i)	10	13	8	6	12	18	14	7	8	4

Commençons par faire quelques remarques. Tout d'abord, il y a des observations qui ont le même temps de travail mais qui n'ont pas la même note. Ceci signifie que la corrélation ne sera pas égale à 1 (ou -1). Ceci signifie aussi qu'il y a d'autres facteurs qui interviennent dans la note que le temps de

travail, comme par exemple: la concentration pendant ce temps de travail, les facilités de l'étudiant pour la statistique, la chance, le nombre de cafés bus le matin de l'examen, les conditions de travail, etc. Lorsqu'on étudie une corrélation, il ne s'agit pas d'étudier l'ensemble des facteurs qui affectent la note à l'examen mais d'évaluer à quel point deux variables, ici temps passé à étudier et note finale, sont reliés.

Une bonne façon d'apprécier la force de cette relation sans pour autant faire le calcul de la corrélation est de représenter le nuage de points (insérer graphique). Dans ce cas, on observe clairement une relation croissante entre la note et le temps passé à travailler.

Passons maintenant au calcul de la corrélation. L'application de la formule conduit à une corrélation sur l'échantillon égale ($\hat{\rho}$) à 0.83, c'est l'estimation ponctuelle de la corrélation entre les deux variables sur l'ensemble de la population. La relation entre les deux variables est donc assez forte et positive, bien sûr.

4.2 Test d'hypothèse sur une corrélation

La taille de l'échantillon étant relativement petite, un élève préfère tester si la corrélation est significativement différente de 0 (ou provenant simplement de l'aléa de l'échantillonnage) avant de faire l'effort de passer du temps à étudier ses stats. Nous allons donc décrire ici comment tester si une corrélation est significativement différente de 0.

Nous disposons déjà de la distribution de l'estimateur de la corrélation puisque nous savons que:

$$T = \frac{\hat{\rho} - \rho}{\sqrt{\frac{1-\hat{\rho}^2}{n-2}}} \rightsquigarrow \mathcal{T}_{n-2}.$$

L'hypothèse à tester est la suivante:

$$H_0 : \rho = 0$$

contre l'hypothèse alternative qu'elle est positive (il est peu probable que temps passé à étudier et note soient négativement reliés). Sous l'hypothèse H_0 , la statistique T se réécrit:

$$T = \frac{\hat{\rho}}{\sqrt{\frac{1-\hat{\rho}^2}{n-2}}}$$

et on considèrera que l'hypothèse H_0 doit être rejetée au seuil α si et seulement si

$$t \geq t_{n-2;\alpha}.$$

On peut vouloir transcrire cette condition en une condition sur $\hat{\rho}$. La valeur critique r_C est telle que:

$$t_C = \frac{r_C}{\sqrt{\frac{1-r_C^2}{n-2}}} = t_{n-2;\alpha}.$$

Après un réarrangement des termes, il vient que

$$r_C = \frac{t_{n-2;\alpha}}{\sqrt{n-2 + t_{n-2;\alpha}^2}}$$

et que la règle de décision est la suivante:

$$\text{Rejeter } H_0 \text{ ssi } \hat{\rho} \geq r_C$$

Application numérique de l'exemple précédent.

Bien entendu, nous avons fait un exemple avec un test unilatéral. Il est toujours possible de retenir comme hypothèse alternative que $\rho \neq 0$. La procédure est la même et r_C s'écrira simplement en fonction de $t_{n-2;\alpha/2}$:

$$r_C = \frac{t_{n-2;\alpha/2}}{\sqrt{n-2 + t_{n-2;\alpha/2}^2}}$$

On peut ensuite résumer les conclusions d'un test bilatéral de la façon suivante:

- si $\hat{\rho} < -r_C$, on rejette H_0 et on a identifié une corrélation linéaire négative;
- si $r_C < \hat{\rho} < r_C$, on ne peut pas rejeter H_0 et on conclue donc à une absence de corrélation entre les deux variables;
- si $r_C < \hat{\rho}$, on rejette H_0 et il existe une corrélation linéaire positive.