Statistiques et informatique

UV PSY33A1

Présentation du cours 2000/2001

Organisation matérielle

Cours magistral : 13 heures. Statistiques et Informatique

A peu près 8 heures de statistiques et
5 heures d'informatique
jeudi 8h15-9h15 - Amphi 3

Travaux dirigés :

3 groupes de TD de statistiques.
- mercredi 8h15-9h15 - B222
- mercredi 13h45-14h45 - A335
- mercredi 14h45-15h45 - A335

6 sous-groupes de TD en Informatique.
- salles A204 et A206, 2 h. par quinzaine
- 2 sous-groupes - lundi 13h45-15h45
- 2 sous-groupes - mercredi 8h15-10h15
- 2 sous-groupes - jeudi 13h45-15h45

Monitorat informatique

Contrôle des connaissances : (contrôle continu)

70 % Examen écrit (3 heures)
30 % Revue de presse - Dossier

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.

- D. Laveault, J. Grégoire. Introduction aux théories des tests en sciences humaines. De Boeck Université

- J.P. Rossi. La méthode expérimentale en Psychologie

Contenu

Documents fournis :
Copie des transparents en statistiques
Polycopié du cours en Informatique
Fiches de TD
Documents de l'année 1998/99 (sauf fiches de TD d'informatique) disponibles sur internet (au format .pdf lisible par Acrobat Reader):

http://geai.univ-brest.fr/enseignements/psy33b.html

Statistiques :
Echantillonnage
Estimation de paramètres
Tests statistiques : comparaison de moyennes, tests non paramétriques

Informatique :
Comment l'informatique est-elle perçue ?
Homme et ordinateur : quelques idées sur l'inferface utilisateur.
Hypertextes et hypermedias.
Ordinateur et enseignement.

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, appellée erreur standard ou erreur type :

$$Var(\overline{X}) = \frac{\sigma^2}{n}$$

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 $np(1-p) \geq 20$)
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 dans la population).

Estimation ponctuelle de paramètres

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

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

Intervalle de Confiance

Population nombreuse.
Variable : score obtenu à un test
Echantillon : $n = 100$, $\overline{x}_{obs} = 44$, $s^2 = 142.6$, $s_c^2 = 144$

Distribution d'échantillonnage :
Loi normale
Moyenne : μ
Variance estimée : $\frac{144}{100} = 1,44$. Ecart type : 1,2

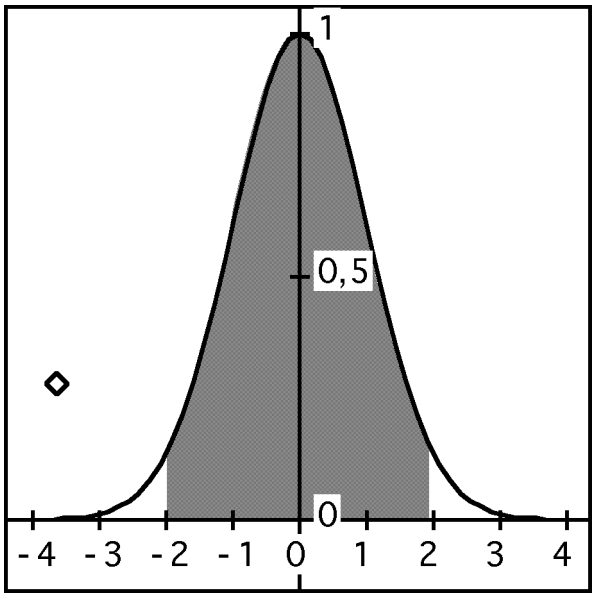
Rappel : Pour une loi normale de paramètres μ et σ, l'intervalle $[\mu - 1.96\sigma, \quad \mu + 1.96\sigma]$ rassemble 95% des observations.

Avec 95% de degré de confiance, on a :

$$\mu - 1.96 \times 1.2 \leq 44 \leq \mu + 1.96 \times 1.2$$

ou encore

$$44 - 1.96 \times 1.2 \leq \mu \leq 44 + 1.96 \times 1.2$$



Intervalle de confiance pour une moyenne
Cas des grands échantillons ($n \geq 30$)

Problème : μ inconnu. Estimer μ avec un degré de confiance $1-\alpha$ connaissant $\overline{x}_{obs}$, s^2, n.

Erreur type : $E^2 = \dfrac{s_c^2}{n}$

On a, avec le degré de confiance $1-\alpha$:

$$\overline{x}_{obs} - z_\alpha E \leq \mu \leq \overline{x}_{obs} + z_\alpha E$$

z_α : valeur lue dans la table de la loi normale centrée réduite, telle que : $P(-z_\alpha \leq Z \leq z_\alpha) = 1-\alpha$.

Intervalle de confiance pour une moyenne
Cas des petits échantillons ($n < 30$)

Avec les mêmes notations, on a avec le degré de confiance $1-\alpha$:

$$\overline{x}_{obs} - t_\alpha E \leq \mu \leq \overline{x}_{obs} + t_\alpha E$$

t_α : valeur lue dans la table de la **loi de Student** pour le seuil α et le nombre de degrés de liberté $ddl = n-1$.

Intervalle de confiance pour une proportion
Cas des grands échantillons ($np(1-p) \geq 20$)

Problème : p inconnu. Estimer p avec un degré de confiance $1-\alpha$ connaissant f_{obs}, n.

Erreur type : $E^2 = \dfrac{f_{obs}(1-f_{obs})}{n}$

On a, avec le degré de confiance $1-\alpha$:

$$f_{obs} - z_\alpha E \leq p \leq f_{obs} + z_\alpha E$$

z_α : valeur lue dans la table de la loi normale centrée réduite, telle que : $P(-z_\alpha \leq Z \leq z_\alpha) = 1-\alpha$.

Introduction aux tests statistiques

Démarche générale d'un test

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

Sujet	1	2	...
Avant	8	13	...
Après	11	11	...

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	8
d_i	3	-2	2	4	-2	1	5	3
Sujet	9	10	11	12	13	14	15	
d_i	4	-1	4	-1	3	3	2	

Caractéristiques de position et de dispersion :
$\overline{d} = 1.87$; $s^2 = 5.05$; $s = 2.25$; $s_c^2 = 5.41$; $s_c = 2.33$

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 : $T = \dfrac{\overline{d}}{E}$ avec $E^2 = \dfrac{s_c^2}{n}$.

Les statisticiens ont montré que, sous l'hypothèse H_0, T suit une loi de Student à $(n-1)$ ddl.

4. Calcul des valeurs critiques (règle de décision)
Pour $ddl = 14$ et $\alpha = .05$, on obtient $t_{crit} = 2.15$.

5. Calcul de la valeur observée de la statistique
Ici : $t_{obs} = \frac{1.87}{0.60} = 3.11$

6. Comparer t_{obs} et t_{crit}. Appliquer la règle de décision
Ici : $t_{obs} > t_{crit}$. t_{obs} est dans la zone de rejet.
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 écarts entre les valeurs observées et les valeurs théoriques sont dues au hasard.
Hypothèse alternative : elle affirme que les écarts sont significatifs (en un sens à préciser).

Les risques d'erreur

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.

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

Tests de comparaison de moyennes

Comparaison de deux moyennes. Groupes appareillés

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 $n > 30$

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

Z suit la loi normale centrée réduite.

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

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

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

Comparaison de deux moyennes. Groupes indépendants

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_{1c}, s_{2c} : écarts types corrigés estimés à partir des échantillon

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. 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.

Statistique de test. Cas où $n_1 \leq 30$ ou $n_2 \leq 30$

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

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

Si $n_1 = n_2 (= n)$, la formule se simplifie en :

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

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

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
$H_0 : p_1 = p_2$
$H_1 : p_1 \neq p_2$ 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", 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
		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!

Conclusion

Tests *paramétriques*

Hypothèses 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

Quel intérêt présentent ces tests ?

– En général, pas d'hypothèse a priori sur les populations parentes

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

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

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
$\leq M$	$N_{1\,1}$	$N_{1\,2}$	$N_{1\cdot}$
$> M$	$N_{2\,1}$	$N_{2\,2}$	$N_{2\cdot}$
Total	$N_{\cdot 1}$	$N_{\cdot 2}$	$N_{\cdot\cdot}$

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

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.

Hypothèses

H_0 : les rangs sont distribués de la même façon dans les deux groupes (ou: les individus des deux populations parentes s'interclassent de manière homogène).

H_1 : Les deux classements sont différents (test bilatéral)
Les individus du premier groupe apparaissent plus fréquemment dans les rangs les moins élevés (test unilatéral).

Construction de la statistique de test

n_1 *et* n_2 *petits : utilisation de tables*

On calcule la somme des rangs du plus petit des deux échantillons : W

On compare W aux valeurs critiques W_s ou W'_s fournies par la table.

$n_1 \geq 10$ *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

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

Z suit une loi normale centrée réduite.

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.

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%
Seuil choisi : α

Cas des petits échantillons ($N < 30$)

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*.

Intervalle d'acceptation, zone de rejet

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

Par exemple, dans le cas d'un test unilatéral tel que H_1 : fréquence inférieure à 50%, on calcule k_c tel que :

$$b(0) + b(1) + \ldots + b(k_c) < \alpha$$
$$b(0) + b(1) + \ldots + b(k_c) + b(k_c + 1) > \alpha$$

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".

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
$T_m = \min(T_+, T_-)$;
$T = \max(T_+, T_-)$

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

$N \leq 15$ *: utilisation de tables spécialisées*

On compare T_m aux valeurs critiques indiquées par la table.

$N > 15$ *: approximation par une loi normale*

$$Z = \frac{T - 0,5 - \frac{N(N+1)}{4}}{E} \; avec \; E^2 = \frac{N(N+1)(2N+1)}{24}$$

Z suit une loi normale centrée réduite.