Tests statistiques
Notes de cours

V. Monbet

L2 S1 - 2009

Table des matières

Chapitre 1

Introduction

1.1 Qu'est ce que la statistique ?

Les statistiques, dans le sens populaire du terme, traitent des populations. Leur objectif consiste à caractériser une population à partir d'une image plus ou moins floue constituée à l'aide d'un échantillon issu de cette population. On peut alors chercher à **extrapoler** une information obtenue à partir de l'échantillon.

Exemple - Répartition par classe d'age d'une population de poissons. Si on veut caractériser la *population* de morue dans une zone donnée de l'Atlantique Nord, on va prélever quelques poissons (ces quelques poissons vont constituer l'*échantillon*). Puis on va mesurer leur age (otolithe), leur poids, leur taille, ... on va enfin chercher à extrapoler ces résultats à toute la population.

Mais on peut aussi chercher à **synthétiser** une information trop dense.

Exemple - Acheteurs potentiels (prospects) d'un certain forfait de téléphone portable. On va chercher les principales caractéristiques spécifiques du groupe des clients du forfait afin de mieux les connaître et d'être capable d'identifier des prospects.

Ou encore à **vérifier** une hypothèse.

Exemple - Contrôle de qualité. Le fabriquant de café fournit des paquets de 250 g. Le remplissage est automatisé. Régulièrement le fabriquant prélève quelques paquets de café ce qui constitue l'échantillon. Il pèse les paquets de l'échantillon afin de vérifier l'hypothèse selon laquelle les paquets de café pèsent bien 250 g en moyenne.

Exemple - Effet d'un traitement. Dans l'industrie pharmaceutique, il est obligatoire de tester l'efficacité d'un traitement avant de le mettre sur le marché. On procède alors de la façon suivante : on sélectionne deux groupes de patients. L'un reçoit le médicament, l'autre un placébo. Il faut alors vérifier que le groupe qui reçoit le médicament voit bien ses symptômes diminuer en moyenne.

On trouve des applications de la statistique dans tous les domaines : industrie, environnement, médecine, finance, marketing, sport, ...

Dans le cadre de ce cours, nous allons nous intéresser principalement aux **tests statistiques**.

1.2 Qu'est ce qu'un test statistique ?

Un test, qu'il soit statistique ou pas, consiste à vérifier une information *hypothétique*. On parle d'ailleurs de *tests d'hypothèses*.

En statistique mathématique, l'information hypothétique concerne la population à laquelle on s'intéresse. C'est une information statistique qui peut être :

- Une distribution qu'une variable d'intérêt quelconque est censée présenter. Exemple : répartition de l'age des poissons.
- Une valeur ponctuelle à laquelle une statistique, par exemple une moyenne, une médiane, une fréquence, etc. serait égale. Exemple : poids des paquets de café.
- Un intervalle de valeurs auquel appartiendrait la valeur d'une statistique, comme ci-dessus (on qualifie un tel intervalle d'hypothèse composite).
- L'indépendance statistique de deux variables.

Un test statistique peut aussi être utilisé pour vérifier le succès (ou l'échec) d'une action entreprise pour modifier la valeur d'une statistique de population. Par exemple,

- On cherche à augmenter le nombre moyen des clients servis à l'heure, qui est actuellement de 10.
- On cherche à faire tomber la proportion des appareils défectueux en dessous de 3%.

Il est généralement impossible de recenser toute la population. On prélève alors un échantillon dont on déduit une statistique (par exemple la moyenne de l'échantillon). Cette statistique est comparée à la valeur à laquelle on peut s'attendre si l'hypothèse est vraie. Cependant, on doit tenir compte du fait qu'on a observé seulement un échantillon de la population. L'observation d'un autre échantillon conduira vraisemblablement à une autre valeur de la statistique. La théorie des tests procure des outils pour bien prendre en compte cette variabilité.

1.3 Exemple

Traitons un exemple [?]. J'ai 114 livres dans ma bibliothèque. J'en extrais un échantillon de 12. Chaque livre doit avoir la même probabilité d'être choisi. Je veux tester l'hypothèse que la médiane du nombre de pages par volume est 220.

Dans mon échantillon, j'observe les nombres de pages suivants :

126	142	156	228	245	246	370	419	433	454	478	503

Je leur associe un signe - si le nombre de pages est inférieur à 220 et un signe + sinon. Si la médiane est 220, il est également probable pour chaque livre sélectionné d'avoir plus ou moins de 220 pages.

En associant un + à un *face* et un - à un *pile*, nous pouvons faire une analogie avec un lancer de pièce. Nous verrons plus loin que le tirage "9 faces et 3 piles" a une probabilité assez forte pour qu'on ne puisse pas rejeter l'hypothèse selon laquelle le nombre de pages médian de mes livres est 220.

Si nous avions observé 12 signes + et pas de signe - (ou inversement 12 signes - et pas de signe +) nous aurions pu raisonnablement rejeter l'hypothèse que la médiane est 220. En effet, on peut vérifier que la probabilité d'obtenir un de ces 2 résultats est seulement de $\frac{1}{2048}$, de sorte qu'un tel résultat dans une expérience signifierait soit que nous avons observé un évènement fortement improbable soit que notre hypothèse d'une pièce équilibrée est incorrecte.

Exercice : Calculer la probabilité d'observer 3 piles parmi 12 lancés d'une pièce équilibrée. On note que si X suit une loi binomiale $B(n,p)$, on a $P(X=k) = C_n^k p^k (1-p)^{n-k}$.

1.4 Rappels de probabilité

Soit X une variable aléatoire. Dans la première partie du cours, nous utiliserons essentiellement des lois discrètes.

1.4.1 Loi de Bernouilli

La loi de Bernouilli modélise un tirage à pile ou face.
- Notation : X suit la loi $B(1,p)$
- Univers : $X(\Omega) = \{0,1\}$
- Loi : $P(X=0) = p$
- Espérance et variance : $E(X) = p$, $Var(X) = p(1-p)$

1.4.2 Loi binomiale

La loi binomiale modélise un tirage avec remise parmi un ensemble de n objets de deux types (ex : boules blanches et noires) avec une probabilité de succès p à chaque tirage (ex : succès = tirer une boule blanche).
- Notation : X suit la loi $B(n,p)$
- Univers : $X(\Omega) = \{0,1,\cdots,n\}$
- Loi : $P(X=k) = C_n^k p^k (1-p)^{n-k}$
- Espérance et variance : $E(X) = np$, $Var(X) = np(1-p)$

Exercice Calculer la probabilité que parmi les 12 livres, 9 d'entre eux aient plus de plus de 220 pages, sous l'hypothèse que la médiane du nombre de pages des livres de ma bibliothèque est égale à 220.

r	0	1	2	3	4	5	6
P	0.000	0.003	0.016	0.054	0.121	0.193	0.226
		7	8	9	10	11	12
		0.193	0.121	0.054	0.016	0.003	0.000

TABLE 1.1 – Probabilités binomiales P, pour r signes +, $n = 12$, $p = \frac{1}{2}$

Chapitre 2

Tests d'hypothèses, généralités

Dans ce chapitre nous énonçons (ou rappelons) un certain nombre de généralités autour des tests d'hypothèse, l'objectif étant d'être capable de bien formuler un test.

2.1 Hypothèses de test

En premier lieu, nous devons formuler les hypothèses. L'hypothèse que nous voulons vérifier sera appelée *hypothèse nulle* et on la notera H_0. Dans l'exemple concernant le nombre de pages des livres de ma bibliothèque, nous poserons alors

$$H_0 : \theta = 220$$

où θ représente ici la médiane du nombre de page. Nous rassemblerons d'autre part l'ensemble des *hypothèses alternatives* sous H_1 :

$$H_1 : \theta \neq 220$$

Et nous parlerons de tester H_0 contre les alternatives bilatérales H_1 (sous H_1, θ peut être inférieur ou supérieur à 220).

2.2 Statistique de test

Une fois les hypothèses de test posées, nous devons choisir la statistique de test. C'est en comparant la valeur de cette statistique observée dans l'échantillon à la sa valeur sous l'hypothèse H_0 que nous pourrons prendre une décision (ie donner la conclusion du test).

Dans l'exemple de nombre de pages des livres tel que nous l'avons traité jusqu'à présent, la statistique de test est par exemple le nombre de signes + observé. On a alors que la loi de probabilité de la statistique de test sous H_0 est ici une loi binomiale $B(12, 1/2)$. Nous aurions pu choisir de manière équivalente le nombre de signes -.

D'après la table de la loi binomiale, nous constatons que si H_0 est vraie, la probabilité est maximale pour 6 signes +.

2.3 Région de rejet et niveau de signification

En suivant une procédure formelle en test d'hypothèse, nous séparons les résultats possibles en deux sous-ensembles. Le premier regroupe les résultats les plus vraisemblables sous l'hypothèse nulle, de façon que la somme de leurs probabilités soit au moins égale à l'une des valeurs conventionnelles 0.90, 0.95 (valeur la plus souvent choisie), 0.99 ou 0.999.

On peut vérifier facilement dans le tableau 1.1 que la probabilité de l'ensemble allant de 3 à 9 signes + est 0.962. Et on ne peut éliminer de l'ensemble aucun de ces résultats sans réduire la probabilité à une valeur inférieure à 0.95. On remarque que dans ce cas symétrique, on doit éliminer les résultats par paire.

Les résultats restants c'est à dire $\{0, 1, 2, 10, 11, 12\}$ forment un ensemble de probabilité 0.038 appelée **région de rejet** (ou région critique) **de niveau de signification nominal** α ou encore de **de niveau de signification réel** (ou degré de signification [1]) 0.038.

La règle des tests d'hypothèse consiste à rejeter H_0 au niveau de signification 0.05 si et seulement si le résultat tombe dans la région de rejet.

La région complémentaire de tous les résultats hors de la région de rejet est appelée **région de non rejet** (ou **d'acceptation**) de l'hypothèse nulle.

En choisissant une région de rejet de probabilité inférieure au égale au niveau de signification on adopte une attitude dite **conservatrice**.

2.4 Les deux espèces d'erreur

Lorsque l'on fait un test d'hypothèse, deux sortes d'erreur sont possibles. On peut rejeter l'hypothèse nulle alors qu'elle est vraie. Ceci se produit si la valeur de la statistique de test tombe dans la région de rejet alors que l'hypothèse H_0 est vraie.

La probabilité de cet évènement est le niveau de signification. On dira aussi que le niveau de signification est la probabilité de rejeter l'hypothèse nulle à tort.

Rejeter l'hypothèse nulle à tort constitue une **erreur de première espèce**.

Si nous ne rejetons pas l'hypothèse nulle alors qu'elle est fausse nous commettons une **erreur de seconde espèce**. C'est le cas si la valeur de la statistique de test tombe dans la région de non rejet (ou d'acceptation) alors que H_0 est fausse (c'est à dire si H_1 est vraie).

Lorsque l'alternative H_1 est de la forme $\theta \neq \theta_0$, notre θ peut prendre une infinité de valeurs ; et la probabilité de rejeter H_0 lorsqu'elle est fausse dépend beaucoup de la vraie valeur de θ (qui est inconnue !).

1. En anglais : p-value

Par exemple, en lançant une pièce de monnaie 12 fois, on a plus de chances d'obtenir 10, 11 ou 12 faces si la probabilité de face est $p = 0.99$ que si $p = 0.55$. Or dans les deux cas, H_0 est fausse.

Lorsque la vraie valeur de θ est dans H_1, la probabilité d'obtenir un résultat dans la région de rejet est appelée **puissance** du test de H_0 contre H_1. La puissance d'un test dépend de plusieurs facteurs :

- le niveau de signification du test
- la vraie valeur du paramètre testé
- la taille de l'échantillon
- la nature du test utilisé

De manière générale, plus on tient compte d'informations pertinentes dans un test plus sa puissance est élevée.

2.5 Test unilatéral ou bilatéral

Dans l'exemple du nombre de pages dans les livres de la bibliothèque, nous avons posé des hypthèses de tests telles que l'alternative est bilatérale. C'est à dire que si l'on rejette l'hypothèse nulle, la médiane du nombre de pages peut-être supérieure ou inférieure à 220.

Dans certains problèmes, il est plus pertinent de considérer une hypothèse alternative unilatérale. On pose alors

$$H_0 : \theta \leq \theta_0 \text{ contre } H_1 : \theta > \theta_0$$

ou

$$H_0 : \theta \geq \theta_0 \text{ contre } H_1 : \theta < \theta_0$$

La définition de la région de rejet du test dépend de la forme de l'hypothèse alternative (voir TD 1, ex. 2).

Le choix d'un test unilatéral ou bilatéral dépend de la logique de la situation expérimentale et doit être fait avant d'inspecter les données.

2.6 Estimation

2.6.1 Intervalle de confiance

Bien que ce problème soit souvent formulé différemment, une des façons de spécifier un intervalle de confiance à $100(1-\alpha)$ pour un paramètre de position θ consiste à le définir comme l'ensemble de toutes les valeurs qui seraient acceptées par un test de niveau α.

2.6.2 Intervalle de tolérance

L'intervalle de tolérance est un autre concept utile. Il s'agit d'intervalles ayant la propriété suivante : étant donnés p_1 et α, l'intervalle contient $p_1\%$ de la population avec une probabilité α. ...

Chapitre 3

Tests non paramétriques - Estimation de la position pour un échantillon isolé

Dans ce chapitre, nous allons décrire plusieurs tests pour la position d'un échantillon isolé. La position d'un échantillon peut être caractérisée par différents paramètres. Les plus usuels sont la moyenne et la médiane.

3.1 Le test du signe

Dans le chapitre précédent, nous avons déjà introduit le test du signe à titre d'exemple et nous ne reviendrons pas ici sur sa théorie. Mais ajoutons ici quelques remarques.

3.1.1 Quelques remarques

Il peut arriver que, dans un échantillon, une ou plusieurs observations soient exactement égales à la valeur θ_0 du paramètre θ sous H_0. Dans ce cas, il est recommandé d'ignorer ces observations.

Les tables habituellement utilisées pour construire la région de rejet sont les tables des probabilités binomiales cumulées qui correspondent aux probabilités d'observer au plus r succès (c'est à dire r signes plus). Nous voyons par exemple, dans le tableau de la loi $B(16, \frac{1}{2})$, que pour un test bilatéral de niveau nominal 5% (la partie inférieure de la région de rejet doit représenter une probabilité d'au plus 0.025), le plus petit des deux nombres de signes "plus" et "moins" ne doit pas dépasser 3. Pour un test unilatéral de

$$H_0 : \theta \geq \theta_0 \text{ contre } H_1 : \theta < \theta_0$$

au niveau nominal 5%, la région de rejet contient les valeurs de 0 à 4.

Exercice - Quel est le niveau de signification réel du test du signe unilatéral

$$H_0 : \theta \geq \theta_0 \text{ contre } H_1 : \theta < \theta_0$$

dans le cas où l'on a 24 observations et que le niveau nominal est 5% ?

Exercice - Quel est la region de rejet du test du signe unilatéral

$$H_0 : \theta \geq \theta_0 \text{ contre } H_1 : \theta < \theta_0$$

dans le cas où l'on a 24 observations et que le niveau de signification nominal est fixé à 1% ?

On remarque que la table des probabilités cumulées de la loi binomiale ne donne des valeurs que pour $n \leq 20$. Plus loin, nous verrons que pour des échantillons plus grands, nous utilisons des approximations.

3.1.2 Intervalle de confiance

Vu en travaux dirigés.

3.1.3 Approximation pour les grands échantillons

Si $n > 20$, une approximation basée sur la loi Gauss est généralement satisfaisante. Lorsque n est assez grand et p pas trop petit (typiquement $np > 10$), si X suit la loi binomiale $B(n, p)$, alors la variable

$$Z = \frac{X - np}{\sqrt{np(1-p)}}$$

suit une loi de Gauss de moyenne égale à zéro et de variance égale à un. Dans le cas du test du signe, $p = \frac{1}{2}$, et on utilise alors

$$Z = \frac{X - n/2}{\sqrt{n}/2}$$

3.1.4 Test du signe modifié : test d'un quantile

On peut adapter le test du signe pour tester des hypothèses sur un quantile d'une distribution.

On définit le k-ième quantile de la distribution continue de la variable aléatoire X comme la valeur q_k telle que

$$P(X < q_k) \leq k \text{ et } P(X > q_k) \leq 1 - k$$

On remarque que $q_{1/2}$ est la médiane.

Cas particuliers : si $k = r/10$ avec $r \in \{1, 2, \cdots, 9\}$ alors q_k est appelé **décile** et si $k = r/4$ avec $r \in \{1, 2, 3\}$ alors q_k est appelé **quartile**.

Test du signe modifié : voir exercices.

3.2 Inférence à base de rangs

Le test du signe utilise seulement une petite partie de l'information contenue dans un jeu de données comme les nombres de pages de l'exemple des livres : pour chaque observation nous avons noté si elle était supérieure ou inférieure à la médiane spécifiée dans H_0.

Si maintenant, nous postulons de plus que la distribution de la population est symétrique, le centre de symétrie est alors la médiane de la population (ou sa moyenne puisque dans ce cas elles coïncident) et nous pouvons mieux tenir compte des valeurs des observations pour nos déicisions (inférences) statistiques.

3.2.1 Test des signes et rangs de Wilcoxon

Hypothèses : Nous supposons que la distribution de la variable dans la population est *symétrique* et *continue*.

Définition : Soient X une variable aléatoire de ***distribution symétrique*** *et μ le centre de symétrie, on a*

$$P(X \leq \mu - x) = P(X \geq \mu + x)$$

Etant donné un échantillon de n mesures indépendantes, nous pouvons au lieu de noter seulement les signes des écarts à la médiane spécifiée dans H_0, relever aussi la grandeur de chaque écart. Si H_0 est vraie, les écarts d'une grandeur donnée ont autant de chance, pour une distribution symétrique, d'être positifs que négatifs; et une valeur dépassant θ de 4 ou 5 unités a la même probabilité d'être observée qu'une valeur inférieure à θ de 4 à 5 unités. C'est sur cette idée que se base le test des **signes et rangs** de Wilcoxon [1]

Reprenons l'exemple du nombre de pages dans les livres de ma bibliothèque. En notant θ le nombre de pages médian, les hypothèses de test sont

$$H_0 : \theta = 220 \text{ contre } H_1 : \theta \neq 220$$

Nous rappelons que nous avions observé l'échantillon suivant :

126	142	156	228	245	246	370	419	433	454	478	503

TABLE 3.1 – Nombre de pages des 12 livres tirés au hasard dans ma bibliothèque

Formulation et postulat. Nous rangeons par ordre croissant les écarts à 220 (écarts en valeurs absolue), puis nous associons à chaque écart son signe (c'est à dire un signe + si l'observation correspondante est supérieure à la médiane spécifiée sous H_0 et un signe - sinon). On calculons la somme S_p des rangs des écarts positifs et la somme S_n des rangs des écarts négatifs. Si H_0 est vraie, on s'attend à ce que ces deux sommes soit presque égales. La statistique de test est la plus petite des deux sommes. Pour évaluer la signification, nous utilisons la table des signes et rangs de Wilcoxon qui donne le seuil de la région de rejet.

Exercice

1. Combien y a t'il de façons différentes d'attribuer des signes + et - à un ensemble de 12 valeurs ?
2. Quelle est la probabilité que tous les signes soient positifs (ie $S_n = 0$) ?

1. En anglais, on dit *signed ranks test* ce qui est aussi traduit **test des rangs signés**.

3. Si seul le rang 1 est négatif, que vaut S_n ? Quelle est la probabilité associée ?
4. Utiliser `excel` ou `openoffice` pour construire la loi de la statistique de test du test des signes et rangs de Wilcoxon dans le cas où le nombre d'observations est égal à 11. En déduire la probabilité que la statistique de test soit inférieure ou égale à 15, à 10.

Procédure. Dans l'exemple des livres, nous rangeons par ordre de valeur absolue croissante les écarts à 220. En conservant le signe, nous obtenons

8, 25, 26, -64, -78, -94, 150, 199, 213, 234, 258, 283

Les signes et rangs correspondants sont

1,2,3,-4,-5,-6,7,8,9,10,11

La somme des rangs négatifs est $S_n = 15$. Or dans la table, nous voyons que si $n = 11$, le test bilatéral de niveau 5% rejette H_0 si la plus petite des deux sommes, S_n et S_p est inférieur ou égale à 10.

En conclusion, nous ne rejetons pas H_0 au niveau nominal 5%.

Discussion

1. Hypothèse de symétrie ?
2. Hypothèse de continuité ?

3.2.2 Le problème des ex aequo

Nous avons supposé que la distribution de la variable d'intérêt est continue dans la population. Or pour une distribution continue, la probabilité d'obtenir des observations égales est nulle de même que celle d'obtenir des observations égales à la médiane de la population. Cependant, en pratique, les observations ne sont pas strictement continues (arrondis ou précision limitée des appareils de mesure).

Si une ou plusieurs valeurs coïncident avec la médiane spécifiée sous H_0, nous leur attribuons le rang 0.

Si plusieurs écarts ont le même rang (en valeur absolue) ; nous leur attribuons le rang moyen. Par exemple, si les écarts signés sont :

3, 4.7, -5.2, 5.2, 7,7,-7,

nous leur attribuons les rangs suivants :

1,2,-3.5,3.5,6,6,-6.

3.2.3 Approximation pour les grands échantillons

Pour des tailles d'échantillon $n > 20$, on peut approcher la statistique de test du test des signes et rangs de Wilcoxon par une variable aléatoire de loi de Gauss. Soit S la statistique de test, on vérifie que la moyenne de S est $n(n+1)/4$ et que sa variance est $n(n+1)(2n+1)/24$ et on a que la variable

$$Z = \frac{S - \frac{1}{2} - \frac{n(n+1)}{4}}{\sqrt{n(n+1)(2n+1)/24}}$$

suit approximativement une loi de Gauss de moyenne 0 et de variance 1 si n est plus grand que 20. Le $1/2$ au numérateur est une correction de continuité.

Si un grand échantillon comporte des valeurs égales à la médiane sous H_0 ou des ex aequo, on modifie Z de la façon suivante

$$Z = \frac{S - \frac{n(n+1)}{4} - d_0(d_0+1)}{\sqrt{n(n+1)(2n+1)/24 - d_0(d_0+1)(2d_0+1)/24 - \sum_{i=1}^{n_{ge}}(d_i^3 - d_i)/48}}$$

où d_0 est le nombre de valeurs égales à la médiane spécifiée sous H_0, n_{ge} est le nombre de groupes d'ex aaequo et d_i le nombre d'ex aequo dans le $i$ème groupe.

Chapitre 4

Tests paramétriques - Estimation de la position pour un échantillon isolé

Dans le cadre des tests non paramétriques tels que le test du signe ou le test des signes et rangs, on ne fait aucune hypothèse sur la distribution de la variable observée. On n'utilise que la position des observations les unes par rapport aux autres. Ceci est un avantage, car ça permet d'appliquer ces tests dans un grand nombre de situations. Cependant l'inconvénient est une perte de puissance liée au fait qu'on utilise peu d'information.

Dans les tests paramétriques, on utilise d'avantage d'information sur la distribution de la variable étudiée ou sur celle des estimateurs des paramètres considérés.

4.1 Eléments de probabilité

L'idée dans les tests apramétriques est d'jouter de l'information structurante. En pratique, on va modéliser la loi de la variable d'intérêt et/ou des estimateurs considérer. On a donc besoin de disposer d'outils de modélisation qui sont ici des lois de probabilités.

4.1.1 Quelques lois de probabilité continues

1. Loi de Gauss (ou loi normale)

   On sait écrire la densité de probabilité ϕ de la loi de Gauss de moyenne μ et de variance σ^2

   $$\phi(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \text{ pour } x \in \mathbb{R}$$

   La fonction de répartition Φ de la loi de Gauss n'admet pas d'expression analytique simple. On l'obtient par l'approximation numérique de l'intégrale

   $$\Phi(x) = \int_{-\infty}^{x} \phi(t)dt$$

   On parle de *loi de Gaus centrée réduite* si la moyenne est nulle (*centrée*) et la variance est égale à 1 (*réduite*).

Pour simplifier l'écriture, on notera parfois $X \sim \mathcal{N}(\mu, \sigma)$ pour signifier que la v.a. X suit une loi de Gauss de moyenne μ et de variance σ^2.

Proposition 1 *Toute combinaison linéaire de variables aléatoires de loi de Gauss suit une loi de Gauss.*

Exercice : Soient X et Y deux variables aléatoires indépendantes de loi de Gauss. Notons respectivement μ_X et μ_Y leurs moyennes et σ_X^2 et σ_Y^2 leurs variances. Quelles est la loi de la variable aléatoire $Z = X + 2Y$? Donner ses paramètres et écrire sa fonction de densité de probabilité.

Exemple important : Soient $X_1, \cdots, X_n$ n variables aléatoires de loi de Gauss de moyenne μ et de variance σ^2. Alors l'estimateur empirique $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ de laa moyenne μ est une variable aléatoire de loi de Gauss de moyenne μ et de variance $\frac{\sigma^2}{n}$.

2. Loi du chi2
   La loi du chi2 permet de modéliser la loi d'une somme de carrés de variables aléatoires gaussiennes centrées réduites : soient $X_1, \cdots, X_k$ k variables aléatoires gaussiennes indépendantes et de même variance σ^2 alors, $Z = \sum_{i=1}^{k} X_i^2$ suit une loi du chi 2 à k degrés de liberté.

   Remarque : $Z = \sum_{i=1}^{k}(X_i - \bar{X})$ suit une loi du chi 2 à $(k-1)$ degrés de liberté.

3. Loi de Student
   La loi de student permet de modéliser la loi du rapport d'une variable aléatoire gaussienne centrée réduite sur la racine carrée d'une variable aléatoire de loi chi 2 normalisée par le nombre de degrés de liberté : soient U une variable gaussienne centrée réduite et Z une variable aléatoire du chi 2 à k degrés de liberté, alors $\frac{U}{\sqrt{Z/k}}$ soit une loi de student à k degrés de liberté.

4. Loi de Fisher
   La loi de Fisher permet de modéliser le rapport de deux variables distribuées suivant des lois du chi 2. Soient Z_1 et Z_2 deux variables de loi de chi 2 à k_1 et k_2 degrés de libertés et d'écart-types σ_1 et σ_2 alors

   $$\frac{Z_1/\sigma_1}{Z_2/\sigma_2}$$

   suit une loi de Fisher à (k_1, k_2) degrés de liberté.

Loi	Prob. ou ddp	Moyenne	Variance
0-1	$P(X=0)=1-p$ et $P(X=1)=p$	p	$p(1-p)$
Uniforme	$P(X=x)=\frac{1}{n}, x\in[1,n]$	$\frac{n+1}{2}$	$\frac{n^2-1}{12}$
Binomiale	$P(X=x)=C_n^x p^x(1-p)^{n-x}$ pour $x\in[0,n]$	np	$np(1-p)$
Géométrique	$P(X=x)=p(1-p)^{x-1}$ pour $x=1,2,\ldots$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
Pascal	$P(X=x)=C_{x-1}^{n-1}p^n(1-p)^{x-n}$ $\frac{n}{p}$	$\frac{n(1-p)}{p^2}$	
Poisson	$P(X=x)=\frac{e^{-\lambda}\lambda^x}{x!}$ pour $\lambda\geq 0$ et $x=1,2,\ldots$	λ	λ
Uniforme	$f(x)=\frac{1}{b-a}$ avec $a\leq x\leq b$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Gauss	$f(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ pour $x\in\mathbb{R}$	μ	σ^2
Cauchy	$f(x)=\frac{a}{\pi(a^2+x^2)}$	non défini	non défini
Gamma	$f(x)=\frac{\lambda^k x^{k-1}e^{-\lambda x}}{\Gamma(k)}$	$\frac{k}{\lambda}$	$\frac{k}{\lambda^2}$
Exponentielle	$f(x)=\frac{1}{a}e^{-\frac{x}{a}}$ pour $x>0$ et $a>0$	a	a^2
Rayleigh	$f(x)=\frac{x}{\sigma^2}e^{-\frac{x^2}{2\sigma^2}}$ pour $x>0$	$\sigma\sqrt{\frac{\pi}{2}}$	$\sigma^2(2-\frac{\pi}{2})$
Laplace	$f(x)=\frac{a}{2}e^{-a\lvert x\rvert}$	0	$\frac{2}{a^2}$
χ^2	$f(x)=\frac{1}{2^{\frac{m}{2}}\ \Gamma(\frac{m}{2})}x^{\frac{m}{2}-1}\ e^{-\frac{x}{2}}$	m	$2m$
Student	$f(x)=\frac{\frac{n+1}{2}}{\sqrt{n\pi}\Gamma(\frac{n}{2})\left(1+\frac{x^2}{n}\right)^{(n+1)/2}}$	0	$\frac{n}{n-2}$; $n>2$

4.1.2 Convergence en loi

On s'intéresse à la loi d'une suite de v.a. identiquement distribuées, et plus particulièrement à la convergence à l'infini. Pour étudier cette convergence, il existe de nombreux outils ; nous utiliserons ici uniquement la notion de convergence en loi.

Définition 1 - ***Convergence en loi.***
Soit une suite de v.a. X_n de fonction de répartition $F_n(x)$, et soit X une v.a. de fonction de répartition $F(x)$. On dit que la suite X_n converge en loi vers la v.a. X si et seulement si $F_n(x)$ converge vers $F(x)$.

C'est ce type de convergence qu'on utilise quand on dit abusivement qu'une statistique de test est *approximativement distribuée* suivant une loi de Gauss. On devrait toujours dire que *la statistique de test converge en loi vers une variable aléatoire de loi normale*.

4.1.3 Théorème de limite centrale

Le théorème de limite centrale est l'un des résultats les plus importants de la théorie des probabilités. De façon informelle, ce théorème donne une estimation très précise de l'erreur que l'on commet en approchant l'espérance mathématique par la moyenne arithmétique. Ce phénomène a d'abord été observé par Gauss qui l'appelait loi des erreurs ; mais ce dernier n'en a pas donné de démonstration rigoureuse. La preuve du théorème a été apportée par Moivre et Laplace ; le théorème porte donc parfois leurs noms.

Ce théorème est fondamental car il justifie toutes les approximations par la loi normale.

Théorème 1 - ***Théorème de limite centrale***
Soit X_n une suite de v.a. de même loi d'espérance μ et d'écart type σ. Alors la v.a. $\frac{1}{\sqrt{n}}(\frac{X_1+X_2+\ldots+X_n-n\mu}{\sigma})$ converge en loi vers une v.a. normale centrée réduite $\aleph(0,1)$ quand n tend vers l'infini.

Exemples

1. La moyenne expérimentale ou arithmétique $\left(\frac{X_1+X_2+\ldots+X_n}{n}\right)$ est de moyenne μ, la moyenne théorique, et d'écart-type $\frac{\sigma}{\sqrt{n}}$. Et d'après de théorème de limite centrale,

$$\frac{(X_1 + X_2 + \ldots + X_n)/n - \mu}{\sigma/\sqrt{n}}$$

converge vers une variable aléatoire de loi normale centrée et réduite quand n tend vers l'infini. Exercice :

   (a) Vérifier que la moyenne et l'écart-type de la moyenne arithmétique sont bien μ et $\frac{\sigma}{\sqrt{n}}$

   (b) Montrer que si Y suit une loi de Gauss de moyenne μ et de variance σ^2, alors $\frac{Y-\mu}{\sigma}$ suit une loi de Gauss de moyenne 0 et de variance 1.

2. Une proportion F_n admet pour moyenne la proportion théorique p et pour écart-type $\sqrt{\frac{p(1-p)}{n}}$. Ainsi d'après le théorème de limite centrale

$$\frac{F_n - p}{\sqrt{p(1-p)/n}}$$

tend vers une variable aléatoire de loi normale centrée et réduite quand n tend vers l'infini.

3. Comme cas particulier de ce théorème, on retrouve également la convergence d'une suite de variables aléatoires de loi binomiale vers une variable aléatoire de loi normale (théorème de Bernoulli). Ce théorème justifie l'utilisation de la loi normale lorsqu'il y a répétition d'expériences identiques.

4.2 Test de la moyenne (ou Test de Student)

Un contrôle anti-dopage a été effectué sur 16 sportifs. On a mesuré la variable X de moyenne μ, qui est le taux (dans le sang) d'une certaine substance interdite. Voici les données obtenues :

0.35	0.4	0.65	0.27	0.14	0.59	0.73	0.13
0.24	0.48	0.12	0.70	0.21	0.13	0.74	0.18

La variable X est supposée gausienne et de variance $\sigma^2 = 0.04$. On veut tester, au niveau de signification nominal 5% l'hypothèse selon laquelle le taux moyen dans le sang de la population des sportifs est égal 0.4.

On pose des hypothèses de test unilatérales :

$$H_0 : \mu = \mu_0 = 0.4 \text{ contre } H_1 : \mu > 0.4$$

La statistique de test est la moyenne empirique (enocre appelée moyenne arithmétique). Si on note $X_1, \cdots, X_n$ un échantillon de variables aléatoires de même loi que X, la moyenne empirique est donnée par

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

Intuitivement, on comprend bien qu'on va rejeter H_0 si $\bar{X}_n - \mu_0$ est trop grand en valeur absolue c'est à dire si la moyenne empirique est trop éloignée de la moyenne sous H_0.

D'après le théorème de limite centrale, sous H_0, $Z = \frac{\bar{X}_n - \mu_0}{\sigma/\sqrt{n}}$ converge vers une variable aléatoire de loi de Gauss de moyenne 0 et de variance 1 quand n tend vers l'infini. D'autre part, d'après la remarque faite plus haut on comprend qu'on rejette H_0 si $|Z| > z_0$. Pour construire la région de rejet de H_0, on cherche donc z_0 tel que

$$P(|Z| > z_0) = \alpha$$

soit encore

$$P\left(Z > z_0 \text{ ou } Z < z_0\right) = P(Z > z_0) + P(Z < -z_0) = \alpha$$

or on a par symétrie de la loi de Gauss de moyenne 0 et de variance 1

$$P(Z > z_0) = P(Z < -z_0) = \Phi(-z_0) = 1 - \Phi(z_0)$$

où on note Φ la fonction de répartition de la loi Gauss de moyenne 0 et de variance 1. Ainsi z_0 est tel que

$$1 - \Phi(z_0) = \alpha/2$$

ce qui s'écrit encore

$$z_0 = \Phi^{-1}(1 - \alpha_2)$$

D'après la table de la fonction de répartition inverse de la loi normale, on en déduit que $z_0 = 1.96$ car $\alpha = 0.05$.

Finalement, on rejette donc H_0 si

$$|\bar{X}_n - \mu_0| > 1.96\frac{\sigma}{\sqrt{n}}$$

.

Remarques

- On peut aussi conclure le test en calculant son degré de signification soit

$$pv = P(Z >) = \cdots$$

- Lorsque le nombre d'observations n est grand (supérieur à 30), d'après le théorème de limite centrale on a que la statistique de test

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$$

suit approximativement une loi de Gauss quelque soit la loi de la variable X considérée.

4.2.1 Si la variance est inconnue

Dans le cas où la variance n'est pas connue, on doit l'estimer en utilisant les observations. La statistique de test du test de la moyenne est alors donnée par

$$Z = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$$

où s est l'entimateur de la variance défini de la façon suivante

$$S^2 = \frac{1}{n-1}\sum_i (X_i - \bar{x})^2$$

Dans ce cas, Z ne suit plus une loi de Gauss car le dénominateur n'est plus une constante mais une réalisation de l'estimateur de la moyenne de la variable X. L'écart-type s Par construction, S^2 suit une loi du χ^2 à $(n-1)$ degrés de liberté **si X suit une loi de Gauss**. Y est alors une v.a. suivant une de Student à $(n-1)$ degrés de libertés. Et on utilise une table de la loi de Student pour conclure le test.

Remarque : Lorsque le nombre d'observations n est grand (supérieur à 30), on peut utiliser le théorème de limite centrale pour approcher la loi de la statistique Z.

4.2.2 Calcul de la puissance du test

Dans le cas d'un test de Student, on peut calculer la puissance du test si on peut donner une valeur de la moyenne sous l'hypothèse alternative.

$$H_0 : \mu = \mu_0 \text{ contre } H_1 : \mu = \mu_1$$

La puissance est définie par

$$\mathcal{P} = P(\text{rejeter } H_0 | H_0 \text{ est fausse})$$

Ainsi la puissance est la probabilité de la région de rejet de H_0 sous la loi de H_1.

$$\begin{aligned}
\mathcal{P} &= P\left(Z > z_0 | \frac{Z-\mu_1}{\sigma/\sqrt{n}} \text{ suit une loi } \mathcal{N}(0,1)\right) \\
&= P\left(\tilde{Z} > \frac{z_0 - \mu_1}{\sigma/\sqrt{n}}\right) \\
&= 1 - \Phi\left(\frac{z_0 - \mu_1}{\sigma/\sqrt{n}}\right)
\end{aligned}$$

4.3 Test pour une proportion

Soit une population très grande où la proportion d'individus possédant le caractère A est égale à p. On pense que cette proportion ne peut avoir que deux valeurs p_0 ou p_1. Au vu d'un échantillon de taille n, on désire prendre une décision quant à la valeur de cette proportion, avec une signification α.

A partir de l'échantillon, l'estimateur de la proportion théorique sera la fréquence empirique $f_n = \frac{n_A}{n}$ où n_A est le nombre d'individus possédant le caractère A dans l'échantillon.

Les hypothèses de test sont donc

$$\begin{cases} H_0 : p = p_0 \\ H_1 : p = p_1 \end{cases}$$

La règle de décision est donnée par

$$\begin{cases} si\ f_n \geq \pi\ alors\ H_1 \\ si\ f_n < \pi\ alors\ H_0 \end{cases}$$

où π désigne la borne de la région critique.

f_n est une réalisation d'une v.a. F_n dont la loi de probabilité peut être déterminée grâce au théorème central limite. Si la taille de l'échantillon est suffisamment grande (en pratique, $nf_n > 5$ et $n(1-f_n) > 5$), on admet que la loi de F_n tend vers une loi normale de moyenne p et d'écart-type $\sqrt{\frac{p(1-p)}{n}}$. Ce qui nous conduit à

$$\alpha = P(F_n \geq \pi | H_0\ vraie)$$

avec F_n : $\aleph[p, \sqrt{\frac{p(1-p)}{n}}]$.

Sous l'hypothèse H_0, on obtient

$$\alpha = P\left[\frac{(F_n - p_0)\sqrt{n}}{\sqrt{p_0(1-p_0)}} \geq \frac{(\pi - p_0)\sqrt{n}}{\sqrt{p_0(1-p_0)}}\right] = P\left[Y \geq \frac{(\pi - p_0)\sqrt{n}}{\sqrt{p_0(1-p_0)}}\right]$$

où $Y = \frac{(F_n - p_0)\sqrt{n}}{\sqrt{p_0(1-p_0)}}$ est une v.a. normale centrée réduite. La valeur du seuil critique est lue dans une table de la loi normale.

L'erreur de seconde espèce est donnée par :

$$\beta = P\left[Y \leq \frac{(\pi - p_1)\sqrt{n}}{\sqrt{p_1(1-p_1)}}\right]$$

où $Y = \frac{(F_n - p_1)\sqrt{n}}{\sqrt{p_1(1-p_1)}}$ est une v.a. normale centrée réduite. On en déduit la puissance du test.

Remarque : voir le test du signe pour les grands échantillons !

Chapitre 5

Tests sur la position et la dispersion pour deux échantillons indépendants

5.1 Introduction

Le problème qui consiste à comparer la position de deux échantillons est un problème très courant. Il se pose, par exemple, losque l'on veut verifier l'efficacité d'un traitement médical. Dans ce cas, on considère deux groupes de patients, l'un recevant le traitement et l'autre un placebo. Si on note respectivement μ_T et μ_P les positions des populations sous traitement et sous placebo, on pose les hypothèses de test suivantes :

$$H_0 : \mu_T = \mu_P \text{ contre } H_1 : \mu_T \neq \mu_P$$

On remarque que l'hypothèse nulle traduit toujours l'absence d'effet (c'est à dire un effet nul).

Lorsque l'on veut comparer les positions (médiane ou moyenne) de deux échantillons indépendants, on doit tenir compte de la dispersion des deux échantillons et non plus d'un seul. On ne peut pas simplement se ramener aux tests étudiés précédemment.

Nous introduisons ci-dessous deux types de tests :
- des tests non paramétriques (ou libres de distribution) reposant sur des propriétés des statistiques d'ordre comme dans le test du signe ou le test des rangs signés de Wilcoxon ; ces tests sont utilisés quand on étudie des petits échantillons pour lesquels on ne peut/veut pas faire d'hypothèse sur la distribution de la variable d'intérêt.
- des tests praramètriques basés sur une hypothèse de normalité de la variable d'intérêt ou de l'estimateur considéré.

5.2 Tests non paramétriques

Considérons l'exemple suivant. Un psychologue note le temps (en s) mis par des enfants, dont 7 sont considérés comme normaux et 8 comme mentalement retardés, pour accomplir une série de tâches manuelles simples. Les temps sont

Enfants normaux	204	218	197	183	227	233	191	
Enfants retardés	243	228	261	202	270	242	220	239

On se demande alors si les populations d'où proviennent ces deux séries d'observations sont significativement différentes. Notons μ_1 et μ_2 les temps médians des deux groupes d'enfants. On pose les hypothèses de tests :

$$H_0 : \mu_1 = \mu_2 \text{ contre } H_1 : \mu_1 < \mu_2$$

5.2.1 Test de la médiane

Le test de la médiane généralise le test du signe. L'idée est que si les deux échantillons proviennent de deux populations ayant la même médiane, alors chacune des deux médianes empiriques est un estimateur raisonnable de la médiane commune.

Soient deux suites d'observations de tailles respectives n_1 et n_2 et issues de deux populations de médianes respectives μ_1 et μ_2. Si les médianes des deux populations coïncident, on s'attend à ce que la médiane M de toutes les observations regroupées soit proche de la médiane de chacun des échantillons.

Pour définir la statistique de test, nous construisons le tableau de contingence suivant

	Ech. 1	Ech. 2
$> M$	a_1	a_2
$< M$	$n_1 - a_1$	$n_2 - a_2$

où a_1 est le nombre d'observations de l'échantillon 1 qui sont supérieurs à la médiane.

On définit alors la statistique de test par

$$T = (2a_1 - n_1)^2 \frac{n_1 + n_2}{n_1 n_2}$$

On peut montrer que sous H_0, T suit une loi du chi2 à un degré de liberté (voir test du chi2). Ainsi, si la statistique de test observée est supérieure à 3.84, on rejette H_0 au risque 5

5.2.2 Test de Mann-Whitney-Wilcoxon

Le test de la médiane utilise très peu d'information et comme le test du signe il est peu puissant. On introduit alors le test de Mann-Whitney-Wilcoxon qui est une extension du test des signes et rangs. Ce test est utilisé pour comparer deux échantillons qui ne peuvent se distinguer que par un glissement de leur position. Aussi, pour utiliser ce test, on fait l'hypothèse que la dispersion des deux échantillons est comparable.

Considérons l'exemple suivant. Un psychologue note le temps (en s) mis par des enfants, dont 7 sont considérés comme normaux et 8 comme mentalement retardés, pour accomplir une série de tâches manuelles simples. Les temps sont

Enfants normaux	204	218	197	183	227	233	191	
Enfants retardés	243	228	261	202	270	242	220	239

On se demande alors si les populations d'où proviennent ces deux séries d'observations sont significativement différentes. Notons μ_1 et μ_2 les temps médians des deux groupes d'enfants. On pose les hypothèses de tests :

$$H_0 : \mu_1 = \mu_2 \text{ contre } H_1 : mu_1 < mu_2$$

On remarque ici que les écart-types estimés des deux groupes d'observations sont respectivement 18.9 et 21.8. Il est en pratique insuffisant de comparer ces deux valeurs et nous proposerons un test plus loin. Cependant, on convient qu'elles sont du même ordre de grandeur.

Si les deux échantillons ont la même médiane, on s'attend à ce qu'ils se répartissent de façon homogène autour de cette médiane. Autrement dit, on s'attend à ce que les rangs des deux échantillons regroupés soient bien mélangés.

Valeurs	183	191	197	*202*	204	218	*220*	227
Rangs	1	2	3	*4*	5	6	*7*	8
Valeurs	*228*	233	*239*	*242*	*243*	*261*	*270*	
Rangs	*9*	10	*11*	*12*	*13*	*14*	*15*	

On fait alors la somme des rangs de chacun des échantillons et on obtient $S_1 = 35$ et $S_2 = 85$. On en déduit la valeur des statistiques de test qui sont

$$U_1 = S_1 - \frac{n_1(n_1+1)}{2} \text{ et } U_2 = S_2 - \frac{n_2(n_2+1)}{2}$$

Ici, $U_1 = 14$ et $U_2 = 42$.

Dans le test de Mann-Whitney-Wicoxon, on rejette H_0 si la plus petite des deux statistiques (test bilatéral) ou celle qui est appropriée (test unilatéral), supérieure ou égale à la valeur lue dans la table (table A6 de P. Sprent).

Ici, la valeur seuil correspondant au niveau de signification de 5% est égale à 10. Donc on rejette H_0.

Cas des grands échantillons

Dans le cas où l'un des échantillons est de taille supérieure à 20, on donne une approximation gaussienne de la loi de la statistique de test :

$$Z = \frac{U + 1/2 - n_1 n_2/2}{\sqrt{n_1 n_2 (n_1 + n_2 + 1)/12}}$$

suit approximativement une loi de Gauss centrée et réduite.

Le problème des ex aequo

Dans le cas où il y a peu d'ex aequo, on peut utiliser la méthode des rangs moyens.

5.3 Tests paramétriques

Les tests non paramétriques ont l'inconvénient d'être souvent peu puissants. Ceci vient du fait que l'on n'utilise que la position des observations dans les échantillons et non leur valeur. Quand on sait faire des hypothèses sur la distribution de la variable d'intérêt, il est préférable de construire un test paramétrique qui sera plus puissant. Pour comparer deux moyennes de la variables aléatoires gaussiennes, on utilise le test de Student pour deux échantillons.

5.3.1 Comparaison de deux moyennes - Test de Student

Soient X_1 et X_2 deux variables aléatoires indépendantes de lois normales de moyennes μ_1 et μ_2, et d'écart types σ_1 et σ_2. On dispose de deux échantillons indépendants $\{X_1^{(1)}, \cdots, X_{n_1}^{(1)}\}$ et $\{X_1^{(2)}, \cdots, X_{n_2}^{(2)}\}$ tels que $X_i^{(1)}$ (resp. $X_i^{(2)}$) suit la même loi que X_1 (resp. X_2).

Sachant les échantillons, on cherche à décider si les moyennes μ_1 et μ_2 sont significativement différentes ou non. On teste alors $H_0 : \mu_1 = \mu_2$ contre $H_1 : \mu_1 \neq \mu_2$ au risque α

On utilise le test de Student pour deux échantillons indépendants.

Si les écart types σ_1 et σ_2 sont connus, on calcule

$$z = \frac{m_1 - m_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

On rejette H_0 au risque α si $z \notin [-t_{1-\frac{\alpha}{2}}, t_{1-\frac{\alpha}{2}}]$ où la valeur $t_{1-\frac{\alpha}{2}}$ est lue dans la table de la loi normale centrée réduite.

Si les écart types σ_1 et σ_2 sont inconnus, il faut tenir compte de la taille des échantillons

a) Si n_1 et n_2 sont tous les deux supérieurs à 30, on calcule

$$z = \frac{m_1 - m_2}{\sqrt{\frac{s_1^2}{n_1 - 1} + \frac{s_2^2}{n_2 - 1}}}$$

On rejette H_0 au risque α si $z \notin [-t_{1-\frac{\alpha}{2}}, t_{1-\frac{\alpha}{2}}]$ où la valeur $t_{1-\frac{\alpha}{2}}$ est lue dans la table de la loi normale centrée réduite.

b) Si n_1 ou n_2 est inférieur à 30 et $\sigma_1 = \sigma_2$ on calcule

$$z = \frac{m_1 - m_2}{\hat{\sigma}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

où

$$\hat{\sigma} = \sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}}$$

On rejette H_0 au risque α si $z \notin [-t_{1-\frac{\alpha}{2};n_1+n_2-2}, t_{1-\frac{\alpha}{2};n_1+n_2-2}]$ où la valeur $t_{1-\frac{\alpha}{2};n_1+n_2-2}$ est lue dans la table de Student à $n_1 + n_2 - 2$ degrés de liberté.

c) Si n_1 ou n_2 est inférieur à 30 et $\sigma_1 \neq \sigma_2$ on calcule

$$z = \frac{m_1 - m_2}{\sqrt{\frac{s_1^2}{n_1 - 1} + \frac{s_2^2}{n_2 - 1}}}$$

On rejette H_0 au risque α si $z \notin [-t_{1-\frac{\alpha}{2};\nu}, t_{1-\frac{\alpha}{2};\nu}]$ où la valeur $t_{1-\frac{\alpha}{2};\nu}$ est lue dans la table de Student à ν degrés de liberté ; ν est l'entier le plus proche de

$$\frac{\left[\frac{s_1^2}{n_1 - 1} + \frac{s_2^2}{n_2 - 1}\right]^2}{\frac{s_1^4}{(n_1 - 1)n_1^2} + \frac{s_2^4}{(n_2 - 1)n_2^2}}$$

Le test de Student est assez robuste mais si l'on s'éloigne trop des conditions de normalité, il est préférable d'utiliser un test non paramétrique.

5.3.2 Comparaison de deux variances - Test de Fisher

Avec les mêmes notations que précédemment, on teste $H_0 : \sigma_1 = \sigma_2$ contre $H_1 : \sigma_1 \neq \sigma_2$ au risque α

La statistique de test est définie par $S = \frac{\hat{s}_1^2}{\hat{s}_2^2}$ avec $\hat{s}_1^2 = \frac{n_1 s_1^2}{n_1 - 1}$.

D'après les propriétés des variables aléatoires de loi de Gauss, S suit une loi de Fisher à $(n_1 - 1, n_2 - 1)$ degrés de liberté.

On rejette H_0 au risque α si $z \notin [F_{\frac{\alpha}{2}}(n_1 - 1, n_2 - 1), F_{1-\frac{\alpha}{2}}(n_1 - 1, n_2 - 1)]$ où la valeur F_α est la valeur de l'inverse de la fonction de répartition de la loi de Fisher de $(n_1 - 1, n_2 - 1)$ degrés de liberté au point α. Cette valeur est lue dans la table de Fisher-Snédécor à $n_1 - 1$ et $n_2 - 1$ degrés de liberté.

Remarque : $F_{\frac{\alpha}{2}}(n_1 - 1, n_2 - 1) = \frac{1}{F_{1-\frac{\alpha}{2}}(n_2 - 1, n_1 - 1)}$

5.3.3 Comparaison de deux proportions

Soit p_1 (respectivement p_2) la proportion d'individus d'une certaine modalité A dans la population mère M_1 (resp. M_2). On extrait un échantillon de taille n_1 (resp. n_2) dans la population M_1 (resp. M_2). On teste à partir de ces échantillons, on dispose d'une estimation f_1 (resp. f_2) de p_1 (resp. p_2) qui suit une loi F_1 (resp. F_2). $H_0 : p_1 = p_2$ contre $H_1 : p_1 \neq p_2$ au risque α. On suppose que $n_1 F_1$ et $n_2 F_2$ suivent approximativement des lois normales. On calcule $\hat{p} = \frac{n_1 f_1 + n_2 f_2}{n_1 + n_2}$ puis $z = \frac{f_1 - f_2}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}}$ On rejette H_0 au risque α si $z \notin [-t_{1-\frac{\alpha}{2}}, t_{1-\frac{\alpha}{2}}]$ où la valeur $t_{1-\frac{\alpha}{2}}$ est lue dans la table de la loi normale centrée réduite.

Chapitre 6

Tests d'adéquation et comparaison de distributions

6.1 Introduction

Jusqu'à présent nous avons étudié des méthodes permettant de tester la position et la dispersion d'un ou deux échantillons. Cependant des populations peuvent présenter d'autres caractéristiques importantes. Nous considérons ici toute la distribution de la population.

La distribution (ou la loi) d'une variable aléatoire X est décrite par sa fonction de répartition [1] c'est à dire par la fonction

$$F(x) = P(X \leq x) \tag{6.1}$$

On observe que la fonction de répartition est une fonction monotone croissante comprise entre 0 et 1.

Exemple - soit X une variable aléatoire de loi uniforme sur l'intervalle $[0, 1]$. n réalisation de la variable X à la même probabilité de prendre toute valeur $x \in [0, 1]$ et sa fonction de répartition est donnée par

$$F(x) = x\mathbf{1}_{[0,1]}(x)$$

Étant données des observations $x_1, \cdots, x_n$ on peut se demander si ces valeurs sont cohérentes avec un échantillonnage d'une distribution continue spécifiées. Le test d'adéquation [2] de Kolmogorov permet de répondre à cette question.

6.2 Test d'adéquation de Kolmogorov

Soient $x_1, \cdots, x_n$, n réalisations d'une variable aléatoire X. On se demande s'il est raisonnable de supposer que X suit la loi caractérisée par la fonction de répartition F et on pose les hypothèses

1. En anglais : cumulative distribution function
2. En anglais : goodness-of-fit test

de tests :

$$H_0 : X \text{suit la loi } F \text{ contre } H_1 : X \text{suit une autre loi}$$

On propose de construire une statistique de test basée sur la distance entre les fonction F et une estimation de la fonction de répartition de X obtenue à partir des observations.

6.2.1 Estimer la fonction de répartition

On construit naturellement un estimateur de la fonction de répartition de X d'après l'équation (6.1).

$$F_n(x) = \frac{\text{Card}(\{i|x_i \leq x\}}{n})$$

Exercice - On considère les données d'un essai visant à déterminer la solidité d'une corde d'escalade. Un morceau de 1 m corde est mis sous tension jusqu'à cassure. On se demande si la corde pour casser à n'importe endroit. On obtient les résultats suivants :

0.1 0.4 0.4 0.6 0.7 0.7 0.8 0.9 0.9 0.9

1. Tracer l'estimation de la fonction de répartition.
2. Ajouter sur le graphique la fonction de répartition théorique pour ce problème.

Exercice :

1. Tracer la fonction de répartition empirique de l'échantillon ci-dessus.
2. Superposer, sur le même graphique, la fonction de répartition de la loi uniforme sur $[0, 1]$.

6.2.2 Statistique de test

Kolmogorov propose d'utiliser la statistique de test suivante :

$$D_n = \sup_{x \in \mathbb{R}} |F_n(x) - F(x)|$$

La loi de cette statistique de test est données dans la table de Kolmogorov.

Si on considère de nouveau les données de l'exercice, on obtient

observations	0.1	0.4	0.4	0.6	0.7	0.7	0.8	0.9	0.9	0.9
F_n	1/10	3/10	3/10	4/10	6/10	6/10	7/10	1	1	1
F	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
$F - F_n$	0	0.1	0	0	0.1	0	0.2	0.1	0	

d'où $D_n = 0.2$ or on rejette H_0 au risque $\alpha = 5\%$ si $D_n > 0.369$. Ainsi, ici on peut supposer que les observations sont issues d'une loi uniforme sur $[0, 1]$.

6.2.3 Cas de la loi normale

La version du test de Kolmogorov adaptée pour la loi de Gauss s'appelle le test de Lilliefors. Ce test est peu puissant et on lui préfère le test de Shapiro-Wilk. Ce dernier est basé sur une comparaison de

deux estimateurs de la variance qui ne peuvent conduire à la même estimation que si les observations sont issues d'une loi de Gauss.

Pour vérifier qu'une série d'observation suit une loi normale, on peut en première approche utiliser une méthode graphique : la droite de Henry (*quantile-quantile plot* ou *qqplot*).

Soit $\{x_1, \cdots, x_n\}$ une suite d'observations. Si cette suite constitue une suite de réalisation d'une variable gaussienne, alors les points de coordonnées $(x_i, \Phi^{-1}((i-1/2)/n))$ sont alignés sur la droite d'équation

$$y = \frac{x - \overline{x}}{\hat{\sigma}}$$

Cette droite est appelée *droite de Henry*.

Exercice - Les observations ci-dessous correspondent à la hauteur de 7 arbres dans une forêt. Peut-on considérer que la distribution de ces hauteurs est gaussienne ?

23.4 24.6 25.0 26.3 26.8 27.0 27.6

On tracera la droite de Henry pour les données centrées et réduites de façon à pouvoir utiliser la table de la loi de Gauss ($\bar{x} = 25.8$, $s = 1.5$).

Données centrées réduites	-1.5	-0.8	-0.5	0.3	0.6	0.7	1.1
$\Phi^{-1}((i-1/2)/7)$	-1.46	-0.79	-0.36	0	0.36	0.79	1.46

6.2.4 Test d'adéquation du chi 2 : loi discrète

Pour une distribution discrète on utilise le test d'adéquation du chi 2.

Exemple : On suppose que le nombre de pièces défectueuses produites en un jour par une machine suit une loi de Poisson, de paramètre inconnu. Rappelons les caractéristiques de cette loi : si une variable aléatoire X suit une loi de Poisson de paramètre λ , alors $E(X) = \lambda$, $Var(X) = \lambda$, et pour tout $k \in \mathbb{N}$,

$$P(X = k) = \frac{\lambda^k e^{\lambda}}{k!}$$

On pose les hypothèses de test

$$H_0 : X suit une loi de Poisson de paramtre 1.2 \text{ contre } H_1 : X \text{ suit une autre loi}$$

On observe 100 jours de production de cette machine, voici les résultats, regroupés en 5 classes.

Nombre de pièces défectueuses	0	1	2	3	4 et plus
Nombre d'observations	27	41	21	7	4
Fréquence empirique	0,27	0,41	0,21	0,07	0,04

On utilise une statistique de test du chi2 donnée par

$$T = \sum_{k=1}^{K} \frac{(\hat{n} f_k - n p_k)^2}{n p_k}$$

avec f_k et p_k les fréquences empirique et théorique de la classe k et K le nombre de classes. T suit une loi du chi 2 à $K-1$ degrés de liberté.

Pour l'exemple considéré, les fréquences théoriques sont données ci-dessous.

Nombre de pièces défectueuses	0	1	2	3	4 et plus
Fréquence théorique	0,30	0,36	0,22	0,09	0,03

La statistique de test $T = 0,0112$; or d'après la table du chi 2 on rejette H_0 au risque 5% si $T > 5.99$.

6.3 Test d'identité de deux distributions de deux distributions

On peut généraliser le test de Kolmogorov au cas de deux échantillons afin de comparer leurs distributions. Le test s'appelle alors test de Kolmogorov-Smirnov.

6.3.1 Test de Kolmogorov-Smirnov

L'hypothèse nulle est que les deux échantillons proviennent de la même distribution ; l'alternative est qu'ils proviennent de distributions ayant des répartitions différentes. On ne spécifie aucune forme particulière pour leur différence. Et la statistique de test est basée sur un écart en valeur absolue entre la fonctions de répartition empiriques des deux suites d'observations.

Exemple - Un psychologue fait passer un test de rapidité à des enfants normaux et d'autres considérés comme mentalement retardés. Les temps qu'ils mettent pour accomplir une série de tâches sont les suivants :

Enfants normaux	183	191	197	204	218	227	233	
Enfants retardés	202	220	228	239	242	243	261	270

On pose les hypothèses de test

$$H_0 : F_{EN} = F_{ER} \text{ contre } H_1 : F_{EN} \neq F_{ER}$$

Et on estime les fonctions de répartition des deux groupes :

obs.	*183*	*191*	*197*	202	*204*	*218*	220	*227*	228	*233*	239	242	243	261	270
$\hat{F}_{EN}$	1/7	2/7	3/7	3/7	4/7	5/7	5/7	6/7	6/7	1	1	1	1	1	1
$\hat{F}_{ER}$	0	0	0	1/8	1/8	1/8	2/8	2/8	3/8	3/8	4/8	5/8	6/8	7/8	1
Ecart	8/56	16/56	...												

L'écart en valeur absolue le plus grand $D_{n_{EN},n_{ER}} \sup_{x\in\mathbb{R}} |F_{EN} - F_{ER}|$ est |1-3/8| = 0.62.

La loi du supremum des écarts en valeur absolue est tabulée dans la table de Smirnov. On rejette H_0 si $D_{n_{EN},n_{ER}} > 0.71$. Donc ici, on ne peut pas rejeter l'hypothèse selon laquelle les distributions sont différentes pour les deux groupes d'enfants.

6.3.2 Test de Cramér-von Mises

Il existe d'autres tests permettant de comparer des distributions. Par exemple, le test de Cramér-von Mises repose sur la somme des carrés des écarts en valeurs absolue entre les deux fonctions de répartition. En notant, S_d^2 cette somme, la statistique de test est

$$T = \frac{nmS_d^2}{n+m}$$

avec m et n les nombres d'observation des deux groupes.

Pour un test bilatéral, on rejette H_0 au niveau de signification 5% (resp. 1%) si T est supérieur à 0.461 (resp. 0.743).

Le test de Cramér-von Mises est souvent plus puissant que le test de Kolmogorov-Smirnov et il est plus facile à utiliser grace à la bonne approximation qui évite le recours à des tables.

Références

Fourdrinier D., (2002). Statistique inférentielle. Dunod.
Jolion J.M., (2003). Probabilité et Statistique. Cours de l'INSA. *http ://rfv.insa-lyon.fr/ jolion*
Kaufman P., (1994). Statistique : Information, Estimation, Test. Dunod.
Saporta G., (1990). Probablités, analyse des données et statistique. Edition Technip.
Reau J.P., Chauvat G., (1996). Probabilités et statistiques. Excercices et corrigés, Armand Colin, Collection cursus TD, série économie.
Scherrer B., (1988). Biostatitiques. Edition Gaetan Morin.
Schwartz D., (1984). Méthodes statistiques á l'usage des médecins et des biologistes, Flammarion, Médecine-Sciences, Collection Statistique en biologie et médecine.