LES TESTS D'HYPOTHÈSE

1. *GÉNÉRALITÉS*

1.1. PRINCIPE D'UN TEST D'HYPOTHÈSES

Les tests d'hypothèse constituent un autre aspect important de l'inférence statistique. Le principe général d'un test d'hypothèse peut s'énoncer comme suit :

- On étudie une population dont les éléments possèdent un caractère (mesurable ou qualitatif) et dont la valeur du paramètre relative au caractère étudié est inconnue.

- Une hypothèse est formulée sur la valeur du paramètre : cette formulation résulte de considérations théoriques, pratiques ou encore elle est simplement basée sur un pressentiment.

- On veut porter un jugement sur la base des résultats d'un échantillon prélevé de cette population.

Il est bien évident que la statistique (c'est-à-dire la variable d'échantillonnage) servant d'estimateur au paramètre de la population ne prendra pas une valeur rigoureusement égale à la valeur théorique proposée dans l'hypothèse. Cette variable aléatoire comporte des fluctuations d'échantillonnage qui sont régies par des distributions connues.

Pour décider si l'hypothèse formulée est supportée ou non par les observations, il faut une méthode qui permettra de conclure si l'écart observé entre la valeur de la statistique obtenue dans l'échantillon et celle du paramètre spécifiée dans l'hypothèse est trop important pour être uniquement imputable au hasard de l'échantillonnage.

La construction d'un test d'hypothèse consiste en fait à déterminer entre quelles valeurs peut varier la variable aléatoire, en supposant l'hypothèse vraie, sur la seule considération du hasard de l'échantillonnage.

Les distributions d'échantillonnage d'une moyenne, d'une variance et d'une proportion que nous avons traitées dans un chapitre précédent vont être particulièrement utiles dans l'élaboration des tests statistiques.

1.2. DÉFINITION DES CONCEPTS UTILES A L'ÉLABORATION DES TESTS D'HYPOTHÈSE

Hypothèse statistique

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

Test d'hypothèse

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.

Hypothèse nulle (H_0) et hypothèse alternative (H_1)

L'hypothèse selon laquelle on fixe à priori un paramètre de la population à une valeur particulière s'appelle **l'hypothèse nulle** et est notée H_0. N'importe quelle autre hypothèse qui diffère de l'hypothèse H_0 s'appelle **l'hypothèse alternative** (ou contre-hypothèse) et est notée H_1.

C'est l'hypothèse nulle qui est soumise au test et toute la démarche du test s'effectue en considérant cette hypothèse comme vraie.

Dans notre démarche, nous allons établir des règles de décision qui vont nous conduire à l'acceptation ou au rejet de l'hypothèse nulle H_0. Toutefois cette décision est fondée sur une information partielle, les résultats d'un échantillon. Il est donc statistiquement impossible de prendre la bonne décision à coup sûr. En pratique, on met en oeuvre une démarche qui nous permettrait, à long terme de rejeter à tort une hypothèse nulle vraie dans une faible proportion de cas. La conclusion qui sera déduite des résultats de l'échantillon aura un caractère probabiliste : on ne pourra prendre une décision qu'en ayant conscience qu'il y a un certain risque qu'elle soit erronée. Ce risque nous est donné par le seuil de signification du test.

Seuil de signification du test

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 signification** du test et s'énonce en probabilité ainsi : $\alpha = P(\text{rejeter } H_0 | H_0 \text{ vraie})$.

A ce seuil de signification, on fait correspondre sur la distribution d'échantillonnage de la statistique une **région de rejet** de l'hypothèse nulle (appelée également région critique). L'aire de cette région correspond à la probabilité α. Si par exemple , on choisit $\alpha = 0.05$, cela signifie que l'on admet d'avance que la variable d'échantillonnage peut prendre, dans 5% des cas, une valeur se situant dans la zone de

rejet de H_0, bien que H_0 soit vraie et ceci uniquement d'après le hasard de l'échantillonnage.

Sur la distribution d'échantillonnage correspondra aussi une région complémentaire, dite **région d'acceptation** de H_0 (ou région de non-rejet) de probabilité $1-\alpha$.

Remarques : **1.** Les seuils de signification les plus utilisés sont $\alpha = 0.05$ et $\alpha = 0.01$, dépendant des conséquences de rejeter à tort l'hypothèse H_0.

2. La statistique qui convient pour le test est donc une variable aléatoire dont la valeur observée sera utilisée pour décider du « rejet » ou du « non-rejet » de H_0. La distribution d'échantillonnage de cette statistique sera déterminée en supposant que l'hypothèse H_0 est vraie.

Exemple de formulation d'un test :

Supposons que nous affirmions que la valeur d'un paramètre θ d'une population est égale à la valeur θ_0. On s'intéresse au changement possible du paramètre θ dans l'une ou l'autre direction (soit $\theta > \theta_0$ soit $\theta < \theta_0$). On effectue un test bilatéral.

Les hypothèses H_0 et H_1 sont alors : $\begin{cases} H_0 : & \theta = \theta_0 \\ H_1 : & \theta \neq \theta_0 \end{cases}$

On peut schématiser les régions de rejet et de non-rejet de H_0 comme suit :

Si, suite aux résultats de l'échantillon, la valeur de la statistique utilisée se situe dans l'intervalle $[\theta_{c_1}, \theta_{c_2}]$, on acceptera H_0 au seuil de signification choisi. Si, au contraire, la valeur obtenue est supérieure à θ_{c_2} ou inférieure à θ_{c_1}, on rejette H_0 et on accepte H_1.

Remarque : Si on s'intéresse au changement du paramètre dans une seule direction, on opte pour un **test unilatéral**, en choisissant comme hypothèse H_1

soit $\theta > \theta_0$ soit $\theta < \theta_0$. La région critique est alors localisée uniquement à droite ou uniquement à gauche de la région d'acceptation.

Dans un souci de simplification, nous nous intéresserons dans ce cours essentiellement aux tests bilatéraux.

2. TESTS PERMETTANT DE DÉTERMINER SI UN ÉCHANTILLON APPARTIENT A UNE POPULATION DONNÉE

2.1. TESTS SUR UNE MOYENNE : COMPARAISON D'UNE MOYENNE EXPÉRIMENTALE A UNE MOYENNE THÉORIQUE DANS LE CAS D'UN CARACTÈRE QUANTITATIF

Nous voulons déterminer si l'échantillon de taille n dont nous disposons appartient à une population de moyenne m_0 au seuil de signification α.
Nous allons dans tous les tests travailler de la même façon, en procédant en quatre étapes.

1ère étape : formulation des hypothèses

L'échantillon dont nous disposons provient d'une population de moyenne m. Nous voulons savoir si $m = m_0$.

On va donc tester l'hypothèse H_0 contre l'hypothèse H_1 : $\begin{cases} H_0 : & m = m_0 \\ H_1 : & m \neq m_0 \end{cases}$.

2ème étape : Détermination de la fonction discriminante du test et de sa distribution de probabilité.

- On détermine la statistique qui convient pour ce test.. Ici, l'estimateur de la moyenne m, c'est-à-dire $\overline{X}$, semble tout indiquée.
- On détermine la loi de probabilité de $\overline{X}$ **en se plaçant sous l'hypothèse H_0**. Deux cas peuvent se produire :

Premier cas: L'échantillon est de grande taille ($n \geq 30$) ou bien la population est normale de variance σ_{pop}^2 connue.

$\overline{X}$ suit alors une loi normale de moyenne m_0 (puisqu'on se place sous H_0) et d'écart-type $\frac{\sigma_{pop}}{\sqrt{n}}$: $\overline{X} \rightsquigarrow N(m_0, \frac{\sigma_{pop}}{\sqrt{n}})$. On pose $T = \frac{\overline{X} - m_0}{\sigma_{pop} / \sqrt{n}}$.

T mesure un écart réduit. T est aussi appelée **fonction discriminante du test.**

$T \rightsquigarrow N(0,1)$.

Deuxième cas : L'échantillon est de petite taille ($n < 30$) prélevé au hasard d'une population normale de variance σ_{pop}^2 inconnue.

Dans ce cas la **fonction discriminante du test** sera : $T = \dfrac{\overline{X} - m_0}{\Sigma_{ech} / \sqrt{n-1}}$.

Ici $T \rightsquigarrow T_{n-1}$ (loi de Student à (n-1) degrés de liberté).

3ème étape : Détermination des valeurs critiques de T délimitant les zones d'acceptation et de rejet

On impose toujours à la zone d'acceptation de H_0 concernant l'écart réduit d'être centrée autour de 0.

Il nous faut donc déterminer dans la table la valeur maximale $t_{\alpha/2}$ de l'écart réduit imputable aux variations d'échantillonnage au seuil de signification α, c'est-à-dire vérifiant : $P(-t_{\alpha/2} \leq T \leq t_{\alpha/2}) = 1 - \alpha$.

4ème étape : Calcul de la valeur de T prise dans l'échantillon et conclusion du test

- On calcule la valeur t_0 prise par T dans l'échantillon.
- → Si la valeur t_0 se trouve dans la zone de rejet, on dira que l'écart-réduit observé est **statistiquement significatif** au seuil α. Cet écart est anormalement élevé et ne permet pas d'accepter H_0. On rejette H_0.
  → Si la valeur t_0 se trouve dans la zone d'acceptation, on dira que l'écart-réduit observé **n'est pas significatif** au seuil α. Cet écart est imputable aux fluctuations d'échantillonnage. On accepte H_0.

2.2. TESTS SUR UNE PROPORTION

Nous nous proposons de tester si la proportion p d'éléments dans la population présentant un certain caractère qualitatif peut être ou non considérée comme égale à une valeur hypothétique p_0. Nous disposons pour ce faire de la proportion d'éléments possédant ce caractère dans un échantillon de taille n. Nous allons procéder comme au paragraphe précédent, en quatre étapes.

1ère étape : formulation des hypothèses

L'échantillon dont nous disposons provient d'une population dont la proportion d'éléments présentant le caractère qualitatif est p. Nous voulons savoir si $p = p_0$.

On va donc tester l'hypothèse H_0 contre l'hypothèse H_1 : $\begin{cases} H_0 : & p = p_0 \\ H_1 : & p \neq p_0 \end{cases}$.

2ème étape : **Détermination de la fonction discriminante du test et de sa distribution de probabilité.**

- On détermine la statistique qui convient pour ce test. Ici, l'estimateur de la proportion p, c'est-à-dire F, semble tout indiquée.
- On détermine la loi de probabilité de F **en se plaçant sous l'hypothèse H_o**.

  On suppose que l'on dispose d'un grand échantillon ($n \geq 30$) et que « p n'est pas trop petit » (de manière que l'on ait $np \geq 15$ et $n(1-p) \geq 15$).

  F suit alors une loi normale de moyenne p_0 (puisqu'on se place sous H_0) et d'écart-type $\sqrt{\frac{p_0(1-p_0)}{n}}$: $F \rightsquigarrow N(p_0, \sqrt{\frac{p_0(1-p_0)}{n}})$.

  On pose $T = \frac{F - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$. T mesure un écart réduit.

  T est aussi appelée **fonction discriminante du test.** $T \rightsquigarrow N(0,1)$.

3ème étape : Détermination des valeurs critiques de T délimitant les zones d'acceptation et de rejet

On impose toujours à la zone d'acceptation de H_0 concernant l'écart réduit d'être centrée autour de 0.

Il nous faut donc déterminer dans la table la valeur maximale $t_{\alpha/2}$ de l'écart réduit imputable aux variations d'échantillonnage au seuil de signification α, c'est-à-dire vérifiant : $P(-t_{\alpha/2} \leq T \leq t_{\alpha/2}) = 1 - \alpha$.

4ème étape : Calcul de la valeur de T prise dans l'échantillon et conclusion du test

- On calcule la valeur t_0 prise par T dans l'échantillon.

- → Si la valeur t_0 se trouve dans la zone de rejet, on dira que l'écart-réduit observé est **statistiquement significatif** au seuil α. Cet écart est anormalement élevé et ne permet pas d'accepter H_0. On rejette H_0.
  → Si la valeur t_0 se trouve dans la zone d'acceptation, on dira que l'écart-réduit observé **n'est pas significatif** au seuil α. Cet écart est imputable aux fluctuations d'échantillonnage. On accepte H_0.

Nous étudierons ces sortes de tests sur des exemples en travaux dirigés.

3. *RISQUES DE PREMIÈRE ET DE DEUXIÈME ESPÈCE*

3.1. DÉFINITIONS

Tous les règles de décision que nous avons déterminées acceptaient un risque α qui était le risque de rejeter à tort l'hypothèse H_0, c'est-à-dire le risque de rejeter l'hypothèse H_0, alors que H_0 est vraie. Ce risque s'appelle aussi le **risque de première espèce.**

La règle de décision du test comporte également un deuxième risque, à savoir de celui de ne pas rejeter l'hypothèse nulle H_0 alors que c'est l'hypothèse H_1 qui est vraie. C'est le **risque de deuxième espèce.**

Les deux risques peuvent se définir ainsi :
α = P(rejeter $H_0 \mid H_0$ vraie) = probabilité de comm ettre une erreur de première espèce
β = P(ne pas rejeter $H_0 \mid H_1$ vraie) = probabilité de comm ettre une erreur de deuxième espèce

Le risque de première espèce α est choisi à priori. Toutefois le risque de deuxième espèce β dépend de l'hypothèse alternative H_1 et on ne peut le calculer que si on spécifie des valeurs particulières du paramètre dans l'hypothèse H_1 que l'on suppose vraie.

Les risques liés aux tests d'hypothèses peuvent se résumer ainsi :

		SITUATION VRAIE			
		H_0 EST VRAIE		**H_1 EST VRAIE**	
		La décision est	probabilité de prendre cette décision avant expérience	La décision est	probabilité de prendre cette décision avant expérience
					β

Conclusion du test	**Accepter H_0**	**Bonne**	$1 - \alpha$	**Fausse**	(risque de deuxième espèce)
	Rejeter H_0	**Fausse**	α (risque de première espèce)	**Bonne**	$1 - \beta$

Remarque : La probabilité complémentaire du risque de deuxième espèce ($1-\beta$) définit la **puissance du test** à l'égard de la valeur du paramètre dans l'hypothèse alternative H_1. La puissance du test représente la probabilité de rejeter l'hypothèse nulle H_0 lorsque l'hypothèse vraie est H_1. Plus β est petit, plus le test est puissant.

3.2. SCHÉMATISATION DES DEUX RISQUES D'ERREUR SUR LA DISTRIBUTION D'ÉCHANTILLONNAGE

A titre d'exemple, regardons ce qu'il se passe à propos d'un test sur la moyenne.
On peut visualiser sur la distribution d'échantillonnage de la moyenne comment sont reliés les deux risques d'erreur associés aux tests d'hypothèses.

Les zones d'acceptation de H_0 ($m = m_0$)
et de rejet de H_0 se visualisent ainsi :

Donnons diverses valeurs à m (autres que m_0)
que l'on suppose vraie et schématisons le risque
de deuxième espèce β.

Hypothèse vraie : $m = m_1$ ($m_1 < m_0$)

La distribution d'échantillonnage de $\overline{X}$
en supposant vraie $m = m_1$ est illustrée en pointillé
et l'aire hachurée sur cette figure correspond
à la région de non-rejet de H_0.
Cette aire représente β par rapport à la valeur m_1.

Hypothèse vraie : $m = m_2$ ($m_2 > m_0$)

Hypothèse vraie : $m = m_3$ ($m_3 > m_0$)

Cette schématisation permet d'énoncer quelques propriétés importantes concernant les deux risques d'erreur :

1. Pour un même risque α et une même taille d'échantillon, on constate que, si l'écart entre la valeur du paramètre posée en H_0 et celle supposée dans l'hypothèse vraie H_1 augmente, le risque β diminue.

2. Une réduction du risque de première espèce (de $\alpha = 0.05$ à $\alpha = 0.01$ par exemple) élargit la zone d'acceptation de H_0. Toutefois, le test est accompagné d'une augmentation du risque de deuxième espèce β. On ne peut donc diminuer l'un des risques qu'en consentant à augmenter l'autre.

3. Pour une valeur fixe de α et un σ déterminé, l'augmentation de la taille d'échantillon aura pour effet de donner une meilleure précision puisque $\sigma(\overline{X}) = \frac{\sigma}{\sqrt{n}}$ diminue. La zone d'acceptation de H_0 sera alors plus restreinte, conduisant à une diminution du risque β. Le test est alors plus puissant.

4. *TESTS PERMETTANT DE DÉTERMINER SI DEUX ÉCHANTILLONS APPARTIENNENT A LA MÊME POPULATION*

4.1. INTRODUCTION

Il existe de nombreuses applications qui consistent, par exemple, à comparer deux groupes d'individus en regard d'un caractère quantitatif particulier (poids, taille, rendement scolaire, quotient intellectuel,....) ou à comparer deux procédés de fabrication selon une caractéristique quantitative particulière (résistance à la rupture, poids, diamètre, longueur,...) ou encore de comparer les proportions d'apparition d'un caractère qualitatif de deux populations (proportion de défectueux, proportion de gens favorisant un parti politique,...).

Les variables aléatoires qui sont alors utilisées pour effectuer des tests d'hypothèses (ou aussi calculer des intervalles de confiance) sont la différence des moyennes d'échantillon, le quotient des variances d'échantillon ou la différence des proportions d'échantillon.

4.2. ON ÉTUDIE UN CARACTÈRE QUANTITATIF

4.2.1. Comparaison de deux moyennes d'échantillon : « test T »

Nous nous proposons de tester si la moyenne de la première population (m_1) peut être ou non considérée comme égale à la moyenne de la deuxième population (m_2). Nous allons alors comparer les deux moyennes d'échantillon $\overline{x}_1$ et $\overline{x}_2$. Il est évident que si $\overline{x}_1$ et $\overline{x}_2$ diffèrent beaucoup, les deux échantillons n'appartiennent pas la même population. Mais si $\overline{x}_1$ et $\overline{x}_2$ diffèrent peu, il se pose la question de savoir si l'écart

$d = \overline{x}_1 - \overline{x}_2$ peut être attribué aux hasards de l'échantillonnage. Afin de donner une réponse rigoureuse à cette question, nous procéderons encore en quatre étapes.

1ère étape : formulation des hypothèses

Le premier échantillon dont nous disposons provient d'une population dont la moyenne est m_1. Le deuxième échantillon dont nous disposons provient d'une population dont la moyenne est m_2.

Nous voulons savoir si il s'agit de la même population en ce qui concerne les moyennes, c'est-à-dire si $m_1 = m_2$.

On va donc tester l'hypothèse H_0 contre l'hypothèse H_1 : $\begin{cases} H_0 : \ m_1 = m_2 \\ \\ H_1 : \ m_1 \neq m_2 \end{cases}$.

2ème étape : Détermination de la fonction discriminante du test et de sa distribution de probabilité.

- On détermine la statistique qui convient pour ce test. Ici, la différence $D = \overline{X}_1 - \overline{X}_2$ des deux moyennes d'échantillon, semble tout indiquée.
- On détermine la loi de probabilité de D **en se plaçant sous l'hypothèse H_0**.
  On suppose que l'on dispose de grands échantillons ($n_1 \geq 30$ et $n_2 \geq 30$)et que les deux variances d'échantillon σ_{ech1}^2 et σ_{ech2}^2 sont connues.

  $\Rightarrow \overline{X}_1$ suit alors une loi normale de moyenne m_1 et d'écart-type $\frac{\sigma_{pop1}}{\sqrt{n_1}}$ que l'on peut sans problème estimer par $\frac{\sigma_{ech1}}{\sqrt{n_1 - 1}}$ (car $n_1 \geq 30$).

  $$\overline{X}_1 \rightsquigarrow N(m_1, \frac{\sigma_{ech1}}{\sqrt{n_1 - 1}}).$$

  $\Rightarrow$ De même $\overline{X}_2$ suit alors une loi normale de moyenne m_2 et d'écart-type $\frac{\sigma_{pop2}}{\sqrt{n_2}}$ que l'on peut sans problème estimer par $\frac{\sigma_{ech2}}{\sqrt{n_2 - 1}}$ (car $n_2 \geq 30$).

  $$\overline{X}_2 \rightsquigarrow N(m_2, \frac{\sigma_{ech2}}{\sqrt{n_2 - 1}}).$$

  $\Rightarrow$ On en déduit, puisque $\overline{X}_1$ et $\overline{X}_2$ sont indépendantes que $D = \overline{X}_1 - \overline{X}_2$ suit également une loi normale.

  $E(D) = E(\overline{X}_1) - E(\overline{X}_2) = m_1 - m_2 = 0$ puisqu'on se place sous H_0.

$V(D) = V(\overline{X}_1) + V(\overline{X}_2) = \frac{\sigma_{ech1}^{\ 2}}{n_1 - 1} + \frac{\sigma_{ech2}^{\ 2}}{n_2 - 1}$ puisque les variables sont indépendantes.

On pose $T = \dfrac{D}{\sqrt{\dfrac{{\sigma_{ech1}}^2}{n_1 - 1} + \dfrac{{\sigma_{ech2}}^2}{n_2 - 1}}}$. T mesure un écart réduit.

T est la **fonction discriminante du test.** $T \rightsquigarrow N(0,1)$.

<u>3ème étape</u> : Détermination des valeurs critiques de T délimitant les zones d'acceptation et de rejet

On impose toujours à la zone d'acceptation de H_0 concernant l'écart réduit d'être centrée autour de 0.

Il nous faut donc déterminer dans la table la valeur maximale $t_{\alpha/2}$ de l'écart réduit imputable aux variations d'échantillonnage au seuil de signification α, c'est-à-dire vérifiant : $P(-t_{\alpha/2} \leq T \leq t_{\alpha/2}) = 1 - \alpha$.

<u>4ème étape</u> : Calcul de la valeur de T prise dans l'échantillon et conclusion du test

- On calcule la valeur t_0 prise par T dans l'échantillon.
- → Si la valeur t_0 se trouve dans la zone de rejet, on dira que l'écart-réduit observé est **statistiquement significatif** au seuil α. Cet écart est anormalement élevé et ne permet pas d'accepter H_0. <u>On rejette H_0</u>.
  → Si la valeur t_0 se trouve dans la zone d'acceptation, on dira que l'écart-réduit observé **n'est pas significatif** au seuil α. Cet écart est imputable aux fluctuations d'échantillonnage. <u>On accepte H_0</u>.

<u>Remarque</u> : Si on travaille sur de petits échantillons, si la loi suivie par la grandeur est une loi normale et si on ignore les écarts-type des populations, on doit utiliser la loi de Student.

4.2.2. <u>Comparaison de deux variances d'échantillon : « test F »</u>

<u>1ère étape</u> : formulation des hypothèses

Le premier échantillon dont nous disposons provient d'une population dont l'écart-type est σ_{pop1}. Le deuxième échantillon dont nous disposons provient d'une population dont l'écart-type est σ_{pop2}. Nous voulons savoir si il s'agit de la même population en ce qui concerne les écarts-type, c'est-à-dire si $\sigma_{pop1} = \sigma_{pop2}$.

On va donc tester l'hypothèse H_0 contre l'hypothèse H_1 : $\begin{cases} H_0 : \ \sigma_{pop1} = \sigma_{pop2} \\ \\ H_1 : \ \sigma_{pop1} \neq \sigma_{pop2} \end{cases}$.

<u>2ème étape</u> : **Détermination de la fonction discriminante du test et de sa distribution de probabilité.**

- On détermine la statistique qui convient pour ce test.. Ici, la variable aléatoire dont on connaît la loi est le rapport $F_0 = \frac{S_1^{\,2}}{S_2^{\,2}}$ où $S_1^{\,2}$ et $S_2^{\,2}$ sont les variables aléatoires variances d'échantillon.
- On détermine la loi de probabilité de F_0 **en se plaçant sous l'hypothèse H_0**.

On suppose ici que les deux populations dont nous avons tiré les échantillons sont **normales.** Il en découle que :

$\Rightarrow \frac{(n_1 - 1)S_1^{\,2}}{\sigma_{pop1}^{\,2}}$ suit la loi du khi-deux à $(n_1 - 1)$ ddl.

$\Rightarrow$ De même $\frac{(n_2 - 1)S_2^{\,2}}{\sigma_{pop2}^{\,2}}$ suit la loi du khi-deux à $(n_2 - 1$ ddl).

On considère alors le quotient $F_0 = \frac{\frac{S_1^{\,2}}{\sigma_{pop1}^{\,2}}}{\frac{S_2^{\,2}}{\sigma_{pop2}^{\,2}}}$ qui est distribué suivant la loi de Fisher avec $\nu_1 = n_1 - 1$ et $\nu_2 = n_2 - 1$ degrés de liberté.

Lorsqu'on se place sous l'hypothèse H_0, c'est le rapport : $F_0 = \frac{S_1^{\,2}}{S_2^{\,2}}$ qui suit la loi de Fisher avec $\nu_1 = n_1 - 1$ et $\nu_2 = n_2 - 1$ degrés de liberté puisque $\sigma_{pop1} = \sigma_{pop2}$.

Ici la **fonction discriminante du test** est F_0.

<u>3ème étape</u> : Détermination des valeurs critiques de F_0 délimitant les zones d'acceptation et de rejet

On impose maintenant à la zone d'acceptation de H_0 concernant le quotient des deux variances d'échantillon d'être centrée autour de 1.
On détermine dans les tables les deux valeurs $F_{\alpha/2,\nu_1,\nu_2}$ et $F_{1-\frac{\alpha}{2},\nu_1,\nu_2}$ telles que : $P(F_{\alpha/2,\nu_1,\nu_2} < F_0 < F_{1-\frac{\alpha}{2},\nu_1,\nu_2}) = 1 - \alpha$.
On rejettera H_0 si la valeur f_0 prise par F_0 dans l'échantillon

se trouve à l'extérieur de l'intervalle $[F_{\alpha/2,\nu_1,\nu_2}, F_{1-\frac{\alpha}{2},\nu_1,\nu_2}]$.

Remarque : On notera que pour obtenir la valeur critique inférieure de F_0, on doit utiliser la relation : $F_{1-\frac{\alpha}{2},\nu_1,\nu_2} = \frac{1}{F_{\alpha/2,\nu_2,\nu_1}}$

4ème étape : Calcul de la valeur de F_0 prise dans l'échantillon et conclusion du test

- On calcule la valeur f_0 prise par F_0 dans l'échantillon.

- → Si la valeur f_0 se trouve dans la zone de rejet, on dira que la valeur observée pour F est **statistiquement significative** au seuil α. Ce quotient est éloigné de 1 et ne permet pas d'accepter H_0. On rejette H_0.
  → Si la valeur f_0 se trouve dans la zone d'acceptation, on dira que la valeur observée pour F **n'est pas significative** au seuil α. L'écart constaté par rapport à la valeur 1 attendue est imputable aux fluctuations d'échantillonnage. On accepte H_0.

4.3. ON ÉTUDIE UN CARACTÈRE QUALITATIF : COMPARAISON DE DEUX PROPORTIONS ÉCHANTILLON

Il y a de nombreuses applications (échéances électorales, expérimentations médicales...) où nous devons décider si l'écart observé entre deux proportions échantillonnales est significatif où s'il est attribuable au hasard de l'échantillonnage. Pour répondre à cette question, nous procéderons comme d'habitude en quatre étapes.

1ère étape : formulation des hypothèses

Le premier échantillon dont nous disposons provient d'une population 1 dont les éléments possèdent un caractère qualitatif dans une proportion inconnue p_1. Le deuxième échantillon dont nous disposons provient d'une population 2 dont les éléments possèdent le même caractère qualitatif dans une proportion inconnue p_2.

Nous voulons savoir si il s'agit de la même population en ce qui concerne les proportions, c'est-à-dire si $p_1 = p_2$.

On va donc tester l'hypothèse H_0 contre l'hypothèse H_1 : $\begin{cases} H_0 : \ p_1 = p_2 \\ \\ H_1 : \ p_1 \neq p_2 \end{cases}$.

2ème étape : Détermination de la fonction discriminante du test et de sa distribution de probabilité.

Nous traiterons uniquement le cas où nous sommes en présence de grands échantillons.

- On détermine la statistique qui convient pour ce test. Ici, la différence $D = F_1 - F_2$ des deux proportions d'échantillon, semble tout indiquée, puisque F_1 est un estimateur sans biais de p_1 et F_2 un estimateur sans biais de p_2..

- On détermine la loi de probabilité de D **en se plaçant sous l'hypothèse H_o**.

  $\Rightarrow F_1$ suit alors une loi normale de moyenne p_1 et d'écart-type $\sqrt{\frac{p_1(1-p_1)}{n_1}}$

  $\Rightarrow$ De même, F_2 suit alors une loi normale de moyenne p_2 et d'écart-type $\sqrt{\frac{p_2(1-p_2)}{n_2}}$

  $\Rightarrow$ On en déduit, puisque F_1 et F_2 sont indépendantes que $D = F_1 - F_2$ suit également une loi normale.

  $E(D) = E(F_1) - E(F_2) = p_1 - p_2 = 0$ puisqu'on se place sous H_0.

  $V(D) = V(F_1) + V(F_2) = \frac{p(1-p)}{n_1} + \frac{p(1-p)}{n_2}$ puisque les variables sont indépendantes. Ici, on a posé $p_1 = p_2 = p$ puisque l'on se place sous H_0.

  Mais comment trouver p puisque c'est justement sur p que porte le test ?

  Puisque nous raisonnons en supposant l'hypothèse H_0 vraie, on peut considérer que les valeurs de F_1 et F_2 obtenues sur nos échantillons sont des approximations de p. De plus , plus la taille de l'échantillon est grande, meilleure est l'approximation (revoir le chapitre sur les intervalles de confiance). Nous allons donc pondérer les valeurs observées dans nos échantillons par la taille respective de ces échantillons.

  On approchera p dans notre calcul par : $\hat{p} \approx \frac{n_1 f_1 + n_2 f_2}{n_1 + n_2}$

  On pose $T = \frac{D}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$. T mesure un écart réduit.

  T est la **fonction discriminante du test.** $T \rightsquigarrow N(0,1)$.

<u>3ème étape</u> : Détermination des valeurs critiques de T délimitant les zones d'acceptation et de rejet

On impose toujours à la zone d'acceptation de H_0 concernant l'écart réduit d'être centrée autour de 0.

Il nous faut donc déterminer dans la table
la valeur maximale $t_{\alpha/2}$ de l'écart réduit
imputable aux variations d'échantillonnage
au seuil de signification α, c'est-à-dire
vérifiant : $P(-t_{\alpha/2} \leq T \leq t_{\alpha/2}) = 1-\alpha$.

4ème étape : Calcul de la valeur de T prise dans l'échantillon et conclusion du test

- On calcule la valeur t_0 prise par T dans l'échantillon.

- → Si la valeur t_0 se trouve dans la zone de rejet, on dira que l'écart-réduit observé est **statistiquement significatif** au seuil α. Cet écart est anormalement élevé et ne permet pas d'accepter H_0. On rejette H_0.
  → Si la valeur t_0 se trouve dans la zone d'acceptation, on dira que l'écart-réduit observé **n'est pas significatif** au seuil α. Cet écart est imputable aux fluctuations d'échantillonnage. On accepte H_0.

5. *UNE DISTRIBUTION STATISTIQUE OBÉIT-ELLE A UNE LOI DE PROBABILITÉ DONNÉE ? : TEST D'AJUSTEMENT DE DEUX DISTRIBUTIONS (TEST DU KHI-DEUX)*

5.1. INTRODUCTION

Dans le chapitre 1 de ce cours, nous avons traité de diverses distributions expérimentales dans lesquelles on présentait la répartition des fréquences (absolues ou relatives) pour divers caractères. Lorsque nous avons accumulé suffisamment de données sur une variable statistique, on peut alors examiner si la distribution des observations semble s'apparenter à une distribution théorique connue (comme une loi binomiale, de Poisson, normale...). Un outil statistique qui permet de vérifier la concordance entre une distribution expérimentale et une distribution théorique est le **test de Pearson**, appelé aussi le test du **Khi-deux.**

On cherche donc à déterminer si un modèle théorique est susceptible de représenter adéquatement le comportement probabiliste de la variable observée, comportement fondé sur les fréquences des résultats obtenus sur l'échantillon.

Comment procéder ?

- **<u>Répartitions expérimentales</u>**

  On répartit les observations suivant k classes (si le caractère est continu) ou k valeurs (si le caractère est discret); On dispose alors des effectifs des k classes : n_1, n_2,....., n_k.

  On a bien sûr la relation : $\sum_{i=1}^{i=k} n_i = N$ (N = nombre total d'observations effectuées).

  <u>Remarque</u> : Dans la pratique, on se placera dans le cas où $N \geq 50$ et où chaque n_i est supérieur ou égal à 5. Si cette condition n'est pas satisfaite, il y a lieu de regrouper deux ou plusieurs classes adjacentes. Il arrive fréquemment que ce regroupement s'effectue sur les classes aux extrémités de la distribution. k représente donc le nombre de classes après regroupement.

- **Répartitions théoriques**

En admettant comme plausible une distribution théorique particulière, on peut construire une répartition idéale des observations de l'échantillon de taille N en ayant recours aux probabilités tabulées (ou calculées) du modèle théorique : p_1, p_2, ..., p_k.
On obtient alors les effectifs théoriques n_{t_i} en écrivant : $n_{t_i} = p_i N$. On doit disposer également de la relation : $\sum_{i=1}^{k} n_{t_i} = N$.

- **L'écart entre les deux distributions**

⇒ Définition de l'écart

Pour évaluer l'écart entre les effectifs observées n_i et les effectifs théoriques n_{ti}, on utilise la somme des écarts normalisés entre les deux distributions, à savoir :

$$\chi^2 = \frac{(n_1 - n_{t_1})^2}{n_{t_1}} + \frac{(n_2 - n_{t_2})^2}{n_{t_2}} + \ldots\ldots + \frac{(n_k - n_{t_k})^2}{n_{t_{k1}}}.$$

Plus le χ^2 ainsi calculé est grand, plus la distribution étudiée différera de la distribution théorique.

⇒ Quelques considérations théoriques à propos de cet écart :

Le nombre d'observations n_i parmi l'échantillon de taille N susceptible d'appartenir à la classe i est la réalisation d'une variable binomiale N_i de paramètres N et p_i (chacune des N observations appartient ou n'appartient pas à la classe i avec une probabilité p_i).
Si N est suffisamment grand (on se place dans le cas d'échantillons de taille 50 minimum) et p_i pas trop petit (on a effectué des regroupements de classes pour qu'il en soit ainsi), on peut approcher la loi binomiale par la loi normale, c'est-à-dire B(N, p_i) par $N(Np_i, \sqrt{Np_i(1-p_i)})$.

Or $Np_i(1-p_i) = Np_i - Np_i^2 \approx Np_i$.

$$\text{Donc } T_i = \frac{N_i - Np_i}{\sqrt{Np_i}} \text{ suit la loi N(0, 1).}$$

Lorsqu'on élève au carré toutes ces quantités et qu'on en fait la somme, on obtient une somme de k lois normales centrées réduites indépendantes. Nous avons vu au chapitre 3 que cette somme suivait une loi du khi-deux.

⇒ Mais quel est le nombre de degrés de liberté de cette variable du khi-deux ?

Il y a k carrés indépendants, donc à priori k degrés de liberté. Mais on perd toujours un degré de liberté à cause des restrictions sur les probabilités p_i :

$$\sum_{i=1}^{k} p_i = 1.$$

On peut perdre d'autres degrés de liberté si certains paramètres de la loi théorique doivent être estimés à partir de l'échantillon.

1. Si la distribution théorique est entièrement spécifiée, c'est-à-dire si on cherche à déterminer si la distribution observée suit une loi dont les paramètres sont connus avant même de choisir l'échantillon, on a (k - 1) degrés de liberté (k carrés indépendants moins une relation entre les variables).
2. S'il faut d'abord estimer r paramètres de la loi à partir des observations de l'échantillon (par exemple on cherche si la distribution est normale mais on ne connaît d'avance ni sa moyenne ni son écart-type), il n'y a plus que (k - 1 - r) degrés de liberté.

Dans un cas général, on dira que la loi du khi-deux suivie par l'écart entre les deux distributions a (k - 1 - r) degrés de liberté lorsqu'on a estimé r paramètres de la loi théorique à partir des observations de l'échantillon (avec la possibilité pour r de valoir 0).

5.2. LE TEST D'AJUSTEMENT DE PEARSON

Il nous faut maintenant décider, à l'aide de cet indicateur qu'est le χ^2, si les écarts entre les effectifs théoriques et ceux qui résultent des observations sont significatifs d'une différence de distribution ou si ils sont dus aux fluctuations d'échantillonnage. Nous procéderons comme d'habitude en quatre étapes.

1ère étape : formulation des hypothèses

On va donc tester l'hypothèse H_0 contre l'hypothèse H_1 :

H_0 : Les observations suivent la distribution théorique spécifiée.
H_1 : : Les observations ne suivent pas la distribution théorique spécifiée

2ème étape : Détermination de la fonction discriminante du test et de sa distribution de probabilité.

On utilise la variable aléatoire

$$\chi^2 = \frac{(N_1 - n_{t_1})^2}{n_{t_1}} + \frac{(N_2 - n_{t_2})^2}{n_{t_2}} + \ldots\ldots + \frac{(N_{k-} n_{t_k})^2}{n_{t_k}}$$

3ème étape : Détermination des valeurs critiques de χ^2 délimitant les zones d'acceptation et de rejet

On impose à la zone d'acceptation de H_0 concernant la valeur du χ^2 d'être un intervalle dont 0 est la borne inférieure (car un χ^2 est toujours positif).

Il nous faut donc déterminer dans la table

la valeur maximale $\chi^2_{\alpha,\nu}$ de l'écart entre
les deux distributions imputable aux variations
d'échantillonnage au seuil de signification α,
c'est-à-dire vérifiant : $P(\chi^2 > \chi^2_{\alpha,\nu}) = \alpha$.
$\chi^2_{\alpha,\nu}$ représente donc la valeur critique pour un test sur la concordance entre deux distributions et le test sera toujours unilatéral à droite.

4ème étape : Calcul de la valeur de χ^2 prise dans l'échantillon et conclusion du test.

- On calcule la valeur χ_0^2 prise par χ^2 dans l'échantillon.

- → Si la valeur χ_0^2 se trouve dans la zone de rejet, on dira que l'écart observé entre les deux distributions est **statistiquement significatif** au seuil α. Cet écart est anormalement élevé et ne permet pas d'accepter H_0. On rejette H_0.
  → Si la valeur χ_0^2 se trouve dans la zone d'acceptation, on dira que l'écart-réduit observé **n'est pas significatif** au seuil α. Cet écart est imputable aux fluctuations d'échantillonnage. On accepte H_0.

6. *PLUSIEURS DISTRIBUTIONS SONT-ELLES COMPARABLES ?: TEST D'HOMOGÉNÉITÉ DE PLUSIEURS POPULATIONS*

6.1. INTRODUCTION

On prélève au hasard k échantillons de taille n_1, n_2,...., n_k de k populations. Les résultats du caractère observé dans chaque population sont ensuite classés selon r modalités. Dans ce cas, les totaux marginaux (les ni) associés aux k échantillons sont fixés et ne dépendent pas du sondage. Il s'agit de savoir comparer les k populations entre elles et de savoir si elles ont un comportement semblable en regard du caractère étudié (qualitatif ou quantitatif). On rassemble les données dans un tableau à double entrée appelé **tableau de contingence** :

	POPULATIONS ÉCHANTILLONNÉES						
		j = 1	j = 2	...	j	...	j = k
CARACTÈRE OBSERVE SELON r MODALITÉS	i = 1	n_{11}	n_{12}		n_{1j}		n_{1k}
	i = 2	n_{21}	n_{22}		n_{2j}		n_{2k}
	...						
	i	n_{i1}	n_{i2}		n_{ij}		n_{ik}
	...						

	i = r	n_{r1}	n_{r2}		n_{rj}		n_{rk}
		$n_1 = \sum_{i=1}^{r} n_{i1}$	$n_2 = \sum_{i=1}^{r} n_{i2}$		$n_j = \sum_{i=1}^{r} n_{ij}$		$n_k = \sum_{i=1}^{r} n_{ik}$

6.2. TEST D'HOMOGÉNÉITÉ

Il s'agit de comparer les effectifs observés pour chaque modalité du caractère avec les effectifs théoriques sous l'hypothèse d'un répartition équivalente entre les k populations et ceci pour chaque modalité du caractère. Si nous notons p_{ij} la probabilité théorique pour qu'une unité statistique choisie au hasard dans la population j présente la modalité i du caractère étudié, on peut alors préciser les hypothèses de la façon suivante :

1ère étape : formulation des hypothèses

H_0 : $p_{i1} = p_{i2} = ... = p_{ik}$ pour i = 1, 2, ..., r
Soit encore : les proportions d'individus présentant chaque modalité du caractère sont les mêmes dans les k populations.

H_1 : : $p_{ij_1} \neq p_{ij_2}$ pour au moins un i parmi 1, 2, ..., r et pour au moins deux j_{i1} et j_{i2} différents choisis parmi 1, 2, ..., k
Soit encore : les proportions d'individus présentant chaque modalité du caractère ne sont pas identiques pour toutes les populations pour au moins une modalité du caractère.

2ème étape : **Détermination de la fonction discriminante du test et de sa distribution de probabilité.**

Sous l'hypothèse d'homogénéité des populations, on doit comparer les effectifs observés aux effectifs théoriques.
Pour calculer les effectifs théoriques, il nous faut déterminer $p_{i.}$ la proportion d'individus associée à la modalité i et que l'on suppose identique dans les k populations. On obtiendra une estimation de cette proportion en utilisant l'ensemble des données collectées.

On choisit donc : $p_{i.} \cong \frac{\sum_{j=1}^{k} n_{ij}}{\sum_{j=1}^{k} n_j}$; On en déduit les effectifs théoriques de chaque classe grâce à la relation : $n_{t_{ij}} = p_{i.} n_j$.

Pour comparer les écarts entre ce qu'on observe et ce qui se passe sous l'hypothèse H_0, on considère la somme des écarts réduits de chaque classe, à savoir la quantité :

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{k}\frac{(N_{ij} - n_{t_{ij}})^2}{n_{t_{ij}}}$$

Cette variable aléatoire suit une loi du khi-deux (voir paragraphe précédent), mais quel est donc son nombre de degrés de liberté ?

Calcul du nombre de degrés de liberté du khi-deux :

- A priori, on a kr cases dans notre tableau donc (kr) degrés de liberté. Mais il faut retirer à cette valeur, le nombre de paramètres estimés ainsi que le nombre de relations entre les différents éléments des cases.
- On a estimé les probabilités théoriques à l'aide des valeurs du tableau $(p_{1.}, p_{2.}, ..., p_{r.})$, mais seulement (r - 1) sont indépendantes puisqu'on impose la restriction : $\sum_{i=1}^{r} p_{i.} = 1$. Par ces estimations, on a donc supprimé (r - 1) degrés de liberté.
- Les effectifs de chaque colonne sont toujours liés par les relations : $\sum_{i=1}^{r} n_{ij} = n_j$ (puisque les n_j sont imposés par l'expérience) et ces relations sont au nombre de k.

Finalement, le nombre de degrés de liberté du khi-deux est :

$$\nu = kr - (r - 1) - k = (r - 1)(k - 1)$$

3ème étape : Détermination des valeurs critiques de χ^2 délimitant les zones d'acceptation et de rejet

On impose à la zone d'acceptation de H_0 concernant la valeur du χ^2 d'être un intervalle dont 0 est la borne inférieure (car un χ^2 est toujours positif).

Il nous faut donc déterminer dans la table la valeur maximale $\chi^2_{\alpha,\nu}$ de l'écart entre les deux distributions imputable aux variations d'échantillonnage au seuil de signification α, c'est-à-dire vérifiant : $P(\chi^2 > \chi^2_{\alpha,\nu}) = \alpha$.

4ème étape : Calcul de la valeur de χ^2 prise dans l'échantillon et conclusion du test.

- On calcule la valeur χ_0^2 prise par χ^2 dans l'échantillon.
- → Si la valeur χ_0^2 se trouve dans la zone de rejet, on dira que l'écart observé entre les k distributions est **statistiquement significatif** au seuil α. Cet

écart est anormalement élevé et ne permet pas d'accepter H_0. On rejette H_0 : les populations n'ont pas un comportement homogène.

→ Si la valeur χ_0^2 se trouve dans la zone d'acceptation, on dira que l'écart-observé **n'est pas significatif** au seuil α. Cet écart est imputable aux fluctuations d'échantillonnage. On accepte H_0.

CONCLUSION : Nous avons appris à effectuer un certain nombre de tests. Il en existe d'autres. Tous fonctionnent sur le même principe. Si vous avez compris ce qui précède, vous serez capables de les appréhender correctement lorsque vous les rencontrerez : suivez le modèle.