STATISTIQUES

(résumé du cours)

Christian MICHEL

Université Louis Pasteur Strasbourg
Département Informatique

michel@dpt-info.u-strasbg.fr

PLAN

CS1: MESURES STATISTIQUES

1. INTRODUCTION

1.1. Types de mesure

Trois types de mesure pour l'analyse des données

(i) Mesures de tendance centrale: la moyenne arithmétique, la moyenne géométrique, la moyenne harmonique, la moyenne quadratique, la médiane et le mode.

(ii) Mesures de dispersion: la variance, l'écart-type, l'étendue et l'écart-moyen absolu.

(iii) Mesures de concentration (non étudiées dans ce cours).

1.2. Notation

Soit X_i une donnée d'un échantillon de taille n, $i \in \{1, \dots, n\}$.

2. MESURES DE TENDANCE CENTRALE

2.1. Moyenne (arithmétique)

Déf

La moyenne (arithmétique) $\bar{X}$ est la somme des n valeurs divisée par n.

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

2.1.1. Moyenne pondérée

Si aux valeurs X_i de X sont associées des poids w_i alors

$$\bar{X} = \frac{1}{\sum_{i=1}^{n} w_i}\sum_{i=1}^{n} w_i X_i$$

2.1.2. Propriétés

P1

$$\sum_{i=1}^{n}\left(X_i - \bar{X}\right) = 0$$

La somme algébrique des écarts d'un ensemble de nombres à leur moyenne est nulle.

P2

La somme des carrés des écarts d'un ensemble de nombres à un nombre a donné $\sum_{i=1}^{n}\left(X_i - a\right)^2$ est minimal si et seulement si $a = \bar{X}$.

P3

Si les X_i ont un écart $d_i = X_i - a$ avec un nombre a donné alors

$$\bar{X} = a + \frac{1}{n}\sum_{i=1}^{n} d_i$$

2.2. Moyenne géométrique

Déf

La moyenne géométrique g est la racine n-ième du produit des n valeurs.

$$g = \sqrt[n]{\prod_{i=1}^{n} X_i}$$

2.3. Moyenne harmonique

Déf

La moyenne harmonique h est l'inverse de la moyenne des inverses des n valeurs.

$$h = \frac{1}{\frac{1}{n}\sum_{i=1}^{n} \frac{1}{X_i}}$$

2.4. Moyenne quadratique

Déf

La moyenne quadratique q est la racine carrée de la moyenne des carrés des n valeurs.

$$q = \sqrt{\frac{1}{n}\sum_{i=1}^{n} X_i^2}$$

2.5. Médiane

Déf

La médiane est la valeur de la variable statistique du milieu, l'ensemble des valeurs étant rangées par ordre de grandeur croissante (par convention):

(i) si n est impair, la médiane est $X_{(n+1)/2}$

(ii) si n est pair, la médiane est $X_{n/2}$ (par convention)

2.6. Mode

Déf

Le mode est la valeur de la variable statistique la plus fréquente. Le mode peut ne pas exister et s'il existe, il peut ne pas être unique.

3. MESURES DE DISPERSION

3.1. Variance

Déf

La variance S^2 est la moyenne des carrés des écarts d'un ensemble de nombres à leur moyenne.

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$$

3.1.1. Variance pondérée

Si aux valeurs X_i de X sont associées des poids w_i alors

$$S^2 = \frac{1}{\sum_{i=1}^{n} w_i}\sum_{i=1}^{n} w_i\left(X_i - \bar{X}\right)^2$$

3.1.2. Méthode rapide de calcul de la variance

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i^2 - 2\bar{X}X_i + \bar{X}^2\right) = \frac{1}{n}\sum_{i=1}^{n}X_i^2 - 2\bar{X}\frac{1}{n}\sum_{i=1}^{n}X_i + \bar{X}^2$$

$$= \frac{1}{n}\sum_{i=1}^{n}X_i^2 - \bar{X}^2$$

3.1.3. Propriétés

P1

Si la variance est définie pour un nombre quelconque a, c'est-à-dire $S^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - a\right)^2$ alors parmi toutes les variances possibles, la plus petite est celle pour laquelle $a = \bar{X}$.

P2

Si on transforme les données X_i telles que $X'_i = aX_i + b$ où a et b sont 2 constantes, alors

$$\bar{X}' = a\bar{X} + b$$

$$S'^2 = a^2 S^2$$

P3: Autre expression de S^2

$$S^2 = \frac{1}{2n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(X_i - X_j\right)^2$$

3.2. Ecart-type

Déf

L'écart-type S est la racine carrée de la variance S^2

$$S = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}$$

3.2.1. Ecart-type pondéré

Si aux valeurs X_i de X sont associées des poids w_i alors

$$S = \sqrt{\frac{1}{\sum_{i=1}^{n} w_i}\sum_{i=1}^{n} w_i\left(X_i - \bar{X}\right)^2}$$

3.2.2. Méthode rapide de calcul de l'écart-type

$$S = \sqrt{\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2}$$

3.2.3. Propriétés

P1

L'écart-type s'exprime avec les mêmes unités que la variable statistique (contrairement à la variance qui est un carré).

P2

Si on transforme les données X_i telles que $X'_i = \frac{X_i - \bar{X}}{S}$, alors

$$\bar{X}' = 0$$

$$S'^2 = S' = 1$$

X'_i est appelée variable centrée réduite.

P3

Si l'écart-type est défini pour un nombre quelconque a, c'est-à-dire $S = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - a)^2}$ alors parmi tous les écarts-types possibles, le plus petit est celui pour lequel $a = \bar{X}$.

3.3. Etendue

Déf

L'étendue est la différence entre les valeurs extrêmes, c'est-à-dire $X_n - X_1$.

3.4. Ecart moyen absolu

Déf

L'écart moyen absolu EMA est la moyenne de la valeur absolue des écarts d'un ensemble de nombres à leur moyenne

$$EMA = \frac{1}{n}\sum_{i=1}^{n}\left|X_i - \bar{X}\right|$$

P1

La somme des valeurs absolues des écarts d'un ensemble de nombres à un nombre a donné est minimale si et seulement si a=médiane

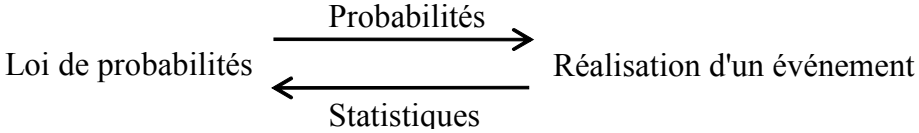
$$\text{Min}\left\{\sum_{i=1}^{n}\left|X_i - a\right|\right\} = \text{Médiane}$$

CS2: ESTIMATION

1. INTRODUCTION

1.1. Présentation du concept

Les probabilités s'intéressent à la structure d'une population, c'est-à-dire à une distribution théorique en précisant ses propriétés, la valeur de ses paramètres, etc. Cette approche permet de calculer les probabilités d'obtenir un échantillon donné de la population. Les statistiques s'intéressent à l'approche inverse qui consiste à déterminer la distribution théorique à partir d'un échantillon issu de la population.



L'estimation est un problème statistique qui consiste à se servir des données fournies par un échantillon issu d'une population pour attribuer certaines valeurs aux paramètres inconnus de la distribution théorique de la population. L'estimation ponctuelle consiste à attribuer une valeur unique aux paramètres inconnus. L'estimation par intervalle consiste à déterminer un intervalle (région de confiance) dans lequel se situeront les paramètres inconnus. Dans ce cas, il faudra également chiffrer la confiance (la crédibilité) attachée à cet intervalle. L'estimation ponctuelle et par intervalle sera limitée dans ce cours aux paramètres inconnus (théoriques) concernant la moyenne et la variance.

1.2. Notation

	Paramètres théoriques (inconnus) de la population	Paramètres expérimentaux (connus) de l'échantillon
Moyenne	μ	$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$
Variance	σ^2	$S^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$

où n représente la taille de l'échantillon et où X_i est la variable aléatoire associée à la valeur de l'échantillon. Donc, les paramètres expérimentaux X_i, $\bar{X}$ et S^2 (connus) de l'échantillon sont donc des variables aléatoires (appelés estimateurs) qui permettent d'estimer les paramètres théoriques (inconnus) de la population. Les symboles de l'échantillon sont représentés par des lettres majuscules, les symboles de la population, par des lettres minuscules grecques. Les symboles μ (resp. σ^2) sont notés $E(X)$ (resp. $V(X)$) dans le cours de Probabilités.

1.3. Théorèmes (rappel du cours de Probabilités)

Th1

$$\text{Si } X_i \sim N(0,1) \text{ et si } X_i \text{ indépendants alors } \sum_{i=1}^{n} X_i^2 \sim \chi_n^2$$

$$X_i \sim N(\mu,\sigma) \text{ et si } X_i \text{ indépendants alors } \sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma} \right)^2 \sim \chi_n^2$$

Th2

$$\text{Si } X_i \sim N(\mu,\sigma) \text{ et si } X_i \text{ indépendants alors } \sum_{i=1}^{n} \left(\frac{X_i - \bar{X}}{\sigma} \right)^2 \sim \chi_{n-1}^2$$

Th3

$$\text{Si } X_i \sim N(\mu,\sigma) \text{ et si } X_i \text{ indépendants alors } \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right) \text{ et } \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$$

Th4

$$\text{Si } X \sim N(0,1) \text{ et } X' \sim \chi_n^2 \text{ alors } \frac{X}{\sqrt{\frac{X'}{n}}} \sim t_n$$

Th5

$$E\left(\chi_n^2\right) = n \text{ et } V\left(\chi_n^2\right) = 2n$$

Th6

$$E\left(t_n\right) = 0 \text{ et } V\left(t_n\right) = \frac{n}{n-2} \quad n > 2$$

2. ESTIMATION PONCTUELLE

2.1. Propriétés d'un estimateur

2.1.1. Estimateur non biaisé

Un estimateur T, c'est-à-dire la v.a. $\bar{X}$ ou S^2, d'un paramètre θ, c'est-à-dire μ ou σ^2, est dit non biaisé si son espérance $E(T)$ est égale au paramètre θ : $E(T) = \theta$. Un estimateur T d'un paramètre θ est dit biaisé si il existe un biais b égal à la différence entre son espérance $E(T)$ et le paramètre θ : $b = E(T) - \theta$.

2.1.2. Estimateur convergent

Un estimateur non biaisé T est dit convergent si sa variance $V(T)$ tend vers 0 quand la taille n de l'échantillon tend vers l'infini

$$\lim_{n \to \infty} V(T) = 0$$

2.1.3. Estimateur efficace

Un estimateur non biaisé T est dit d'autant plus efficace que sa variance est petite. Pour un échantillon de taille n fixée, l'efficacité d'une estimation se mesure par sa variance. La variance d'un estimateur non biaisé donné est toujours supérieure ou égale à une limite appelée borne de Cramer-Rao. Quand cette limite est effectivement atteinte, l'estimateur est de variance minimale relativement à tous les autres estimateurs non biaisés possibles du même paramètre. En règle générale, la variance d'un estimateur est inversement proportionnelle à la taille de l'échantillon étudié. Plus n est grand plus l'estimation sera précise puisque la variance de l'estimateur diminue.

2.1.4. Conclusion

Un estimateur sera d'autant meilleur qu'il sera non biaisé, convergent et efficace.

2.2. Estimateur de la moyenne avec une variance connue

Le meilleur estimateur de $\theta = \mu$ avec σ connu est

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \text{ avec } E(\bar{X}) = \mu \text{ et } V(\bar{X}) = \frac{\sigma^2}{n}$$

$\bar{X}$ est un estimateur non biaisé, convergent et efficace.

2.3. Estimateur de la moyenne avec une variance inconnue

Le meilleur estimateur de $\theta = \mu$ avec σ inconnu est

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \text{ avec } E(\bar{X}) = \mu \text{ et } V(\bar{X}) = \frac{S^2}{n-1}$$

$\bar{X}$ est un estimateur non biaisé, convergent et efficace.

2.4. Estimateur de la variance avec une moyenne connue

Le meilleur estimateur de $\theta = \sigma^2$ avec μ connu est

$$S^2 = \frac{1}{n}\sum_{i=1}^{n} (X_i - \mu)^2 \text{ avec } E(S^2) = \sigma^2 \text{ et } V(S^2) = \frac{2\sigma^4}{n}$$

S est un estimateur non biaisé, convergent et efficace.

Ne pas confondre $S^2 = \frac{1}{n}\sum_{i=1}^{n} (X_i - \bar{X})^2$ et $S^2 = \frac{1}{n}\sum_{i=1}^{n} (X_i - \mu)^2$

2.5. Estimateur de la variance avec une moyenne inconnue

Le meilleur estimateur de $\theta = \sigma^2$ avec μ inconnu est

$$\frac{n}{n-1}S^2 \text{ avec } E\left(\frac{n}{n-1}S^2\right) = \sigma^2 \text{ et } V\left(\frac{n}{n-1}S^2\right) = \frac{2\sigma^4}{n-1}$$

$\frac{n}{n-1}S^2$ est un estimateur non biaisé, convergent et efficace.

Rem

(i) $\frac{n}{n-1}S^2 = \frac{n}{n-1} \times \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$ est l'estimateur de σ^2.

(ii) Si n est grand $(n \geq 30)$ alors $\frac{n}{n-1} \sim 1$ et $\frac{n}{n-1}S^2 = S^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$ peut également être considéré comme un estimateur de σ^2.

3. ESTIMATION PAR INTERVALLE

3.1. Niveau de confiance de l'intervalle de confiance

Au lieu d'attribuer à un paramètre inconnu θ à estimer une valeur unique comme dans l'estimation ponctuelle, l'estimation par intervalle donne un ensemble de valeurs susceptibles d'être prises par ce paramètre. Une borne inférieure et une borne supérieure délimitent cet intervalle de valeurs appelé intervalle de confiance. Cet intervalle de confiance est affecté d'un coefficient de crédibilité appelé niveau de confiance. Ce niveau de confiance est quantifié par une probabilité $1-\alpha$ où α est la probabilité d'erreur appelé risque d'erreur; le plus souvent $\alpha = 5\%$ ou $\alpha = 1\%$. Ainsi, un paramètre inconnu θ se trouve dans un intervalle de confiance avec un niveau de confiance $1-\alpha$ signifie que si l'on prélevait un grand nombre d'échantillons de la même population, $1-\alpha$ des intervalles de confiance calculés de la même manière contiendraient effectivement le paramètre inconnu θ.

3.2. Estimation de la moyenne d'une v.a. Normale de variance connue

Au niveau de confiance $1-\alpha$

$$\mu = \bar{X} \pm b\frac{\sigma}{\sqrt{n}}$$

écrit également sous forme d'intervalle $\mu \in \left[\bar{X} - b\frac{\sigma}{\sqrt{n}}, \bar{X} + b\frac{\sigma}{\sqrt{n}}\right]$

μ : moyenne théorique de la population à estimer

$\bar{X}$: moyenne expérimentale de l'échantillon de valeur X_i et de taille n

σ : écart-type théorique de la population

b : valeur lue dans la table Normale Centrée Réduite $N(0,1)$ telle que

$$P(N(0,1) > b) = \frac{\alpha}{2}$$

3.3. Estimation de la moyenne d'une v.a. Normale de variance inconnue

Au niveau de confiance $1-\alpha$

$$\mu = \overline{X} \pm b\sqrt{\frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}{n(n-1)}}$$

μ : moyenne théorique de la population à estimer

$\overline{X}$: moyenne expérimentale de l'échantillon de valeur X_i et de taille n

b: valeur lue dans la table de Student t_{n-1} à n-1 degrées de libertés telle que $P\left(t_{n-1} > b\right) = \alpha/2$

3.4. Estimation de la variance d'une v.a. Normale de moyenne connue

Au niveau de confiance $1-\alpha$

$$\sigma^2 \in \left[\frac{1}{b}\sum_{i=1}^{n}(X_i - \mu)^2, \frac{1}{a}\sum_{i=1}^{n}(X_i - \mu)^2\right]$$

μ : moyenne théorique de la population

X_i : valeur de l'échantillon de taille n

σ^2 : variance théorique de la population à estimer

a et b: valeurs lues dans la table du Khi-Deux χ_n^2 à n degrés de libertés telles que

$$\begin{cases} P\left(\chi_n^2 < a\right) = \alpha/2 \\ P\left(\chi_n^2 > b\right) = \alpha/2 \end{cases}$$

3.5. Estimation de la variance d'une v.a. Normale de moyenne inconnue

Au niveau de confiance $1-\alpha$

$$\sigma^2 \in \left[\frac{1}{b}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2, \frac{1}{a}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2 \right]$$

$\bar{X}$: moyenne expérimentale de l'échantillon de valeur X_i et de taille n

σ^2: variance théorique de la population à estimer

a et b: valeurs lues dans la table du Khi-Deux χ^2_{n-1} à $n-1$ degrés de libertés telles que

$$\begin{cases} P\left(\chi^2_{n-1} < a\right) = \alpha/2 \\ P\left(\chi^2_{n-1} > b\right) = \alpha/2 \end{cases}$$

3.6. Cas des grands échantillons (n≥30)

Propriétés quand la taille n de l'échantillon tend vers l'infini

P1

$$t_n \sim N\left(0, \sqrt{\frac{n}{n-2}}\right) \sim N(0,1)$$

P2

$$\chi^2_n \sim N\left(n, \sqrt{2n}\right)$$

3.6.1. Estimation de la moyenne d'une v.a. Normale de variance connue

Test identique au test 3.2.

3.6.2. Estimation de la moyenne d'une v.a. Normale de variance inconnue

Test identique au test 3.3 mais b est lue dans la table Normale Centrée Réduite $N(0,1)$ (et non dans la table t_{n-1}) telle que $P(N(0,1) > b) = \alpha/2$.

3.6.3. Estimation de la variance d'une v.a. Normale de moyenne connue

Au niveau de confiance $1-\alpha$

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 \times \frac{1}{1 \pm b\sqrt{\frac{2}{n}}}$$

μ : moyenne théorique de la population

X_i : valeur de l'échantillon de taille n

σ^2 : variance théorique de la population à estimer

b: valeur lue dans la table Normale Centrée Réduite $N(0,1)$ telle que $P(N(0,1) > b) = \frac{\alpha}{2}$

3.6.4. Estimation de la variance d'une v.a. Normale de moyenne inconnue

Au niveau de confiance $1-\alpha$

$$\sigma^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \times \frac{1}{1 \pm b\sqrt{\frac{2}{n-1}}}$$

$\bar{X}$: moyenne expérimentale de l'échantillon de valeur X_i et de taille n

σ^2 : variance théorique de la population à estimer

b: valeur lue dans la table Normale Centrée Réduite $N(0,1)$ telle que $P(N(0,1) > b) = \alpha/2$

3.6.5. Conséquence du théorème central limite

Le théorème central limite prouve que $\bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$ quelque soit la distribution de la population dont l'échantillon de variable aléatoire X_i est extrait, mais à condition que n soit grand et les X_i indépendants. Pour n grand, les X_i ne doivent pas nécessairement suivre une loi Normale (condition qui est nécessaire dans le Th3). Cette conséquence du théorème central limite permet de nombreuses applications.

Application: Calcul de l'intervalle de confiance du paramètre p d'une v.a. de Bernoulli X au niveau de confiance $1-\alpha$

$$p = \frac{\left(2n\bar{X} + b^2\right) \pm \left(b\sqrt{b^2 + 4n\bar{X} - 4n\bar{X}^2}\right)}{2\left(n + b^2\right)}$$

3.7. Estimations par intervalle non étudiées

La différence des moyennes de 2 v.a. Normales (loi de Student) et le quotient des variances de 2 v.a. Normales (loi de Fischer-Snédécor) peuvent être estimés par intervalle. Les résultats seront donnés dans le cadre des tests d'hypothèse.

CS3: TESTS D'HYPOTHESE

1. INTRODUCTION

Il y a une analogie entre les problèmes d'estimation et particulièrement ceux de l'estimation par intervalle, et les problèmes de tests d'hypothèse. Dans les deux cas
- on dispose d'échantillons issus d'une population
- on fixe une probabilité d'erreur α
- on étudie des moyennes, des variances, des coefficients de corrélation, etc.
- on distingue entre échantillons de taille élevée $n \geq 30$ et de taille petite $n < 30$.

Toutefois, la nature du problème est différente. En effet, les tests d'hypothèse se proposent de vérifier des hypothèses se rapportant à des paramètres inconnus d'une population, par exemple que tel paramètre prend telle valeur, que 2 paramètres de 2 échantillons proviennent de la même population, etc. Les tests d'hypothèse ont donc pour objet de se prononcer sur la validité de l'hypothèse. Cette approche constitue la théorie de la décision limitée dans ce cours aux 2 alternatives suivantes: l'hypothèse est soit vraie soit fausse et l'hypothèse est soit acceptée soit rejetée. L'hypothèse étudiée, appelée hypothèse nulle, est symbolisée par H_0. La contre-hypothèse, appelée hypothèse alternative, est symbolisée par H_1.

Tableau récapitulant les erreurs possibles à la vue de ces 4 situations

Décision Hypothèse	On accepte H_0	On rejette H_0 = On accepte H_1
H_0 vraie	Parfait Probabilité $= 1-\alpha$	Erreur de 1ère espèce Probabilité $= \alpha$
H_0 fausse = H_1 vraie	Erreur de 2ème espèce Probabilité $= \beta$	Parfait Probabilité $= 1-\beta$

Il y a 2 types d'erreur possibles

$P(\text{Rejeter } H_0 \,|\, H_0 \text{ vraie}) = \alpha$: erreur de 1ère espèce

$P(\text{Accepter } H_0 \,|\, H_0 \text{ fausse}) = \beta$: erreur de 2ème espèce

Il n'y a pas de lien entre les 2 probabilités d'erreur α et β. L'idéal serait d'obtenir des valeurs aussi proches de 0 que possible. En général, il est impossible de minimiser à la fois ces 2 probabilités d'erreur. On dissymétrise alors le problème en tenant compte du fait que dans la pratique l'une des erreurs est plus grave que l'autre. L'erreur de 1ère espèce sera choisie comme étant la plus grave. On prendra donc un α très petit puis on minimisera β. En conclusion, on choisit comme hypothèse H_0 celle dont le rejet entraîne les conséquences les plus graves.

Ex

Un homme accusé d'un délit comparaît devant un juge. Cet homme est soit innocent (H_0 vraie) soit coupable (H_0 fausse = H_1 vraie). Le juge a le choix entre 2 décisions: il accepte l'innocence (H_0) ou il accepte la culpabilité (H_1).

L'innocence est l'hypothèse H_0 parce que son rejet entraîne les conséquences les plus graves:

(i) Si le juge accepte l'innocence alors que l'homme est innocent: c'est parfait.

(ii) Si le juge accepte la culpabilité (rejette l'innocence) alors que l'homme est innocent est une erreur très grave: il condamne un innocent.

(iii) Si le juge accepte la culpabilité alors que l'homme est coupable: c'est parfait.

(iv) Si le juge accepte l'innocence alors que l'homme est coupable est une erreur moins grave que l'erreur précédente: il est plus grave de condamner un innocent que de gracier un coupable.

Rem

(i) la probabilité $1-\beta$ détermine la puissance du test.

(ii) l'ensemble des valeurs dans lequel l'hypothèse H_0 est acceptable peut être considéré comme un intervalle de confiance constituant un lien entre estimation par intervalle et test d'hypothèse.

2. METHODOLOGIE

On se restreint dans ce cours

- au test bilatéral: l'hypothèse à tester est susceptible d'être fausse dans 2 directions, par surévaluation ou sous-évaluation du paramètre inconnu théorique θ de la population. Par opposition, le test unilatéral étudie soit la surévaluation soit la sous-évaluation du paramètre θ.
- aux tests concernant la comparaison (l'égalité) de 2 moyennes.
- la minimisation de β déterminant la puissance du test n'est pas étudiée (cf. livre).

Tout test d'hypothèse comporte les 7 étapes suivantes

1. Données: Définir la loi de probabilités de la population.
2. Test: Formulation de l'hypothèse H_0.
3. Sous H_0 (H_0 vraie): Choix de la variable aléatoire et de sa loi de probabilités.
4. Région de rejet R par rapport à la valeur critique c.
5. Calcul de la valeur critique c ou seuil α d'après la loi de probabilités définie en 3.
6. Calcul de la variable aléatoire définie en 3 avec les données de l'échantillon.
7. Décision.

3. COMPARAISON DE DEUX MOYENNES

3.1. Comparaison de deux moyennes de variances connues avec des échantillons de tailles n et n' quelconques

1. Données

$X_i \sim N(\mu,\sigma), i = 1,\dots,n$

$X'_j \sim N(\mu',\sigma'), j = 1,\dots,n'$

2. Test

$H_0 : \mu = \mu'$

$H_1 : \mu \neq \mu'$

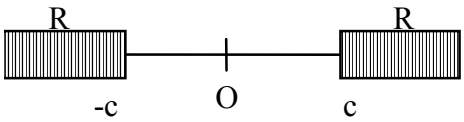
3. Sous H_0

$$Y = \frac{\bar{X} - \bar{X}'}{\sqrt{\frac{\sigma^2}{n} + \frac{\sigma'^2}{n'}}} \sim N(0,1)$$

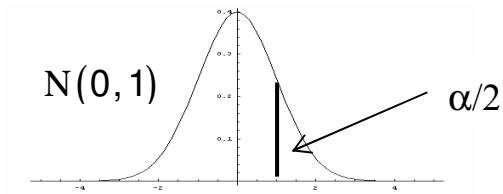
Rq: Si $\sigma = \sigma'$

$$Y = \frac{\bar{X} - \bar{X}'}{\sigma\sqrt{\frac{1}{n} + \frac{1}{n'}}}$$

4. Région de rejet R



5. Calcul de c



6. Calcul de Y

7. Décision

On accepte H_0 si $-c \leq Y \leq c$

On rejette H_0 autrement

3.2. Comparaison de deux moyennes de même variance inconnue avec au moins un échantillon de taille n<30

1. Données

$X_i \sim N(\mu,\sigma), i = 1,\ldots,n$

$X'_j \sim N(\mu',\sigma), j = 1,\ldots,n'$

Le test n'est théoriquement applicable que si les 2 échantillons ont la même variance.

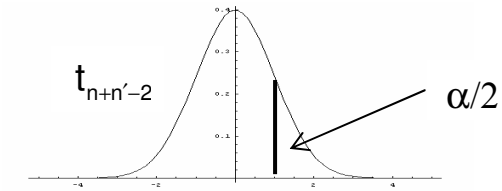
2. Test: cf. Test 3.1

3. Sous H_0

$$Y = \frac{\bar{X} - \bar{X}'}{\sqrt{\left(\frac{1}{n} + \frac{1}{n'}\right)\frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2 + \sum_{j=1}^{n'}\left(X'_j - \bar{X}'\right)}{n + n' - 2}}} \sim t_{n+n'-2}$$

4. Région de rejet R: cf. Test 3.1

5. Calcul de c



6 Calcul de Y

7. Décision: cf. Test 3.1

3.3. Comparaison de deux moyennes de même variance inconnue avec des échantillons de tailles n≥30 et n'≥30

1. Données: cf. Test 3.2

2. Test: cf. Test 3.1

3. Sous H_0

$$Y = \frac{\bar{X} - \bar{X}'}{\sqrt{\frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}{n(n-1)} + \frac{\sum_{j=1}^{n'}\left(X'_j - \bar{X}'\right)^2}{n'(n'-1)}}} \sim N(0,1)$$

4. Région de rejet R: cf. Test 3.1

5. Calcul de c: cf. Test 3.1

6. Calcul de Y

7. Décision: cf. Test 3.1

3.4. Comparaison de deux moyennes avec des échantillons appariés

1. Données

Soient X_i' et X_i'' les données des 2 échantillons appariés de taille n. On pose

$$X_i = X_i' - X_i'' \sim N(\mu, \sigma), i = 1, \ldots, n$$

2. Test

$H_0 : \mu = 0$

$H_1 : \mu \neq 0$

3. Sous H_0

(i) σ connu et $\forall n$

$$Y = \frac{\bar{X}}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$$

Remarquer l'analogie entre cette formule Y (test d'hypothèse) et la formule $\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$ (§ 3.2 de CS2 concernant l'estimation par intervalle ou Th3) où sous H_0 $\mu = 0$.

(ii) σ inconnu et $n < 30$

$$Y = \frac{\bar{X}}{\sqrt{\frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n(n-1)}}} \sim t_{n-1}$$

(iii) σ inconnu et $n \geq 30$

$$Y = \frac{\bar{X}}{\sqrt{\frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n(n-1)}}} \sim N(0,1)$$

4. Région de rejet R: cf. Test 3.1

5. Calcul de c: cf. Test 3.1 pour (i) et (iii) et Test 3.2 pour (ii) avec $n-1$ degrés de libertés.

6. Calcul de Y

7. Décision: cf. Test 3.1

3.5. Comparaison de deux moyennes avec un test unilatéral

Les tests bilatéraux peuvent être transformés en tests unilatéraux en posant

$H_0 : \mu \leq \mu'$

$H_1 : \mu > \mu'$

La procédure dans les tests unilatéraux est la même que celle des tests bilatéraux, en particulier le choix de la v.a. et de sa loi de probabilités.

Les 2 seules différences concernent la région de rejet qui est unilatérale dans les tests unilatéraux (bilatérale dans les tests bilatéraux) et le calcul de c qui est associé à α dans les tests unilatéraux ($\alpha/2$ dans les tests bilatéraux).

CS4: TEST DU KHI-DEUX

1. INTRODUCTION

Le test du Khi-Deux χ^2 est un test permettant de comparer une distribution théorique à une distribution observée (expérimentale). En pratique, ce test peut également être utilisé pour comparer 2 distributions observées.

2. PRINCIPE

Soit une population contenant k classes de boules de couleur différente en probabilités $p_1, \ldots, p_k$ avec $\sum_{i=1}^{k} p_i = 1$. La composition de cette population est parfaitement définie par $k-1$ de ses probabilités, $k-1$ représente le nombre de degrés de liberté. Soit un échantillon de n boules extrait de cette population. Si la composition de cet échantillon est la même que celle de la population, elle contiendrait $np_1 = T_1, \ldots, np_k = T_k$ boules des k couleurs. T_i est l'effectif théorique. En fait, on observe en réalité les effectifs observés $O_1, \ldots, O_k$ qui diffèrent plus ou moins des effectifs théoriques. Il existe donc un écart à mesurer entre la composition de la population et celle de l'échantillon.

(i) Pour $k = 2$

Cet écart peut être mesuré en faisant la différence entre les pourcentages observé et théorique pour l'une des classes, les pourcentages observé et théorique de l'autre classe étant complémentaire à 100.

(i) Pour $k > 2$

Cet écart ne peut pas être mesuré par la somme ou la moyenne qui est évidemment nulle: $\sum_{i=1}^{k} (O_i - T_i) = 0$. La somme des valeurs absolues des écarts est peu commode en calcul des probabilités. La somme des carrés des écarts évite les inconvénients précédents mais est une mesure encore imparfaite car elle donne le même poids à tous les écarts qu'ils se rapportent à des effectifs de petite ou de grande taille. Des

considérations théoriques ont conduit à la mesure suivante dont la loi de probabilités ne dépend pas de la loi de la population (c'est la définition d'un test non paramétrique) mais uniquement du nombre de classes

$$\sum_{i=1}^{k} \frac{(O_i - T_i)^2}{T_i} \sim \chi^2_{k-1} \quad \text{quand} \quad n \to \infty$$

O_i : effectif observé de la classe i

T_i : effectif théorique de la classe i

n: taille de l'échantillon avec $n = \sum_{i=1}^{k} O_i = \sum_{i=1}^{k} T_i$

Rem

Ce test s'applique à des effectifs non à des pourcentages.

3. TEST DE COMPARAISON DE DEUX DISTRIBUTIONS

1. Données

- Pas de condition sur la loi de probabilités de la population (test non paramétrique)
- Echantillon de taille n à k classes
- O_i : effectif observé de la classe i
- T_i : effectif théorique de la classe i
- Condition d'application du test: $T_i \geq 5 \quad \forall i$ (satisfaisant avec $T_i \geq 1.5 \quad \forall i$). Si cette condition d'application du test n'est pas vérifiée pour tout i, on peut regrouper certaines classes.

2. Test

$H_0 : O_i = T_i, i = 1, \ldots, k$, c'est-à-dire la distribution observée est identique à la distribution théorique

$H_1 : O_i \neq T_i$ pour au moins une valeur de i, c'est-à-dire la distribution observée diffère de la distribution théorique.

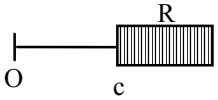
3. Sous H_0

$$Y = \sum_{i=1}^{k} \frac{(O_i - T_i)^2}{T_i} = \sum_{i=1}^{k} \frac{O_i^2}{T_i} - n \sim \chi^2_{k-1}$$

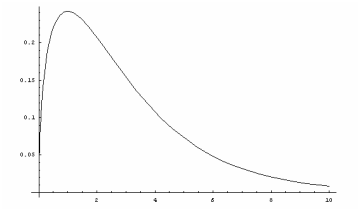
quand $n \to \infty$.

En pratique, on utilise la condition d'application.

4. Région de rejet R



5. Calcul de c



6. Calcul de Y

7. Décision

On accepte H_0 si $Y \leq c$

On rejette H_0 autrement

4. TEST DE L'INDEPENDANCE DE DEUX VARIABLES QUALITATIVES

1. Données

- Pas de condition sur la loi de probabilités de la population (test non paramétrique)
- 2 variables qualitatives (par ex: couleur des yeux et personnalité d'un individu)
- Echantillon associé à un tableau à l lignes et c colonnes
- $O_{i,j}$: effectif observé en ligne i et colonne $j, 1 \le i \le l, 1 \le j \le c$
- $T_{i,j}$ effectif théorique en ligne i et colonne $j, 1 \le i \le l, 1 \le j \le c$
- Condition d'application du test: $T_{i,j} \ge 5, \forall i,j$.

2. Test

H_0: les 2 variables sont indépendantes

H_1: les 2 variables sont dépendantes (liés)

3. Sous H_0

$$Y = \sum_{i=1}^{l} \sum_{j=1}^{c} \frac{\left(O_{i,j} - T_{i,j}\right)^2}{T_{i,j}} \sim \chi^2_{(l-1)(c-1)}$$

avec $T_{i,j} = \frac{1}{\sum_{i=1}^{l} \sum_{j=1}^{c} O_{i,j}} \sum_{i=1}^{l} O_{i,j} \sum_{j=1}^{c} O_{i,j}$

4. Région de rejet R: cf. Test 3

5. Calcul de c: cf. Test 3

6. Calcul de Y

7. Décision: cf. Test 3

CS5: TEST DE WILCOXON

1. INTRODUCTION

On se limitera au test de rang signé de Wilcoxon pour 2 échantillons appariés. Ce test, en ne présupposant aucune loi de probabilités pour la population (test non paramétrique), permet de comparer la moyenne (ou la loi) de 2 échantillons appariés en comparant le rang moyen (ce qui est équivalent à la somme des rangs) des observations d'un échantillon avec le rang moyen des observations de l'autre échantillon.

2. TEST

1. Données et procédure

Soient X_i' et X_i'' les données (observations) des 2 échantillons appariés de taille n.

(i) On calcule $|d_i| = |X_i' - X_i''|$. On suppose que $d_i \neq 0, \forall i$.

Rem: il existe des tests de Wilcoxon qui tiennent compte de l'existence de $d_i = 0$.

(ii) On classe $|d_i|$ par ordre croissant des valeurs.

(iii) On assigne un rang de 1 à n sur les $|d_i|$ classés. On suppose que tous les $|d_i|$ sont différents.

Rem: il existe des tests de Wilcoxon qui tiennent compte de l'existence de $|d_i|$ identiques.

(iv) On associe à leur rang le signe de $X_i' - X_i''$.

(v) Soit T^+ (resp. T^-) la somme des rangs positifs (resp. négatifs).

Il existe une relation entre T^+ et T^-

$$T^+ + T^- = \sum_{i=1}^{n} i = \frac{n(n+1)}{2}$$

Il suffit donc de considérer l'une des 2 sommes. La statistique de Wilcoxon est basée sur T^+.

2. Test

$H_0 : T^+ = T^-$, c'est à dire les 2 échantillons ont même loi

$H_1 : T^+ \neq T^-$, c'est à dire les 2 échantillons ont des lois différentes

3. Sous H_0

(i) $n < 30$

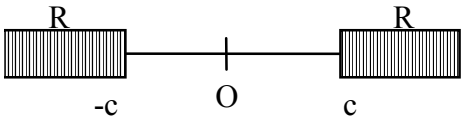
$T^+ \sim$ loi de distribution nulle de Wilcoxon

Rem: cf. § 3

(ii) $n \geq 30$

$$Y = \frac{T^+ - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}} \sim N(0,1)$$

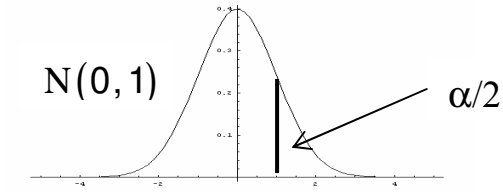
4. Région de rejet R



5. Calcul de c

(i) $n < 30$: Table de Wilcoxon

(ii) $n \geq 30$



6. Calcul

(i) $n < 30$: Calcul de T^+

(ii) $n \geq 30$: Calcul de Y

7. Décision

(i) $n < 30$

On accepte H_0 si $-c \leq T^+ \leq c$

On rejette H_0 autrement

(ii) $n \geq 30$

On accepte H_0 si $-c \leq Y \leq c$

On rejette H_0 autrement.

3. LOI DE DISTRIBUTION NULLE DE WILCOXON

3.1. Définition

$$P\left(T^{+}=x\right)=\frac{N(n,x)}{2^{n}}$$

avec $x=0,1,2,\ldots,\frac{n(n+1)}{2}$

$N(n,x=0)=1$

$N(n,x\neq 0)=$ nombre de sous-ensembles de $\{1,\ldots,n\}$ tels que la somme de leurs éléments soit égale à x.

Cette loi de Wilcoxon est donnée dans des tables.

3.2. Espérance et variance

Soit W_i la v.a. de Bernoulli égale à 1 si le rang i est positif, à 0 si le rang i est négatif. Sachant que pour un rang i il y a autant de chance qu'il soit positif que négatif, alors la loi de probabilités de W_i

$$P(W_i=0)=P(W_i=1)=\frac{1}{2}$$

Espérance d'une v.a. de Bernoulli: p

$$E(W_i)=\frac{1}{2}$$

Variance d'une v.a. de Bernoulli: $p(1-p)$

$$V(W_i)=\frac{1}{4}$$

La statistique T^{+} peut être mis en fonction de W_i

$$T^{+}=\sum_{i=1}^{n} iW_i$$

Espérance de T^+

$$E\left(T^+\right) = \frac{n(n+1)}{4}$$

Variance de T^+

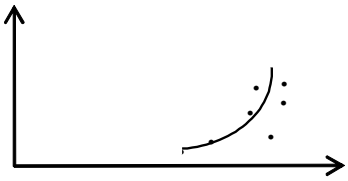
$$V\left(T^+\right) = \frac{n(n+1)(2n+1)}{24}$$

CS6: AJUSTEMENT D'UNE COURBE

1. INTRODUCTION

Soient 2 variables X et Y prenant les couples de valeurs $(x_1,y_1),...,(x_n,y_n)$. Cette approche générale peut être appliquée à des variables aléatoires X et Y de valeurs $(X_1,Y_1),...,(X_n,Y_n)$. L'ensemble des points ainsi obtenu forme un nuage de points. Un problème fréquent consiste à rechercher une courbe continue approchant au mieux ce nuage de points. Une telle courbe est appelée courbe d'ajustement.

Courbe d'ajustement linéaire



Courbe d'ajustement non linéaire

2. EQUATION DES COURBES D'AJUSTEMENT

Equation	Fonction
$Y = a_0 + a_1X$	Linéaire (droite)
$Y = a_0 + a_1X + a_2X^2$	Quadratique (parabole)
$Y = a_0 + a_1X + a_2X^2 + a_3X^3$	Cubique
$Y = a_0 + a_1X + a_2X^2 + a_3X^3 + a_4X^4$	Du 4ème degré
$Y = a_0 + a_1X + ... + a_nX^n$	Du nième degré
$Y = \frac{1}{a_0 + a_1X}$ ou $\frac{1}{Y} = a_0 + a_1X$	Hyperbole
$Y = ab^X$ ou $\log Y = \log a + (\log b)X = a_0 + a_1X$	Exponentielle
$Y = aX^b$ ou $\log Y = \log a + b\log X$	Puissance
$Y = ab^X + g$	Exponentielle modifiée
$Y = aX^b + g$	Puissance modifiée
$Y = pq^{b^X}$ ou $\log Y = \log p + b^X \log q = ab^X + g$	Fonction de Gompertz
$Y = pq^{b^X} + h$	Fonction de Gompertz modifiée
$Y = \frac{1}{ab^X + g}$ ou $\frac{1}{Y} = ab^X + g$	Fonction logistique

3. DROITE DE 2 POINTS QUELCONQUES

La droite de 2 points quelconques (x_1,y_1), (x_2,y_2) a pour équation $Y = a_0 + a_1 X$.

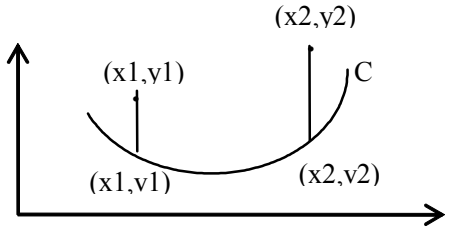
$$\begin{cases} y_1 = a_0 + a_1 x_1 \\ y_2 = a_0 + a_1 x_2 \end{cases} \Leftrightarrow \begin{cases} a_0 = \dfrac{y_1 x_2 - y_2 x_1}{x_2 - x_1} \\ a_1 = \dfrac{y_2 - y_1}{x_2 - x_1} \end{cases}$$

La constante a_0 est l'ordonnée à l'origine. La constante a_1 est la pente.

4. COURBE DES MOINDRES CARRES (COURBE D'AJUSTEMENT)

4.1. Introduction

La courbe (droite, parabole, etc.) des moindres carrés, notée C, permet un ajustement d'un nuage de points $(x_1,y_1),...,(x_n,y_n)$. On considère les projections verticales des points $(x_1,y_1),...,(x_n,y_n)$ sur la courbe C, c'est-à-dire les points $(x_1,v_1),...,(x_n,v_n)$.



Rem
Les projections horizontales ou orthogonales sont rarement utilisées.

L'ajustement sera d'autant plus efficace que les différences (positives ou négatives) $v_i - y_i$, $i \in [1,n]$, sont les plus faibles possibles, c'est-à-dire si

$$S = \sum_{i=1}^{n} (v_i - y_i)^2 \text{ est minimal}$$

4.2. Droite des moindres carrés

La droite des moindres carrés ajustant le nuage de n points $(x_1,y_1),...,(x_n,y_n)$ a pour équation $Y = a_0 + a_1 X$.

Equations normales

$$\begin{cases} \sum_i y_i = a_0 n + a_1 \sum_i x_i \\ \sum_i x_i y_i = a_0 \sum_i x_i + a_1 \sum_i x_i^2 \end{cases}$$

A partir des équations normales de la droite des moindres carrés, on calcule les constantes a_0 et a_1

$$a_0 = \frac{\sum_i x_i^2 \sum_i y_i - \sum_i x_i \sum_i x_i y_i}{n\sum_i x_i^2 - \left(\sum_i x_i\right)^2}$$

$$a_1 = \frac{n\sum_i x_i y_i - \sum_i x_i \sum_i y_i}{n\sum_i x_i^2 - \left(\sum_i x_i\right)^2}$$

Ex: Calcul de la droite des moindres carrés pour 2 points (x_1,y_1) et (x_2,y_2)

$$a_0 = \frac{x_1 y_2 - x_2 y_1}{x_1 - x_2}$$

$$a_1 = \frac{y_1 - y_2}{x_1 - x_2}$$

On retrouve les formules du § 3.

P1

La droite des moindres carrés passe par le centre de gravité $(\overline{x},\overline{y})$ du nuage des points avec $\overline{x} = \frac{1}{n}\sum_i x_i$

Rem

De la même façon qu'il existe une droite des moindres carrés ajustant le nuage de points $(x_1,y_1),...,(x_n,y_n)$, il existe un plan des moindres carrés ajustant le nuage de points $(x_1,y_1,z_1),...,(x_n,y_n,z_n)$, c'est-à-dire avec une équation linéaire de 3 variables de la forme: $Z = a_0 + a_1X + a_2Y$.

On peut généraliser à des espaces de dimension supérieure, on parlera alors d'hyperplan.

Equations normales

$$\begin{cases} \sum_i z_i = a_0 n + a_1 \sum_i x_i + a_2 \sum_i y_i \\ \sum_i x_i z_i = a_0 \sum_i x_i + a_1 \sum_i x_i^2 + a_2 \sum x_i y_i \\ \sum_i y_i z_i = a_0 \sum_i y_i + a_1 \sum_i x_i y_i + a_2 \sum y_i^2 \end{cases}$$

4.3. Parabole des moindres carrés

La parabole des moindres carrés ajustant le nuage de points $(x_1,y_1),...,(x_n,y_n)$ a pour équation $Y = a_0 + a_1X + a_2X^2$.

Equations normales

$$\begin{cases} \sum_i y_i = a_0 n + a_1 \sum_i x_i + a_2 \sum_i x_i^2 \\ \sum_i x_i y_i = a_0 \sum_i x_i + a_1 \sum_i x_i^2 + a_2 \sum_i x_i^3 \\ \sum_i x_i^2 y_i = a_0 \sum_i x_i^2 + a_1 \sum_i x_i^3 + a_2 \sum_i x_i^4 \end{cases}$$

Rem1

Cette technique peut être étendue aux équations normales de courbes des moindres carrés du 3ème, ..., nième degré.

Rem2

Avec des équations non linéaires de 3 variables, on parle de surface des moindres carrés et non de plan des moindres carrés. Avec des équations non linéaires de plus de 3 variables, on parle d'hypersurface des moindres carrés.

Rem3

Il est important de transformer les données par changement de variable pour simplifier les calculs. La transformation la plus intéressante est de changer les x_i de façon à ce que $\sum_i x_i = 0$ et de les affecter avec des valeurs simples. On utilisera alors les équations normales pour calculer les constantes.