COURS DE STATISTIQUES
L2 - SOCIOLOGIE - Université de Bourgogne

Contents

1 Introduction

L'objet principal du cours de Statistiques en L1 est la ***statistique descriptive*** : il s'agit de la branche des statistiques regroupant les techniques permettant de décrire, aussi précisément que possible, un nombre important de données. A partir d'un ***échantillon*** d'une (ou plusieurs) population(s), on calcule un certain nombre de quantités permettant de le décrire de façon synthétique.

Exemple 1.1. On s'intéresse au salaire des employés de catégorie CSP+. Pour cela, comme il est impossible de les répertorier dans leur totalité, on sélectionne un *échantillon* suffisamment grand, par exemple de *taille* $n = 1000$, pour lesquels on répertories le salaire annuel de la dernière année écoulée. Comme il n'est pas commode de transmettre, comme information, un millier de nombres (!), on calcule certaines quantités à partir de ces données qui peuvent servir de base à une réflexion sociologique. Par exemple :

1. La ***moyenne*** m des salaires de cet échantillon.
2. Les différents ***quartiles*** dont le principal, appelé ***médiane***, donne une information sur la répartition des salaires dans cet échantillon.
3. L'***écart-type*** s qui mesure à quel point les salaires de l'échantillon sont *dispersés*, c'est à dire situé loin de la moyenne de l'échantillon.
4. etc...

L'objet principal du cours de Statistiques en L2 est l'introduction à la ***statistique inférentielle***. Il s'agit d'un ensemble de méthodes permettant de déduire des conclusions *vraisemblables* concernant des populations en se basant sur des données *descriptives* observées sur des échantillons prélevés au hasard dans ces populations. Naturellement, plus les échantillons sur lequels on travaille sont de grande taille, plus il est naturel de penser que les quantités calculées (moyenne, écart-type, ...) seront proches des quantités réelles. Ces effectivement le cas (bien que ce fait ne soit pas facile à démontrer), mais ces conclusions ne sont pas des affirmations certaines. Elles comportent des risques ou ***probabilités d'erreur***.

La maîtrise de ces probabilités nécessite la connaissance de ce qu'on appelle un ***modèle probabiliste***. En effet, on travaille en Statistiques et Probabilités avec des quantités dont les quantités ne peut être connues avec précision. C'est pourquoi on appelle ces quantités des ***variables aléatoires***. La plupart des temps, dans les problèmes intéressants en sociologie, ces variables suivent une ***loi de probabilité*** connue. Cela signifie que, à défaut de connaître les valeurs exactes d'une variable aléatoire X, on sait déterminer les probabilités $\mathbb{P}(X \leq x)$ (probabilité que les valeurs de la variable X soient inférieures au nombre x) pour tous les nombres x.

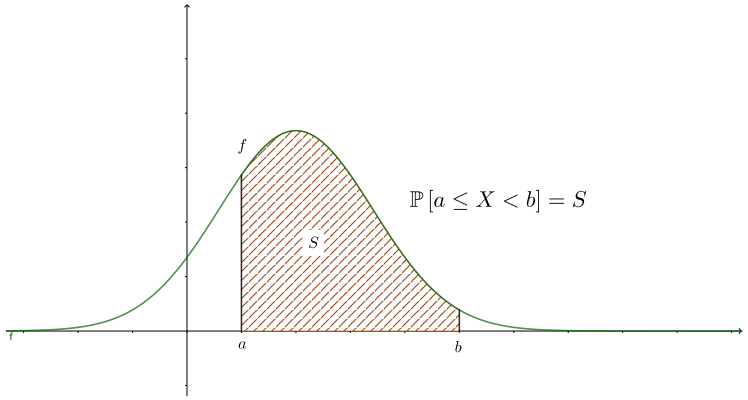
Il s'avère que les plus importants de ces modèles sont construits à partir des lois de probabilité bien particulières :

1. la ***loi normale***. On l'utilise en particulier pour de grands échantillons. En raison de son importance, elle est décrite en détail dans la section suivante.
2. la ***loi de Student***. Cette loi, particulièrement utile lorsque la taille des échantillons étudiés n'est pas très grande, a un fonction très similaire à la loi normale.
3. la ***loi du χ^2 (chi-deux)***. Cette dernière est utile, entre autres, quand il s'agit de comparer deux populations à partir de deux échantillons.

2 Variables aléatoires, lois de probabilité et loi normale.

2.1 Variables aléatoires et lois de probabilité.

Définition 2.1. On considère une variable aléatoire X et une fonction positive $f : \mathbb{R} \to \mathbb{R}_+$. On dit que X suit la ***loi de probabilité de densité*** f, ou plus simplement que f est la densité de X, si, pour tout couple de nombres réels $a < b$, la probabilité que les valeurs de X soient comprises entre a et b est égale à la surface délimitée sous le graphe de la fonction f entre les points a et b. Ce que l'on peut représenter graphiquement comme suit :



Remarque 2.2. On a, dans le cas de la définition précédente,

$$\mathbb{P}[a \leq X < b] = \mathbb{P}[a \leq X \leq b] = \mathbb{P}[a < X \leq b] = \mathbb{P}[a < X < b].$$

On peut également avoir $a = -\infty$ et $b = +\infty$.

Enfin, la "surface totale" doit être égale à 1, c'est à dire $\mathbb{P}[-\infty < X < +\infty] = 1$.

Définition 2.3. Soit X une variable aléatoire de densité f. Alors :

1. La ***moyenne*** $m(X)$ de la variable X est la surface totale sous le graphe de la fonction $xf(x)$.
2. La ***variance*** $V(X)$ de la variable X est la surface totale sous le graphe de la fonction $(x - m(X))^2 f(x)$. On a la formule :
$$V(X) = m\left(X^2\right) - m(X)^2.$$
3. L'***écart-type*** $s(X)$ de la variable X est le nombre :
$$s(x) = \sqrt{V(X)}.$$

Remarque 2.4. On voit que le calcul des probabilités, de la moyenne et de la variance d'une variable aléatoire qui a une densité f revient à savoir calculer les surfaces sous le graphe de certaines fonctions liées à f. Ça n'est pas un calcul facile. Heureusement, ces calculs, pour les fonction de densité principales, sont déjà données dans des tables, ou intégrées dans les calculatrices classiques.

2.2 Loi normale centrée réduite

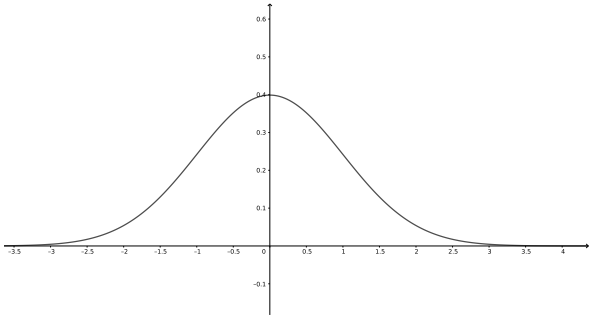
C'est loi loi de probabilité "royale" dans les applications. Elle est définie comme suit :

Définition 2.5. Si une variable aléatoire Z admet pour densité la fonction :

$$f(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-x^2/2\right),$$

on dit que Z suit la ***loi normale centrée réduite***.

Remarque 2.6. Le graphe de la fonction f est le suivant :



On note que ce graphe est symétrique par rapport à l'axe vertical, ce qui sera important dans les calculs.

Proposition 2.7. *Soit Z une variable aléatoire suivant la loi normale centrée réduite. Alors :*

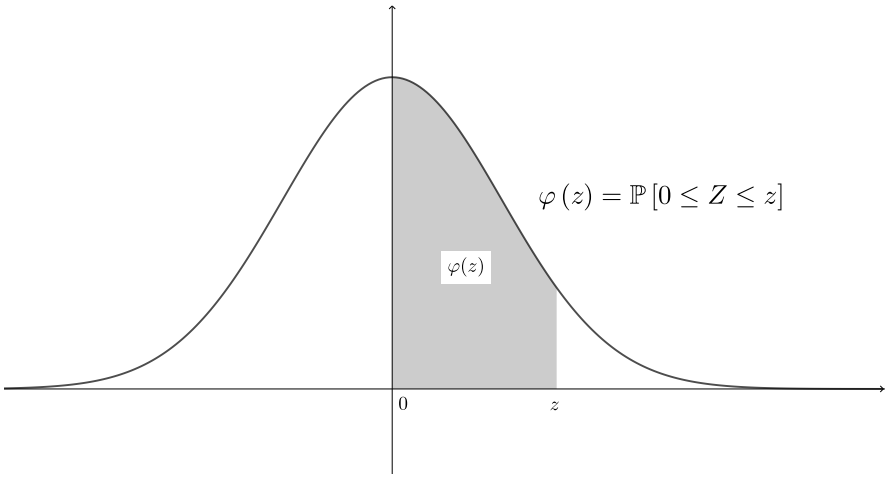
1. *La moyenne de Z est nulle : $m(Z) = 0$.*
2. *La variance et l'écart-type de Z sont égaux à 1 : $V(Z) = 1$, $s(Z) = 1$.*

Pour ces raisons, on désigne par $\mathcal{N}(0,1)$ la loi normale centrée réduite, et on note :

$$Z \hookrightarrow \mathcal{N}(0,1).$$

Calcul avec la loi normale centrée réduite. On peut bien sûr utiliser les calculatrices standard. A défaut, on peut utiliser les tables contenues dans le formulaire de L2. Ces tables donnent les valeur de la fonction

$$\varphi(z) = \mathbb{P}[0 \leq Z \leq z] \text{ pour } z \geq 0.$$



Il est bon dans ces calculs, de s'aider avec un dessin, le dessin du graphe de f, qu'on peut faire sur un brouillon.

On voit par exemple dans la table que, si $z = 0.47$, à l'intersection de la ligne 0.4 et de la colonne 7, on trouve $\varphi(0.47) = 0.1808$.

Si on demande $\mathbb{P}[-1.23 \leq Z \leq 0]$, on utilise la symétrie par rapport à l'axe vertical : pour $z \leq 0$, on a $\varphi(z) = \varphi(|z|)$. Donc :

$$\varphi(-1.23) = \varphi(1.23) = 0.3907.$$

Si on demande $\mathbb{P}[0.24 \leq Z \leq 1.5]$, on décrit la surface que l'on voit sur le dessin :

$$\mathbb{P}[0.24 \leq Z \leq 1.5] = \varphi(1.5) - \varphi(0.24) = 0.4332 - 0.0948 = 0.3384.$$

Si on demande $\mathbb{P}[-1.61 \leq Z \leq 2.31]$, on écrit :

$$\mathbb{P}[-1.61 \leq Z \leq 2.31] = \varphi(-1.61) + \varphi(2.31) = \varphi(1.61) + \varphi(2.31) = 0.4463 + 0.4896 = 0,9359.$$

On peut aussi calculer aussi $\mathbb{P}[1.72 \leq Z < +\infty]$. Pour cela, il faut se souvenir que la *demi-surface totale*, à droite ou à gauche de l'axe vertical, vaut 0.5. Donc :

$$\mathbb{P}[1.72 \leq Z < +\infty] = 0.5 - \mathbb{P}[0 \leq Z \leq 1.72] = 0.5 - \varphi(1.72) = 0.5 - 0.4573 = 0,0427.$$

De même, si on demande $\mathbb{P}[-\infty < Z \leq 2.33]$, on écrit :

$$\mathbb{P}[-\infty < Z \leq 2.33] = 0.5 + \mathbb{P}[0 \leq Z \leq 2.33] = 0.5 + \varphi(2.33) = 0.5 + 0.4901 = 0,9901.$$

2.3 Les lois normales générales

Définition 2.8. Si la variable aléatoire X admet pour densité la fonction

$$f_{\mu,\sigma}(x) = \frac{\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)}{\sqrt{2\pi\sigma^2}},$$

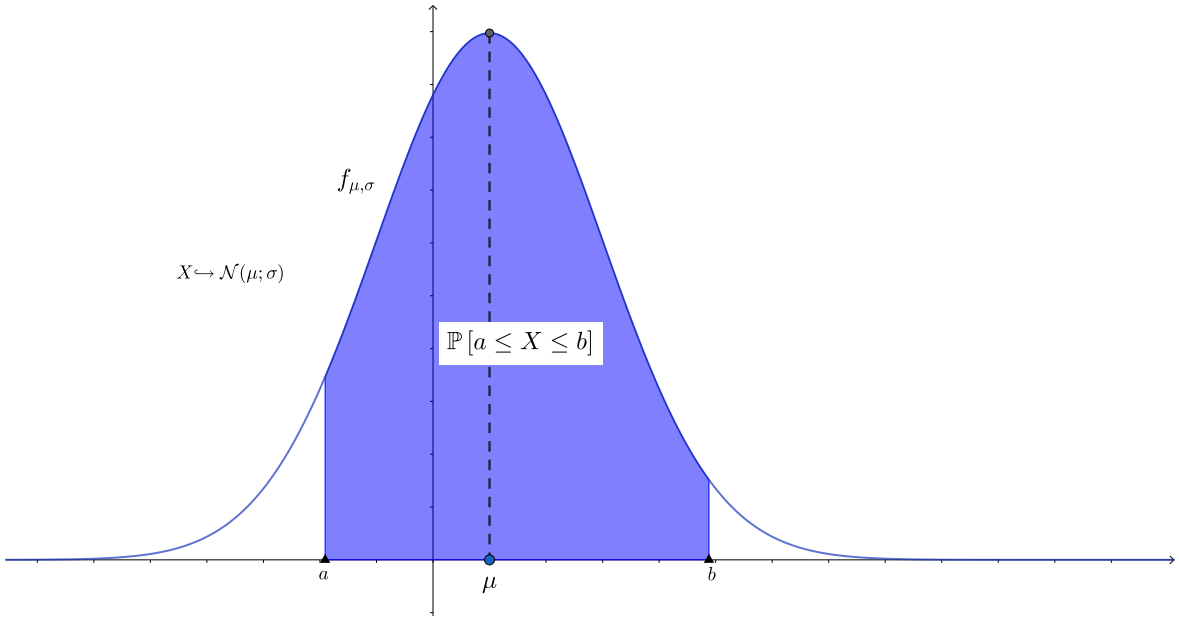
on dit qu'elle suit la ***loi normale de moyenne*** μ ***et d'écart-type*** σ, et on note :

$$X \hookrightarrow \mathcal{N}(\mu, \sigma).$$

Proposition 2.9. *Si la variable X suit la loi normale $\mathcal{N}(\mu, \sigma)$, on a :*

1. $m(X) = \mu$.
2. $s(X) = \sigma$, *et donc* $V(X) = \sigma^2$.

Le graphe de la fonction $f_{\mu,\sigma}$ et l'analyse d'une variable aléatoire X telle que $X \hookrightarrow \mathcal{N}(\mu, \sigma)$ sont résumés dans la figure suivante :



Calculs avec une loi normale générale. A nouveau, on peut tout à fait utiliser les calculatrices. On peut également se ramener à la loi normale centrée réduite $\mathcal{N}(0,1)$, en utilisant la propriété suivante :

Proposition 2.10. *Si la variable aléatoire X suit la la loi normale $\mathcal{N}(\mu,\sigma)$ de moyenne μ et d'écart-type σ, alors la variable aléatoire $Z = \dfrac{X-\mu}{\sigma}$ suit la loi normale centrée réduite $\mathcal{N}(0,1)$:*

$$\text{Si } X \hookrightarrow \mathcal{N}(\mu,\sigma), \text{ alors } Z = \frac{X-\mu}{\sigma} \hookrightarrow \mathcal{N}(0,1).$$

On peut utiliser la propriété ci-dessus pour calculer des probabilités pour la variable X. En effet, on a par exemple :

$$\mathbb{P}[a \leq X \leq b] = \mathbb{P}\left[\frac{a-\mu}{\sigma} \leq Z \leq \frac{b-\mu}{\sigma}\right],$$

avec $Z = \dfrac{X-\mu}{\sigma}$.

2.4 Problèmes inverses avec $\mathcal{N}(0,1)$ ou $\mathcal{N}(\mu,\sigma)$

On est souvent amené en Statistique Inférentielle à se poser le problème suivant :

> Étant donné un nombre réel α compris entre 0 et 1 et une variable aléatoire $X \hookrightarrow \mathcal{N}(\mu,\sigma)$, déterminer la valeur a telle que $\mathbb{P}[X \geq a] = \alpha$ ou $\mathbb{P}[X \leq a] = \alpha$.

On appelle cela ***problème inverse***, car cette fois la probabilité α est donnée à l'avance, et qu'on souhaite connaître le nombre a. Il ne s'agit donc plus de calculer la probabilité d'un évènement, mais de déterminer la valeur de la variable aléatoire X pour que l'évènement $\{X \leq a\}$ ou $\{X \geq a\}$ ait une probabilité fixée.

A nouveau, pour traiter ce type de problème, on peut utiliser la calculatrice.

1. Sur *Casio Graph 35+*, pour trouver a tel que $\mathbb{P}[X \leq a] = \alpha$, dans le menu **STAT**, on choisit **DIST**, puis **NORM** puis **invN**, et on complète les paramètres. Par exemple, pour trouver a tel que $\mathbb{P}[X \leq a; \mathcal{N}(10, 3.2)] = 0.7568$, on choisit Tail : Left, Area : 0.7568, σ : 3.2, μ : 10, et on exécute.
2. Sur *Texas TI 82 Stats*, on choisit **DIST**, puis **invNorm**, puis complète les paramètres :

$$\textbf{invNorm}(0.7568, 10, 3.2).$$

On trouve $a = 12.2273$.

Mais on peut également utiliser la table de la loi normale inverse. Il s'avère que certaines valeurs de α sont particulièrement utiles. Elles sont données dans les tables pour la loi $\mathcal{N}(0,1)$.

Exemple 2.11. On demande de déterminer le nombre $z > 0$ tel que $\mathbb{P}[Z \geq z] = 0.05$, sachant que $Z \hookrightarrow \mathcal{N}(0,1)$. La table du formulaire utilise la fonction $\varphi(z) = \mathbb{P}[0 \leq Z \leq z]$. On voit que $\mathbb{P}[Z \geq z] = 0.5 - \varphi(z)$. On cherche ainsi z pour lequel $0.5 - \varphi(z) = 0.05$, c'est-à-dire $\varphi(z) = 0.5 - 0.05$. On lit donc la table pour la valeur $\alpha = 0.05$, ce qui donne $z = 1.645$.

Exemple 2.12. On considère une variable aléatoire $X \hookrightarrow \mathcal{N}(10,2)$ (c'est à dire qui suit la loi normale de moyenne 10 et d'écart-type σ), et on demande le nombre x tel que $\mathbb{P}[X \geq x] = 0.05$. Pour cela, on se ramène à une variable aléatoire $Z \hookrightarrow \mathcal{N}(0,1)$, en posant

$$Z = \frac{X - \mu}{\sigma} = \frac{X - 10}{2}.$$

On sait alors (voir exemple précédent) que, pour $x = 1.645$, on a $\mathbb{P}[Z \geq z] = 0.05$. On revient aux valeurs de la variable aléatoire X en écrivant :

$$z = \frac{x - 10}{2} \Longrightarrow 2z = x - 10 \Longrightarrow x = 2z + 10.$$

On trouve ainsi $x = 2 \times 1.645 + 10 = 13.29$. Si on fait l'exercice avec la fonction invNorm de Texas ou invN de Casio, on trouve $x = 13.2897$.

Remarque 2.13. ***A RETENIR !*** Le valeurs de z telles que $\mathbb{P}[Z \geq z = 0.05; \mathcal{N}(0,1)]$ et $\mathbb{P}[Z \geq z = 0.025; \mathcal{N}(0,1)]$. On a :

$$\mathbb{P}[Z \geq 1.645; \mathcal{N}(0,1)] = 0.05 \text{ et } \mathbb{P}[Z \geq 1.96; \mathcal{N}(0,1)] = 0.025.$$

3 Problèmes d'estimation

C'est une question importante en Statistique inférentielle. Sur la base d'un échantillon de taille n, on calcule les valeurs indicateurs classiques : proportion, moyenne, écart-type. On veut savoir s'il est possible d'en déduire la valeur de ces indicateurs pour la population totale, en autorisant une certaine erreur.

Si l'erreur autorisée est α, on dit que le résultat est donné avec une ***confiance*** $1 - \alpha$.

3.1 Estimation d'une proportion

On considère un échantillon de taille n. On suppose que la proportion d'individus de cet échantillon présentant un caractère donné est p_e (on note p_e pour *proportion expérimentale*). On veut en déduire, avec une *confiance* $1 - \alpha$, la proportion p d'individus de la population totale présentant le même caractère.

On note $q_e = 1 - p_e$.

Exercice 3.1. Dans la population $\mathcal{P}$ des élèves qui passeront leur bac en 2019, on s'intéresse au caractère : "l'élève n'a jamais redoublé". Après un sondage auprès de 200 élèves de terminale d'un lycée, il s'avère que 110 d'entre eux n'a jamais redoublé. La proportion expérimentale est don $p_e = \frac{110}{200} = 55\%$.

Estimer avec une confiance $1 - \alpha = 95\%$ la proportion p d'élèves qui n'ont jamais redoublé sur la population totale des élèves qui vont passer le bac.

On procède comme suit :

1. On vérifie les hypothèses $n > 30$, $np_e \geq 5$ et $nq_e \geq 5$. Si ces hypothèses ne sont pas satisfaites, on ne peut pas estimer la proportion p dans la population totale.
2. On cherche la valeur z_α telle que $\mathbb{P}[Z \geq z_\alpha; \mathcal{N}(0,1)] = \alpha/2$. Cette étape est donc un *problème inverse* (voir Section 2.4) pour la loi normale centrée réduite $\mathcal{N}(0,1)$.
3. On calcule $a_\alpha = z_\alpha \sqrt{\frac{p_e q_e}{n}}$. Alors on peut affirmer que la proportion cherchée p se trouve dans l'*intervalle de confiance* $I_\alpha(p) = [p_e - a_\alpha, p_e + a_\alpha]$:

$$p \in I_\alpha(p) = [p_e - a_\alpha, p_e + a_\alpha].$$

Remarque 3.2. On se doute que plus l'échantillon est grand, plus l'estimation est fiable. On voit en effet que le nombre a_α qui donne la taille de l'intervalle de confiance $I_\alpha(p)$ est d'autant plus petit que n est grand.

Remarque 3.3. Pour avoir une confiance, de 95%, on a $z_\alpha = 1.96$.

3.2 Estimation d'une moyenne

On considère une variable aléatoire X sur une population $\mathcal{P}$. A partir d'un échantillon de taille n, on calcule la moyenne *expérimentale* m_{e} et l'écart-type *expérimental* s_{e} de la variable X. On veut en déduire, avec une confiance $1-\alpha$, la moyenne μ et l'écart-type σ de la variable aléatoire X pour toute la population.

Exercice 3.4. On veut connaître la taille moyenne des femmes adultes au Pays-Bas, afin de la comparer avec la taille moyenne des femmes d'autres pays européens. Le tableau suivant donne la répartition des tailles de 100 femmes adultes aux Pays-Bas :

Taille en cm	140/148	148/156	156/164	164/172	172/180	180/188	188/196	Total	m_{e}	s_{e}
Effectifs	2	6	19	28	32	10	3	100	169.92	9.935

Figure 1: Tailles d'un échantillon de femmes adultes aux Pays-Bas

Dans quelle fourchette, avec une confiance $1-\alpha = 95\%$, peut-on situer la taille moyenne μ de la population des femmes aux Pay-Bas ?

Il y a deux façons de procéder, suivant que l'échantillon est *petit* ($n \leq 30$) ou *grand* ($n > 30$).

1. **Petit échantillon** : $n \leq 30$. *On doit supposer que la variable X suit une loi normale.* Alors :
   1. On considère la ***loi de Student à*** $n-1$ ***degrés de liberté*** (ddl) $\mathrm{St}\,(\mathrm{n}-1)$, et on cherche t_α tel que $\mathbb{P}\left[T > t_\alpha\right] = \dfrac{\alpha}{2}$.
   2. On calcule $a_\alpha = t_\alpha \dfrac{s_{\mathrm{e}}}{\sqrt{n-1}}$. On peut alors affirmer que la valeur de μ se trouve dans l'intervalle de confiance $I_\alpha(\mu) = [m_{\mathrm{e}} - a_\alpha, m_{\mathrm{e}} + a_\alpha]$:
   $$\mu \in I_\alpha(\mu) = [m_{\mathrm{e}} - a_\alpha, m_{\mathrm{e}} + a_\alpha].$$
2. **Grand échantillon** : $n > 30$.
   (a) On considère la ***loi normale*** $\mathcal{N}(0,1)$ et on cherche z_α tel que $\mathbb{P}\left[Z \geq z_\alpha; \mathcal{N}(0,1)\right] = \dfrac{\alpha}{2}$.
   (b) On calcule $a_\alpha = z_\alpha \dfrac{s_{\mathrm{e}}}{\sqrt{n-1}}$. On peut alors affirmer que la valeur de μ se trouve dans l'intervalle de confiance $I_\alpha(\mu) = [m_{\mathrm{e}} - a_\alpha, m_{\mathrm{e}} + a_\alpha]$:
   $$\mu \in I_\alpha(\mu) = [m_{\mathrm{e}} - a_\alpha, m_{\mathrm{e}} + a_\alpha].$$

3.3 Estimation d'un écart-type (ou d'une variance)

On considère une variable aléatoire X sur une population $\mathcal{P}$, *qui suit une loi normale.* On travaille cette fois avec la ***loi du*** χ^2. Connaissant la valeur expérimentale de la moyenne m_{e} et de l'écart-type s_{e} de la variable X sur un échantillon de taille n, on veut en déduire une estimation de l'écart-type σ de X sur toute la population $\mathcal{P}$, avec une confiance $1-\alpha$.

Exercice 3.5. Dans une population $\mathcal{P}$, on sait que le Q.I. est régi par une loi normale. Pour un échantillon de taille 32 de la population $\mathcal{P}$, on observe un écart-type expérimental $s_{\mathrm{e}} = 7.4$. Donner une estimation, avec une confiance $1-\alpha = 95\%$, de l'écart-type du Q.I. pour la population $\mathcal{P}$.

On procède comme suit :

1. A l'aide de la ***table du*** χ^2 ***à*** $n-1$ ***ddl***, on cherche les valeurs de x_1 et x_2 telles que
$$\mathbb{P}\left[Y \leq x_1\right] = \frac{\alpha}{2} \text{ et } \mathbb{P}\left[Y \geq x_2\right] = \frac{\alpha}{2}.$$
2. On peut alors affirmer que l'écart-type σ vérifie :
$$\sigma \in I_\alpha(\sigma) = \left[s_{\mathrm{e}}\sqrt{\frac{n}{x_2}}, s_{\mathrm{e}}\sqrt{\frac{n}{x_1}}\right],$$
et donc que la variance $V = \sigma^2$ vérifie :
$$V \in I_\alpha(V) = \left[s_{\mathrm{e}}^2\frac{n}{x_2}, s_{\mathrm{e}}^2\frac{n}{x_1}\right].$$

4 Tests paramétriques d'ajustement

4.1 Ajustement d'une proportion

4.1.1 Méthode générale

On considère une population $\mathcal{P}$ d'individus, et on veut comparer la proportion p d'individus de cette population qui présentent un certain caractère, avec une proportion fixée p_0. Pour cela, on étudie un échantillon de n individus de cette population, et on note p_{exp} ou p_e la proportion d'individus de cet échantillon qui présentent ce caractère. On procède alors en suivant les cinq étapes suivantes :

Etape 1 : Formulation des deux hypothèses, en français, puis en termes statistiques. On formule deux hypothèses :
1. *L'hypothèse nulle* H_0 : elle affirme que la situation respecte un *statu quo*, souvent dû au hasard.
2. *L'hypothèse alternative* H_1 : elle met en évidence un phénomène dont on veut tester le bien fondé.

On formule maintenant ces deux hypothèses en termes statistiques, en les traduisent à l'aide des deux proportions p et p_0 :
1. H_0 :$p = p_0$;
2. H_1 : $p > p_0$ (*cas unilatéral*), ou bien $p < p_0$ (*cas unilatéral*), ou bien $p \neq p_0$ (*cas bilatéral*).

Etape 2 : Choix du modèle statistique. Pour un échantillon aléatoire de taille n, on considère la variable aléatoire P_n, qui représente la proportion d'individus de l'échantillon aléatoire qui présentent le caractère donné. Alors, si l'échantillon est grand ($n > 30$, $pn > 5$, $qn > 5$ avec $q = 1 - p$), sous l'hypothèse H_0, on a :

$$P_n \hookrightarrow \mathcal{N}\left(p_0, \sqrt{\frac{p_0 q_0}{n}}\right)$$

.

Etape 3 : Détermination de la région critique. On fixe un *niveau de signification* $\alpha = 0.05$. La région critique K_α est l'ensemble des valeurs de P_n favorables à l'hypothèses H_1 de probabilité égale à α (sous l'hypothèse H_0) :$\mathbb{P}\left[K_\alpha; \mathcal{N}\left(p_0, \sqrt{\frac{p_0 q_0}{n}}\right)\right]$.

Etape 4 : Décision du test. Si $p_e \in K_\alpha$, on accepte H_1 au niveau α. Si $p_e \notin K_\alpha$, on rejette H_1 au niveau α.

Etape 5 : Questions subsidiaires, sur la signification et la puissance du test.
1. *p-value* ou *signification du test* : c'est la probabilité (sous H_0) l'ensemble des valeurs de P_n favorables à H_1 bordé par le résultat expérimental.
2. *Puissance du test* : c'est la probabilité d'accepter H_1 alors que H_1 est vraie. Elle dépend du paramètre p, qui représente la proportion *réelle* de la population qui satisfait le caractère demandé. C'est la fonction $\eta(p) = \mathbb{P}\left[K_\alpha; \mathcal{N}\left(p, \sqrt{\frac{pq}{n}}\right)\right]$. Une puissance faible représente généralement un échantillon de taille n trop petite.

4.1.2 Exemple : activités sportives ou musicales de enfants de 6 ans.

Dans la population $\mathcal{P}$ des enfants de 6 ans, on se demande si un enfant préfère pratiquer une activité sportive ou musicale. Sur un échantillon de 300 enfants, 180 ont répondu qu'ils préfèrent une activité sportive, soit une proportion expérimentale de $p_e = 60\%$. Donc on teste si les enfants préfèrent pratiquer une activité sportive.

Etape 1. Les deux hypothèses sont :
1. *Hypothèse nulle* : les enfants ne pas de préférence particulière entre la musique et le sport.
2. *Hypothèse alternative* : Les enfants préfèrent le sport.

On désigne par p la proportion d'enfants qui préfèrent le sport à la musique. Alors :
1. H_0 : $p = p_0 = 1/2$.
2. H_1 : $p > 1/2$.

Etape 2. On appelle P_n la proportion d'enfants qui préfèrent le sport à la musique parmi un échantillon aléatoire de taille n. Sous l'hypothèse nulle, on a :

$$P_n \hookrightarrow \mathcal{N}\left(p_0, \sqrt{\frac{p_0 q_0}{n}}\right) = \mathcal{N}\left(0.5, \sqrt{\frac{0.5 \times 0.5}{300}}\right) = \mathcal{N}(0.5, 0, 029).$$

Etape 3. On niveau de signification $\alpha = 0.05$, on cherche a tel que $\mathbb{P}[P_n \geq a; \mathcal{N}(0.5, 0, 029)] = 0.05$. On trouve $a = 0.546$. Donc $K_\alpha = [P_n \geq 0.546]$.

Etape 4. La valeur expérimentale est $p_e = 0.6$. Donc $p_e \in K_\alpha$: on accepte l'hypothèse H_1 au niveau de signification 0.05.

Etape 5 : questions subsidiaires. $p - \text{value} = \mathbb{P}[P_n \geq 0.6; \mathcal{N}(0.5, 0, 028)] = 0.0003$. Elle est très petite : le test est significatif.

Puissance $\eta(0.6) = \mathbb{P}\left[K_\alpha; \mathcal{N}\left(0.6, \sqrt{\frac{0.6 \times 0.4}{300}}\right)\right] = 0.97$. Forte puissance !

4.1.3 Exemple : le point de vue bilatéral, le téléphone portable chez les enfants de 7 à 12 ans

On utilise souvent un test bilatéral pour confirmer H_0 . Par exemple, une information lue sur internet affirme que 55% des enfants de 7 à 12 ans ont un téléphone portable. Une enquête réalisée auprès de 300 enfants a donné un pourcentage de expérimental $p_e = 59\%$. Cette valeur expérimentale est-elle significativement différente de la valeur annoncée ?

Etape 1 : formulation des hypothèses Soit p la proportion théorique d'enfants de 7 à 12 ans possédant un téléphone portable.

1. Hypothèse nulle H_0 : "la valeur trouvée sur internet est valable" soit $p = 55\%$.
2. Hypothèse alternative H_1 : "la valeur trouvée sur internet n'est pas valable" soit $p \neq 55\%$.

On opte pour un test bilatéral.

Etape 2 : statistique du test. Avec $n = 300$, sous l'hypothèse H_0, $P_n \hookrightarrow \mathcal{N}(0.55, 0, 0287)$. (on vérifie qu'avec $n = 300$, on a bien un grand échantillon).

Etape 3 : région critique. Les valeurs de P_n favorables à H_1 sont celles qui sont "loin" de la valeur théorique $p_0 = 0.55$. La région critique, au niveau $\alpha = 0.55$, est donc du type $K_\alpha = [P_n \leq b_1] \cup \left[P_n \geq b_2\right]$, ou encore du type $K_\alpha = |P_n - 0.55| \geq a_\alpha$. Donc $K_\alpha = \{P_n \leq 0.4937\} \cup \{P_n \geq 0.6063\} = \{P_n \leq 49.37\%\} \cup \{P_n \geq 60.63\%\}$.

Etape 4 : décision du test. La valeur expérimentale est $p_e = 59\%$. On voit que $p_e \notin K_\alpha$. Donc au niveau $\alpha = 0.05$, on rejette l'hypothèse alternative H_1, et on conserve l'hypothèse H_0 : il y a bien 55% d'enfants entre 7 et 12 ans qui possèdent un téléphone portable.

Etape : puissance du test. On veut calculer $\eta(0.59)$. On rappelle qu'il s'agit de la probabilité d'accepter H_1 alors que H_1 est vraie. Dans le cas bilatéral, il est plus commode de calculer l'erreur de seconde espèce $\beta(0.59)$, qui est la probabilité d'accepter H_0 alors que H_1 est vrai, et d'en déduirer $\eta(0.59) = 1 - \beta(0.59)$. On a :

$$\begin{aligned}
\beta(0.59) &= \mathbb{P}\left[P_n \in [0.4937, 0.6063]; \mathcal{N}\left(0.59, \sqrt{\frac{0.59 \times (1 - 0.59)}{300}}\right)\right] \\
&= \mathbb{P}[P_n \in [0.4937, 0.6063]; \mathcal{N}(0.59, 0.0284)] \\
&= 0.7166, \text{ et donc :} \\
\eta(0.59) &= 1 - \beta(0.59) = 0.2834 = 28.23\%.
\end{aligned}$$

On constate que cette puissance est extrêmement faible, il conviendrait sans doute de travailler avec un plus grand échantillon.

4.2 Ajustement d'une moyenne

On considère une variable aléatoire quantitative X sur une une population $\mathcal{P}$ (âge au moment du décès, salaire à l'embauche, mesure de l'intelligence (QI,...)). On veut comparer la moyenne μ de la variable X à une moyenne théorique μ_0.

On ne travaille que sur de *grands* échantillons ($n > 30$).

4.2.1 Méthode générale.

On calcule la moyenne expérimentale μ_e de X sur un échantillon de n individus.

Etape 1. On formule deux hypothèses H_0 et H_1, qui, exprimées en termes statistiques, deviennent :

$$H_0 : \mu = \mu_0; \quad H_1 : \mu > \mu_0, \text{ ou } \mu < \mu_0, \text{ ou } \mu \neq \mu_0.$$

Etape 2. On établit la statistique du test sous l'hypothèse nulle. Sur un grand échantillon ($n > 30$), la variable aléatoire M_n, qui représente la moyenne de X sur un échantillon aléatoire de taille n, vérifie :

$$M_n \hookrightarrow \mathcal{N}\left(\mu_0, \frac{\sigma_e}{\sqrt{n-1}}\right),$$

où σ_e est l'écart-type de la variable X sur l'échantillon expérimental.

Etape 3. En ayant fixé un niveau α, on détermine la région critique K_α.

Etape 4. On décide selon que μ_e appartient ou non à la région critique K_α.

Etape 5. On peut également calculer la $p-$value et la puissance $\eta(\mu)$ du test pour des valeurs données de μ.

4.2.2 Exemple : âge moyen des français au moment du décès.

On se demande si l'*âge moyen* des français a augmenté depuis 30 ans. Il y a 30 ans, il était de $\mu_0 = 72$ ans. Sur un échantillon de 100 décès pris au hasard en 2010, on obtient une moyenne expérimentale $\mu_e = 74$ ans, et un écart-type $\sigma_e = 8.9$ ans. Que conclure ?

Etape 1 : formulation des hypothèses. H_0 : l'âge moyen de décès des français est le même qu'il y a 30 ans.
H_1 : l'âge moyen de décès des français a *augmenté* en 30 ans (test *unilatéral*).
En termes statistiques, on désigne par μ l'âge moyen (inconnu) de décès des français. Alors :
$H_0 : \mu = \mu_0$, $H_1 : \mu > \mu_0$.

Etape 2 : détermination de la statistique du test. On a bien un grand échantillon, $n = 100 > 30$. Sous l'hypothèse nulle, la *moyenne aléatoire* M_n de l'âge du décès des français sur un échantillon de n individus vérifie :

$$M_n \hookrightarrow \mathcal{N}\left(\mu_0, \frac{\sigma_e}{\sqrt{n-1}}\right) = \mathcal{N}\left(72, \frac{8.9}{\sqrt{99}}\right) = \mathcal{N}(72, 0.8945).$$

Etape 3 : détermination de la région critique. Les grandes valeurs de M_n sont favorables à H_1. Au niveau de signification $\alpha = 0.05$, on trouve $K_\alpha = [M_n \geq 73.47]$.

Etape 4 : décision du test. On a $m_e = 74 \in K_\alpha$, donc, au niveau 0.05, on accepte H_1 (l'âge moyen du décès des français a augmenté en 30 ans).

Etape 5 : p-value et puissance. Si la moyenne réelle de l'âge de décès est 74 ans (et donc que H_1 est vraie), la probabilité que le test accepte H_1 est :

$$\eta(74) = \mathbb{P}[M_n \geq 73.47; \mathcal{N}(74, 0.89)] = 0.7242,$$

ce qui est une forte puissance.

5 Tests paramétriques de comparaison

5.1 Comparaison de deux proportions

On considère deux populations $\mathcal{P}_1$ et $\mathcal{P}_2$ sur lesquelles on étudie le même caractère, et on note p_1 et p_2 les proportions d'individus des populations $\mathcal{P}_1$ et $\mathcal{P}_2$ qui satisfont ce caractère. On souhaite comparer ces deux proportions. Pour cela, on prend un échantillon $\mathcal{E}_1$ de n_1 individus de la population $\mathcal{P}_1$ et un échantillon $\mathcal{E}_2$ de n_2 individus de la population $\mathcal{P}_2$, et on observe la proportion p_1^{e} de l'échantillon $\mathcal{E}_1$ et la proportion p_2^{e} de l'échantillon $\mathcal{E}_2$ d'individus vérifiant le caractère donné.

5.1.1 Comparaison de deux proportions : méthode générale.

Etape 1 : formulation des hypothèses.

$$H_0 \ : p_1 = p_2; \quad H_1 \ : p_1 < p_2 \text{ ou } p_1 > p_2 \text{ ou } p_1 \neq p_2.$$

Etape 2 : statistique du test. Pour des tailles n_1 et n_2 d'échantillons, qu'on suppose grands (c'est à dire $n_1 > 30$, $n_2 > 30$), on considère les proportions aléatoires P_{n_1} et P_{n_2} d'individus d'échantillons (aléatoires) de $\mathcal{P}_1$ et $\mathcal{P}_2$, à n_1 et n_2 éléments respectivement. On pose $p_0 = \dfrac{n_1 p_1^{\mathrm{e}} + n_2 p_2^{\mathrm{e}}}{n_1 + n_2}$ et $q_0 = 1 - p_0$. Sous l'hypothèse nulle, on a :

$$P_{n_1} - P_{n_2} \hookrightarrow \mathcal{N}\left(0, \sqrt{p_0 q_0 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}\right).$$

Etape 3 : détermination de la région critique K_α, au niveau α.

Etape 4 : décision du test.

5.1.2 Comparaison de deux proportions, exemple : embauche suivant la formation suvie.

Parmi 250 diplômés du secteur industriel, 180 ont trouvé un emploi correspondant à leur qualification pendant la première année après obtention de leur diplôme. Parmi 300 diplômés du secteur commercial 196 ont trouvé un emploi pendant la première année après obtention de leur diplôme. On se demande si les jeunes diplômés du secteur industriel trouvent plus rapidement un emploi que leurs pairs du secteur commercial.

On nomme $\mathcal{P}_1$ la population des diplômés du secteur industriel, et $\mathcal{P}_2$ la population des diplômés du secteur commercial. On a ainsi $n_1 = 180$ et $n_2 = 196$. On en déduit les proportions expérimentales $p_1^{\mathrm{e}} = \frac{180}{250} = 0.72$ et $p_2^{\mathrm{e}} = \frac{196}{300} = 0.65$.

Etape 1. On désigne par p_1 la proportion (inconnue) de diplômés du secteur industriel qui trouvent du travail après une année, et par p_2 la proportion (inconnue) de diplômés du secteur commercial qui trouvent du travail après une année. On a :

$$H_0 \ : p_1 = p_2; \quad H_1 \ : p_1 > p_2.$$

Etape 2. Les deux échantillons sont grands ($n_1, n_2 > 30$). On a $p_0 = \frac{180+196}{250+300} = 0.684$. Sous l'hypothèse H_0, on a :

$$P_{n_1} - P_{n_2} \hookrightarrow \mathcal{N}(0, 0.04).$$

Etape 3. On pose $\alpha = 0.05$. Les grandes valeurs de $P_{n_1} - P_{n_2}$ sont favorables à H_1. Donc on cherche a tel que $K_\alpha = \{P_{n_1} - P_{n_2} \geq a\}$. On trouve $K_\alpha = \{P_{n_1} - P_{n_2} \geq 0.066\}$.

Etape 4. On a $p_1^{\mathrm{e}} - p_2^{\mathrm{e}} = 0.07 \in K_\alpha$. Donc, au niveau $\alpha = 0.05$, on accepte l'hypothèse H_1.

5.2 Comparaison de deux moyennes

Dans cette partie on considère encore deux variables *quantitatives* X_1 et X_2 définies sur deux populations, de moyennes respectives μ_1 et μ_2 inconnues, et d'écarts-type respectifs σ_1 et σ_2 . On souhaite comparer ces deux moyennes en se basant sur les données de deux échantillons. Dans ce cours, nous supposerons toujours, ce qui une hypothèse raisonnable quand on travaille en sociologie, que les échantillons sont de grande taille, c'est à dire plus grande que 30. On se poser alors deux questions, qui orienteront le choix du test à effectuer.

1. Les échantillons sont-ils ***appariés*** ou ***indépendants*** ? On dit que deux échantillons sont ***appariés*** lorsque chaque individu de des deux échantillons détermine de façon unique un individu de l'autre échantillon. Les deux échantillons sont alors nécessairement de *même taille*. Souvent il s'agit des mêmes individus que l'on a observés à deux reprises, avant ou après un évènement particulier. C'est le cas lorsqu'on veut par exemple tester l'évolution d'une moyenne au cours du temps.
2. Si les cas contraire, on dit que les échantillons sont ***indépendants***. On doit alors se demander si les *variances* des variables aléatoires X_1 et X_2 sont comparables. Hélas, on ne connaît pas en général la valeur exacte de ces variances. ***On doit donc précéder le test de comparaison des moyennes par un test de comparaison des variances !***

5.2.1 Comparaison de deux moyennes pour deux (petits) échantillons appariés : méthode générale.

Pour que ce test soit valide, il faut supposer que les deux variables aléatoires X_1 et X_2 suivent une loi normale (inconnue).

Comme d'habitude, les deux hypothèses du test peuvent se formuler comme suit :

$$H_0 : \mu_1 = \mu_2;\ H_1 : \mu_1 > \mu_2 \text{ ou } \mu_1 < \mu_2 \text{ ou } \mu_1 \neq \mu_2.$$

On réécrit ces hypothèses sous la forme :

$$H_0 : \mu_1 - \mu_2 = 0;\ H_1 : \mu_1 - \mu_2 > 0 \text{ ou } \mu_1 - \mu_2 < 0 \text{ ou } \mu_1 - \mu_2 \neq 0.$$

On introduit la variable aléatoire $D = X_1 - X_2$. Les hypothèses deviennent :

$$H_0 \colon \mu_D = 0;\ H_1 : \mu_D > 0 \text{ ou } \mu_D < 0 \text{ ou } \mu_D \neq 0.$$

On introduit la moyenne aléatoire $M_n(D)$ et l'écart-type aléatoire $S_n(D)$. Alors, si les données de D suivent une loi normale, on a :

$$T_n = \frac{M_n(D)}{S_n(D)/\sqrt{n-1}} \hookrightarrow \mathrm{St}(n-1).$$

Puis on procède comme à l'accoutumée.

5.2.2 Comparaison de deux moyennes pour deux échantillons appariés : exemple.

On veut mesurer l'effet du bruit dans les conditions de travail, et à quel point le bruit affecte l'efficacité. On constitue un groupe de 10 sujets, et on mesure le nombre d'erreurs commises lors d'une tâche répétitive. Dans un premier temps, le groupe n'est pas soumis au bruit, et dans un deuxième temps, le groupe effectue ses tâches en présence d'un bruit continu. On suppose que les variables aléatoires "nombre d'erreurs avec bruit" et "nombre d'erreurs sans bruit" suivent une loi normale. On obtient les résultat suivants :

Numéro du sujet	1	2	3	4	5	6	7	8	9	10
nombre d'erreurs pour les conditions "sans bruit"	2	1	4	6	3	2	0	1	2	3
nombre d'erreurs pour les conditions "avec bruit"	2	2	5	8	5	1	3	2	0	3
$D = X_1 - X_2$	0	-1	-1	-2	-2	1	-3	-1	2	0

On constate que $m_{\mathrm{e}}^s = 2.4$, $m_{\mathrm{e}}^a = 3.1$, et donc que $m_{\mathrm{e}}^s < m_{\mathrm{e}}^a$. On veut savoir si on peut en déduire que le nombre moyen d'erreurs lors d'un travail sans bruit est *en moyenne* inférieur au nombre moyen d'erreurs commises en présence de bruit. On procède donc au test de comparaison des moyennes, *avec des échantillons appariés*, suivant.

Etape 1.

$$H_0 : \mu^s = \mu^a;\ H_1 : \mu^s \neq \mu^a.$$

Etape 2. On travaille avec un petit échantillon, de taille $n = 28$. On note $D = X_1 - X_2$. Sous l'hypothèse nulle, la variable aléatoire $T_n = \dfrac{M_n(D)}{S_n(D)/\sqrt{n-1}}$, où $M_n(D)$ est la moyenne de D et $S_n(D)$ est l'écart-type de D sur un échantillon aléatoire de taille n, vérifie :

$$T_n \hookrightarrow \mathrm{St}(9).$$

Etape 3. On travaille au niveau d'erreur $\alpha = 5\% = 0.05$. Les petites valeurs de T_n sont favorables à H_1, donc $K_\alpha = \{T_n \leq t_\alpha\}$, avec $\mathbb{P}[K_\alpha, \mathrm{St}(9)] = 0.05$. On trouve $t_\alpha = -1.833$.

Etape 4. On a $m_e^D = -0.7$, $s_e^D = 1.418$, donc $t_e = -1.48 \notin K_\alpha$. Donc au niveau $\alpha = 0.05$, on reste avec l'hypothèse H_0 : il n'y a le même nombre d'erreurs commises en présence de bruit que sans bruit.

5.2.3 Comparaison de moyennes pour deux (grands) échantillons indépendants : méthode générale

On note μ_1 et σ_1 la moyenne et l'écart-type de la variable X_1, et μ_2 et σ_2 la moyenne et l'écart-type de la variable X_2. Selon que l'on peut supposer que $\sigma_1 = \sigma_2$ ou bien que $\sigma_1 \neq \sigma_2$, le test de comparaison des moyennes se conduira différemment.

On doit déterminer au préalable si on peut suppose $\sigma_1 = \sigma_2$ ou bien $\sigma_2 \neq \sigma_2$, à l'aide d'un *test de comparaison des écart-types*. On étudie pour cela deux échantillons de taille respective n_1 et n_2.

a) Test de comparaison des variances (ou écart-types) de X_1 et X_2 Cette année, faute de temps, on ne pourra sans doute pas étudier ce test. Donc, dans les exercices de comparaison de moyennes, dans le cas de deux grands échantillons indépendants, les exercices mentionneront s'il faut suppose $\sigma_1 = \sigma_2$ ou $\sigma_1 \neq \sigma_2$.

On procède à un test classique :

$$H_0 : \sigma_1 = \sigma_2,\ H_1 : \sigma_1 \neq \sigma_2.$$

C'est donc un test ***bilatéral***. A partir des deux échantillons, on calcule les écart -types expérimentaux s_1^e et s_2^e, ainsi que les écarts-types corrigés, $\hat{s}_1^e$ et $\hat{s}_2^e$. Supposons que $\hat{s}_1^e > \hat{s}_2^e$. Alors, *sous l'hypothèse nulle*, la variable aléatoire $F = \dfrac{\widehat{S}_{n_1}}{\widehat{S}_{n_2}}$, où $\widehat{S}_{n_1}$ et $\widehat{S}_{n_2}$ représentent les écart-types corrigés de X_1 et X_2 sur des échantillons aléatoires de tailles respectives n_1 et n_2, vérifie :

$$F \hookrightarrow \mathrm{FS}(n_1 - 1, n_2 - 1),$$

où $\mathrm{FS}(n_1 - 1, n_2 - 1)$ est la loi de ***Fisher-Snedecor avec les degrés de liberté*** $n_1 - 1$ et $n_2 - 1$.

Si on prend $\alpha = 0.05$, on détermine sur les tables de la loi de Fisher-Snedecor la région critique $K_{0.025} = \{F \geq f_\alpha\}$, avec $\mathbb{P}[K_{0.025}; \mathrm{FS}(n_1 - 1, n_2 - 1)] = 0.025$. On décide comme d'habitude selon que la quantité expérimentale $f_e = \frac{s_1^e}{s_2^e}$ appartient ou non à K_α.

b) Test de comparaison des moyennes, grands échantillons, premier cas : $\sigma_1 \neq \sigma_2$. On suppose que le test de comparaison des écart-types a permis de conclure que $\sigma_1 \neq \sigma_2$. Les hypothèses du test sont :

$$H_0 : \mu_1 = \mu_2;\ H_1 : \mu_1 < \mu_2,\ \text{ou } \mu_1 > \mu_2 \text{ ou } \mu_1 \neq \mu_2.$$

On travaille sur deux échantillons de taille respective $n_1 \geq 30$, $n_2 \geq 30$, sur lesquels les variables X_1 et X_2 ont les moyennes expérimentales respectives m_1^e et m_2^e et les écart-types expérimentaux respectifs s_1^e et s_2^e. Sous l'hypothèse nulle, la variable aléatoire $M = M_{n_1} - M_{n_2}$, où M_{n_1} et M_{n_2} sont les moyennes de X_1 et X_2 sur des échantillons aléatoires de taille respectives n_1 et n_2, vérifie :

$$M = M_{n_1} - M_{n_2} \hookrightarrow \mathcal{N}\left(0, \sqrt{\frac{(s_1^e)^2}{n_1 - 1} + \frac{(s_2^e)^2}{n_2 - 1}}\right).$$

c) Test de comparaison des moyennes, grands échantillons, deuxième cas : $\sigma_1 = \sigma_2$. On suppose que le test de comparaison des écart-types a permis de conclure que $\sigma_1 = \sigma_2$. Les hypothèses du test sont :

$$H_0 : \mu_1 = \mu_2;\ H_1 : \mu_1 < \mu_2,\ \text{ou } \mu_1 > \mu_2 \text{ ou } \mu_1 \neq \mu_2.$$

On travaille sur deux échantillons de taille respective $n_1 \geq 30$, $n_2 \geq 30$, sur lesquels les variables X_1 et X_2 ont les moyennes respectives m_1^{e} et m_2^{e} écart-types expérimentaux respectifs s_1^{e} et s_2^{e}. Sous l'hypothèse nulle, la variable aléatoire $M = M_{n_1} - M_{n_2}$, où M_{n_1} et M_{n_2} sont les moyennes de X_1 et X_2 sur des échantillons aléatoires de taille respectives n_1 et n_2, vérifie :

$$M = M_{n_1} - M_{n_2} \hookrightarrow \mathcal{N}\left(0, s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right), \text{ avec } s = \sqrt{\frac{n_1 \left(s_1^{\mathrm{e}}\right)^2 + n_2 \left(s_2^{\mathrm{e}}\right)^2}{n_1 + n_2 - 2}}.$$

5.2.4 Test de comparaison des moyennes : exemple

Les résultats suivants résument les résultats d'un test de développement intellectuel sur deux groupes professionnels :109 travailleurs manuels et 77 cadres. Les résultats sont les suivants :

Résultats	Effectifs Travailleurs Manuels : groupe 1	Effectifs Cadres : groupe 2
$[0, 8[$	6	1
$[8, 16[$	37	7
$[16, 24[$	50	33
$[24, 32[$	16	30
$[32, 40[$	0	6

Peut-on en déduire que les scores de développement intellectuel des cadres sont en moyenne différents de ceux des travailleurs manuels ?

On teste (rapidement) l'égalité des écart-types.

On a $n_1 = 109$, $m_1^{\mathrm{e}} = 17.58$, $s_1^{\mathrm{e}} = 6.2725$, $\hat{s}_1^{\mathrm{e}} = 6.3015$, et $n_2 = 77$, $m_2^{\mathrm{e}} = 23.4286$, $s_2^{\mathrm{e}} = 6.5016$, $\hat{s}_2^{\mathrm{e}} = 6.544$.

Donc $f_{\mathrm{e}} = 6.544/6.3015 = 1,038$. Or pour $\alpha = 0.05$, $K_\alpha = \{F \geq 1.507\}$. Donc $f_{\mathrm{e}} \notin K_\alpha$, on considère $\sigma_1 = \sigma_2$. On procède donc au test de comparaison des moyennes, dans le cas de grands échantillons ($n_1, n_2 \geq 30$) et d'égalité des écart-types.

Etape 1.

$$H_0 : \mu_1 = \mu_2,\ H_2 : \mu_1 \neq \mu_2.$$

Etape 2. $s = 6,4028593$, et donc $s_{\mathrm{e}} = s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = 0,953$. Ainsi, sous l'hypothèse nulle, $M = M_{n_1} - M_{n_2} \hookrightarrow \mathcal{N}(0, 0.953)$.

Etape 3. On fait un test bilatéral avec la loi $\mathcal{N}(0, 0.953)$. Donc, avec $\alpha = 0.05$, on trouve $K_\alpha = \{M \leq -1.868\} \cup \{M \geq 1.868\}$.

Etape 4. On voit que $m_{\mathrm{e}} = m_1^{\mathrm{e}} - m_2^{\mathrm{e}} = 21.8 - 23.4286 = -1,6286 \notin K_\alpha$, donc on conserve l'hypothèse H_0 : les moyennes de scores des travailleurs manuels et des cadres sont significativement égales.

6 Test d'indépendance du χ^2

6.1 Problème

On étudie ici deux variables X et Y sur une même population. On cherche à déterminer si X et Y sont ***liées*** ou ou ***indépendantes***. Par *indépendantes*, on veut dire que le fait d'appartenir à une modalité de la première variable n'a pas d'influence sur la modalité d'appartenance de la deuxième variable.

Si par exemple X est la taille et Y le poids des individus, il est clair que X et Y sont ***liées***.

En revanche, si X est la taille et Y le salaire (sauf peut-être dans certains métiers), on s'attend à ce que X et Y soient ***indépendantes***.

Il s'agit d'évaluer si la répartition des effectifs dans une *table de contingence* est significativement différente de celle de la table calculée sous l'hypothèse d'indépendance des deux variables croisées (et dont les valeurs sont dites *valeurs théoriques*).

6.2 Exemple et méthode générale

On illustre la méthode générale sur l'exemple suivant.

On mène une étude sur le rapport éventuel chez les hommes entre la situation maritale et l'emploi. Sur un échantillon de 1074 hommes, on obtient les résultats suivants :

		1	2	3
		marié	séparé ou veuf	jamais marié
1	avec emploi	679	103	114
2	sans emploi	63	10	20
3	hors statistiques	42	18	25

Dans cet échantillon, les profils d'emploi semblent différents selon la situation maritale. Par exemple, être marié semble être lié au fait d'avoir un emploi. On veut donc savoir si cette différence est significative.

On introduit donc les variables aléatoires suivantes :

i) La variable *qualitative* X, qui décrit la situation maritale.

ii) La variable *qualitative* Y, qui décrit la situation par rapport à l'emploi.

Hypothèses. Pour ce type de test, les hypothèses seront toujours :

$$H_0 : \text{ les deux variables sont indépendantes}$$
$$H_1 : \text{ les deux variables sont ne le sont pas (donc sont dépendantes)}$$

Attention : dire que deux variables sont dépendantes ne signifie pas que l'une est la cause de l'autre !

Statistique du test. Pour chaque case (i, j) du tableau, qui est le croisement de la ligne i et de la colonne j, on calcule l'***effectif théorique*** : c'est le nombre :

$$n_{ij}^{\text{th}} = \frac{n_i m_j}{n},$$

où n_i est la somme des effectifs de la ligne i, n_j est la somme des effectifs de la colonne j, et n est la taille totale de l'échantillon. On reporte tous ces nombres sur le tableau :

		1		2		3		
		marié		séparé ou veuf		jamais marié		
1	avec emploi	679	654	103	109	114	133	896
2	sans emploi	63	68	10	11	20	14	93
3	hors statistiques	42	62	18	10	25	13	85
		784		131		159		1074

Alors, sous hypothèse nulle, la variable aléatoire $Y = \sum_{i,j} \frac{\left(n_{ij} - n_{ij}^{\text{th}}\right)^2}{n_{ij}^{\text{th}}}$ suit ***une loi du*** χ^2 ***à*** $(\ell - 1) \times (c - 1)$ ***degrés de liberté***, où ℓ est le nombre de lignes et c le nombre de colonnes, et n_{ij} la variable aléatoire qui compte le nombres d'individus dans la case (i, j) sur un échantillon aléatoire de taille n. On note :

$$Y = \sum_{i,j} \frac{\left(n_{ij} - n_{ij}^{\text{th}}\right)^2}{n_{ij}^{\text{th}}} \hookrightarrow \chi^2\left((\ell - 1) \times (c - 1)\right).$$

Dans l'exemple, on a $Y \hookrightarrow \chi^2(4)$.

Région critique. On fixe le niveau d'erreur $\alpha = 0.05$. Les grandes valeurs de Y sont favorables à H_1, donc la région critique est faite des grandes valeurs de Y. Dans le formulaire, on trouve $y_\alpha = 9.488$, et donc :

$$K_\alpha = \{Y \geq 9.488\}.$$

Décision. Un calcul direct montre que $y_{\text{e}} \sim 31$. Donc $y_{\text{e}} \in K_\alpha$: au niveau $\alpha = 0.05$, on admet l'hypothèse H_1, qui dit que les variables "situation maritale" et "situation d'emploi" sont significativement liées.