Part I

CHAPITRE 4

Part II

LES TESTS STATISTIQUES

1 Introduction

Un test statistique est appelé à dégager un résultat significatif au milieu d'un ensemble de données expérimentales aléatoires.

La méthodologie des tests consiste à répondre à l'aide de résultats expérimentaux à une question concernant les paramétres de la loi de probabilité des variables aléatoires.

Quatre conditions préalables au calcul d'un test doivent être réunies :

– la question doit être posée de telle sorte qu'il n'y ait que deux réponses possibles : oui et non ;

– on doit avoir des données chiffrées résultant d'un échantillon ou d'une expérimentation ;

– ces données doivent pouvoir être considérées comme la réalisation de variables aléatoires dont la forme de la loi de probabilité est connue ;

– la question doit concerner un ou plusieurs paramétres de cette loi.

Une fois posée cette dernière, la réponse du test est :

– soit l'acceptation de l'hypothèse, ce qui signifie que les donneés ne sont pas en contradiction avec l'hypothèse ;

– soit le rejet de cette hypothèse, ce qui signifie qu'il est très peu probable d'obtenir les résultats que l'on a trouvés si l'hypothèse est vraie, ou encore que les données sont en contradiction avec elle.

En un sens, le test d'hypothèse est une généralisation probabiliste du raisonnement par l'absurde, mais alors que ce dernier met en contradiction logique deux affirmations formelles,

le premier oppose une affirmation formelle (l'hypothèse) avec des résultats du monde réel (les résultats de l'expérience).

De plus, le premier ne donne pas une certitude logique (l'hypothèse est fausse), mais seulement une forte présomption mesurée par une probabilité.

Enfin les deux formes du raisonnement ont en commun qu'elles ne peuvent que prouver (ou donner une présomption de preuve de) la fausseté de l'hypothèse et non sa vérité : ce n'est que parce qu'une expérience ne conduit pas au rejet de l'hypothèse que cette dernière est vraie : on peut imaginer d'autres expériences qui pourraient peut-être la rejeter.

Remark 1 *Il s'agira de prendre une décision; Elle consistera à accipter ou non une hypothèse de départ, formulée soit à partir de connaissances théhoriques soit à partir de présomptions suggérées par le ou les échantillons étudiés. Cependant, on ne pourra jamais conclure avec une certitude absolue, puisque la base de l'information qui perment de mener un test statistique provient de sous-ensembles de la population sur laquelle sont formulées les hypothèses.*

Remark 2 *Il faudra donc se fixer un certain risque d'erreur qui n'est autre que la probabilité de se tromper en prenant la décision retenue.*

2 Principe général

2.1 L'interprétation statistique

Dans les chapitres précédent on a pu tirer un certains nombre de conclusins à partir d'un nombre limité d'observations. Ces conclusions ont permis d'estimer certains caractéristiques inconnues de la population.

Dautres méthodes, regroupées sous la dénomination de tests statistiques qui constituent la théorie de la décision, vont permettre de résoudre des problèmes pratiques tels que:

les tests de nouvelles thérapeutiques,

les comparaisons de méthodes de cultures,

les tests d'emploi de tel ou tel engrais.

2.2 La formulation des hypothèses

Un test statistique est un mécanisme qui permet de trancher entre deux hypothèses à partir de résultats observés sur un ou plusieurs échantillons.

Soit H_0 et H_1 ces deux hypothèses. La première appelée **hypothèse nulle**, joue un rôle particulier; elle prétendra que les différences observées entre valeurs calculées et valeurs

théoriques sont dûes au hasard. Si on doit rejeter l'hypothèse nulle H_0, on dira que les écarts obsevés sont significatifs et on choisira H_1 appelée **hypothèse alternative.** Les tests statistiques permettent de retenir ou de rejeter H_0 qui est la seule hypothèse testée et celle qui permet les calculs pour conduire à la conclusion.

On a

H_0 vraie et H_1 fausse
ou
H_0 fausse et H_1 vraie

Il ya 4 solutions dont seulement les deux premières son justes:
a)-H_0 est vraie et on a choisi H_0
b)-H_0 est fausse et on a rejeté H_0
c)-H_0 est vraie et on a rejeté H_0
d)-H_0 est fausse est on a choisi H_0

2.3 Le risque d'erreur

Soit un test qui aboutit à chgoisir H_0 ou H_1. Seule une de ces deux hypothèses est vraie et on peut résumer les différents cas de décision et de validité de cette décision par le tableau suivant:

	Hypothèse vraie		
Hy pothèse retenue		H_0	H_1
	H_0	$1-\alpha$	β
	H_1	α	$1-\beta$

De ce tableau on tire les définitions suivantes:

2.3.1 1-3-1-Le risque de première espèce α

On appelle risque de première espèce et on note α la probabilité de rejeter l'hypothèse nulle H_0 alors qu'elle sest vraie.

Dans la pratique des tests statistiques, il est d'usage de choisir α a priori ($\alpha = 1\%$ ou 5% dans la plupart des cas), cette probabilité est aussi appelée **seuil de signification du test.**

2.3.2 1-3-2-Le risque de deuxième espèce β

On appelle risque de seconde espèce et on note β la probabilité d'accipter l'hypothèse nulle H_0 alors qu'elle sest fausse.

α etant fixé, β est déterminé par un calcul de pribabilité si H_1 est précisément définie.

On appelle puissance du test la probabilité $(1-\beta)$ de rejeter H_0 en ayant raison.

3

4 Les différents types de tests

1-Les tests de conformité
2-les tests de comparaison
3-Les tests d'ajustement à une loi théorique
4-les tests d'indépendance

4.1 I- Les tests de conformité

Dans cette partie nous traiterons un premier typpe de test d'hypothèse en nous limitant au cas des grands échantillons (en pratique des échantillons de taille $n \geq 50$).

Nous disposons d'une distribution statistique expérimentale se présentant sous la forme d'un tableau d'effectifs ou des fréquences du caractère étudié.

Nous voulons savoir si ces effectifs ou ces fréquences sont compatible avec une distribution théorique connue. Il s'agit de déterminer si les différences constatées entre la distribution théorique et la distribution expérimentale sont liées à la constitution de l'échantillon ou si elle sont trop importantes.

4.1.1 1-Etude des moyennes

Nous nous proposon d'étudier la conformité d'un échantillon par rapport à une norme préalablement définie.

4.1.2 position du problème

Dans un laboratoire pharmaceutique, une machine automatique fabrique en grande quantité des suppositoires contenant du paracétamol.

On désigne par X la variable aléatoire, qui à tout suppositoire pris au hasard dans la production, associe la masse (en mg) de paracitamol qu'il contient.

On admet que X suit la loi normale de moyenne m et d'écart-type $\sigma = 8$.

On veut contrôler la qualité de fabrication sur une période donnée. Dans ce but, pendant le fonctionnement de la machine, on prélève d'un temps à l'autre un suppositoire dont on mesure la masse du paracétamole. On constitue ainsi un échantillon de 100 suppositoires. Les tirages sont supposéss indépendants.

On se propose de construire un test bilatéral permettant d'accepter ou de refuser, au seuil de significationde 5%, l'hypothèse selon laquelle la masse moyenne de paracétamolcontenue dans un suppositoire est égale à 170 mg.

l'hypothèse nulle (H_0) est $m = 170\ mg$ et lhypothèse alternative est (H_1) $m \neq 170\ mg$.

1°) Sou H_0 quelle est la loi de la variable alétoire $\overline{X}$? préciser ces paramètres.

2°) Enoncer clairement la règle de décision du test

3°) Les résultats des mesures de l'échantion prélevé sont donnés dans le tableau:

Masse (mg)	$[145;155[$	$[155;165[$	$[165;175[$	$[175;185[$	$[185;195[$
Effectifs	7	30	43	16	4

Peut-on accepter l'hypothèse H_0 au seuil de signification de 5%?.

Lois d'échantillonnage Puisque $n \geq 30.$, le théhrème de la limite centrée nous permet de dire que la variable aléatoire $\overline{X}$ qui à chaque échantillon de taille n associe sa moyenne, suit approximativement la loi normale $N(u; \frac{\sigma}{\sqrt{n}})$.

Alors la variable aléatoire $T = \dfrac{\overline{X} - u}{\dfrac{\sigma}{\sqrt{n}}}$ suit la loi normale centrée réduite.

Construction d'un test bilatéral: l'hypothèse nulle (H_0) est $m = 170\ mg$ et l'hypothèse alternative est (H_1) $m \neq 170\ mg$

Règle de décision: Fixon, à priori, le risque maximal que nous acceptons de prendre en refusant H_0 alors qu'elle est vraie. Ce risque dit de première espèce, et noté α.

Puisque T suit la loi normale centrée réduite, il existe un unique réel strictyement positif t_α tel que: $p(|T| > t_\alpha) = \alpha$. $t_\alpha = \Pi^{-1}\left(1 - \dfrac{\alpha}{2}\right)$

Si $|T| > t_\alpha$ on rejette H_0 avec le risque α de se tromper.

Si $|T| \leq t_\alpha$ on accepte H_0 avec le risque de se tromper.(risque β de seconde espèce non quantifié).

4.1.3 Application numérique

Sous l'hypothèse H_0 la variable aléatoire $\overline{X}$ suit la loi normale $N(170; 0,8)$ donc la variable aléatoire $T = \dfrac{\overline{X} - 170}{0,8}$ suit la loi normale centrée réduite.

au seuil de risque $\alpha = 0,05$ on rejette H_0 si $|T| > 1,96$.

Pour l'échantillon proposé, en utilisant les centres des classes, on trouve $\overline{x} = 168$

On en déduit $t = -2,5$ donc $|t| > 1,96$ et on rejette H_0 au risque de 5% de se tromper.

Test unilatéral droit l'hypothèse nulle (H_0) est $m = 170\ mg$ et l'hypothèse alternative est (H_1) $m \geq 170\ mg$

La démarche ne diffère du précédente que sur deux points:

Hypothèse alternative H_1: est selon le problème posé $m > 170$ ou $m < 170$.

Le risque α n'est plus symétriquement répartie.

Pour fixer les idées, supposons que l'hypothèse alternative H_1: est $m > 170$ alors T est nécessairement positve.

Il existe un unique réel strictement positif u_α tel que $P(t > u_\alpha) = \alpha$ ou, ce qui équivalent tel que $P(t > u_\alpha) = 1 - \Pi(u_\alpha)$.

On a donc $\Pi(u_\alpha) = 1 - \alpha$ soit $u_\alpha = \Pi^{-1}(1 - \alpha)$

La règle de décision en résulte:

Si T$> u_\alpha$ on rejette H_0 avec un risque α de se tromper

Si T$\leq u_\alpha$ on accepte H_0 avec un risque β (non quantifié) de se tromper..

4.1.4 Etude des fréquences

Position du problème On étudie ici un caractère quantitatif C et on dispose de deux grands échantillons indépendants

A d'effectif n_A , où la fréquence du caractère est f_A.

B d'effectif n_B , où la fréquence du caractère est f_B.

A quelles condition peut-on conclure, qu'à un risque doné, ces deux échantillons proviennent de la même population?

Lois d'échantillonnage Supposons que l'échantillon A provienne de la population P , où la fréquence du caractère C et p.

Supposons que l'échantillon B provienne de la population P$'$, où la fréquence du caractère C et p$'$.

On sait que si $N_A \geq 30$, La variable aléatoire F_A qui à tout échantillon de taiile n_A associe la fréquence f_A du caractère C dans cette échantillon suit apprximativement la loi normale $N(p; \sqrt{\frac{p(1-p)}{n_A}})$

Même si $N_B \geq 30$, La variable aléatoire F_B qui à tout échantillon de taiile n_B a fréquence f_B du caractère C dans cette échantillon suit apprximativement la loi normale $N(p'; \sqrt{\frac{p'(1-p')}{n_B}})$

Les variables alétoires F_A et F_B etant indépendantes et La variable aléatoire $F_A - F_B$ suit apprximativement la loi normale $N(p-p'; \sqrt{\frac{p(1-p)}{n_A} + \frac{p'(1-p')}{n_B}})$.

Tests d'hypothèse bilatéral

Hypotèse à tester Nous nous proposons de tester l'hypothèse nulle, notée H_0 "p et p' ne sont pas significativement différentes"

Hypothèse alternative H_1: le test étant bilateral H_1 est p et p' sont significativement différentes"

Règle de décision: Sous l'hypothèse H_0, la variable alétoire $F_A - F_B$ suit apprximativement la loi normale $N(p-p'; \sqrt{\frac{p(1-p)}{n_A} + \frac{p'(1-p')}{n_B}})$.

Donc la variable alétoire $T = \dfrac{F_A - F_B}{\sqrt{p(1-p)(\frac{1}{n_A} + \frac{1}{n_B})}}$ suit apprximativement la loi normale $N(0;1)$.

Fixons alors un seuil de risque α (donc un seil de confiance $1-\alpha$), on sait qu'il existe un réel unique t_α strictement positif tel que $P(|T| \leq t_\alpha) = 1-\alpha$

$P(|T| \leq t_\alpha) = 1-\alpha$ é quivaut à $t_\alpha = \Pi^{-1}\left(1 - \dfrac{\alpha}{2}\right)$

La règle de décision du test en résulte:

Si $|T| \leq t_\alpha$ on a aucune raison de rejeter H_0 donc on l'accepte.avec un risque β (non contifié) de se tromper

Si $|T| > t_\alpha$ on rejette H_0 un risque α de se tromper

Mise on oeuvre du test: $t = \dfrac{|f_A - f_B|}{\sqrt{p(1-p)(\frac{1}{n_A} + \frac{1}{n_B})}}$ On compare le nombre t avec t_α et on utilise la règle de décision pur conclure.

En général p est inconnu et, sous l'hypothèse (H_0) on réunit les deux échantillons.

Alors en estime p par $\widehat{p} = \dfrac{n_A f_A - n_B f_B}{n_A + n_B}$.

4.1.5 Tests d'hypothèse unilatéral

La démarche ne diffère du précédente que sur deux points:

Hypothèse alternative H_1: est selon le problème posé $p > p'$ ou $p < p'$.

Le risque α n'est plus symétriquement répartie.

Pour fixer les idées, supposons que l'hypothèse alternative H_1: est $p < p'$ alors T est nécessairement négative.

Il existe un unique réel strictement positif v_α tel que $P(t < -v_\alpha) = \alpha$ ou, ce qui équivalent tel que .

$1 - \Pi(v_\alpha) = \alpha$ On a donc $v_\alpha = \Pi^{-1}(1-\alpha)$

La règle de décision en résulte:

Si $T < -v_\alpha$ on rejette H_0 avec un risque α de se tromper.

Si $T \geq u_\alpha$ on accepte H_0 avec un risque β (non quantifié) de se tromper.

$\frac{1}{\sqrt{6.28}} \exp(-\frac{x^2}{2})$

Test unilatéral gauche l'hypothèse nulle (H_0) est $m = 170\ mg$ et l'hypothèse alternative est (H_1) $m \leq 170\ mg$

Règle de décision: Fixon, à priori, le risque maximal que nous acceptons de prendre en refusant H_0 alors qu'elle est vraie. Ce risque dit de première espèce, et noté α.

Puisque T suit la loi normale centrée réduite, il existe un unique réel strictyement positif t_α tel que: $p(|T| > t_\alpha) = \alpha$. $t_\alpha = \Pi^{-1}\left(1 - \frac{\alpha}{2}\right)$

Si $|T| > t_\alpha$ on rejette H_0 avec le risque α de se tromper.

Si $|T| \leq t_\alpha$ on accepte H_0 avec le risque de se tromper.(risque β de seconde espèce non quantifié).

4.1.6 Application numérique

Sous l'hypothèse H_0 la variable aléatoire $\overline{X}$ suit la loi normale $N(170; 0,8)$ donc la variable aléatoire $T = \frac{\overline{X} - 170}{0,8}$ suit la loi normale centrée réduite.

au seuil de risque $\alpha = 0,05$ on rejette H_0 si $|T| > 1,96$.

Pour l'échantillon proposé, en utilisant les centres des classes, on trouve $\overline{x} = 168$

On en déduit $t = -2,5$ donc $|t| > 1,96$ et on rejette H_0 au risque de 5% de se tromper.

4.1.7 2-Test de conformité d'une fréquence

Position du problème

Dans la population algerienne, 15 personnes sur 100 ont un facreur rhésus négatif. Un laboratoire d'analyses médicales a contrôlé le facteur rhésus de 459 personnes. Il aconstaté que 75 d'entre elles avaient un facteur rhésus négatif.

Construire un test bilatéral permittant de dire, aua risque de 5%, si ce résultat est compatible, ou non, avac la norme dans la population algérienne.

Construction d'un test l'hypothèse nulle (H_0) est: le résultat est compatible, ou non, avac la norme dans la population algérienne.

Le test etant bilatéral:

l'hypothèse alternative est (H_1) est: "le résultat est significativement différent de la norme habituelle".

Lois d'échantillonnage Puisque $n \geq 30.$, le théhrème de la limite centrée nous permet de dire que la variable aléatoire F qui à chaque échantillon de taille n associe la fréquence du caractère dans cet échantillon, suit approximativement la loi normale $N(0,15; 0,017)$. Alors la variable aléatoire $T = \frac{F - 0,15}{0,017}$ suit la loi normale centrée réduite.

Règle de décision (condition de rejet): on rejette H_0 au risque $\alpha = 0,05$ si $|T| > 1,96$

Mise on oeuvre du test: $t = \dfrac{f-p}{\sqrt{\dfrac{p(1-p)}{n}}} = \dfrac{1,63-0,15}{0,017} \approx 0,76 < 1,96$ (on ne peut pas rejeter H_0).

Conclusion: on constat que lerésultat obtenu est conforme à la norme dans la population avec un risque de de 5% de se tromper.

4.1.8 II-les tests de d'homogeneité (grands échantillon)

Nous disposons de deux échantillons indépendants donnés sous la forme d'un tableau d'éffectifs ou de fréquencesdu caractère étudié.

Nous désirons savoir si les différences observées sur la moyenne ou sur le fréquence sont dues u'niquement au hasard de l'échantillonage ou si elle sont trop importantes est doivent être atribuées à d'autres causes.

4.1.9 Etude des moyennes

Position du problème On étudie ici un caractère quantitatif C et on dispose de deux grands échantillons indépendants

A d'effectif n_A , de moyenne m_A et d'écrt-type σ_A

B d'effectif n_B , de moyenne m_B et d'écrt-type σ_B

A quelles condition peut-on conclure, qu'à un risque doné, ces deux échantillons proviennent de la même population?

Lois d'échantillonnage Supposons que l'échantillon A provienne de la population P , d'effectif N, de moyenne u et d'écart-type σ.

Supposons que l'échantillon B provienne de la population P$'$, d'effectif N$'$, de moyenne u$'$ et d'écart-type σ'.

On sait que si $N_A \geq 30$, La variable aléatoire $\overline{X}$ qui à tout échantillon de taile n_A associe sa moyenne m_A suit apprximativement la loi normale $N(u; \frac{\sigma}{\sqrt{n_A}})$.

Même si $N_B \geq 30$, La variable aléatoire $\overline{X}$ qui à tout échantillon de taiile n_B associe sa moyenne m_B suit apprximativement la loi normale $N(u'; \frac{\sigma'}{\sqrt{n_B}})$

Les variables alétoires $\overline{X}_A$ et $\overline{X}_B$ etant indépendantes et La variable aléatoire $\overline{X}_A - \overline{X}_B$ suit apprximativement la loi normale $N(u - u'; \sqrt{\frac{\sigma^2}{n_A} + \frac{\sigma'^2}{n_B}})$.

Tests d'hypothèse bilatéral

Hypotèse à tester Nous nous proposons de tester l'hypothèse nulle, notée H_0 "u et u' ne sont pas significativement différentes"

Hypothèse alternative H_1: le test étant bilateral H_1 est u et u' sont significativement différentes"

Règle de décision: Sous l'hypothèse H_0, la variable alétoire $\overline{X}_A - \overline{X}_B$ suit apprximativement la loi normale $N(0; \sqrt{\frac{\sigma^2}{n_A} + \frac{\sigma'^2}{n_B}})$.

Donc la variable alétoire $\dfrac{\overline{X}_A - \overline{X}_B}{\sqrt{\dfrac{\sigma^2}{n_A} + \dfrac{\sigma'^2}{n_B}}}$ suit apprximativement la loi normale $N(0;1)$.

Fixons alors un seuil de risque α (donc un seil de confiance $1-\alpha$), on sait qu'il existe un réel unique t_α strictement positif tel que $P(|T| \leq t_\alpha) = 1-\alpha$

$P(|T| \leq t_\alpha) = 1-\alpha$ é quivaut à $t_\alpha = \Pi^{-1}\left(1 - \dfrac{\alpha}{2}\right)$

La règle de décision du test en résulte:

Si $|T| \leq t_\alpha$ on a aucune raison de rejeter H_0 donc on l'accepte.avec un risque β (non contifié) de se tromper

Si $|T| > t_\alpha$ on rejette H_0 un risque α de se tromper

Mise on oeuvre du test: $t = \dfrac{m_A - m_B}{\sqrt{\dfrac{\sigma^2}{n_A} + \dfrac{\sigma'^2}{n_B}}}$ On compare alors $|t|$ avec t_α et on utilise la règle de décision pur conclure.

En général σ et σ' sont inconnus et remplacés dans cette formule par $\widehat{\sigma}_A = \sigma_A\sqrt{\dfrac{n_A}{n_A - 1}}$ et $\widehat{\sigma}_B = \sigma_A\sqrt{\dfrac{n_B}{n_B - 1}}$

4.1.10 Tests d'hypothèse unilatéral

La démarche ne diffère du précédente que sur deux points:

Hypothèse alternative H_1: est selon le problème posé $u > u'$ ou $u < u'$.

Le risque α n'est plus symétriquement répartie.

Pour fixer les idées, supposons que l'hypothèse alternative H_1: est $u > u'$ alors T est nécessairement positve.

Il existe un unique réel strictement positif u_α tel que $P(t > u_\alpha) = \alpha$ ou, ce qui équivalent tel que $P(t \leq u_\alpha) = 1-\alpha$.

On a donc $\Pi(u_\alpha) = 1-\alpha$ soit $u_\alpha = \Pi^{-1}(1-\alpha)$

La règle de décision en résulte:

Si T$\leq u_\alpha$ on accepte H_0 avec un risque β (non quantifié) de se tromper.

Si T$> u_\alpha$ on rejette H_0 avec un risque α de se tromper.

4.1.11 Etude des fréquences

Position du problème On étudie ici un caractère quantitatif C et on dispose de deux grands échantillons indépendants

A d'effectif n_A , où la fréquence du caractère est f_A.

B d'effectif n_B , où la fréquence du caractère est f_B.

A quelles condition peut-on conclure, qu'à un risque doné, ces deux échantillons proviennent de la même population?

Lois d'échantillonnage Supposons que l'échantillon A provienne de la population P , où la fréquence du caractère C et p.

Supposons que l'échantillon B provienne de la population P$'$, où la fréquence du caractère C et p$'$.

On sait que si $N_A \geq 30$, La variable aléatoire F_A qui à tout échantillon de taiile n_A associe la fréquence f_A du caractère C dans cette échantillon suit apprximativement la loi normale $N(p; \sqrt{\frac{p(1-p)}{n_A}})$

Même si $N_B \geq 30$, La variable aléatoire F_B qui à tout échantillon de taiile n_B a fréquence f_B du caractère C dans cette échantillon suit apprximativement la loi normale $N(p'; \sqrt{\frac{p'(1-p')}{n_B}})$

Les variables alétoires F_A et F_B etant indépendantes et La variable aléatoire $F_A - F_B$ suit apprximativement la loi normale $N(p-p'; \sqrt{\frac{p(1-p)}{n_A} + \frac{p'(1-p')}{n_B}})$.

Tests d'hypothèse bilatéral

Hypotèse à tester Nous nous proposons de tester l'hypothèse nulle, notée H_0 "p et p' ne sont pas significativement différentes"

Hypothèse alternative H_1: le test étant bilateral H_1 est p et p' sont significativement différentes"

Règle de décision: Sous l'hypothèse H_0, la variable alétoire $F_A - F_B$ suit apprximativement la loi normale $N(p-p'; \sqrt{\frac{p(1-p)}{n_A} + \frac{p'(1-p')}{n_B}})$.

Donc la variable alétoire $T = \dfrac{F_A - F_B}{\sqrt{p(1-p)(\frac{1}{n_A} + \frac{1}{n_B})}}$ suit apprximativement la loi normale $N(0;1)$.

Fixons alors un seuil de risque α (donc un seil de confiance $1-\alpha$), on sait qu'il existe un réel unique t_α strictement positif tel que $P(|T| \leq t_\alpha) = 1 - \alpha$

$P(|T| \leq t_\alpha) = 1 - \alpha$ é quivaut à $t_\alpha = \Pi^{-1}\left(1 - \dfrac{\alpha}{2}\right)$

La règle de décision du test en résulte:

Si $|T| \leq t_\alpha$ on a aucune raison de rejeter H_0 donc on l'accepte.avec un risque β (non contifié) de se tromper

Si $|T| > t_\alpha$ on rejette H_0 un risque α de se tromper

Mise on oeuvre du test: $t = \dfrac{|f_A - f_B|}{\sqrt{p(1-p)(\frac{1}{n_A} + \frac{1}{n_B})}}$ On compare le nombre t avec t_α et on utilise la règle de décision pur conclure.

En général p est inconnu et, sous l'hypothèse (H_0) on réunit les deux échantillons.

Alors en estime p par $\widehat{p} = \dfrac{n_A f_A - n_B f_B}{n_A + n_B}$.

4.1.12 Tests d'hypothèse unilatéral

La démarche ne diffère du précédente que sur deux points:

Hypothèse alternative H_1: est selon le problème posé $p > p'$ ou $p < p'$.

Le risque α n'est plus symétriquement répartie.

Pour fixer les idées, supposons que l'hypothèse alternative H_1: est $p < p'$ alors T est nécessairement négative.

Il existe un unique réel strictement positif v_α tel que $P(t < -v_\alpha) = \alpha$ ou, ce qui équivalent tel que .

$1 - \Pi(v_\alpha) = \alpha$ On a donc $v_\alpha = \Pi^{-1}(1 - \alpha)$

La règle de décision en résulte:

Si $T < -v_\alpha$ on rejette H_0 avec un risque α de se tromper.

Si $T \geq u_\alpha$ on accepte H_0 avec un risque β (non quantifié) de se tromper.

$\frac{1}{\sqrt{6.28}} \exp(-\frac{x^2}{2})$

5

Le test chi − deux(χ^2)

5.1 INTRODUCTION

Problème 1:

Partant des races pures, un sélectionneur a croisé de mufliers ivoires avec des mufliers rouges, il a obtenu en $F1$ des mufliers pâles, puis en $F2$ après autofécondation des plantes de la génération F1: 22 mufliers rouges, 52 mufliers pâles et 23 mufliers ivoires.

La couleur des fleurs est-elle gérée par un couple d'allèles?.

Le test chi-deux est fréquemment utilisé par les biologistes. A la différence des autres test, ce test ne s'appuie pas sur un modèle probabiliste rigoureux, mais sur une loi asymptotique; il est donc délicat à utiliser et il est parfois préférable de le remplacer, lorsque c'est possible, par un test non parmétrique plus adapté.

Le test du χ^2 est le plus célèbre des tests dits **non paramètriques** qui n'exigent aucune condition sur la distriution de la population mère. C'est un test globale qui porte sur l'ensemble des effectifs ou fréquences observées après expérience et calculés à partir de l'hypothèse testée. On pourra comparer:

Une distribution expérimentale et une distribution théorique. Les caractéristiques de cette distribution théorique sont connues ou estimées à partir des observations. Selon le cas, on parlera de test de **conformité ou d'ajustement** à une loi théorique.

Plusieurs distributions pour savoir si on peut accépter l'hypothèse qu'elles proviennent de la même population parente, dans ce cas on mènera un test **d'homogéneité** ou

d'indépendance. On a en fait généraliser le cas précédent en comparant chaque distribution empirique à une même distribution théorique.

Le mécanisme du test du χ^2 permet de savoir si les écarts constatés entre les distributions à comparer sont imputables ou non au hasard.

Definition 3 *Soit X une v.a de loi.N(0;1), alors la v.a X^2 est dite v.a de chi-deux à 1 degé de liderté.*

Definition 4 *Soient X_1, X_2,...,X_n n v.a indépendantes suivent toutes loi N(0;1), alors la v.a $Z = X_1^2 + X_2^2 + ... + X_n^2$ est une v.a de chi-deux à n degés de liderté, avec E(Z)=n et Var(Z)=2n*

Remark 5 *Si Z suit la loi du χ^2 à n degrés de liberté, la table du chi-deux donne pour un risque α choisi, le nombre χ^2_α tel que*

$$P(Z \geqslant \chi^2_\alpha) = \alpha.$$

5.2 2-2-COMPARAISON ET AJUSTEMENT A UNE LOI THEORIQUE

5.2.1 2-2-1-Construction du test

On considère une distribution expérimentale donnée par un échantillon de taille n.

Les individus de cet échantillon son classés et on a dénombré la fréquence absolue ou effectif de chaque classe. On note n_i l'effectif observé pour la classe N° i. Si on connait (ou croit connaître) la loi théorique que suit cette distribution, on est alors capable de calculer les effectifs théoriques de chaque classe. En effet la loi théorique est connue dès lors que les probabilités attachées à chaque classe le sont. On note P_i la probabilité qu'un individu tiré au hasard appartienne à la classe N° i. L'effectif théorique associé est alors nP_i.

6 Application du test chi-deux

On expliquera d'abord les principes du test sur une loi multinomiale puis dans ses applications les plus courantes, la méthode non paramétrique qui en découle.

6.1 test sur une loi multinomiale

6.1.1 Distribution à deux classes.

Soit une expérience aléatoire E susceptible

d'entraîner la réalisation d'un évenement E_1 de probabilité $P(E_1)$, ou d'un évenement E_2 de probabilité $P(E_2)$, E_1 et E_2 formant un système complet c-à-d $P(E_1) + P(E_2) = 1$ et $P(E_1 \cap E_2) = 0$.

Soit un ensemble de n expériences identiques à E et indépendantes. On lui associe les variables X_1 et X_2 représentant respectivement le nombre d'événement de E_1 et de E_2 que l'on peut obsever $(X_1 + X_2 = n)$, la réalisatin effective des n expérience entraîne

l'observation des valeurs x_1 de X_1 et x_2 de X_2 $(x_1 + x_2 = n)$, On dit que les résultats sont reparties en deux classes. On désire tester l'hypothèse H_0 "$P(E_1) = P_1$ et $P(E_2) = P_2$ contre l'hypothèse H_1

"$P(E_1) \neq P_1$ et $P(E_2) \neq P_2$".

Compte-tenue de la relation $P_1 + P_2 = 1$, il suffit de tester "$P(E_1) = P_1$" contre

"$P(E_1) \neq P_1$". Ce que l'on peut faire à l'aide de la variable $\mathrm{X}_1 \rightarrow B(n, P_1)$.

X_1 admet pour loi asymptotique, lorsque n augmente indéfiniment, la loi

$$N(nP_1, nP_1(1 - P_1)).$$

Alors un test avec la variable

$Y = \dfrac{X_1 - nP_1}{\sqrt{nP_1(1-P_1)}}$ considéré comme pratiquement normale centrée et réduite sou H_0.

Soit maintenant la variable

$Z = \dfrac{(X_1 - nP_1)^2}{nP_1} + \dfrac{(X_2 - nP_2)^2}{nP_2}$,

on a $Z = \dfrac{(X_1 - nP_1)^2}{nP_1(1-P_1)} = Y^2$

etant donné le comportement asymptotique de Y, il est clair que Z admet pour loi asymptotique la loi de χ_1^2 sous H_0.

Pour un niveau α on peut ecrire $1-\alpha = P\left(-y_{\frac{\alpha}{2}} \leq Y \leq y_{\frac{\alpha}{2}}\right) = P\left(0 \leq Y^2 \leq y_{\frac{\alpha}{2}}^2\right) = P\left(0 \leq Z \leq z_{\frac{\alpha}{2}}\right)$ avec $z_{\frac{\alpha}{2}} = y_{\frac{\alpha}{2}}^2$,

La borne supérieur de l'intervalle d'acceptation ($3.481=(1.96)^2$ au niveau 5%; $6.635=(2.576)^2$ au niveau 1%) étant lue dans les tables de χ^2..

6.1.2 Distribution à r classes.

Plus généralement soit une expérience aléatoire E susceptible d'entraîner la réalisation de r évenements E_1,E_2, ...,E_r de probabilité $P(E_1)$, $P(E_2)$, ..., $P(E_r)$, E_1 E_2 ...E_r,formant un système complet c-à-d $P(E_1) + P(E_2) + ... + P(E_r) = 1$ et $P(E_i \cap E_j) = 0$ *pour* $i \neq j$.

Les résultats de n expériences identiques à E et indépendantes sont donc réparties en r classes. A un tel ensemble d'expériences, On associe les variables X_1 X_2 ...X_r représentant respectivement les effectifs des classes que l'on peut obsever,

Le système $(X_1, X_2, ..., X_r)$, forme un sytème multinomiale, on veut tester l'hypothèse

"$p(E_1) = p_1$ et $p(E_2) = p_2$ *et...* $p(E_r) = p_r$"

contre l'hypothèse H_1 :

"$P(E_1) \neq P_1$ ou $P(E_2) \neq P_2$...ou $P(E_r) \neq P_r$"

En fait il n'y a parmi r variables que $(r-1)$ variables indépendantes; En effet les variables sont liées par la relation $X_1 + X_2 + ... + X_r = n$, dès que le hasard attribué une valeur numérique à r-1 variables, la valeur de la dernière est imposée.

6.1.3 APPLICATION:

Problème 1 (solutions):

Solution: Soient p_1, p_2, p_3 les probabilités pour q'une plante de la génération F2 ait respectivement des fleurs rouges, pâles ou ivoires, soient X_1, X_2 et X_3 les variables représentant les plantes à fleurs rouges, pâles ou ivoires que l'on peut observer sur 97 plantes.

On est amené à tester, après un raisonnement génétique élémentaire, l'hypothèse H_0 :

$p_1 = \frac{1}{4}$, $p_2 = \frac{1}{2}$, $p_3 = \frac{1}{4}$ contre l'hypothèse $H_1 : p_1 \neq \frac{1}{4}$ *ou* $p_2 \neq \frac{1}{2}$ *ou* $p_3 \neq \frac{1}{4}$.

D'où le tableau:

phénotypes	rouge	pâle	ivoir	total
probabilité	1/4	1/2	1/4	1
effectif théorique	24.25	48.5	24.25	97
effectif observé	22	52	23	97

-Les conditions d'application de χ^2 sont s'atisfaites, à savoir:
-Les classes constituent un systèmes complet d'événements;
-Les 97 expériences sont identiques et indépendantes;
-Leur nombre est assez élvé;
-Les effectifs théoriques sont suffisament élevés.

Dans ces conditions, sous H_0, la variable $Z = \sum_{i=1}^{3} \frac{(X_i - 97p_i)^2}{97p_i}$ est pratiquement une variable X^2, on effectue un test. L'intervalle d'acceptation de H_0 est, au niveau

5%: $[0\ ;\ 5,991]$.

On a observé la valeur $Z_0 = \frac{(2,25)^2}{24.25} + \frac{(3,50)^2}{48,5} + \frac{(1,25)^2}{24,25} \simeq 0.52$.

Conclusion:

Au niveau 5% on peut accepter l'hypothèse que La couleur des fleurs est gérée par un couple d'allèles.

6.2

Tests d/hmogénéité

Principe

Le test χ^2 est également utilisé pour la comparaison de plusieurs échantillons. Le principe du test va être exposé dans un exemple à deux échantillons. on le généralise sans peine pour plusieurs échantillons.

Problème 2:

On a étudié sur deux échantillons prevenant de deux populations différentes la répartition des quatre groupes sanguins: O, A, B,AB les résultats obtenus sont réparties dans un tableau dit tableau de contingence, à deux lignes et à quatre colonnes:

Groupe	O	A	B	AB	tot
1^{er} éch	121	120	79	33	353
2^{em} éch	118	95	121	30	364
total	239	215	200	63	717

On veut tester l'hypotèse H_0 " les quatre groupes sanguins sont réparties de la même manière sur les deux populations"

contre l'hypothèse H_1 "les répartitions sont différentes".

Sous H_0. la probabilité, pour un individu prélevé au hasard, d'être d'un groupe donné est la même dans les deux populations, on ne connaît pas cette probabilité, sinon le problème serait résolu; on peut cependant l'estimer et, toujours sous H_0. La meilleure estimation que l'on puise en donner est la proportion des indiviudus de ce groupe observée sur l'ensemble des deux échantillons. C'est ainsi que l'on obtient les estimations:

Pour le groupe O	$p_1 = 239/717 \simeq 0,333$
Pour le groupe A	$p_2 = 215/717 \simeq 0,300$
Pour le groupe B	$p_3 = 200/717 \simeq 0,249$
Pour le groupe AB	$p_4 = 63/717 \simeq 0,088$

$p_1 + p_2 + p_3 + p_4 = 1$. La relation $p_1 + p_2 + p_3 + p_4 = 1$ montre qu'en fait il suffit de trois paramètres pour déterminer complètement le modèle. On déduit de l'estimation précédente les éffectifs théoriques de chaque classe pour un échantillon de taille 353 d'une part et pour un échantillon de taille 364 d'autre part. D'où le tableau:

Groupe	O	A	B	AB	total
1^{er} éch	121 (117,7)	120 (105,9)	79 (98,5)	33 (31)	353
2^{em} éch	118 (121,3)	95 (109,1)	121 (101,5)	30 (32)	364
total	239	215	200	63	717

les effectifs théhoriques sont entre parenthèses, on a par exemple, 117=0,333.353.

Soient maintenant les variables X_1, X_2, X_3, X_4 représentant les effctifs des classes du premiers échantillon et Y_1, Y_2, Y_3, Y_4 représentant les effctifs des classes du deuxième échantillon.

On pose:

$$Z = \frac{(X_1 - 117,7)^2}{117,7} + \frac{(X_2 - 105,9)^2}{105,9} +$$
$$\frac{(X_3 - 98,5)^2}{98,5} + \frac{(X_4 - 31,0)^2}{31,0} + \frac{(Y_1 - 121,3)^2}{121,3} +$$
$$\frac{(Y_1 - 109,1)^2}{109,1} + \frac{(Y_1 - 101,5)^2}{101,5} + \frac{(Y_4 - 32,0)^2}{32}.$$

Les conditions d'application du test χ^2 étant satisfaites pour chaque échantillon, sous H_0, la variable Z peut être considérée comme la somme de deux variables χ^2, l'indépendance des deux séries d'observations permet de considérer la varible Z comme une variable χ^2. On est tenté de dire qu'il s'agit d'une variable χ^2 à $2(4-1) = 6$ degrés de liberté; cependant, l'estimation, à partir des observations des trois paramètres qui déterminent coplètement le modèle probabiliste baisse le nombre de degrés de liberté de 6 à 3. D'où $Z \rightarrow \chi^2_3$.

Les valeurs élevées de Z étant plus probables sous H_1 que sous H_0.

Au niveau 5% l'intervalle d'acceptation est $[0;\ 7,815]$, et comme $Z \simeq 11.74 > 7,815$ donc

on peut conclure au rejet de H_0.

C'st-à-dire les quatre groupe sanguins sont réparties différemment sur les deux populations d'où proviennent les deux échantillons.

Même au niveau 1% on rejetterait H_0.

EXERCICES

exercice 1:

en désire intérpréter les résultats suivants: le nombre de guérisons du concer de la peau à été de 1712 individus sur 2015 patient pour un traitement A et de 757 individus sur 1010 patients pour un traitement B.

Tester l'hypothèse H_0"un individu à la même probabilté d'être guéri dans les deux traitements" contre l'hypothèse H_1"les deux traitements sont caractérisés par deux probabilités de guérison différentes".

exercice 2:

Une enqûete a été effectué en milieu hospitalier pour déterminer si l'usage du tabac favorise l'apparition du concer broncopulmonaire. Cette enqûete a été menée de la manière suivante:

Les individus intérrogés sont répartis en quatre catégories selon leur consommation journalière en cigarette: A (non fumeurs) , B(de 1 à 9), C (de 10 à 19), D(de 20 ou plus); il s'agit d'une consommation moyenne évaluée sur les deux dernières années précédant l'enqûete.

Un premier échantillon est constitué de concéreux. Un échantillon témoin a ensuite été choisi parmi les accidentés, c'est-à-dire les patients hospitalisés pour des raisons qui n'ont rien à voir avec le tabac, de plus pour éliminer tout autre facteur, àchaque concéreu correspond un témoin de même sexe, de même âge et intérrogé par le même enquiteur.

A partir des résultats ci-dessous, peut-on conclure à l'influence du tabac?.

Ca	A	B	C	D	TOT
Co	25	66	177	334	602
T	130	136	165	171	602
TOT	155	202	342	505	1204

Ca=Catégorie, Co=Concéreux, T=Témoins.

exercice 3: On a vacciné contre la grippe 300 personnes réparties en deux groupes A et B en fonction de l'âge:

Le groupe A comporte 120 individus de 55 ans au plus.

Le groupe B comporte 180 individus de plus de 55 ans.

On a constaté que, dans le groupe A, 38 individus ont eu la grippe l'hiver suivant la vaccination, tandis que 73 individus du groupe B ont eu la grippe ce même hiver.

Pet-on, au risque 10%, considérer qu'il existe un liaison entre l'éfficacité du vaccin et l'âge de la personne vaccinée?.

exercice 4: On a vacciné contre la grippe 300 personnes réparties en deux groupes A et B en fonction de l'âge:

Le groupe A comporte 120 individus de 55 ans au plus.

Le groupe B comporte 180 individus de plus de 55 ans.

On a constaté que, dans le groupe A, 38 individus ont eu la grippe l'hiver suivant la vaccination, tandis que 73 individus du groupe B ont eu la grippe ce même hiver.

Pet-on, au risque 10%, considérer qu'il existe un liaison entre l'éfficacité du vaccin et l'âge de la personne vaccinée?.

exercice 5: On a croisé deux races de plantes différant par deux caractères: la couleur (rouge ou blanche) et la taille (grande ou petite) des fleurs qu'elle produisent.

La première génération est homogène et donne de grandes fleurs rouges. La seconde génération fait apparaître quatre type de plantes en fonction des fleurs qu'elles produisent: grandes fleurs rouges, grandes fleurs blanches, petites fleurs rouges et petites fleurs blanches.

Sur un échantillon de 320 plantes on a observé les résultats suivants:

phénotypes	GR	GB	PR	PB
effectifs	202	59	45	14

Peut-on considérer, au risque 5% , que les deux caractères étudiés se transmettent selon les lois de MENDEL?.

exercice 6: Il est admet qu'en Algerie les groupes sanguins sont réparties de la façon suivante: O:40%, A:43% , B:12%, AB:5%.

Un échantillon de 300 étudiants à l'université de jijel a fourni les résultats:

Groupes	O	A	B	AB
effectifs	112	123	44	21

Peut-on affirmer, au risque 5% , que la répartition des groupes sanguins à l' université de jijel ne diffère pas sensiblement de celle de l'Algerie?.

exercice 7: Dan une population de 500 personnes (300 hommes et 200 femmes) on a mesuré la tension artérielle dechaque individu, ce qui a donné les résultats suivants:

	Hypert	TN	Hypot
H	72	192	36
F	38	118	44

Peut-on, au risque 5%, émettre l'hypothèse H_0 d'une liaison entre le sexe de l'individu et la tension artérielle?.

INDICATION: Le nombre de degrés de liberté est le nombre minimum des case du tableau dont il faut connaître l'effectif pour déterminer l'ensemble du tableau où les sommes de chaque ligne et chaque conlonne sont données.

Dans l'exercice précédent le nombre de degrés de liberté est 2.

exercice 8: Un médicament a été expérimenté sur 200 malades dévisés en deux groupes M_1 et M_2 indépendants:

-le groupe M_1 composé de 110 malades a aborbé le médicament étudié.

-le groupe M_2 composé de 90 malades a aborbé un placebo.

Les résultats sont les suivants: 60 malades guéris dans le groupe M_1, 36 malades guéris dans le groupe M_2.

1°) Calculer le pourcentage de guérisons et l'écart-type de ce pourcentage pour chacun des échantillons M_1 et M_2.

2°) En admettant que le phénomène étudié suit une loi normale, construir un test permettant d'accepter ou de rejeter l'hypothèse de l'éfficacité du médicament au risque de 5%.

exercice 9: On veut savoir si une maladie M modifie le taux de certaines protéines dans le song. On a mesuré leurs concentrations dans un échantillon de sujets atteints pa M et dans un autre échantillons formé de sujets en bonne santé (sujets témoins). Les résultats (dans une unité convenable) sont les suivants:

	effectifs	moyenne échantillon	variance échantillon
Malades	77	141	40
Témoins	33	131	32

Tester l'hypothèse "taux identiques chez les malades et les témoins" contre l'hypothèse:

a) "taux différent chez les malades et les témoins".

b) "taux supérieur chez les malades".

exercice 10: On a mesuré les dimensions d'une tumeur chez les souris traitées ou non par une substance anti-tumorale et on a obtenu:

Surface (cm^2)	5	5,5	6	6,5	7	7,5	8
Nombre de témoins	0	0	2	3	8	4	3
Nombre traités	4	4	8	3	0	1	0

La différence observée est-elle significative?.

exercice 11: Dans une maternité, on a compré les poids à la naissance des des bébés de mères primipares et multipares. On a obtenu les résultats suivants:

primipares $n_1 = 100$ $\overline{x_1} = 3180g$ $\sigma^2_{e1} = 214400$

multipares $n_2 = 110$ $\overline{x_2} = 3400g$ $\sigma^2_{e2} = 243300$

Peut-on admettre au coefficient de confiance de 99% que les enfants nés de mères multipares sont plus lourds que ceux nés de mères primipares?.