Statistiques et informatique

E.C. PSY54AA

Présentation du cours 2013/2014

Organisation matérielle

Cours magistral : 12 heures
Mercredi 11h30-12h30 - Amphi 2

Travaux dirigés :
TD de statistiques
- Gr 1 : Merc. 14h45-15h45 - A220
- Gr 2 : Mardi 9h15-10h15 - B315
- Gr 3 - Lun. 8h15-10h15 - A223 - sem. A

TD d'informatique en sous-groupes
salle info. (A204 ou A206), 2 h. par quinzaine
- Gr 1-1 : Mardi 10h30-12h30 - sem. B
- Gr 1-2 : Jeudi 13h45-15h45 - sem. B
- Gr 2-1 : Mardi 10h30-12h30 - sem. A
- Gr 2-2 : Merc. 16h-18h - sem. B
- Gr 3-1 : Vend. 13h45-15h45 - sem B
- Gr 3-2 : Vend. 13h45-15h45 - sem A

Monitorat informatique
- en alternance avec les TD

Contrôle des connaissances :

contrôle continu - 1ère session
70 % Examen écrit (2 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
- A. Méot. Les tests d'hypothèses en psychologie expérimentale. De Boeck
- J. Navarro, L'essentiel de la statistique en psychologie, Ellipses, 2012

Documents fournis :
Transparents du cours de statistiques
Fiches de TD de statistiques et d'informatique

Documents 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

Adresse Web

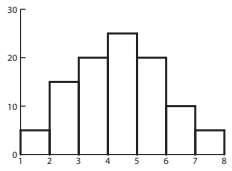
`http://geai.univ-brest.fr/~carpenti/`

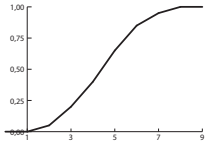
Contenu

Statistiques :
Echantillonnage. Notion de test statistique.
Tests paramétriques : loi de Student et tests d'égalité de deux moyennes sur des groupes indépendants ou appariés ; 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.

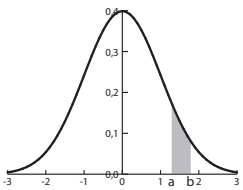
Lois théoriques continues

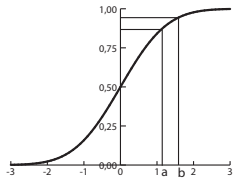
En statistiques descriptives, on rencontre des variables numériques continues, que l'on peut représenter à l'aide d'un histogramme ou d'une fonction de répartition.





De façon analogue, une loi théorique de distribution statistique est donnée par sa densité $f(x)$ ou sa fonction de répartition : $F(x)$





La fréquence (le pourcentage d'observations) vérifiant $a \leq X \leq b$ est donnée par l'aire hachurée ou par la valeur $F(b) - F(a)$.

Loi Normale ou loi de Laplace Gauss

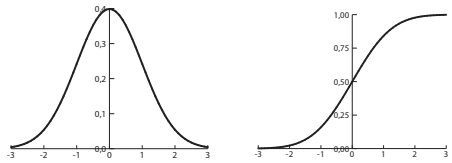
Problème : trouver une loi théorique modélisant la distribution d'une variable dont les valeurs résultent d'une combinaison d'effets *nombreux, indépendants entre eux, additifs et de même ordre de grandeur.*

Réponse : La loi normale.

Loi normale centrée réduite

Moyenne : $\mu = 0$. Ecart type : $\sigma = 1$

Densité : $f(x) = \frac{1}{\sqrt{2\pi}}\, e^{-\frac{x^2}{2}}$.



Loi normale, cas général : transformation en Z

La variable X suit une loi normale de paramètres μ et σ si la variable Z définie par :

$$Z = \frac{X - \mu}{\sigma}$$

suit une loi normale centrée réduite.

"Transformation en Z" ou "centrage réduction" de la variable.

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 tiré au hasard, 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 appelée 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).

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 : $\hat{\sigma}^2 = s_c^2 = \frac{n}{n-1} s^2$

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

Estimer une moyenne par un intervalle de confiance. Cas des grands échantillons ($n \geq 30$)

Exemple : Sur un échantillon de taille $n = 100$, on a observé une moyenne $\overline{x} = 44$ et un écart type corrigé $s_c = 12$.

Problème. Estimer la moyenne sur la population par un intervalle : "la moyenne au test sur la population est comprise entre a et b".

Nécessité d'introduire un *degré de confiance*, par exemple : $\beta = 95\%$.

La distribution d'échantillonnage de la variable $\overline{X}$, moyenne observée sur un échantillon de taille 100 a pour caractéristiques :
- distribution normale,
- moyenne μ,
- écart type $E = \frac{12}{\sqrt{100}} = 1.2$.

On introduit la variable normale centrée réduite
$Z = \frac{\overline{X} - \mu}{1.2}$ et $z_{obs} = \frac{44 - \mu}{1.2}$

On sait (lecture des tables) que, dans 95% des cas, on a : $-1.96 \leq Z \leq 1.96$.

On affirme alors, avec le degré de confiance 95%, que :

$$-1.96 \leq \frac{44 - \mu}{1.2} \leq 1.96.$$

Finalement, on obtient : $41.65 \leq \mu \leq 46.35$.

Synthèse

Problème : μ moyenne inconnue sur une population.

Estimer μ avec un degré de confiance $\beta = 1-\alpha$ connaissant $\overline{x}_{obs}$, s_c, n sur un grand échantillon ($n \geq 30$) tiré au hasard dans la population.

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

On a, avec le degré de confiance β :

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

z_α : valeur lue dans la table de la loi normale centrée réduite, telle que :

$$P(|Z| > z_\alpha) = \alpha$$

ou

$$P(-z_\alpha \leq Z \leq z_\alpha) = \beta$$

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.

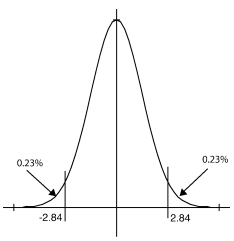
Raisonner en termes de "niveau de significativité"

Avec un logiciel de traitement statistique, les étapes 4, 5 et 6 sont remplacées par :

4'. Calcul de la valeur observée de la statistique
Comme ci-dessus : $z_{obs} = 2.84$

5'. Calcul de la *p-value* correspondante
On évalue, sous l'hypothèse H_0, la fréquence (ou probabilité) d'apparition de tous les protocoles *au moins aussi extrêmes que celui observé.*
Ici : $p = P(Z \leq -2.84) + P(Z \geq 2.84) = 1 - 2 \times 0.4977 = 0.0046 = 0.46\%$.

Autre formulation : si H_0 est vraie, on a seulement 0.46% de chances de tirer un échantillon conduisant à $Z \leq 2.84$ ou $Z \geq 2.84$.



6'. Comparaison du seuil et de la p-value ; conclusion
Ici : $p = 0.46\%$ et $\alpha = 5\%$. D'où $p < \alpha$. Au seuil de 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	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 $n > 30$

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

Sous H_0, Z suit la loi normale centrée réduite.

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

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

Sous H_0, 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 :

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

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

On conclut donc sur H_1.