Statistiques et informatique

E.C. PSY54A

Présentation du cours 2007/2008

Organisation matérielle

Cours magistral : 12 heures
Mercredi 8h15-9h15 - Amphi 1
Travaux dirigés :
TD de statistiques.
- Gr 1 - Vendredi 9h15-10h15 - A220
- Gr 2 - Mardi 14h45-15h45 - A001
- Gr 3 - Mercredi 9h15-10h15 - A212
- Gr 4 - Mercredi 13h45-14h45 - B317

TD d'informatique (en sous-groupes).
salle info., 2 h. par quinzaine
- Gr 1A et 1B - Mercredi 16h-18h
- Gr 2A et 2B - Jeudi 13h45-15h45
- Gr 3A et 3B - Mercredi 13h45-15h45
- Gr 4A et 4B - Vendredi 16h-18h

Monitorat informatique
- en alternance avec les TD

Contrôle des connaissances :

contrôle continu - 1ère session
70 % Examen écrit (3 heures)
30 % Note de TD

2ème session
100 % Examen écrit (2 heures)

Bibliographie

- B. Cadet Méthodes statistiques en psychologie. P.U. de Caen
- G. Mialaret. Statistiques appliquées aux sciences humaines. PUF
- B. Beaufils. Statistiques appliquées à la psychologie. Tomes 1 et 2. LEXIFAC Bréal
- N. Guéguen. Manuel de statistiques pour psychologues
- D.C. Howell. Méthodes statistiques en sciences humaines
- M. Reuchlin. Précis de Statistiques. PUF Coll. Le Psychologue.
- P. Rateau, Méthode et statistique expérimentales en sciences humaines, Ellipses
- N. Gauvrit. Stats pour Psycho - 500 exercices corrigés. De Boeck
- A. Méot. Introduction aux statistiques inférentielles. De Boeck

Documents fournis :

Transparents du cours de statistiques
Polycopié du cours en Informatique
Fiches de TD de statistiques et d'informatique

Documents de l'année 2006/07 disponibles sur internet
– Au format .pdf lisible par Acrobat Reader :
Transparents du CM de Stats, fiches de TD de Stats
– Au format .pdf ou .doc (format Word) :
fiches de TD d'informatique

Adresses Web

`http://geai.univ-brest.fr/~carpenti/`
`http://infolettres.univ-brest.fr/~carpenti/`

Contenu

Statistiques :
Echantillonnage. Notion de test statistique.
Tests paramétriques : loi de Student et tests d'égalité de deux moyennes sur des groupes indépendants ou appareillés ; tests d'égalité de deux proportions ; introduction à l'analyse de variance ; loi de Fisher Snedecor.
Tests non paramétriques : test d'indépendance du khi-2 ; test de la médiane, test du signe ; protocoles de rangs et tests non paramétriques.

Echantillonnage

Echantillonnage - cas d'une moyenne

μ : moyenne sur la population
σ^2 : variance sur la population

Distribution d'échantillonnage de $\overline{X}$, moyenne observée sur un échantillon de taille n.
Loi normale (si $n \geq 30$)
Moyenne : $Moy(\overline{X}) = \mu$
Variance : $Var(\overline{X}) = \frac{\sigma^2}{n}$
La racine carrée de cette variance est appellée erreur standard ou erreur type.

Echantillonnage - cas d'une proportion

p : proportion dans la population
$\sigma^2 = p(1-p)$

Distribution d'échantillonnage de $\overline{F}$, proportion observée sur un échantillon de taille n.
Loi normale (si $np \geq 15$ et $n(1-p) \geq 15$)
Moyenne : $Moy(\overline{F}) = p$
Variance : $Var(\overline{F}) = \frac{p(1-p)}{n}$

Estimation de paramètres

Statistiques inférentielles

Raisonnement de type inductif : à partir de *conséquences* (ce qui est observé sur un (ou des) échantillons de taille n), remonter aux *causes* les plus probables (valeurs des paramètres dans la population).

Rappel : Estimation ponctuelle de paramètres

Population : μ, σ^2 inconnus.
Echantillon de taille n : $\overline{x}, s^2$ observés.

Estimation de μ : $\hat{\mu} = \overline{x}$
Estimation de σ^2 : $\hat{\sigma}^2 = s_c^2 = \dfrac{n}{n-1} s^2$

s_c^2 est appelée **variance corrigée**.

Introduction aux tests statistiques

Démarche générale d'un test

35 sujets soumis à un apprentissage. Deux tests l'un avant, l'autre après l'apprentissage.

Sujet	1	2	3	4	5	6	7	...
Avant	8	13	12	17	14	9	10	...
Après	11	11	14	21	12	10	15	...

Problème : L'apprentissage a-t-il un effet sur la performance ?

Remarques :
Raisonner en termes "d'échantillon tiré d'une population"
Variable pertinente : différence individuelle $d_i = y_i - x_i$

Protocole dérivé des différences individuelles

Sujet	1	2	3	4	5	6	7	...
d_i	3	-2	2	4	-2	1	5	...

Caractéristiques de position et de dispersion :
$\overline{d} = 1.08$; $s^2 = 5.05$; $s = 2.25$; $s_c^2 = 5.20$; $s_c = 2.28$

Construction d'un test statistique
Sujets observés : échantillon tiré dans une population
δ : moyenne des effets individuels dans la population.

1. Formulation des hypothèses
H_0 : hypothèse nulle : $\delta = 0$
H_1 : hypothèse alternative : $\delta \neq 0$

2. Choix d'un risque, ou seuil de signification
Par exemple : $\alpha = 5\%$

3. Choix d'une statistique de test
Une statistique est une variable qui peut être évaluée sur chaque échantillon tiré, et dont la distribution théorique, sous l'hypothèse H_0, est connue.

Ici, on prend : $Z = \dfrac{\overline{d}}{E}$ avec $E^2 = \dfrac{s_c^2}{n}$.

Les statisticiens ont montré que, sous l'hypothèse H_0, Z suit approximativement une loi normale centrée réduite.

4. Calcul des valeurs critiques (règle de décision)
Pour $\alpha = .05$, on obtient $z_{crit} = 1.96$.

5. Calcul de la valeur observée de la statistique
Ici : $z_{obs} = \frac{1.08}{0.38} = 2.84$

6. Comparer z_{obs} et z_{crit}. Appliquer la règle de décision
Ici : $z_{obs} > z_{crit}$. z_{obs} est dans la zone de rejet de H_0.
Sous H_0, l'échantillon tiré a une fréquence d'apparition inférieure à 5%. On refuse donc H_0 et on choisit H_1.

Remarques générales

Test : mécanisme permettant de trancher entre deux hypothèses à partir des résultats observés sur un ou plusieurs échantillons.

Hypothèses
Hypothèse nulle : elle joue un rôle particulier ; elle affirme que les différences observées sont dues au hasard.
Hypothèse alternative : elle affirme que les différences sont significatives (en un sens à préciser).

Les risques d'erreur

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

• α : seuil de significativité. C'est aussi la probabilité de rejeter H_0 alors que H_0 est vraie (risque de première espèce ou risque de commettre une erreur de type I)

• β : risque de seconde espèce. C'est la probabilité d'accepter H_0 alors que H_0 est fausse (risque de commettre une erreur de type II).

$1-\beta$: probabilité de détecter correctement un cas où H_0 doit être rejetée. Puissance du test.

Illustrations Commettre une ...

Erreur de type I : c'est voir une différence entre deux groupes alors qu'en fait, il n'y en a pas.
Exemples :
– Affirmer qu'un programme d'apprentissage coûteux a un effet sur le comportement des sujets, alors que c'est inexact
– "Mettre en évidence" une différence imaginaire entre les sexes, ou les races...
Comment diminuer ce risque : prendre α petit, veiller à neutraliser les autres variables, etc

Erreur de type II : c'est ne pas voir de différence, alors qu'il y en a réellement une.
C'est souvent un moindre mal, mais...
Exemples :
– Ne pas mettre en évidence un risque de somnolence lié à l'absoption d'un médicament.
– L'usine de La Hague est-elle réellement inoffensive pour les riverains ?
Comment diminuer ce risque : augmenter la taille de l'échantillon, ne pas prendre α trop petit, veiller à neutraliser les autres variables, bien choisir le test...

Test de comparaison d'une moyenne à une norme

Notations

X : variable numérique définie sur une population
μ_0 : moyenne (connue) de X sur la population de référence

$\overline{x}$: moyenne observée sur un échantillon
s_c : écart type corrigé observé sur l'échantillon
n : taille de l'échantillon.

On introduit μ : moyenne (inconnue) de X sur la population d'où est tiré l'échantillon.

Hypothèses du test

$H_0 : \mu = \mu_0$

H_1 : A choisir parmi : $\mu \neq \mu_0$ ou $\mu < \mu_0$ ou $\mu > \mu_0$

Statistique de test. Cas où $n > 30$

$$Z = \frac{\overline{x} - \mu_0}{E} \; avec \; E^2 = \frac{s_c^2}{n}$$

Z suit la loi normale centrée réduite.

Statistique de test. Cas où $n \leq 30$

$$T = \frac{\overline{x} - \mu_0}{E} \; avec \; E^2 = \frac{s_c^2}{n}$$

T suit la loi de Student à $n-1$ ddl.

Exemple

Une enquête nationale a montré que le score moyen à un test de niveau à l'entrée au collège est de 40.
Sur un groupe de 50 élèves, on observe une moyenne de 37, avec un écart type corrigé de 9.2.

Peut-on considérer que ce groupe a été tiré au hasard dans la population de l'ensemble des collégiens ?

μ : moyenne (inconnue) dans la population d'où a été tiré l'échantillon.

$H_0 : \mu = 40$

$H_1 : \mu \neq 40$ (test bilatéral)

Seuil choisi : $\alpha = 5\%$

Valeur critique de la statistique de test : $z_c = 1.96$.

Règle de décision : si $|z_{obs}| \leq 1.96$, on conclut sur H_0, sinon, on conclut sur H_1.

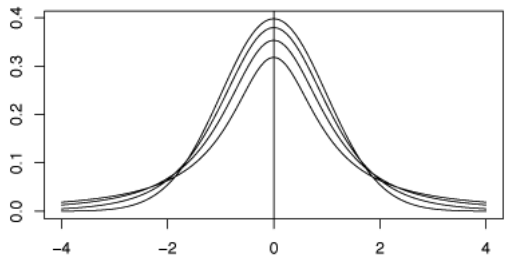
Calcul de la valeur observée de la statistique de test :

$$E^2 = \frac{9.2^2}{50} = 1.6928 \quad ; \quad E = 1.3011$$

$$z_{obs} = \frac{37 - 40}{1.30} = -2.31$$

On conclut donc sur H_1.

Loi de Student



Densité de la loi de Student pour ddl=1, ddl=2, ddl=5 et ddl=100

Cas particulier : Ecart type σ connu

Si la variable X est distribuée selon une loi normale dans la population parente, et si on connaît son écart type σ, la statistique à utiliser est :

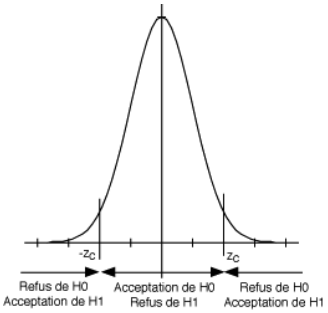
$$Z = \frac{\overline{x} - \mu_0}{E} \; avec \; E^2 = \frac{\sigma^2}{n}$$

Z suit alors une loi normale, même si $n \leq 30$.

Exemple : test de QI sur un échantillon.

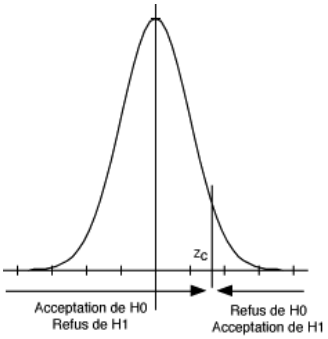
Détermination de la règle de décision selon la forme de H_1

• Test bilatéral. $H_1 : \mu \neq \mu_0$



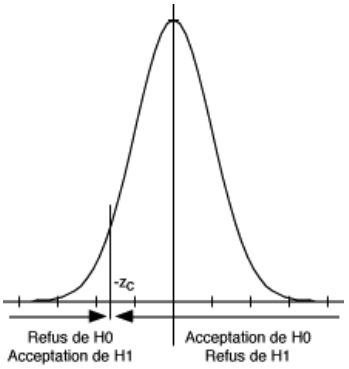
z_c définie par $P(-z_c \leq Z \leq z_c) = 1 - \alpha$

• Test unilatéral. $H_1 : \mu > \mu_0$



z_c définie par $P(Z \leq z_c) = 1 - \alpha$

• Test unilatéral. $H_1 : \mu < \mu_0$



z_c définie par $P(Z \leq z_c) = 1 - \alpha$

Remarque : raisonner en termes de niveau de significativité

Comment interpréter les résultats fournis par les logiciels ?

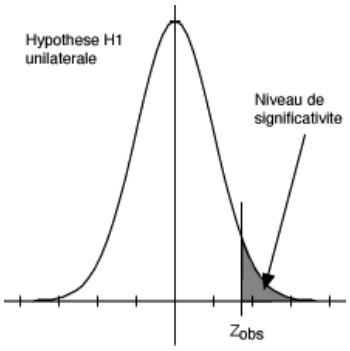
Niveau de significativité (NivSig ou p) : probabilité d'obtenir, sous H_0, une valeur de la statistique de test "au moins aussi extrême" que la valeur observée.

Pour un test unilatéral "à droite" :
$p = P(Z > z_{obs})$

Pour un test unilatéral "à gauche" :
$p = P(Z < z_{obs})$

Pour un test bilatéral :
$p = P(|Z| > |z_{obs}|)$

Exemple pour une hypothèse H_1 unilatérale "à droite" :



Règle :

Soit α le seuil et p le niveau de significativité.

– Si $p < \alpha$, on accepte H_1, et on refuse H_0 au seuil choisi

– Si $p \geq \alpha$, on refuse H_1, et on accepte H_0 au seuil choisi

Test de comparaison d'une proportion à une norme

Notations

X : variable dichotomique définie sur une population
p_0 : fréquence (connue) de la modalité "1" sur la population de référence

f : fréquence observée sur un échantillon
n : taille de l'échantillon.

On introduit p : fréquence (inconnue) de la modalité "1" sur la population d'où est tiré l'échantillon.

Hypothèses du test

$H_0 : p = p_0$

H_1 : A choisir parmi : $p \neq p_0$ ou $p < p_0$ ou $p > p_0$

Statistique de test.
Grands échantillons : $np_0 \geq 15$ et $n(1 - p_0) \geq 15$

$$Z = \frac{f - p_0}{E} \; avec \; E^2 = \frac{p_0(1 - p_0)}{n}$$

Z suit la loi normale centrée réduite.

Tests de comparaison de moyennes

Comparaison de deux moyennes.
Groupes appareillés

Expérience menée selon un plan $\mathcal{S}_n * \mathcal{A}_2$.

On introduit le protocole dérivé des différences individuelles.

Notations
μ_1, μ_2 : moyennes respectives des deux variables étudiées

δ : moyenne des différences individuelles sur la population ($\delta = \mu_1 - \mu_2$) (distribution normale)

n : taille de l'échantillon

$\overline{x_1}, \overline{x_2}$: moyennes respectives des deux variables sur un échantillon de taille n

$\overline{d}$: moyenne des différences individuelles sur un échantillon de taille n ($\overline{d} = \overline{x_1} - \overline{x_2}$)

s_c : écart type corrigé estimant l'écart type des différences individuelles sur la population parente

Hypothèses du test

H_0 : $\mu_1 = \mu_2$, c'est-à-dire $\delta = 0$

H_1 : A choisir parmi : $\mu_1 \neq \mu_2$ ou $\mu_1 < \mu_2$ ou $\mu_1 > \mu_2$

Statistique de test. Cas où $n > 30$

$$Z = \frac{\overline{d}}{E} \; avec \; E^2 = \frac{s_c^2}{n}$$

Sous H_0, Z suit la loi normale centrée réduite.

Statistique de test. Cas où $n \leq 30$

$$T = \frac{\overline{d}}{E} \; avec \; E^2 = \frac{s_c^2}{n}$$

Sous H_0, T suit la loi de Student à $n-1$ ddl.

Exemple :

Temps de réaction de 10 sujets mesuré à jeun ($\overline{x_1}$ = 22.3ms) et sous l'influence d'un tranquilisant ($\overline{x_2}$ = 31.7ms). Ecart type corrigé de la série des différences : $s_c = 11.54$.

H_0 : $\delta = 0$
H_1 : $\delta \neq 0$ (test bilatéral, par exemple).

La statistique de test T suit une loi de Student, et, pour un seuil de 5%, la valeur critique est : $t_{crit} = 2.26$.

Or, $E^2 = \frac{11.54^2}{10}$, $E = 3.65$,
$t_{obs} = \frac{22.3 - 31.7}{3.65} = -2.58$

On conclut donc sur H_1.

Comparaison de deux moyennes.
Groupes indépendants

Expérience menée selon un plan $\mathcal{S}_n < \mathcal{A}_2 >$

Notations
μ_1, μ_2 : moyennes sur les populations parentes respectives (distributions normales de même variance)

n_1, n_2 : tailles respectives des échantillons

$\overline{x_1}, \overline{x_2}$: moyennes respectives sur des échantillons de tailles n_1 et n_2

s_1, s_2 : écarts types des deux échantillons.

s_{1c}, s_{2c} : écarts types corrigés estimés à partir des échantillons.

Hypothèses du test

H_0 : $\mu_1 = \mu_2$

H_1 : A choisir parmi : $\mu_1 \neq \mu_2$ ou $\mu_1 < \mu_2$ ou $\mu_1 > \mu_2$

Statistique de test.

Grands échantillons
Cas où $n_1 > 30$ et $n_2 > 30$

$$Z = \frac{\overline{x_1} - \overline{x_2}}{E} \; avec \; E^2 = \frac{s_{1c}^2}{n_1} + \frac{s_{2c}^2}{n_2}$$

Sous H_0, Z suit la loi normale centrée réduite.

Petits échantillons
Cas où $n_1 \leq 30$ ou $n_2 \leq 30$

Groupes équilibrés : $n_1 = n_2 (= n)$

$$T = \frac{\overline{x_1} - \overline{x_2}}{E} \; avec \; E^2 = \frac{s_{1c}^2 + s_{2c}^2}{n}$$

Sous H_0, T suit la loi de Student à $2(n-1)$ ddl.

Petits échantillons - cas général :

$$T = \frac{\overline{x_1} - \overline{x_2}}{E} \; avec \; E^2 = \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2} \left(\frac{1}{n_1} + \frac{1}{n_2} \right)$$

T suit la loi de Student à $n_1 + n_2 - 2$ ddl.

Exemple :

Un groupe de 30 adultes jeunes, et un groupe de 30 adultes âgés. On soumet les sujets des deux groupes à une épreuve de fluence orthographique. Les paramètres calculés à partir des résultats observés sont les suivants :

Fluence orthographique		
	Jeunes	Agés
n	30	30
$\overline{x}$	11.4	11.0
s_c	3.1	3.2

$H_0 : \mu_1 = \mu_2$
$H_1 : \mu_1 \neq \mu_2$ (on fait ici un test bilatéral).

La statistique de test suit une loi de Student à $30 + 30 - 2 = 58$ ddl.

Pour un seuil de 5%, la valeur critique déduite de la table est $t_c = 2.0017$. La règle de décision est donc :

– si $-2.0017 \leq t_{obs} \leq 2.0017$, on retient H_0.
– si $t_{obs} < -2.0017$ ou $t_{obs} > 2.0017$, on rejette H_0 et on retient H_1.

Or : $E^2 = \frac{3.1^2}{30} + \frac{3.2^2}{30} = 0.6617$ d'où $E = 0.81$ et $t_{obs} = \frac{11.4 - 11.0}{0.81} = 0.4917$.

On retient donc l'hypothèse H_0 : on n'a pas mis en évidence de différence significative de la fluence verbale.

Comparaison de deux proportions. Groupes indépendants

Notations

p_1, p_2 : proportions dans les populations parentes respectives
n_1, n_2 : tailles respectives des échantillons
f_1, f_2 : proportions respectives dans des échantillons de tailles n_1 et n_2

Hypothèses du test
$H_0 : p_1 = p_2$
H_1 : choisir entre : $p_1 \neq p_2$ ou $p_1 < p_2$ ou $p_1 > p_2$

Statistique de test

$$Z = \frac{f_1 - f_2}{E} \; avec$$

$$E^2 = p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right) \; et \; p = \frac{n_1 f_1 + n_2 f_2}{n_1 + n_2}$$

Si $n_1 > 30$, $n_2 > 30$ et p "ni trop grand, ni trop petit" ($np \geq 15$ et $n(1-p) \geq 15$), alors, sous H_0, Z suit la loi normale centrée réduite.

Exemple

Variable *sexe* : deux groupes indépendants
Variable dépendante observée : succès/échec à une épreuve

Résultats

	M	F	Ensemble
Succès	150	120	270
Echec	90	40	130
Total	240	160	400

Calculs

$f_1 = 62,5\%$, $f_2 = 75\%$

$p = 67,5\%$, $n_1 = 240$, $n_2 = 160$

$z_{obs} = -2.615$

Pour un test bilatéral, $z_{crit} = 1.96$

Comparaison avec un test du χ^2

Obs	Théo	Contrib
150	162	0,89
90	78	1.84
120	108	1.33
40	52	2.77
		$\chi^2_{obs} = 6.84$

$ddl = 1$. Au seuil de 5%, $\chi^2_{crit} = 3.84$.

Remarque : $1.96^2 = 3.84$, $(-2.615)^2 = 6.84$, et ce n'est pas un hasard !

Comparaison de deux proportions. Groupes appareillés

Deux groupes appareillés : la même variable dichotomique a été utilisée pour tester un groupe de sujets dans deux conditions A_1 et A_2.
Résultats résumés par le tableau de contingence :

		A_1	
		Réussite	Echec
A_2	Réussite	a	c
	Echec	b	d

L'information utile est alors fournie par les effectifs "de discordance" b et c.

Notations
p_1 : fréquence de la combinaison (réussite en A_1, échec en A_2) par rapport à la discordance totale dans la population.
p_2 : fréquence de la combinaison (échec en A_1, réussite en A_2) par rapport à la discordance totale dans la population.

Hypothèses du test
$H_0 : p_1 = p_2 (= 50\%)$
H_1 : choisir entre : $p_1 \neq p_2$ ou $p_1 < p_2$ ou $p_1 > p_2$

Statistique de test

$$Z = \frac{b - c}{\sqrt{b + c}}$$

Si $b + c > 30$, sous H_0, Z suit la loi normale centrée réduite.

Pour un test bilatéral, on peut aussi utiliser comme statistique de test le χ^2 de Mac Nemar :

$$\chi^2 = \frac{(b-c)^2}{b+c}, \quad ddl = 1$$

ou, avec la correction de Yates (petits effectifs) :

$$\chi^2 = \frac{(|b-c|-1)^2}{b+c}, \quad ddl = 1$$

Exemple

Test de la mémoire à 2 semaines et à un an.

		2 semaines		
		Reconnu	Non reconnu	Total
Un an	Reconnu	81	8	89
	Non reconnu	46	49	95
	Total	127	57	184

$$z_{obs} = \frac{46-8}{\sqrt{46+8}} = 5.17$$

Pour un test unilatéral (Reconnaissance à 1 an $<$ Reconnaissance à 2 semaines), $z_{crit} = 1.645$. La différence est significative.

Conclusion

Tests *paramétriques*

Conditions d'application ou hypothèses *a priori* sur les populations parentes

– Groupes indépendants : distributions normales de même variance pour la variable dépendante,

– Groupes appareillés : distribution normale des effets individuels dans la population parente ou échantillon de grande taille

– Comparaison de fréquences : échantillons de taille suffisante et fréquence "ni trop grande, ni trop petite".

Il est généralement difficile de prouver que ces conditions sont respectées. Mais ces méthodes sont *robustes* et fournissent des résultats corrects même si les hypothèses ne sont qu'approximativement respectées.

Que fait-on quand elles ne le sont visiblement pas ?

Tests non paramétriques

Introduction

Les tests précédents portaient sur des paramètres des distributions observées (moyennes, fréquences). Mais on devait faire l'hypothèse *a priori* de la normalité des distributions parentes.

Au contraire, les tests non paramétriques

– ne nécessitent pas d'hypothèse a priori sur les distributions parentes

– peuvent s'appliquer à des variables ordinales (tests sur les rangs) ou même qualitatives (khi-2)

La mise en œuvre de la plupart de ces tests se fait en deux étapes :

– on construit un protocole dérivé : signes, rangs, etc

– le test proprement dit porte sur les variables dérivées.

Il existe de nombreux tests non paramétriques. Nous n'étudierons que les plus courants.

Indépendance de deux variables nominales - Test du χ^2

Deux variables nominales X et Y observées sur un échantillon de sujets.
Nombre de modalités de X : l
Nombre de modalités de Y : c

Problème : ces deux variables sont-elles indépendantes entre elles ?

Exemple : trois groupes de musiciens : professionnels (MP), en cours de professionnalisation (MCP) et amateurs (MA).
On s'intéresse au niveau d'études des trois groupes. Effectifs observés

	MP	MCP	MA	Total
avant bac.	7	11	4	22
bac.	12	6	5	23
post bac.	17	13	20	50
Total	36	30	29	95

Le niveau d'études et type de professionnalisation sont-ils liés ?

Hypothèses :

H_0 : Les variables X et Y sont indépendantes.
H_1 : Les variables X et Y sont dépendantes.

Statistique de test

Distance du χ^2 entre le tableau des effectifs observés et celui des effectifs théoriques (cf. calcul infra).

Cette statistique suit une loi du χ^2 à $(l-1)(c-1)$ ddl.

Calcul de la distance du χ^2

Données observées : tableau de contingence.

Effectifs attendus (ou théoriques) si indépendance :

Dans chaque case :

$$\text{Effectif théorique} = \frac{\text{total ligne} \times \text{total colonne}}{\text{total général}}$$

Contribution de chaque case au χ^2 :

$$\text{Ctr}_i = \frac{(\text{Eff. Observé} - \text{Eff. Théorique})^2}{\text{Eff. Théorique}}$$

Distance du χ^2 : $\chi^2_{obs} = \sum \text{Ctr}_i$.

Sur l'exemple fourni :

– On choisit un seuil de 5%.
– Le nombre de ddl est : $(3-1) \times (3-1) = 4$.
– Valeur critique : $\chi^2_{crit} = 9.49$

Effectifs théoriques

	MP	MCP	MA
avant bac.	8.34	6.95	6.72
bac.	8.71	7.26	7.02
post bac.	18.95	15.79	15.26

Calcul de la "distance" du χ^2

Mod.	n_{ij}	t_{ij}	$\frac{(n_{ij}-t_{ij})^2}{t_{ij}}$
MP. $<$ Bac	7	8.34	0.21
MP. Bac	...		1.23
MP. $>$ Bac			0.20
MCP. $<$ Bac			2.36
MCP. Bac			0.22
MCP. $>$ Bac			0.49
MA. $<$ Bac			1.09
MA. Bac			0.58
MA. $>$ Bac			1.47
Total			7.85

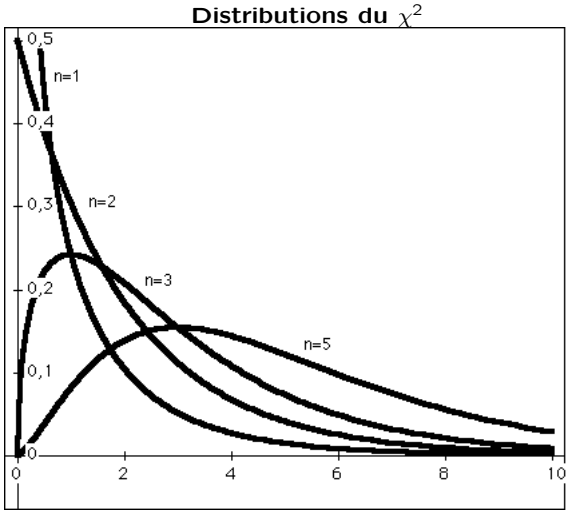
On obtient : $\chi^2_{obs} \leq \chi^2_{crit}$.

– Conclusion : On n'a pas mis en évidence de différence de niveau d'étude selon le type de professionnalisation.

Remarques

– Condition sur les effectifs théoriques minimaux
– Correction de Yates

Distributions du χ^2



Tests non paramétriques sur deux groupes indépendants

Test de la médiane sur des groupes indépendants

Une variable (la variable indépendante) définit deux groupes indépendants.

Une deuxième variable ordinale ou numérique.

Hypothèses

H_0 : Les deux populations parentes ont même médiane.
H_1 : Les deux populations parentes ont des médianes différentes

Construction de la statistique de test

On détermine la médiane M de la série obtenue en réunissant les deux échantillons.

On constitue un tableau de contingence en croisant la variable indépendante et la variable dérivée "position par rapport à M"

	Gr 1	Gr 2	Ensemble
$\leq M$	N_1	N_2	$N_1 + N_2$
$> M$	N_3	N_4	$N_3 + N_4$
Total	$N_1 + N_3$	$N_2 + N_4$	N

On fait un test du χ^2 sur le tableau obtenu.

Exemple

31 basketeurs de 14 ans, répartis en deux groupes d'effectifs $n_1 = 12$ et $n_2 = 19$, selon le jugement porté par l'entraîneur (groupe G_1 : jugement négatif ; groupe G_2 : jugement positif). On a relevé la taille de chaque sujet.

G_1 : 152 163 164 173 174 176 177 177 178 178 181 184

G_2 : 167 171 172 174 175 176 176 177 179 179 180 182 183 186 188 189 189 193 195

Les deux groupes sont-ils significativement différents du point de vue de la taille ?

Détermination de la médiane

152 163 164 167 171 172 *173 174* 174 175 176 176 *176* 177 *177* 177 *178 178* 179 179 180 *181* 182 183 *184* 186 188 189 189 193 195

On obtient : $Md = 177$

Tableau de contingence :

	Gr 1	Gr 2	Ensemble
$\leq Md$	8	8	16
$> Md$	4	11	15
Total	12	19	31

Effectifs théoriques et contributions au χ^2

	Gr 1	Gr 2
$\leq Md$	6.2	9.8
$> Md$	5.8	9.2

	Gr 1	Gr 2
$\leq Md$	0.52	0.33
$> Md$	0.56	0.35

Ici : $\chi^2_{obs} = 1.76$. Pour un seuil de 5%, $\chi^2_{crit} = 3.84$. On retient H_0.

Test de Wilcoxon-Mann-Whitney
Test U de Mann-Whitney

Deux groupes indépendants : deux échantillons tirés de deux populations distinctes.

Variable dépendante : ordinale ou numérique (par exemple, numérique comportant un très grand nombre de modalités).

Construction du protocole des rangs

On classe les n_1+n_2 sujets par valeurs croissantes (par exemple) de la variable. On attribue un rang à chaque sujet, avec la convention du rang moyen pour les ex æquos.

Exemple de construction du protocole des rangs

On reprend l'exemple "basket".

Deux groupes. Taille de chaque sujet.

G_1 : 152 163 164 173 174 176 177 177 178 178 181 184

G_2 : 167 171 172 174 175 176 176 177 179 179 180 182 183 186 188 189 189 193 195

Protocole des rangs :

Groupe	Taille	Rang
1	*152*	1
1	*163*	2
1	*164*	3
2	167	4
2	171	5
2	172	6
1	*173*	7

• *Lorsque $n_1 \geq 10$ et $n_2 \geq 10$: approximation par une loi normale*

$\overline{R_1}$: moyenne des rangs observés sur le premier échantillon

$\overline{R_2}$: moyenne des rangs observés sur le deuxième échantillon

$$Z = \frac{\overline{R_1} - \overline{R_2}}{E} \; avec \; E^2 = \frac{(n_1 + n_2 + 1)(n_1 + n_2)^2}{12 n_1 n_2}$$

Sous H_0, Z suit une loi normale centrée réduite.

Sur l'exemple :

$\overline{R}_1 = 12.13$; $\overline{R}_2 = 18.45$; $E^2 = 11.23$; $Z = -1.88$

Comme précédemment, on conclut sur H_1.

Remarque. Statistica calcule la statistique U de Mann-Whitney, liée aux sommes de rangs W_1 et W_2 par :

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - W_1$$

$$U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - W_2$$

$$U = \min(U_1, U_2)$$

Tests non paramétriques sur deux groupes appareillés

Test du signe

Un échantillon de sujets, placés dans deux conditions expérimentales différentes : groupes appareillés.

Variable dépendante : ordinale ou numérique.

– protocole du signe des différences individuelles
– on élimine les différences nulles

D_+ : nombre de différences positives
D_- : nombre de différences négatives
$N = D_+ + D_-$: nombre total d'observations après élimination des différences nulles.

Hypothèses du test :

H_0 : les différences sont dues au hasard : dans la population parente, la fréquence des différences positives est 50%.

H_1 : Cette fréquence n'est pas 50% (test bilatéral)
ou (tests unilatéraux)
Cette fréquence est inférieure à 50%
Cette fréquence est supérieure à 50%

• *Cas des petits échantillons ($N \leq 30$)*

Sous H_0, la variable statistique "nombre de sujets présentant une différence positive sur un échantillon de taille N" suit une *loi binomiale de paramètres N et 0.5*.

On raisonne en termes de "niveau de significativité".

Par exemple, dans le cas d'un test unilatéral tel que
H_1 : fréquence inférieure à 50%
on calcule la fréquence cumulée $P(X \leq D_+)$ de D_+ pour la loi binomiale $B(N, 0.5)$.
Pour un seuil α donné :
Si $P(X \leq D_+) < \alpha$ on retient H_1
Si $P(X \leq D_+) \geq \alpha$ on retient H_0

Exemple.

14 sujets observés dans deux conditions. 2 différences positives, 10 différences négatives, 2 différences nulles.

La statistique de test D_+ suit une loi binomiale de paramètres $N = 12$ et $p = 0.5$.

Calcul du niveau de significativité de $D_{+,obs}$:

$P(D_+ = 0) = C_{12}^0 \; 0.5^{12} = 0.0002441$

$P(D_+ = 1) = C_{12}^1 \; 0.5^{12} = 0.0029297$

$P(D_+ = 2) = C_{12}^2 \; 0.5^{12} = 0.0161133$

D'où : $P(D_+ \leq 2) = 0.019 = 1.9\%$

Au seuil de 5% unilatéral, on retient donc H_1.

• *Cas des grands échantillons : approximation par une loi normale ($N > 30$)*

$D = \max(D_+, D_-)$

$$Z = \frac{2D - 1 - N}{\sqrt{N}}$$

Z suit une loi normale centrée réduite.

Remarque. Dans le cas d'un test unilatéral, la zone de rejet est toujours située "à droite".

Exemple.

40 sujets observés dans deux conditions. 10 différences positives, 30 différences négatives, 0 différence nulle.

On a ici : $D = 30$ et $Z = \frac{60 - 1 - 40}{\sqrt{40}} = 3.00$

Au seuil de 1% unilatéral, on retient H_1 : les différences négatives sont significativement plus nombreuses que les différences positives.

Test de Wilcoxon sur des groupes appareillés
Test T, ou test des rangs signés

Un échantillon de sujets, placés dans deux conditions expérimentales différentes : groupes appareillés.

Variable dépendante : numérique.

On construit :

– le protocole des effets individuels d_i
– le protocole des valeurs absolues de ces effets $|d_i|$
– le protocole des rangs appliqués aux valeurs absolues, en éliminant les valeurs nulles.

T_+ : somme des rangs des observations tq $d_i > 0$
T_- : somme des rangs des observations tq $d_i < 0$
N = nombre de différences non nulles
$T_m = \min(T_+, T_-)$;
$T_M = \max(T_+, T_-)$

Hypothèses

H_0 : Dans la population parente, les effets individuels positifs et les effets individuels négatifs s'interclassent de manière homogène

H_1 : Les deux classements sont différents (test bilatéral) ou les effets individuels positifs apparaissent plus fréquemment dans les rangs les moins élevés (resp. les plus élevés) (test unilatéral).

Statistique de test

• *Cas des petits échantillons*
$N \leq 15$ *: utilisation de tables spécialisées*

On compare T_m aux valeurs critiques indiquées par la table.

Exemple

On a testé huit sujets dans deux conditions A_1 et A_2. On obtient le protocole suivant :

Suj.	A_1	A_2	d_i	$\lvert d_i \rvert$	r_{i+}	r_{i-}
s1	100	105	5	5	1	
s2	70	63	-7	7		2
s3	40	50	10	10	3	
s4	123	98	-25	25		4
s5	92	60	-32	32		5
s6	120	78	-42	42		6
s7	172	119	-53	53		7
s8	173	101	-72	72		8
T					4	32

On trouve $T_+ = 4$, $T_- = 32$ et donc $T_m = 4$.

Au seuil de 5% unilatéral, on lit dans la table : $T_{crit} = 5$.

Comme $T_m < T_{crit}$, on conclut à une différence significative entre les conditions A_1 et A_2 au seuil de 5% unilatéral.

• *Cas des grands échantillons*
$N > 15$ *: approximation par une loi normale*

$$Z = \frac{T_M - 0.5 - \frac{N(N+1)}{4}}{E}$$

avec

$$E^2 = \frac{N(N+1)(2N+1)}{24}$$

Sous H_0, Z suit une loi normale centrée réduite.

Analyse de Variance à un facteur

Exemple introductif : Test commun à trois groupes d'élèves. Moyennes observées dans les trois groupes : $\overline{x}_1 = 8$, $\overline{x}_2 = 10$, $\overline{x}_3 = 12$.

Question : s'agit-il d'élèves "tirés au hasard" ou de groupes de niveau ?

Première situation :

	Gr1	Gr2	Gr3
	6	8	10.5
	6.5	8.5	10.5
	6.5	8.5	11
	7	9	11
	7.5	9.5	11
	8	10	12
	8	11	13
	10	11	13
	10	12	14
	10.5	12.5	14
$\overline{x}_i$	8	10	12

Deuxième situation :

	Gr1	Gr2	Gr3
	5	4	6
	5.5	5.5	7
	6	7.5	9
	6	9	10
	6.5	9.5	11
	7	10	12
	7.5	11	13
	10	13	14
	12	15	18
	14.5	15.5	20
$\overline{x}_i$	8	10	12

Démarche utilisée : nous comparons la dispersion des moyennes (8, 10, 12) à la dispersion à l'intérieur de chaque groupe.

Comparer a moyennes sur des groupes indépendants

Plan d'expérience : $\mathcal{S} < \mathcal{A}_a >$

Une variable $\mathcal{A}$, de modalités $A_1, A_2, \ldots, A_a$ définit a groupes indépendants.

Variable dépendante X mesurée sur chaque sujet.
x_{ij} : valeur observée sur le i-ème sujet du groupe j.

Problème : La variable X a-t-elle la même moyenne dans chacune des sous-populations dont les groupes sont issus ?

Conditions d'application :
– distribution normale de X dans chacun des groupes
– Egalité des variances dans les populations.

Hypothèses du test :

H_0 : $\mu_1 = \mu_2 = \ldots = \mu_a$
H_1 : Les moyennes ne sont pas toutes égales.

Exemple :

15 sujets évaluent 3 couvertures de magazine. Sont-elles équivalentes ?

	C1	C2	C3	
	14	16	14	
	6	14	16	
	12	8	14	
	10	8	14	
	8	14	12	
$\overline{x}_i$	10	12	14	12

Variation (ou somme des carrés) totale :

$SC_T = (14-12)^2 + (6-12)^2 + \ldots + (12-12)^2 = 144$

Décomposition de la variation totale :

Score d'un sujet = Moyenne de son groupe + Ecart

C1	C2	C3
10	12	14
10	12	14
10	12	14
10	12	14
10	12	14

C1	C2	C3
4	4	0
-4	2	2
2	-4	0
0	-4	0
-2	2	-2

Variation (ou somme des carrés) inter-groupes :

$SC_{inter} = (10-12)^2 + (10-12)^2 + \ldots + (14-12)^2 = 40$

Variation (ou somme des carrés) intra-groupes :

$$SC_{intra} = 4^2 + (-4)^2 + \ldots + (-2)^2 = 104$$

Calcul des carrés moyens :

$$CM_{inter} = \frac{SC_{inter}}{a-1} = 20 \; ; \; CM_{intra} = \frac{SC_{intra}}{N-a} = 8.67$$

Statistique de test :

$$F_{obs} = \frac{CM_{inter}}{CM_{intra}} = 2.31$$

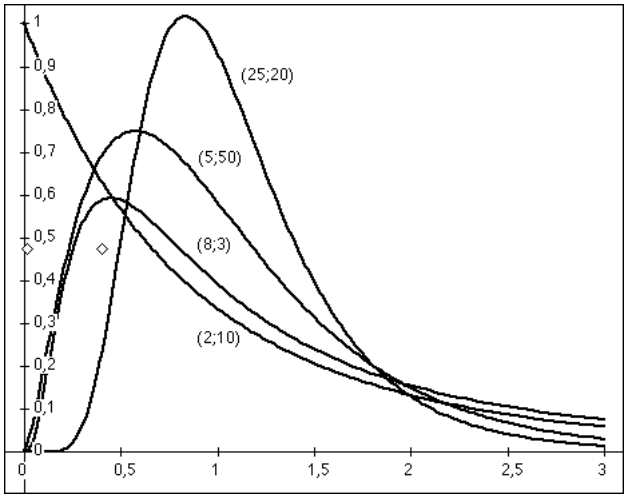
F suit une loi de Fisher avec $ddl_1 = a - 1 = 2$ et $ddl_2 = N - a = 12$.

Résultats

Source	Somme carrés	ddl	Carré Moyen	F
$\mathcal{C}$	40	2	20	2.31
Résid.	104	12	8.67	
Total	144	14		

Pour α=5%, $F_{crit} = 3.88$: H_0 est acceptée

Distributions du F de Fisher



Remarque

Si 2 groupes, équivaut à un T de Student. $F = T^2$

Pour les deux situations proposées en introduction :

Situation 1

Analysis of Variance Table

Response : x1

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
group	2	80.000	40.000	17.008	1.659e-05 ***
Residuals	27	63.500	2.352		

Situation 2

Analysis of Variance Table

Response : x2

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
group	2	80.00	40.00	2.7136	0.08436 .
Residuals	27	398.00	14.74		