

Partie III

Introduction à l'inférence

8. Inférence: échantillonnage et estimation
9. Distribution d'un estimateur
10. Tests statistiques: introduction
11. Tests et intervalles de confiance pour proportions
12. Tests et intervalles de confiance pour moyennes
13. Tests nonparamétriques pour un et deux échantillons
14. Tests d'adéquation et d'indépendance par la méthode du chi-carré
15. Etudes expérimentales, randomisation et causalité
16. Une introduction au bootstrap

Chapitre 8

Inférence: échantillonnage et estimation

La synthèse entre les observations et la population étudiée s'obtient à l'aide des procédés d'inférence statistique, c'est à dire d'estimation de tests statistiques. L'inférence utilise les modèles mathématiques et le calcul des probabilités. Pour que l'échantillon puisse représenter la population ("échantillon représentatif") il doit être pris de façon aléatoire. Ce chapitre, après avoir précisé le concept d'échantillonnage aléatoire, présente les principaux critères qui permettent d'obtenir de bonnes estimations. Le thème des tests sera traité dans d'autres chapitres.

8.1 Echantillonnage aléatoire simple

L'exemple le plus simple d'échantillonnage aléatoire est celui qui consiste à prendre un certain nombre d'individus d'une population de façon aléatoire. L'application typique est le *sondage*, dont le but est de déterminer une certaine caractéristique moyenne de cette population.

Un échantillon est un sous-ensemble de la population. On démontre que, si la population est de *taille finie*, N , il y a

$$\binom{N}{n} = \frac{N!}{n!(N-n)!}$$

échantillons de taille n . Si tout échantillon de taille n a la même probabilité $1/\binom{N}{n}$ d'être extrait, on dit qu'il est obtenu par *échantillonnage aléatoire simple*.

En général, pour réaliser un tel échantillonnage on dispose de la liste de tous les individus (par exemple, l'annuaire du téléphone). On peut alors imaginer d'écrire les noms de ces individus sur N étiquettes identiques, de mettre ces étiquettes dans une urne et de prendre une étiquette l'une après l'autre au hasard et sans remise.

De façon formelle, on décrit les caractéristiques qui seront observées à l'aide de n variables aléatoires indépendantes: X_1 est la caractéristique du premier individu, X_2 celle du deuxième, \dots , X_n celle du n -ième. Il conviendra de penser à X_1, \dots, X_n comme à des "observations futures". Si la population est de *taille infinie* ($N = \infty$), tous les X_i ont la même *distribution de population* F . On dit alors que

" X_1, \dots, X_n sont indépendantes, identiquement distribuées selon la distribution F ".

Par la suite, nous décrirons un *échantillon aléatoire* par des variables aléatoires X_1, \dots, X_n indépendantes et de distribution identique même si elles ne représentent pas des caractéristiques d'individus réels mais, par exemple, les résultats de n jets d'un dé ou n mesures d'une certaine quantité physique. On utilisera l'abréviation i.i.d. pour l'expression "indépendantes et identiquement distribuées" et on utilisera la notation

$$X_1, \dots, X_n \text{ i.i.d. } \sim F.$$

Remarque

Il y a clairement des échantillonnages plus complexes. Parfois, la population étudiée est formée de sous-populations (par exemple, des villes) appelées des *strates*, et il convient d'échantillonner séparément les sous-populations. Si on applique un échantillonnage aléatoire simple à chaque sous-population, on a un *échantillonnage aléatoire stratifié*. Parfois, les unités (individus) échantillonnées ne sont pas des éléments simples mais des

groupes d'éléments (par exemple, une famille). Si les groupes sont échantillonnés par un échantillonnage aléatoire simple, on obtient un *échantillonnage en grappes*. Il y a des situations où les individus de la population sont ordonnés (par exemple, en ordre alphabétique). Nous supposons qu'ils soient numérotés. Dans ce cas, il convient parfois de choisir le premier individu de l'échantillon de façon aléatoire dans une liste de k individus (k préfixé). Supposons qu'il s'agit de l'individu avec le numéro i . Les autres éléments de l'échantillon sont les numéros $i + k, i + 2k, \dots, i + (n - 1)k$. On parle, dans ce cas, d'un *échantillonnage aléatoire systématique (de type 1 dans k)*. En outre, l'hypothèse d'indépendance des observations n'est pas toujours satisfaite. Par exemple, pour des séries d'observations obtenues successivement dans le temps (*séries temporelles*) on admet généralement une certaine corrélation. Toutes ces situations ne sont pas abordées dans ce cours.

8.2 Inférence statistique

Nous supposons maintenant que nous disposons de données effectivement observées x_1, \dots, x_n . L'inférence statistique s'occupe de la relation entre les données et la population. Elle a le but d'*inférer* (induire, extrapoler, transférer) à la population, des résultats statistiques calculés à l'aide des données. En particulier, elle a pour but de déterminer la distribution F à partir de x_1, \dots, x_n .

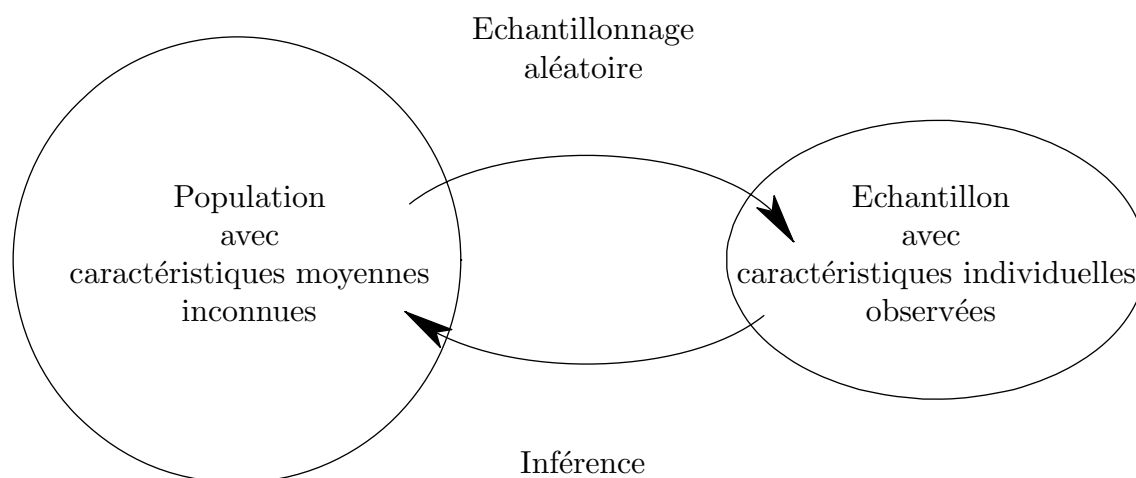


Figure 1. Echantillonnage et inférence statistique.

L'inférence s'appuie sur des modèles mathématiques des observations. En particulier, il conviendra de supposer que x_1 est une valeur observée ou *observation* de X_1 , x_2 est une observation de X_2 , etc. Nous sommes donc confrontés à la situation suivante: *les observations x_1, \dots, x_n sont des issues de variables aléatoires X_1, \dots, X_n , indépendantes et identiquement distribuées selon une certaine distribution F . On appellera l'ensemble x_1, \dots, x_n des observations un échantillon d'observations.*

8.3 Modèle statistique

Un *modèle statistique* est une description mathématique approximative du mécanisme qui génère les observations. Il s'exprime généralement à l'aide d'une *famille de distributions* (ensemble de distributions) et d'hypothèses sur les variables aléatoires X_1, \dots, X_n . Chaque membre de la famille est une approximation possible de F : l'inférence consiste donc à déterminer le membre qui s'accorde le mieux avec les données.

Parfois la famille est presque totalement spécifiée à l'aide d'un modèle paramétrique (Chapitre 7); seuls les paramètres doivent être déterminés à l'aide des données. On parle dans ce cas d'une *approche paramétrique* à l'inférence. Parfois, il vaut mieux de travailler avec une famille moins délimitée, telle que la famille de "toutes les distributions symétriques". On parle dans ce cas d'une *approche non paramétrique*.

Exemple 1. On jette une pièce de monnaie six fois. Un modèle statistique simple des résultats est fourni par l'ensemble de six variables aléatoires indépendantes $X_i, i = 1, \dots, 6$ telles que:

- $X_i = 1$ si le i -ème jet est Pile;
 - $X_i = 0$ si le i -ème jet est Face;
 - les jets, et donc les X_i , sont indépendants;
 - la probabilité de Pile est constante, mais inconnue.
- Donc, $X_i \sim \mathcal{B}(1, p)$ avec $0 \leq p \leq 1$ inconnu.

Supposons qu'on obtient les résultats (Pile, Pile, Face, Pile, Face, Pile). A l'aide de ces données il faut donc déterminer le paramètre p .

Exemple 2. On mesure une quantité physique μ (par exemple le poids d'un objet) n fois. Un modèle fréquemment utilisé dans ce genre de situation est:

$$\text{mesure} = \text{valeur théorique} + \text{erreur de mesure.}$$

Si X_i est la variable aléatoire qui représente la mesure et U_i celle qui représente l'erreur, on peut traduire la formulation précédente du modèle de la façon suivante:

$$X_i = \mu + U_i, \quad i = 1, \dots, n.$$

Parfois on assume uniquement que la distribution des erreurs est symétrique de centre 0 et on cherche donc à déterminer μ selon une approche non-paramétrique. Selon le type de données, il peut être raisonnable de supposer que les erreurs suivent une distribution $\mathcal{N}(0, \sigma^2)$ où σ^2 exprime l'imprécision des mesures. Il faut alors déterminer μ et σ à l'aide d'une approche paramétrique.

8.4 Estimer des paramètres

Intuitivement, le concept d'estimation paramétrique est simple: on observe un échantillon de valeurs de variables aléatoires i.i.d. dont on connaît la distribution paramétrique à l'exclusion de quelques paramètres. A l'aide des observations, on doit déterminer les valeurs des paramètres.

Exemple. Supposons que dans l'Exemple 1 de la section précédente on observe l'échantillon 1, 1, 0, 1, 0, 1. Une façon simple et naturelle de déterminer le paramètre p est:

$$\hat{p} = \frac{\text{nombre de "1"}}{\text{nombre d'essais}} = 4/6 = 0.667.$$

L'exemple nous permet de remarquer que \hat{p} n'est qu'une valeur particulière de la règle générale "nombre de succès/nombre d'essais" applicables à toutes les épreuves qui génèrent un échantillon de succès et d'échecs. On appelle cette règle un estimateur. Plus précisément, on définit un *estimateur* comme une fonction des variables aléatoires X_1, \dots, X_n . Un estimateur est donc une variable aléatoire (qui change de valeur si l'échantillon aléatoire change). Sa valeur particulière obtenue à l'aide de l'échantillon x_1, \dots, x_n est une *estimation* ou *valeur observée* de l'estimateur. Souvent, on utilisera le même symbole pour l'estimation et pour l'estimateur.

Il y a plusieurs critères pour construire des estimateurs. Nous en considérerons quelques uns.

8.5 Le critère du maximum de vraisemblance

Nous considérons un modèle de distribution d'une variable aléatoire X définie par sa distribution de probabilité $P_\theta(X = x_i)$ (cas discret) ou par sa densité $f_\theta(x)$ (cas continu), où θ représente un vecteur de paramètres avec ℓ composantes à déterminer. Voici des exemples.

Modèle binomial:

$$P_\theta(X = k) = \binom{m}{k} p^k (1-p)^{m-k}.$$

D'habitude m est connu. Il y a donc un seul paramètre $\theta = p$ à déterminer (et $\ell = 1$).

Modèle de Poisson:

$$P_\theta(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}.$$

Ici $\theta = \lambda$ et $\ell = 1$.

Modèle de Gauss:

$$f_\theta(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2} \left[\frac{x-\mu}{\sigma}\right]^2\right).$$

En général il faut déterminer μ et σ . Donc $\theta = (\mu, \sigma)$ et $\ell = 2$.

Nous souhaitons déterminer à l'aide de l'échantillon x_1, \dots, x_n une valeur $\hat{\theta}$ de θ , telle que le modèle $P_{\hat{\theta}}$ ou $f_{\hat{\theta}}$ soit bien adapté aux données. Cette valeur est une estimation de θ .

Introduction intuitive du critère du maximum de vraisemblance

Considérons une séquence de 10 épreuves binomiales indépendantes. Chaque épreuve est décrite par une variable aléatoire X_i , $i = 1, \dots, 10$ telle que $X_i \sim \mathcal{B}(1, p)$. Supposons disposer des observations suivantes (échantillon):

$$\{0, 1, 0, 1, 1, 0, 0, 0, 0, 1\}.$$

Selon le modèle, la probabilité d'obtenir une issue identique à celle-ci est

$$P_\theta(X_1 = 0, X_2 = 1, X_3 = 0, \dots, X_{10} = 1) = p^4(1 - p)^6.$$

On appelle cette probabilité la *vraisemblance* de l'échantillon sous le modèle. Selon le critère du maximum de vraisemblance, une valeur plausible de $\theta = p$ est celle qui maximise cette probabilité, c'est-à-dire la valeur de p qui rend l'échantillon observé le plus vraisemblable.

La dérivée de $p^4(1 - p)^6$ est

$$\frac{dP_p(\dots)}{dp} = p^3(1 - p)^5(4 - 10p),$$

et s'annule pour $p = 4/10$. Il est facile de démontrer, en outre, que $P_p(\dots)$ atteint un maximum pour $p = 4/10$.

Le critère du maximum de vraisemblance

Pour une variable aléatoire discrète, nous définissons la *vraisemblance de l'échantillon sous le modèle* $P_\theta(X = x_i)$ comme

$$L(\theta) = P_\theta(X_1 = x_1) \cdot P_\theta(X_2 = x_2) \dots \cdot P_\theta(X_n = x_n).$$

Pour une variable continue, nous définissons, par analogie, la vraisemblance de l'échantillon comme

$$L(\theta) = f_\theta(x_1) \cdot f_\theta(x_2) \cdot \dots \cdot f_\theta(x_n).$$

Le critère du maximum de vraisemblance consiste à estimer θ par la valeur $\hat{\theta}$ telle que

$$L(\hat{\theta}) = \max_{\theta} L(\theta).$$

En général, $\hat{\theta}$ est calculé comme solution de l'équation

$$\frac{dL(\theta)}{d\theta} = 0.$$

On appelle $\hat{\theta}$ l'*estimation du maximum de vraisemblance* de θ . Si on considère $\hat{\theta}$ comme une fonction des variables aléatoires X_1, \dots, X_n , on obtient l'*estimateur du maximum de vraisemblance*.

Exemples

1. Modèle binomial

Soit k le nombre de “succès” (codés 1) dans l'échantillon. Le nombre d'échecs (codés 0) est donc $n - k$. Supposons que $0 < k < n$. Alors

$$\begin{aligned} L(p) &= p^k(1-p)^{n-k}, \\ \ln L(p) &= k \ln(p) + (n-k) \ln(1-p), \\ \frac{d \ln L(p)}{dp} &= k/p - (n-k)/(1-p), \end{aligned}$$

et l'équation $d \ln L(p)/dp = 0$ a comme solution $\hat{p} = k/n$ (la proportion de succès dans l'échantillon). Considérez les cas $k = 0$ et $k = n$ comme des cas particuliers.

2. Modèle de Poisson

Soit x_1, \dots, x_n l'échantillon: noter que les x_i sont des entiers.

$$\begin{aligned} L(\lambda) &= \lambda^{x_1 + \dots + x_n} \exp(-n\lambda) / (x_1! \cdot \dots \cdot x_n!), \\ \ln L(\lambda) &= (x_1 + \dots + x_n) \ln(\lambda) - n\lambda - \ln(x_1! \cdot \dots \cdot x_n!), \\ \frac{d \ln L(\lambda)}{d\lambda} &= (x_1 + \dots + x_n)/\lambda - n, \end{aligned}$$

et l'équation $d \ln L(\lambda)/d\lambda = 0$ a comme solution $\hat{\lambda} = (x_1 + \dots + x_n)/n$ (la moyenne arithmétique des observations).

3. Modèle de Gauss

Il convient de considérer les paramètres μ et σ^2 . Alors

$$L(\mu, \sigma^2) = (2\pi)^{-n/2} (\sigma^2)^{-n/2} \exp\left(-\sum (x_i - \mu)^2 / (2\sigma^2)\right),$$

Les dérivées partielles de L par rapport à μ et à σ^2 s'annulent pour

$$\hat{\mu} = \frac{\sum x_i}{n}, \quad \hat{\sigma}^2 = \frac{1}{n} \sum (x_i - \hat{\mu})^2.$$

La moyenne et la variance des observations sont donc les estimateurs du maximum de vraisemblance.

8.6 Le principe des moindres carrés

Un modèle très général pour des situations où l'on mesure une quantité inconnue μ est

$$X_i = \mu + U_i, \quad i = 1, \dots, n, \quad (1)$$

où X_i est la i -ème mesure, μ la quantité inconnue et U_i est l'erreur de mesure qui a influencé la i -ème observation. Souvent, on assume que les erreurs sont i.i.d. selon une certaine distribution F mais on ne souhaite pas décrire cette distribution de façon plus précise à l'aide d'un modèle de distribution. Dans une modélisation de ce type, la *méthode des moindres carrés* définit un estimateur $\hat{\mu}$ de μ de la façon suivante. On définit

$$Q(\mu) = (x_1 - \mu)^2 + \dots + (x_n - \mu)^2$$

et on cherche la valeur $\hat{\mu}$ de μ telle que cette somme soit minimale. Cette valeur vérifie la relation

$$\frac{dQ(\mu)}{d\mu} = -2(x_1 - \mu) - \dots - 2(x_n - \mu) = 0,$$

et donc

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n x_i.$$

La relation (1) est l'exemple le plus simple d'une multitude de situations. Au Chapitre 3 nous en avons considéré une plus complexe, le modèle de régression simple, qui sera développé au Chapitre 18.

8.7 Estimation non-paramétrique

De manière générale, on peut se poser le problème d'estimer une quantité quelconque dépendante de la distribution d'une variable aléatoire X même s'il ne s'agit pas d'un paramètre d'une famille paramétrique de distributions. Plus précisément, si X_1, \dots, X_n sont des variables aléatoires i.i.d définissant un échantillon, supposons que $X_i \sim F$, où F est la fonction de distribution inconnue de X_i . Des quantités d'intérêt dépendantes de F sont, par exemple:

- $\mu = \int x f(x) dx$, la moyenne de F ;
- $\sigma^2 = \int (x - \mu)^2 f(x) dx$, la variance de F ;
- $q_\alpha = F^{-1}(\alpha)$, le quantile α de F .

Une *estimation non-paramétrique* de F est la fonction de distribution cumulative empirique F_n des observations x_1, \dots, x_n . A l'aide de F_n on estime les quantités μ , σ^2 , p_α par les quantités correspondantes calculées avec F_n . Remarquons, que F_n n'est rien d'autre que la distribution de probabilité qui associe une probabilité $1/n$ à chaque observation. Ainsi on obtient, par exemple:

- $\hat{\mu} = (\sum x_i)/n$, la moyenne arithmétique ou moyenne de F_n ;
- $\hat{\sigma}^2 = \sum (x_i - \hat{\mu})^2/n$, la variance empirique ou variance de F_n ;
- $\hat{q}_\alpha = F_n^{-1}(\alpha)$ le quantile α de F_n (ou d'une version lissée de F_n).

8.8 Evaluation graphique de l'adéquation d'un modèle de distribution

Dans la Table 1, la colonne (2) contient les longueurs (en $\text{mm} \times 10^{-1}$) de 100 ailes de mouches et la colonne (3) les quantités de lait (en $\text{pounds} \times 100$) produites en une année par 100 vaches. La colonne (1) est simplement un numéro de ligne.

Table 1. Longueurs d'ailes de mouche et quantités de lait (Sokal et Rohlf, 1995).

(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)
01	36	51	21	42	58	41	45	61	61	47	67	81	49	76
02	37	51	22	42	58	42	45	61	62	47	67	82	49	76
03	38	51	23	42	58	43	45	61	63	47	68	83	49	79
04	38	53	24	43	58	44	45	61	64	47	68	84	49	80
05	39	53	25	43	58	45	45	61	65	47	69	85	50	80
06	39	53	26	43	58	46	45	62	66	47	69	86	50	81
07	40	54	27	43	58	47	45	62	67	47	69	87	50	82
08	40	55	28	43	58	48	45	62	68	47	69	88	50	82
09	40	55	29	43	58	49	45	62	69	47	69	89	50	82
10	40	56	30	43	58	50	45	63	70	48	69	90	50	82
11	41	56	31	43	58	51	46	63	71	48	70	91	51	83
12	41	56	32	44	59	52	46	63	72	48	72	92	51	85
13	41	57	33	44	59	53	46	64	73	48	73	93	51	87
14	41	57	34	44	59	54	46	65	74	48	73	94	51	88
15	41	57	35	44	60	55	46	65	75	48	74	95	52	88
16	41	57	36	44	60	56	46	65	76	48	74	96	52	89
17	42	57	37	44	60	57	46	65	77	48	74	97	53	93
18	42	57	38	44	60	58	46	65	78	49	74	98	53	94
19	42	57	39	44	60	59	46	67	79	49	75	99	54	96
20	42	57	40	44	61	60	46	67	80	49	76	100	55	98

Pour les ailes de mouche la moyenne arithmétique et la variance empirique sont $\hat{\mu} = 45.5 \text{ mm} \times 10^{-1}$ et $\hat{\sigma}^2 = 15.21 \text{ mm}^2 \times 10^{-2}$. Pour les quantités de lait on trouve $\hat{\mu} = 66.61 \text{ pounds} \times 100$ et $\hat{\sigma}^2 = 124.48 \text{ pounds}^2 \times 100^2$.

L'histogramme des longueurs d'ailes de mouche est représenté dans la Figure 2. Dans le même graphique on voit aussi la densité de Gauss adaptée aux données, c'est-à-dire telle que $\mu = 45.5 \text{ mm} \times 10^{-1}$ et $\sigma^2 = 15.21 \text{ mm}^2 \times 10^{-2}$. L'adéquation est bonne. Dans la Figure 3 on trouve l'histogramme des quantités de lait et la densité de Gauss avec $\mu = 66.61 \text{ pounds} \times 100$ et $\sigma^2 = 124.48 \text{ pounds}^2 \times 100^2$. L'adéquation est mauvaise: les quantités de lait ne peuvent pas être décrites approximativement par le modèle de Gauss. La comparaison d'un histogramme avec la densité ou la distribution de probabilité adaptée aux données permet donc d'évaluer visuellement l'adéquation du modèle. Il y a un procédé graphique d'évaluation plus efficace.

Supposons que la variable aléatoire X a une fonction de distribution cumulative continue F . La distribution cumulative empirique F_n est une estimation de F . Supposons de calculer les centiles c_k , $k = 1, \dots, 100$ de F et les centiles \tilde{c}_k de F_n . La représentation graphique des points (\tilde{c}_k, c_k) ($k = 1, \dots, 100$) est alors un ensemble de points approximativement alignés le long de la droite de pente 1 qui passe par l'origine.

On remarque que les observations ordonnées

$$x_{[1]} \leq x_{[2]} \leq \dots \leq x_{[n]}$$

sont des estimations des quantiles $q_{k/n}$, $k = 1, \dots, n$ de F , mais parfois (en raison des sauts de F_n), on préfère interpréter $x_{[k]}$ comme une estimation de $q_{(k-0.5)/n}$. L'évaluation graphique de l'adéquation d'un modèle de densité f_θ est donc réalisée (à l'aide d'un programme informatique) de la façon suivante.

1. Estimer les paramètres du modèle. Soit $\hat{\theta}$ une estimation de θ .
2. Calculer les quantiles $(k - 0.5)/n$ du modèle estimé $f_{\hat{\theta}}$. En d'autres termes, calculer

$$\hat{q}_{(k-0.5)/n} = F_{\hat{\theta}}^{-1}((k - 0.5)/n), \quad k = 1, \dots, n.$$

3. Représenter les points $(x_{[k]}, \hat{q}_{(k-0.5)/n})$ dans un plan.
4. Vérifier si les points se trouvent à proximité d'une droite (la diagonale). Si ce n'est pas le cas, le modèle n'est pas adéquat.

Cette représentation s'appelle un *probability plot* ou un *qq-plot* (quantile-quantile plot). La Figure 4 montre le qq-plot des longueurs des ailes de mouche: les points se trouvent à proximité de la diagonale. La Figure 5 montre le qq-plot des quantités de lait: les points ne se trouvent pas à proximité d'une droite.

Remarques

1. Les valeurs extrêmes s'éloignent souvent de la droite, même si le modèle est adéquat, car elles ont une "variabilité" élevée. Parfois, il s'agit d'outliers. Pour évaluer l'adéquation du modèle, il faut donc prêter plus d'attention à la partie centrale du diagramme qu'aux extrémités.
2. Pour les modèles dont les paramètres représentent la position et l'échelle il n'est pas nécessaire d'estimer les paramètres pour évaluer l'adéquation. Par exemple, pour le modèle de Gauss, il suffit d'utiliser les quantiles $q_{(k-0.5)/n}$ de la distribution de Gauss standard. En effet, l'alignement des points est maintenu lors d'une transformation de position et d'échelle des données. Toutefois, la droite change de pente et ne passe plus par l'origine. La Figure 6, est obtenue de cette façon avec les longueurs des ailes de mouche. La droite passe par $(0, 45.5)$ et sa pente est $3.90 = 15.21^{0.5}$.

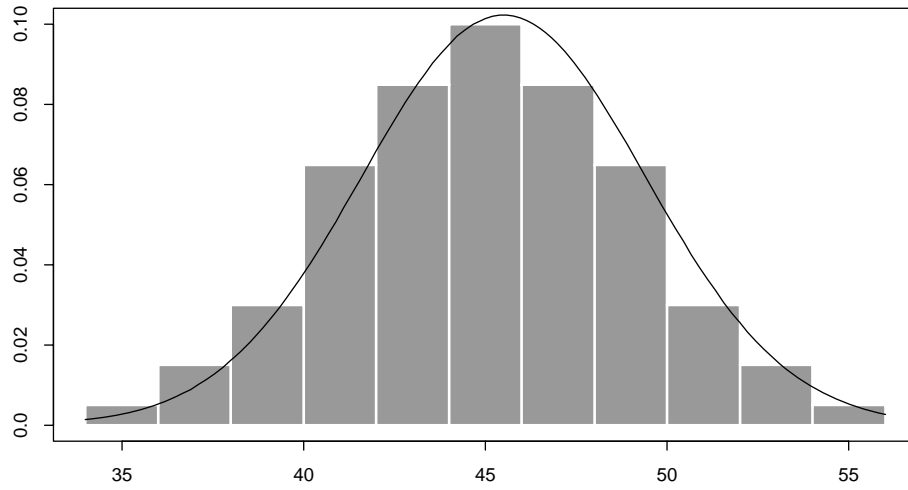


Figure 2 Histogramme des longueurs des ailes de 100 mouches domestiques mesurées en $\text{mm} \times 10^{-1}$ et distribution de Gauss adaptée.

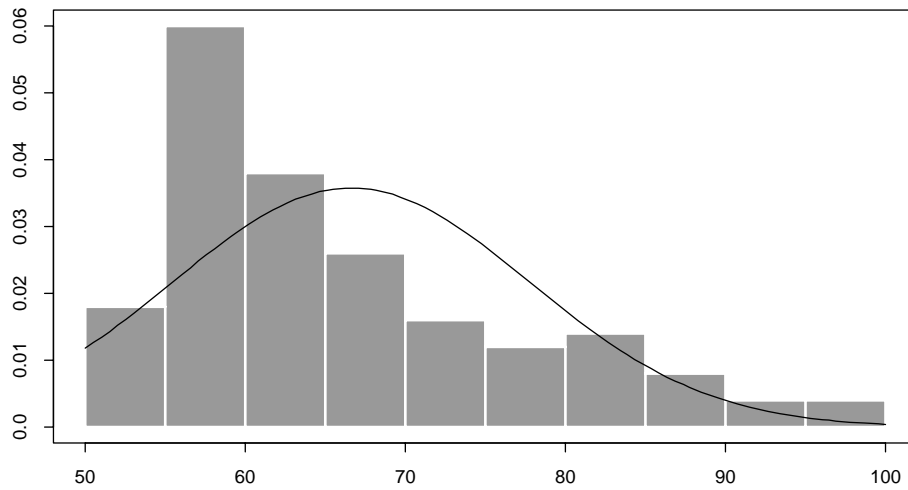


Figure 3 Histogramme des quantités (en $\text{pounds} \times 100$) de lait produites en une année par 100 vaches et distribution de Gauss adaptée.

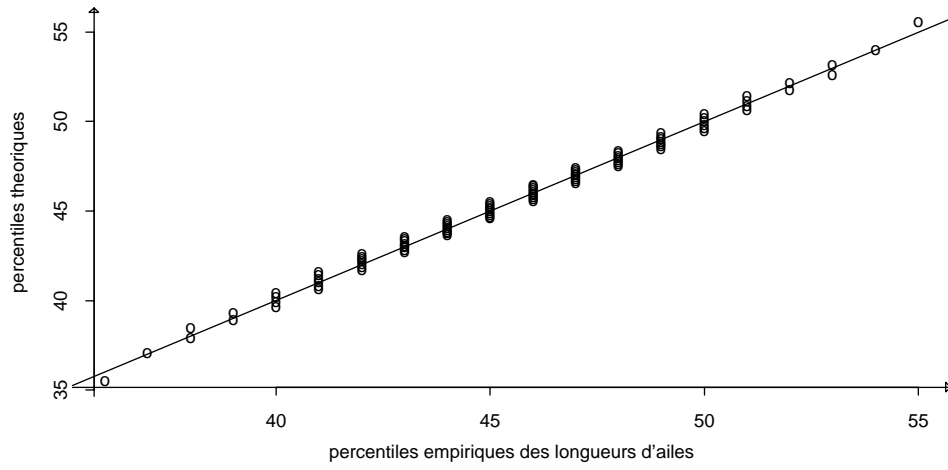


Figure 4

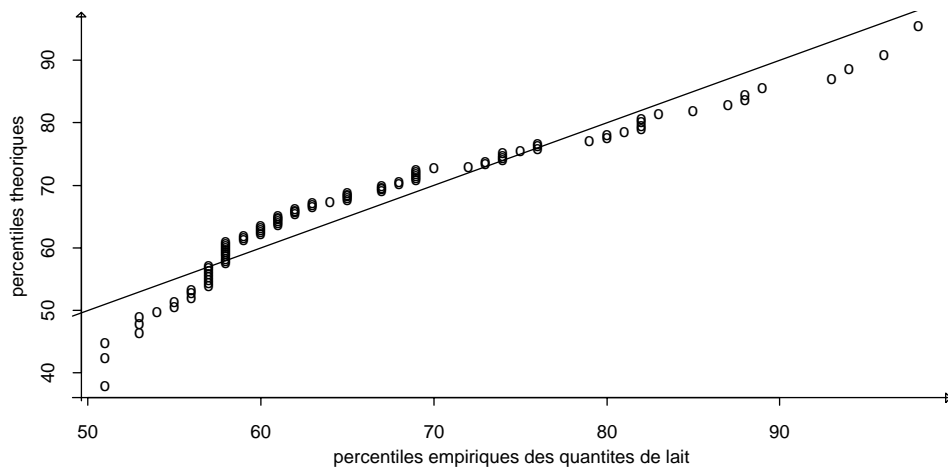


Figure 5

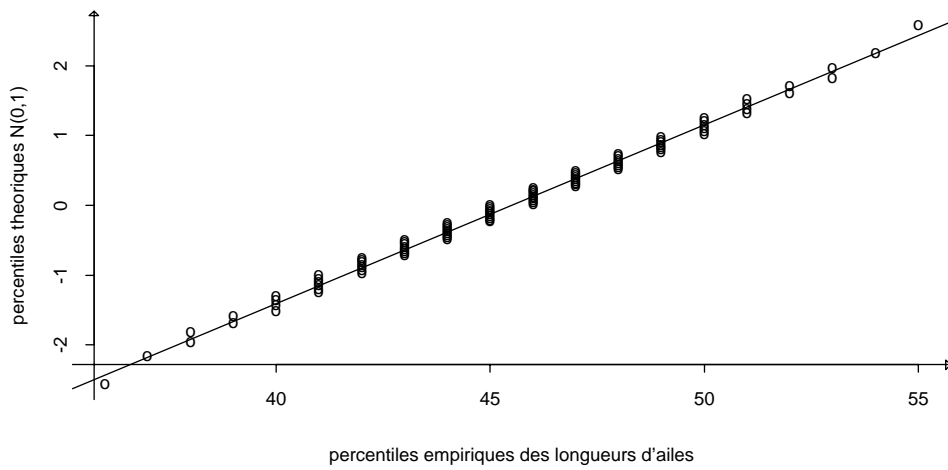


Figure 6

Complément

Dans l'application du critère du maximum de vraisemblance, le problème de maximiser $L(\theta)$ est équivalent à celui de minimiser $-\sum \ln P_\theta(X = x_i)$ (lorsque X est discrète) ou $-\sum \ln f_\theta(x_i)$ (lorsque X est continue). Si on annule les dérivées partielles de cette fonction par rapport aux ℓ composantes $\theta_1, \dots, \theta_\ell$ de θ , on obtient l'équation vectorielle (ou système de ℓ équations)

$$\sum_{i=1}^n \mathbf{s}(x_i, \theta) = \mathbf{0},$$

où

$$\mathbf{s}(x, \theta) = \left(\frac{\partial}{\partial \theta_1} \ln(P_\theta(x)), \dots, \frac{\partial}{\partial \theta_\ell} \ln(P_\theta(x)) \right)^T, \quad \text{si } X \text{ est discrète,}$$

$$\mathbf{s}(x, \theta) = \left(\frac{\partial}{\partial \theta_1} \ln(f_\theta(x)), \dots, \frac{\partial}{\partial \theta_\ell} \ln(f_\theta(x)) \right)^T, \quad \text{si } X \text{ est continue,}$$

est le vecteur (colonne) des *score functions* (le signe T signifie "transposition" d'un vecteur ligne en vecteur colonne).

Chapitre 9

Distribution d'un estimateur

En général, une simple estimation ne suffit pas: il est nécessaire de connaître son degré d'imprécision. L'outil fondamental pour évaluer un estimateur et le comparer à d'autres est sa distribution d'échantillonnage. Par exemple, à égalité entre différents aspects, on préférera l'estimateur avec la plus petite variance. Ce chapitre s'occupe du calcul de la distribution d'un estimateur. Si on suppose que la distribution des données peut être décrite par un modèle paramétrique, on aura une approche paramétrique au calcul de la distribution de l'estimateur; autrement on parlera d'une approche non-paramétrique. Le calcul pourra être effectué à l'aide d'outils mathématiques ou à l'aide de la simulation sur ordinateur.

9.1 La distribution d'échantillonnage

Nous sommes confrontés à la situation suivante: les observations x_1, \dots, x_n (échantillon) sont issues des variables aléatoires X_1, \dots, X_n , i.i.d., qui suivent une distribution F_X , c'est-à-dire,

$$X_1, \dots, X_n \text{ i.i.d. } \sim F_X.$$

Un estimateur est une fonction $S(X_1, \dots, X_n)$. Il est donc une variable aléatoire; sa valeur observée est $s = S(x_1, \dots, x_n)$. Nous nous intéressons à la distribution de $S(X_1, \dots, X_n)$, parfois dite *distribution d'échantillonnage* de S . Nous notons la fonction de distribution cumulative de S par $F_S(s; F_X)$ pour marquer le fait qu'elle dépend de F_X .

Pour calculer $F_S(s; F_X)$ de façon exacte il faudrait connaître F_X de façon exacte. Mais dans tout problème pratique ce n'est pas le cas. (Si F_X était connue, il n'y aurait pas besoin de l'estimer !) Il faut donc recourir à des approximations de F_X : l'*approche paramétrique* utilise des modèles paramétriques ajustés à l'échantillon; l'*approche non-paramétrique* utilise la distribution des données, c'est-à-dire la fonction de distribution cumulative empirique F_n .

Une première façon d'effectuer le calcul de F_S est purement mathématique. Certains résultats concernant la moyenne arithmétique sont donnés dans la section suivante.

9.3 Distribution d'une moyenne

Soit $S(X_1, \dots, X_n) = \sum_{i=1}^n X_i/n$. Deux résultats fondamentaux concernent l'espérance et la variance de la distribution de S .

– On a:

$$E\left(\frac{X_1 + \dots + X_n}{n}\right) = \frac{1}{n} \sum E(X_i).$$

Ainsi, si $E(X_i) = \mu$ pour tous les i ,

$$E\left(\frac{X_1 + \dots + X_n}{n}\right) = \mu.$$

– En outre, si les X_i sont indépendants,

$$\sigma^2\left(\frac{X_1 + \dots + X_n}{n}\right) = \frac{1}{n^2} \sum \sigma^2(X_i).$$

Ans, si $\sigma^2(X_i) = \sigma^2$ (une constante) pour tous les i ,

$$\sigma^2 \left(\frac{X_1 + \dots + X_n}{n} \right) = \frac{\sigma^2}{n}.$$

L'écart type de la moyenne est donc σ/\sqrt{n} et peut être estimé par $\hat{\sigma}/\sqrt{n}$ (l'écart type empirique divisé par \sqrt{n}) où $\hat{\sigma} = (\sum(x_i - m(X))^2/n)^{1/2}$.

Si on peut décrire la distribution F_X de X_1, \dots, X_n à l'aide d'un modèle paramétrique, on obtient les résultats suivants.

- Si les X_i sont indépendants et $X_i \sim \mathcal{N}(\mu_i, \sigma_i)$ alors la distribution de S est une distribution de Gauss de moyenne $\sum \mu_i/n$ et variance $\sum \sigma_i^2/n^2$.
- Si les X_i sont indépendants et $X_i \sim \mathcal{B}(n_i, p)$ alors la distribution de $\sum X_i$ est une distribution binomiale $\mathcal{B}(n, p)$ avec $n = \sum n_i$. La distribution de S (proportion de "1") peut être déduite facilement de ce résultat: les modalités de S sont $0, 1/n, \dots, n/n = 1$ et les probabilités correspondantes sont données par la distribution binomiale.
- Si les X_i sont indépendants et $X_i \sim \mathcal{P}(\lambda_i)$ alors la distribution de S est une distribution de Poisson de moyenne $\sum \lambda_i/n$.

Il est clair que, comme les modèles paramétriques sus-mentionnés ne sont que des descriptions mathématiques de F_X , les distributions de S qui en découlent ne sont que des approximations des distributions réelles !

Parfois il n'est pas possible d'utiliser un modèle paramétrique comme approximation de F_X . On peut alors calculer la distribution de la moyenne arithmétique de façon approximative, à l'aide du théorème suivant.

– *Théorème limite centrale*

Supposons que X_1, \dots, X_n soient i.i.d. selon une distribution F_X inconnue, telle que que $E(X_i) = \mu$ et $V(X_i) = \sigma^2$. Alors, si $n \rightarrow \infty$,

$$P \left(\frac{\sum_{i=1}^n X_i/n - \mu}{\sigma/\sqrt{n}} < s \right) \rightarrow \Phi(s).$$

La distribution de la moyenne arithmétique centrée et réduite est donc approximativement Gaussienne $\mathcal{N}(0, 1)$, indépendamment de la distribution F_X , pourvu que n soit suffisamment élevé. La distribution de la moyenne arithmétique est approximativement normale de moyenne μ et variance σ^2/n et ces paramètres peuvent être estimés. Malheureusement, il n'y a pas en général une règle simple pour déterminer la valeur minimale de n pour que l'approximation soit bonne. Cette valeur dépend de la forme de F_X . En outre, l'approximation peut être sérieusement perturbée par la présence d'outliers.

Exemples

1. La Figure 1 montre six qq-plots. Chaque qq-plot compare la distribution empirique de 1000 moyennes arithmétiques – standardisées à l’aide de leur moyenne et écart type – avec la distribution $\mathcal{N}(0, 1)$.

Pour obtenir le diagramme (a), 1000 échantillons de taille $n = 10$ ont été tirés d’une population uniforme $U[a, b]$ de moyenne 2 et variance 2 ($b = \sqrt{24}/2 + 2 \approx 4.45$, $a = 4 - b \approx -0.45$). Une technique de simulation a été utilisée; voir Complément 2 à la fin de ce chapitre. Le qq-plot indique que la distribution des moyennes de ces 1000 échantillons est approximativement Gaussienne.

Pour obtenir le diagramme (b), 1000 échantillons de taille $n = 10$ ont été tirés d’une population constituée d’un mélange entre la distribution $U[a, b]$ et la distribution $U[10, 50]$. Chaque observation provient de la première avec probabilité 0.9 et de la deuxième avec probabilité 0.1. Le qq-plot indique que la distribution des moyennes est influencée par les outliers provenant de $U[10, 50]$

Pour le diagramme (c), 1000 échantillons de taille $n = 10$ ont été tirés d’une population lognormale de moyenne 2 et variance 2 ($\mu = 0.490$, $\sigma = 0.637$). La distribution des moyennes n’est pas Gaussienne.

Pour le diagramme (d), 1000 échantillons de taille $n = 10$ ont été tirés d’un mélange entre une distribution lognormale ($\mu = 0.490$, $\sigma = 0.637$) et une uniforme $U[10, 50]$ (outliers) dans les proportions 0.9 et 0.1. La distribution des moyennes est influencée par les outliers.

Les diagrammes (e) et (f) ont été obtenus comme les diagrammes (c) et (d) mais avec $n = 100$. La distribution représentée en (e) est approximativement normale; celle en (f) est influencée par les outliers.

2. Soient $X_1, \dots, X_n \sim \mathcal{B}(1, p)$. Dans ce cas, $\hat{p} = \sum X_i/n$ est la proportion de “1” dans la répétition de n épreuves binomiales. La somme $\sum X_i$ suit une distribution $\mathcal{B}(n, p)$ et la distribution de \hat{p} peut être calculée à l’aide de ce résultat. Toutefois, si n n’est pas petit il convient généralement de remarquer que

$$E(\hat{p}) = p \quad \text{et} \quad \sigma(\hat{p}) = \sqrt{p(1-p)/n}$$

et que

$$\frac{\hat{p} - p}{\sigma(\hat{p})} \sim \mathcal{N}(0, 1)$$

si $n \rightarrow \infty$. La variable aléatoire $(\hat{p} - p)/\sigma(\hat{p})$ a donc approximativement une distribution $\mathcal{N}(0, 1)$. Il se trouve que l’approximation est bonne si les deux nombres $p - 2\sigma(\hat{p})$ et $p + 2\sigma(\hat{p})$ sont situés entre 0 et 1.

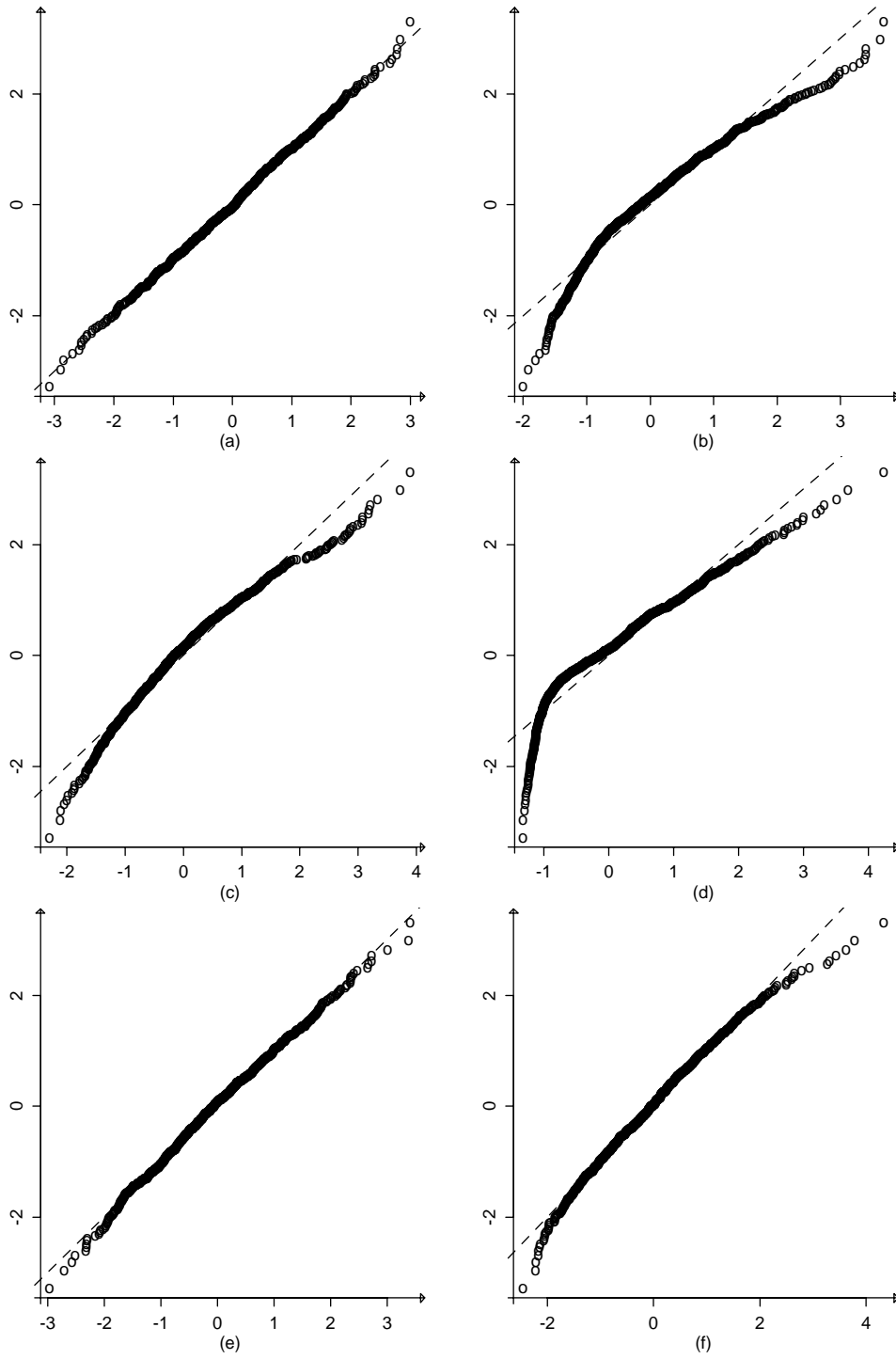


Figure 1. qq-plots.

9.4 Calcul de la distribution d'un estimateur par simulation: le bootstrap

Ici, $S(X_1, \dots, X_n)$ est une statistique quelconque, pas nécessairement la moyenne arithmétique. Supposons d'abord de connaître F_X . Avec un ordinateur, on peut générer n valeurs x_1^*, \dots, x_n^* qui *simulent* n observations indépendantes d'une variable aléatoire $X \sim F_X$ (voir Complément 2). On calcule ensuite une valeur simulée $s^* = S(x_1^*, \dots, x_n^*)$ de S . Si on demande à l'ordinateur de répéter cette simulation un grand nombre de fois, par exemple $k = 1000$ fois, on obtient 1000 valeurs simulées de s^* . Leur distribution cumulative empirique est une approximation de $F_S(s; F_X)$ et l'approximation est d'autant plus précise que k est élevé.

En pratique, la distribution F_X est normalement inconnue et il faut utiliser une estimation à sa place. Par exemple, on peut utiliser un modèle paramétrique adapté à l'échantillon. On dit alors que la simulation est paramétrique ou que l'approximation de F_X est obtenue par *bootstrap paramétrique*.

D'autre part, une estimation non-paramétrique de F_X est la fonction de distribution cumulative empirique $F_n(x) = (\text{nombre de valeurs } x_i \leq x)/n$. On peut donc approcher $F_S(s; F_X)$ par $F_S(s; F_n)$, c'est-à-dire par la distribution de S lorsque $X \sim F_n$. On appelle $F_S(s; F_n)$ l'approximation *bootstrap non-paramétrique* de $F_S(s; F_X)$. On calcule $F_S(s; F_n)$, de façon approximative, en générant k échantillons simulés d'une variable aléatoire $X \sim F_n$: en d'autres termes, on remplace F_X par F_n dans la simulation. On démontre facilement que pour réaliser cette simulation, il suffit de tirer avec remise des éléments de l'ensemble des données $\{x_1, \dots, x_n\}$. On parle, dans ce cas, de *rééchantillonnage* avec remise.

Exemple

Les données suivantes sont les durées de séjour d'un échantillon de patients hospitalisés au CHUV pour des désordres du système nerveux:

1, 1, 2, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 4, 4, 4, 4, 4, 5, 5, 5, 5, 6, 6, 7, 7, 8, 9, 16, 115, 198, 374.

La médiane est de 4 jours. On aimerait connaître la distribution de la médiane pour savoir si 4 jours est une bonne estimation de la durée médiane de l'ensemble des séjours du même type au CHUV. Pour simuler la distribution de $S(X_1, \dots, X_{32})$, la médiane de X_1, \dots, X_{32} , nous avons tiré avec remise 5000 échantillons (x_1^*, \dots, x_{32}^*) à partir des 32 observations disponibles et calculé chaque fois $s^* = S(x_1^*, \dots, x_{32}^*)$. Deux de ces échantillons simulés sont, par exemple,

3, 2, 3, 4, 7, 4, 2, 5, 5, 8, 2, 7, 4, 4, 5, 2, 374, 3, 4, 8, 4, 2, 4, 115, 7, 1, 4, 2, 4, 2, 3, 3;

8, 3, 8, 3, 3, 5, 16, 1, 374, 198, 374, 1, 6, 115, 5, 2, 1, 3, 2, 5, 3, 6, 5, 5, 9, 5, 2, 4, 3, 7, 2, 3

et leurs médianes sont 4 et 5. La distribution empirique des 5000 médianes simulées est représentée dans la Figure 2 à l'aide d'un histogramme. On remarque que la probabilité que la médiane (d'un nouvel échantillon) soit égale à 4 est estimée à 52%; la probabilité qu'elle se situe entre 3 et 5 est estimée à 98%.

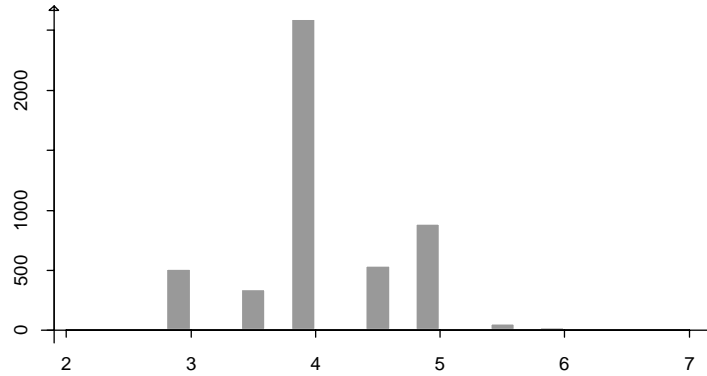


Figure 2. Approximation bootstrap de la distribution de la médiane des durées de séjour.

Compléments

1. Simulation de nombres pseudo-aléatoires

On appelle une *suite de nombres pseudo-aléatoires*, des nombres $x_1, x_2, \dots, x_n, \dots$, générés par un programme informatique qui “se comportent” comme s’ils étaient des observations indépendantes d’une variable aléatoire X . Un procédé courant de génération de nombres pseudo-aléatoires est celui basé sur le *schéma linéaire congruentiel*. Le schéma dépend de trois paramètres m , a et b . Il s’agit de trois nombres entiers positifs préfixés; souvent m est l’entier le plus grand représentable sur l’ordinateur (et dépend donc du type d’ordinateur), tandis que a et b sont choisis convenablement (selon certains critères) en fonction de m . L’utilisateur choisit n et un entier de départ z_0 (*seed*). L’ordinateur calcule alors successivement n entiers z_1, \dots, z_n et n nombres rationnels y_1, \dots, y_n selon la règle

$$\begin{aligned} z_i &= az_{i-1} + b \quad \text{modulo } m, \\ y_i &= z_i/m, \end{aligned}$$

pour $i = 1, \dots, n$. Si les paramètres m , a et b sont bien choisis, les nombres y_i ont une distribution uniforme dans l’intervalle $(0, 1)$.

On trouve facilement des programmes (*uniform random number generators*) pour produire la suite de base $\{y_i\}$ avec distribution uniforme. A partir de celle-ci on peut construire d’autres suites avec des distributions quelconques. Le principe est le suivant: si Y est une variable aléatoire avec distribution uniforme dans $(0, 1)$, alors $X = F^{-1}(Y)$ est une variable aléatoire avec distribution F . Pour obtenir une suite de nombres pseudo-aléatoires avec distribution F , il suffit alors de calculer la suite $\{F^{-1}(y_i)\}$ (voir Figure 3).

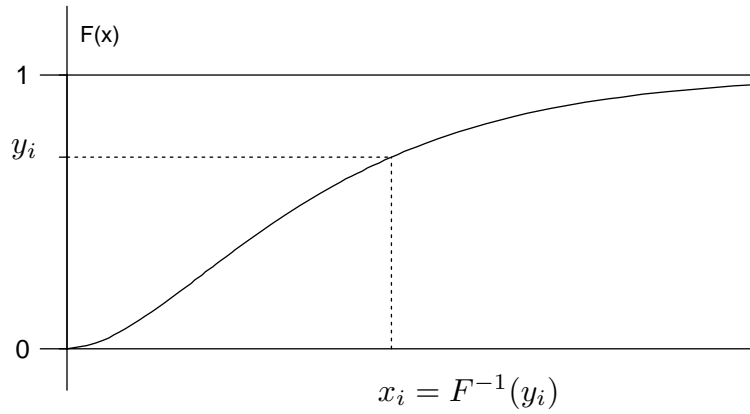


Figure 3. Calcul de nombres pseudo-aléatoires avec distribution F .

2. Biais et carré moyen de l'erreur d'un estimateur

Supposons que $X \sim F_X$ et que F_X dépend d'un paramètre θ inconnu. Pour cette raison, nous remplaçons ici la notation F_X par F_θ . Soit $S(X_1, \dots, X_n)$ un estimateur de θ . La qualité de l'estimateur dépend évidemment de l' *erreur*

$$S(X_1, \dots, X_n) - \theta.$$

Pour mesurer la qualité d'un estimateur S on s'exprime parfois à l'aide des deux quantités suivantes:

– le *biais*

$$b(\theta) = E_\theta(S) - \theta;$$

– le *carré moyen de l'erreur*

$$E_\theta [(S(X_1, \dots, X_n) - \theta)^2].$$

Le suffixe θ indique que l'espérance est calculée en utilisant la distribution F_θ . En général, on démontre que:

$$E_\theta [(S - \theta)^2] = \sigma_\theta^2(S) + (E_\theta(S) - \theta)^2,$$

où nous avons écrit $E_\theta(S)$ et $\sigma_\theta^2(S)$ à la place de $E_\theta(S(X_1, \dots, X_n))$ et $\sigma_\theta^2(S(X_1, \dots, X_n))$. Donc, le carré moyen de l'erreur d'un estimateur S est égal à la variance de l'estimateur plus le carré du biais de l'estimateur. Un estimateur *sans biais* ($b(\theta) = 0$ pour tout θ) donne en moyenne, c'est-à-dire lors d'un usage répété, la bonne réponse.

Exemple 1. Supposons que X_1, \dots, X_n sont i.i.d. selon une distribution qui dépend de deux paramètres μ et σ^2 , notée F_{μ, σ^2} , et que $E_{\mu, \sigma^2}(X_i) = \mu$, $\sigma^2(X_i) = \sigma^2$. Soit $S(X_1, \dots, X_n) = \hat{\mu} = (X_1 + \dots + X_n)/n$ la moyenne arithmétique qu'on utilise comme estimateur de μ . Comme $E(\hat{\mu}) = \mu$, le carré moyen de l'erreur est

$$E_{\mu, \sigma^2} [(\hat{\mu} - \mu)^2] = \sigma^2(\hat{\mu}) = \frac{\sigma^2}{n}.$$

Exemple 2. Supposons que X_1, \dots, X_n sont i.i.d. selon une distribution qui dépend de deux paramètres μ et σ^2 , notée F_{μ, σ^2} , et que $E_{\mu, \sigma^2}(X_i) = \mu$, $\sigma^2(X_i) = \sigma^2$. Soient

$$S_0(X_1, \dots, X_n) = \frac{1}{n} \sum_{i=1}^n (X_i - \hat{\mu})^2$$

et

$$S_1(X_1, \dots, X_n) = \frac{1}{n-1} \sum_{i=1}^n (X_i - \hat{\mu})^2,$$

où

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n X_i.$$

S_0 et S_1 sont les deux versions de variance empirique qu'on utilise d'habitude comme estimateurs de σ^2 . Alors,

$$E_{\mu, \sigma^2}(S_0) = \frac{n-1}{n} \sigma^2$$

et

$$E_{\mu, \sigma^2}(S_1) = \sigma^2.$$

En d'autres termes, S_1 est un estimateur sans biais de σ^2 tandis que S_0 est un estimateur biaisé de σ^2 . Toutefois, le biais s'annule pour $n \rightarrow \infty$.

En effet, on obtient:

$$\begin{aligned} \sum (X_i - \hat{\mu})^2 &= \sum (X_i - \mu + \mu - \hat{\mu})^2 \\ &= \sum [(X_i - \mu)^2 + 2(X_i - \mu)(\mu - \hat{\mu}) + (\mu - \hat{\mu})^2] \\ &= \sum (X_i - \mu)^2 - 2(\hat{\mu} - \mu) \sum (X_i - \mu) + n(\mu - \hat{\mu})^2 \\ &= \sum (X_i - \mu)^2 - 2n(\mu - \hat{\mu})^2 + n(\mu - \hat{\mu})^2 \\ &= \sum (X_i - \mu)^2 - n(\hat{\mu} - \mu)^2. \end{aligned}$$

Donc,

$$\begin{aligned} E_{\mu, \sigma^2} \left(\sum (X_i - \hat{\mu})^2 \right) &= E_{\mu, \sigma^2} \left(\sum (X_i - \mu)^2 - n(\hat{\mu} - \mu)^2 \right) \\ &= n\sigma^2 - n \frac{\sigma^2}{n} \\ &= (n-1)\sigma^2. \end{aligned}$$

Chapitre 10

Tests statistiques: introduction

Un test statistique est un procédé d'inférence: son but est d'énoncer des propriétés de la population en s'appuyant sur un échantillon d'observations. A l'aide d'un test on construit des intervalles de confiance qui expriment le degré d'incertitude associé à une estimation. Ce chapitre introduit les concepts nécessaires pour développer et appliquer les tests et les intervalles de confiance. Les chapitres suivants décrivent des tests et des intervalles de confiance spécifiques pour des situations fréquemment rencontrées en pratique.

10.1 Le concept de test statistique

L'étape préliminaire à la réalisation d'un test est la formulation d'*hypothèses* concernant les caractéristiques moyennes ou "paramètres" de la population en question. Un premier type d'hypothèse est l'*hypothèse nulle*, notée H_0 . En général, elle affirme que les paramètres ont des valeurs données (par exemple, suggérées par des études antérieures). L'utilisateur du test cherche à établir si son activité pourra se fonder sur cette hypothèse ou s'il vaudra mieux d'admettre que cette hypothèse est fautive. Dans ce cas, il préférera une *hypothèse alternative*, notée H_1 , qui nie H_0 .

Exemple

Le coût moyen d'un séjour dans un grand hôpital suisse pour "chirurgie cardiovasculaire" était de 11'000 Fr en 1999. Le directeur des finances doit établir le budget pour 2003; peut-il admettre que le coût moyen sera de 11'000 Fr ? Ne vaut-il pas mieux qu'il suppose qu'il sera supérieur ? En d'autres termes, si μ indique le coût moyen pour 2003, les hypothèses sont:

$$H_0 : \mu = 11'000, \quad H_1 : \mu > 11'000.$$

Pour choisir entre H_0 et H_1 on utilise un *test statistique*. Un test est un procédé qui permet de décider entre H_0 et H_1 sur la base d'un *échantillon d'observations*, d'une *statistique de test* et d'une *règle de décision*. La règle repose sur la statistique; elle doit permettre d'*accepter* (*confirmer*) l'hypothèse nulle ou de la *rejeter* (*infirmer*). Lorsque l'hypothèse nulle est rejetée, l'utilisateur se prononce en faveur de l'hypothèse alternative.

Exemple: continuation

Le directeur obtient un échantillon aléatoire de séjours en "chirurgie cardiovasculaire" du deuxième semestre 2002 ainsi que leurs coûts. Il calcule le coût moyen $\hat{\mu}$ et la différence

$$d_0 = \hat{\mu} - 11'000.$$

Cette différence est la statistique de test et le directeur utilise la règle suivante: "si $d_0 < 1'000$ Fr, accepter H_0 ; si $d_0 > 1'000$ Fr, rejeter H_0 et choisir H_1 ".

Deux types d'erreurs sont possibles (Figure 1):

- rejeter une hypothèse nulle vraie: *erreur de type I*,
- accepter une hypothèse nulle fausse: *erreur de type II*.

		R E A L I T E	
		H_0 vraie	H_0 fausse
D E C I S I O N	rejeter H_0	erreur type I	OK
	ne pas rejeter H_0	OK	erreur type II

Figure 1. Les types d'erreur

La règle de décision (par exemple, la limite de 1'000 Fr) doit être telle que la probabilité de commettre une erreur de type I est plus petite qu'un certain *niveau* ou *seuil* α préétabli (par exemple, $\alpha = 5\%$). Pour atteindre ce but il faudra calculer la distribution de la statistique de test en supposant que H_0 est correcte. Cette distribution s'appelle la *distribution nulle* de la statistique de test.

Pour effectuer ce calcul il est parfois nécessaire d'admettre que les données peuvent être décrites à l'aide d'un modèle ou qu'elles satisfont à certaines conditions. On dira alors que le calcul est effectué en s'appuyant sur des *conditions d'application*.

Exemple: continuation

Conditions d'application: (1) la distribution des coûts est lognormale (c'est-à-dire, il est raisonnable de décrire cette distribution à l'aide du modèle lognormal); (2) la distribution des coûts ne change pas entre le deuxième semestre 2002 et 2003.

La *probabilité d'erreur de type I* est donc

$$P_0(\text{rejeter } H_0),$$

où le suffixe "0" indique que le calcul est effectué en supposant que H_0 est correcte. Comme on l'a déjà dit, la règle de décision est choisie de façon à maintenir cette probabilité sous un certain niveau. D'autre part, si l'hypothèse H_1 est bien spécifiée (voir remarque ci-dessous) on pourra calculer aussi

$$P_1(\text{rejeter } H_0),$$

où le suffixe "1" indique que le calcul est effectué en supposant que H_1 est correcte. Cette probabilité s'appelle la *puissance* du test. C'est la probabilité de rejeter l'hypothèse nulle si l'alternative est correcte. En général, il est souhaitable que la puissance du test soit élevée (par exemple 95%) et on atteindra ce but en prenant un échantillon de "taille suffisamment élevée".

Remarques

1. Il n'est pas possible de calculer la puissance d'un test si on ne spécifie pas précisément H_1 . Par exemple, on ne peut pas effectuer des calculs sous l'alternative $\mu > 11'000$. Il faut spécifier une valeur "simple" de μ , par exemple $\mu = 13'000$ Fr.

2. Lorsqu'un test rejette une hypothèse, il n'est pas certain qu'elle soit fautive ! On ne peut même pas affirmer que, si un test de niveau α rejette H_0 , cette hypothèse est fautive avec probabilité $1 - \alpha$. En effet, il est impossible d'établir si une hypothèse est vraie ou fautive sans examiner la population de façon exhaustive. Toutefois, avant d'appliquer le test, le statisticien sait que la probabilité qu'il commette une erreur de type I est inférieure à α . Pour interpréter cela le statisticien peut imaginer d'appliquer des tests de seuil α (par exemple $\alpha = 5\%$) un grand nombre de fois pendant sa vie. Parmi toutes les hypothèses H_0 correctes qu'il aura testées à 5%, il en aura rejeté environ 5%. Mais il ne connaîtra jamais le nombre d'hypothèses correctes qu'il a testées.

Nous allons considérer deux exemples de tests de niveau α pour des situations très élémentaires. Ces exemples ne représentent pas des tests couramment utilisés en pratique. Leur but est d'illustrer les concepts.

Exemple: tester si une moyenne est égale à une valeur donnée

La moyenne de population $\mu(X)$ d'une certaine variable X est inconnue. Nous écrivons μ à la place de $\mu(X)$ et considérons les hypothèses

$$H_0 : \mu = \mu_0 \quad \text{et} \quad H_1 : \mu > \mu_0.$$

Par exemple, X est la taille des poissons du lac Léman. La taille moyenne μ est inconnue et un pêcheur considère les hypothèses $H_0: \mu = 5$ cm et $H_1: \mu > 5$ cm. Selon le cas, il choisit son filet.

Soit x_1, \dots, x_n un échantillon d'observations de X indépendantes (par exemple, les tailles de 30 poissons pris selon échantillonnage simple). Il est raisonnable d'utiliser la statistique

$$D(X_1, \dots, X_n) = \hat{\mu} - \mu_0$$

où $\hat{\mu}$ est la moyenne arithmétique $\hat{\mu}(X_1, \dots, X_n) = \sum X_i/n$. L'échantillon fournit une valeur observée de D , notée d_0 (par exemple, $\hat{\mu} = 7$ cm et $d_0 = 2$ cm). Est-ce que d_0 est suffisamment élevé pour rejeter H_0 ? Pour répondre à cette question il faut comparer d_0 à une certaine limite que nous allons déterminer. Considérons la statistique standardisée

$$Z(X_1, \dots, X_n) = \frac{\hat{\mu} - \mu_0}{\hat{\sigma}/\sqrt{n}},$$

où $\hat{\sigma}$ est l'écart type de x_1, \dots, x_n , et notons sa valeur observée par z_0 (par exemple, $\hat{\sigma} = 3.94$ et $z_0 = 2.78$). Grâce au théorème centrale limite, sous H_0 (et sous de faibles conditions d'application)

$$Z \sim \mathcal{N}(0, 1)$$

approximativement. En d'autres termes, la distribution nulle de Z est approximativement la distribution de Gauss standard. Soit donc $z_{1-\alpha}$ le quantile $1 - \alpha$ de cette distribution. Par exemple, pour $\alpha = 5\%$ on trouve $z_{0.95} = 1.645$. La règle de décision pourra enfin être formulée:

$$\text{rejeter } H_0 \text{ si } Z > z_{1-\alpha}.$$

Evidemment on applique cette règle à la valeur observée de Z . Par exemple, si $z_0 = 2.78$ il faut rejeter H_0 . Pour cette règle la probabilité d'erreur de type I est

$$P_0(\text{rejeter } H_0) \approx P(Z > z_{1-\alpha}) = \alpha$$

où P est calculé à l'aide de la distribution de Gauss.

Exemple: tester si une proportion est égale à une valeur donnée

Dans une population, le taux p d'individus avec une certaine caractéristique est inconnu: par exemple le taux de fumeurs en Suisse. Nous cherchons un test pour les hypothèses

$$H_0 : p = 50\%, \quad H_1 : p \neq 50\%.$$

Le seuil souhaité est $\alpha \approx 5\%$. Nous disposons d'un échantillon de 10 personnes prises au hasard (selon un plan d'échantillonnage simple).

Nous considérons la variable aléatoire

$$K = \text{nombre d'individus avec la caractéristique.}$$

K sera la statistique de test. Sous l'hypothèse H_0 (et si la taille de la population est très grande – condition d'application) elle suit un modèle binomial $\mathcal{B}(n = 10, p = 0.5)$ et sa distribution est donnée dans le tableau suivant.

$k:$	0	1	2	3	4	5	6	7	8	9	10
$P(K = k):$	0.0010	.0098	.0439	.1172	.2051	.2461	.2051	.1172	.0439	.0098	.0010

Nous remarquons que

$$P(K = 0 \text{ ou } K = 1 \text{ ou } K = 9 \text{ ou } K = 10) = 0.022,$$

tandis que

$$P(K = 0 \text{ ou } K = 1 \text{ ou } K = 2 \text{ ou } K = 8 \text{ ou } K = 9 \text{ ou } K = 10) = 0.109.$$

Nous définissons alors la règle de décision suivante:

$$\text{“Rejeter } H_0 \text{ si } K = 0 \text{ ou } K = 1 \text{ ou } K = 9 \text{ ou } K = 10\text{”}. \quad (10.1)$$

En effet, dans ce cas, la probabilité de rejeter H_0 si elle est vraie est $0.022 < 5\%$, tandis que si on rejetait H_0 pour $K \geq 8$ ou $K \leq 2$ cette probabilité deviendrait $0.109 > 5\%$.

Remarques

1. Le modèle binomial n'est pas applicable si la population a une taille N modérée par rapport à n (car dans ce cas, la probabilité que la deuxième, personne – ainsi que la troisième etc. – soit un fumeur n'est plus p).
2. Le même raisonnement peut servir à construire un test pour l'hypothèse

$$H_0 : \text{le taux } p \text{ est égal à une valeur donnée } p_0,$$

(où p_0 n'est pas nécessairement 50%) avec une taille d'échantillon n quelconque (mais beaucoup plus petit que la taille de la population). Evidemment, sous H_0 , il faudra considérer le modèle binomial $\mathcal{B}(n, p = p_0)$.

3. Souvent, la règle de décision est du type “rejeter H_0 si $S > c_\alpha$ ” où S est la statistique de test et c_α une constante qui dépend du niveau α . On appelle c_α une *valeur critique*.

10.2 Le p-value

Très souvent, surtout lorsqu'on utilise les logiciels statistiques, la fixation du niveau α ne précède pas la détermination de la règle de décision. En effet, un cheminement inverse est utilisé: on calcule, sous l'hypothèse nulle, la probabilité p d'obtenir "une valeur plus extrême que celle qu'on a observée". Si cette probabilité – dite *p-value* – est très petite, l'événement serait surprenant et alors l'hypothèse H_0 est rejetée.

Supposons que dans l'échantillon de l'exemple on ait observé 0 fumeurs. Calculons la probabilité $p = P(K = 0 \text{ ou } K = 10) = 0.002$ d'obtenir une valeur aussi extrême que celle observée, sous l'hypothèse H_0 . Cette probabilité est très petite: elle est inférieure à $\alpha = 5\%$. En d'autres termes, l'événement observé ne soutient pas l'hypothèse H_0 . Nous rejetons alors H_0 en faveur de l'alternative H_1 .

En général on peut directement définir la règle de décision à l'aide du p-value. Dans l'exemple du test pour la moyenne, la règle "rejeter H_0 si $Z > z_{1-\alpha}$ " est équivalente à "rejeter H_0 si le p-value est inférieur à α ".

La limite de ce qu'on appelle surprenant (rare) est arbitraire, mais dans beaucoup de domaines (biologie, médecine) on utilise assez systématiquement 5% (on lit " $p < 5\%$ " dans les publications). La borne de 5% peut être abaissée dans le cas où une erreur de type I pourrait avoir des conséquences jugées graves. Considérons par exemple le problème de comparer la survie moyenne de patients soumis à un procédé opératoire nouveau et très coûteux à la survie obtenue par un procédé traditionnel. Supposons que l'hypothèse nulle d'égalité entre les survies moyennes soit rejetée par un test statistique en faveur du nouveau traitement. Les conséquences financières d'une introduction généralisée du nouveau procédé pourraient être très lourdes en cas d'erreur de type I. Evidemment, tout dépend du point de vue !

10.3 Test unilatéral et test bilatéral

Considérons encore les exemples de la Section 10.1. Dans le cas du taux, nous avons développé un *test bilatéral*, c'est à dire, nous avons cherché à savoir si la fréquence de fumeurs dans l'échantillon était suffisamment "petite ou élevée" pour rejeter H_0 . Le rejet de H_0 était équivalent à l'acceptation de H_1 , c'est à dire, à affirmer que le taux de population est "inférieur ou supérieur" à 50%. Or, le chercheur a souvent en tête une alternative H_1 unilatérale; c'est à dire qu'il cherche à savoir si par exemple le taux p est "supérieur" à 50%. Il peut alors adopter un *test unilatéral* et rejeter H_0 seulement si le nombre de fumeurs dans l'échantillon est élevé, par exemple:

"Rejeter H_0 si $K = 9$ ou $K = 10$ ".

Cette règle est associée à une probabilité d'erreur de type I égale à $P(K = 9 \text{ ou } K = 10) = 0.011$, qui est la moitié de la probabilité d'erreur associée à la règle bilatérale. Evidemment, le chercheur peut aussi adopter la règle "Rejeter H_0 si $K = 8$ ou $K = 9$ ou $K = 10$ ". avec $P(K = 8 \text{ ou } K = 9 \text{ ou } K = 10) = 0.0547 \approx 5\%$. On remarque ainsi, que tout en gardant un seuil à 5%, il devient plus probable que l'alternative soit acceptée si elle est vraie: en d'autres termes, la puissance du test augmente.

Dans le cas du test pour la moyenne nous avons considéré l'hypothèse alternative unilatérale $H_1: \mu > \mu_0$ et nous avons utilisé la règle de décision unilatérale: rejeter H_0 si $Z > z_{1-\alpha}$. Pour tester H_0 contre l'alternative bilatérale $H_1: \mu \neq \mu_0$ au niveau α il faut utiliser la règle bilatérale

"Rejeter H_0 si $Z < z_{\alpha/2}$ ou si $Z > z_{1-\alpha/2}$ ".

10.4 L'intervalle de confiance

Considérons l'exemple du test concernant la moyenne $\mu(X)$ et supposons que $\hat{\mu}$ ait été observée, par exemple 7 cm. Quelle est la qualité de cette estimation ? Quelle mesure d'erreur pouvons-nous associer à cette estimation ? Une réponse nous est fournie par l'intervalle de confiance. Ce concept est fondé sur celui du test. En effet, ayant observé une certaine valeur de $\hat{\mu}$, imaginons de tester les hypothèses nulles

$$H_0 : \mu = \mu_0$$

pour toutes les valeurs de μ_0 comprises entre $-\infty$ et ∞ , et ceci avec probabilité d'erreur de type I inférieure à α (par exemple, 5%). Certaines de ces hypothèses seront rejetées, d'autres acceptées. L'ensemble des valeurs μ_0 acceptées forme un intervalle appelé *intervalle de confiance* pour $\mu(X)$ avec *coefficient de couverture* ou *coefficient de confiance* $1 - \alpha$ (par exemple, 95%).

Plus précisément, supposons que H_1 est bilatérale: $\mu \neq \mu_0$. L'hypothèse $H_0: \mu = \mu_0$ est donc acceptée si

$$z_{\alpha/2} < \frac{\hat{\mu} - \mu_0}{\hat{\sigma}/\sqrt{n}} < z_{1-\alpha/2}$$

c'est à dire, comme $z_{\alpha/2} = -z_{1-\alpha/2}$, si

$$\hat{\mu} - z_{1-\alpha/2}\hat{\sigma}/\sqrt{n} < \mu_0 < \hat{\mu} + z_{1-\alpha/2}\hat{\sigma}/\sqrt{n}.$$

L'ensemble des valeurs μ_0 qui satisfont ces inégalités est un intervalle de confiance pour μ avec coefficient de couverture $1 - \alpha$. Dans l'exemple, avec $\hat{\mu} = 7$ cm, $\hat{\sigma} = 3.94$ cm, $n = 30$, $z_{1-\alpha/2} = z_{0.975} = 1.960$, on trouve l'intervalle (5.59, 8.41).

Notons que les limites de l'intervalle sont des variables aléatoires et que, sous H_0 ,

$$P(\hat{\mu} - z_{1-\alpha/2}\hat{\sigma}/\sqrt{n} < \mu_0 < \hat{\mu} + z_{1-\alpha/2}\hat{\sigma}/\sqrt{n}) \approx 1 - \alpha.$$

En d'autres termes, avant de tirer l'échantillon, on sait que l'intervalle de confiance de coefficient $1 - \alpha$ couvrira le paramètre $\mu(X)$ avec probabilité $1 - \alpha$. Comment interpréter cette affirmation ? Si on observe un nombre très élevé d'échantillons et si chaque fois on calcule l'intervalle de confiance, une proportion $1 - \alpha$ de ces intervalles couvrira la valeur inconnue du paramètre ! De façon moins précise, on interprète un intervalle de confiance comme un ensemble de valeurs plausibles du paramètre étant donné l'échantillon.

Considérons encore l'exemple du test pour le taux p de fumeurs. Supposons que le nombre de fumeurs observé dans un échantillon de 10 personnes est $K = 8$. On peut alors construire un intervalle de confiance pour le paramètre p avec coefficient de couverture $1 - \alpha$ à l'aide d'un programme informatique qui calcule des tests de niveau α de $H_0: p = p_0$ contre $H_1: p \neq p_0$ pour toutes les valeurs p_0 comprises entre 0% et 100%. Certaines de ces hypothèses seront rejetées, d'autres acceptées. L'ensemble des valeurs p_0 acceptées forme un intervalle de confiance (p_g, p_d) . La proportion observée (80%) est approximativement au milieu de l'intervalle. A noter que l'on obtient des intervalles différents selon les valeurs de K ! L'interprétation de l'intervalle reste la même que dans l'exemple précédent.

Remarque

Un test bilatéral fournit une borne inférieure ainsi qu'une borne supérieure d'un intervalle de confiance. Un test unilatéral fournit une seule borne, l'autre est $-\infty$ ou $+\infty$.

Chapitre 11

Tests et intervalles de confiance pour proportions

Nous considérons ici un procédé pour tester l'hypothèse qu'une certaine proportion est égale à une valeur donnée et un autre pour tester l'hypothèse que deux proportions sont identiques. Ces tests sont très utilisés en pratique. Le calcul de la distribution nulle se base sur le théorème limite centrale.

11.1 Tests et intervalles de confiance pour une seule proportion

Comme dans le Chapitre 10 (exemple), nous nous intéressons au taux d'individus ayant une certaine caractéristique A dans une population; soit $p = P(A)$ cette proportion. Nous considérons un échantillon aléatoire de taille n ; soit K la variable aléatoire qui indique le nombre d'individus ayant A dans l'échantillon. La valeur observée de K est k . Soit K/n la proportion d'individus ayant A dans l'échantillon; la variable K/n est un estimateur de p et $\hat{p} = k/n$ est l'estimation de p fournie par cet estimateur.

Les inférences (tests et intervalles de confiance) concernant p peuvent être basées sur la distribution binomiale (voir Chapitres 7 et 10). Toutefois, lorsque n est suffisamment grand on préfère utiliser l'approximation normale fournie par le théorème limite centrale car les calculs sont plus aisés. Plus précisément, si $np > 5$ et $nq > 5$, où $q = 1 - p$, la variable aléatoire

$$Z = \frac{K/n - p}{\sqrt{pq/n}}$$

a approximativement une distribution de Gauss standard.

Pour effectuer un *test bilatéral* au niveau α de

$$H_0 : p = p_0 \quad \text{contre} \quad H_1 : p \neq p_0,$$

où p_0 indique une valeur spécifiée, on peut donc calculer la statistique de test

$$z = \frac{\hat{p} - p_0}{\sqrt{p_0 q_0/n}},$$

où $q_0 = 1 - p_0$, et rejeter H_0 si $|z| > z_{1-\alpha/2}$. La notation $z_{1-\alpha/2}$ indique le quantile $1 - \alpha/2$ de la distribution de Gauss standard. Par exemple, pour $\alpha = 5\%$ on a $z_{0.975} = 1.96$ (voir Table de la distribution de Gauss).

Pour effectuer un *test unilatéral* au niveau α de

$$H_0 : p = p_0 \quad \text{contre} \quad H_1 : p > p_0,$$

il suffira de rejeter H_0 si $z > z_{1-\alpha}$. On procédera de façon similaire pour tester l'alternative $H_1 : p < p_0$.

Exemple. Selon certaines sources, la proportion de patients avec troubles locomoteurs dans les hôpitaux est $p_0 = 0.05$. Dans un certain hôpital, on pense que ce taux est supérieur; on souhaite donc tester l'hypothèse $H_0 : p = 0.05$ contre l'alternative $H_1 : p > 0.05$. Dans un échantillon de 257 patients de cet hôpital on a trouvé 23 patients avec des troubles locomoteurs et donc $\hat{p} = 23/257 = 0.09$. La statistique de test vaut

$$z = \frac{0.09 - 0.05}{\sqrt{0.05 \times 0.95/257}} = 2.90 ,$$

Cette valeur excède le quantile 0.995 de la distribution normale standard (qui vaut 2.576, voir Tables) et la différence est donc significative au niveau (unilatéral) 0.5%.

Intervalle de confiance. Un intervalle de confiance bilatéral à coefficient $1 - \alpha$ pour la proportion p consiste en les valeurs de p qui ne seraient pas rejetées par un test bilatéral au niveau α choisi. Si le test est basé sur la statistique z , alors un intervalle de confiance avec coefficient environ égal à $1 - \alpha$ consiste en toutes les valeurs de p satisfaisant

$$\frac{|\hat{p} - p|}{\sqrt{p(1-p)/n}} \leq z_{1-\alpha/2} .$$

Les bornes de cet intervalle sont obtenues en élevant au carré et en résolvant l'équation du second degré qui en résulte. La borne inférieure est donc

$$p_i = \frac{1}{1+c} \left(\hat{p} + c/2 - \sqrt{c^2/4 + c\hat{p}(1-\hat{p})} \right) \quad (1)$$

avec $c = z_{1-\alpha/2}^2/n$, et la borne supérieure

$$p_s = \frac{1}{1+c} \left(\hat{p} + c/2 + \sqrt{c^2/4 + c\hat{p}(1-\hat{p})} \right) . \quad (2)$$

Evidemment, si la formule de p_i fournit un résultat négatif, on posera $p_i = 0$; analoguement, si la formule de p_s donne un résultat supérieur à 1, on posera $p_s = 1$.

Exemple. Supposons que $n = 257$ et $\hat{p} = 23/257$. Pour calculer un intervalle à 95% de confiance, il faut prendre $\alpha = 0.05$, $z_{1-\alpha/2} = z_{0.975} = 1.96$, d'où on obtient $p_i = 0.060$ et $p_s = 0.131$.

Les formules de p_i et p_s données ci-dessus sont utilisées lorsque la valeur observée de \hat{p} est proche de zéro ou de un. Lorsque \hat{p} est plus centré, disons $0.3 \leq \hat{p} \leq 0.7$, et $n \geq 50$ alors les formules plus simples suivantes peuvent être utilisées.

$$p_i = \hat{p} - z_{1-\alpha/2} \sqrt{\frac{\hat{p}\hat{q}}{n}} , \quad (3)$$

$$p_s = \hat{p} + z_{1-\alpha/2} \sqrt{\frac{\hat{p}\hat{q}}{n}} . \quad (4)$$

Exemple. Si $n = 257$ et $\hat{p} = 23/257$ on obtient $p_i = 0.055$, $p_s = 0.124$. L'intervalle est décalé vers la gauche par rapport à l'intervalle (0.060, 0.131) calculé selon (1)–(2).

11.2 Inférences relatives à deux proportions

Nous considérons maintenant deux populations et nous nous intéressons aux proportions p_1 et p_2 des individus ayant une certaine caractéristique A dans la première et la deuxième population respectivement. Nous utilisons les abréviations $q_1 = 1 - p_1$, et $q_2 = 1 - p_2$. Des échantillons aléatoires de taille n_1 et n_2 sont pris dans la première et la deuxième population. Soient K_1 et K_2 les nombres d'individus avec A dans chacun des échantillons; donc $K_1 \sim \mathcal{B}(n_1, p_1)$ et $K_2 \sim \mathcal{B}(n_2, p_2)$. Nous notons par k_1 et k_2 les valeurs observées de K_1 et K_2 et considérons les estimations $\hat{p}_1 = k_1/n_1$ et $\hat{p}_2 = k_2/n_2$ de p_1 et p_2 .

A l'aide du théorème limite centrale on démontre que, si $p_1 = p_2$ (et si n_1 et n_2 sont suffisamment grands), la variable aléatoire

$$Z = \frac{K_1/n_1 - K_2/n_2}{\sqrt{pq/n_1 + pq/n_2}}$$

a approximativement une distribution de Gauss standard. Ici, p est la valeur commune de p_1 et p_2 et $q = 1 - p$. Ces quantités peuvent être estimées par $\hat{p} = (k_1 + k_2)/(n_1 + n_2)$ et $\hat{q} = 1 - \hat{p}$, respectivement.

Pour effectuer un *test bilatéral* au niveau α de

$$H_0 : p_1 = p_2 \quad \text{contre} \quad H_1 : p_1 \neq p_2$$

on peut donc calculer la statistique de test

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}\hat{q}(1/n_1 + 1/n_2)}}$$

et rejeter H_0 si $|z| > z_{1-\alpha/2}$. Pour effectuer un *test unilatéral* au niveau α de

$$H_0 : p_1 = p_2 \quad \text{contre} \quad H_1 : p_1 > p_2$$

il suffira de rejeter H_0 si $z > z_{1-\alpha}$. On procédera de façon similaire pour tester l'alternative $H_1 : p_1 < p_2$.

Les données utilisées pour la comparaison de deux proportions peuvent être présentées sous forme d'un *tableau* 2×2 habituellement noté de la façon suivante.

Echantillon	Caractère A		Total
	Présent	Absent	
1	n_{11}	n_{12}	$n_{1.}$
2	n_{21}	n_{22}	$n_{2.}$

où $n_{11} = k_1$, $n_{12} = n_1 - k_1$, $n_{1.} = n_{11} + n_{12} = n_1$, $n_{21} = k_2$, $n_{22} = n_2 - k_2$, $n_{2.} = n_{21} + n_{22} = n_2$. On peut alors démontrer que

$$z^2 = \frac{n(n_{11}n_{22} - n_{12}n_{21})^2}{n_{1.}n_{2.}n_{.1}n_{.2}},$$

ayant défini $n_{.1} = n_{11} + n_{21}$, $n_{.2} = n_{12} + n_{22}$, $n = n_{11} + n_{12} + n_{21} + n_{22}$.

En outre, la condition $|z| > z_{1-\alpha/2}$ est équivalente à la condition $z^2 > \chi_{1,1-\alpha}^2$, où $\chi_{1,1-\alpha}^2$ représente le quantile $1 - \alpha$ de la distribution χ^2 à un degré de liberté (voir Tables). Certains quantiles de cette distribution sont donnés dans la table suivante.

α	:	0.10	0.05	0.025	0.01	0.005	0.001
$\chi_{1,1-\alpha}^2$:	2.71	3.84	5.02	6.63	7.88	10.83

Un test basé sur la condition $z^2 > \chi_{1,1-\alpha}^2$ est appelé un *test du chi carré*. Ce type de test est souvent utilisé pour étudier l'association entre deux caractères A et B , comme nous allons voir dans la section suivante.

11.3 Tester l'indépendance entre deux caractères

Supposons que chaque individu d'une population a ou n'a pas un certain caractère A et un certain caractère B . Nous utilisons ici les concepts et les notations du Chapitre 4, en particulier, $P(A)$, $P(B)$, $P(A \cap B)$, $P(A|B)$ et nous nous intéressons à tester l'hypothèse que A et B sont indépendants. Il y a trois façons de tester cette hypothèse; elles se distinguent par la méthode d'échantillonnage et par la formulation mathématique de l'hypothèse.

Etude prospective. L'hypothèse d'indépendance peut être exprimée de la façon suivante:

$$H_0 : P(B|A) = P(B|\bar{A}).$$

Pour la tester on prend deux échantillons aléatoires de tailles fixées n_1 et n_2 ; le premier est pris dans la sous-population avec A , le deuxième dans la sous-population sans A . Dans chaque échantillon on compte les individus avec et sans B . On peut alors présenter les résultats comme indiqué dans le tableau 2×2 suivant.

Nombre d'individus avec et sans B dans les deux échantillons

	B	\bar{B}	Total
A	n_{11}	n_{12}	$n_{1.}$
\bar{A}	n_{21}	n_{22}	$n_{2.}$

Les probabilités $P(B|A)$ et $P(B|\bar{A})$ sont estimées par $\hat{p}_1 = n_{11}/n_{1.}$ et $\hat{p}_2 = n_{21}/n_{2.}$ et le test bilatéral décrit dans la Section 11.2 est appliqué avec $\hat{q}_1 = 1 - \hat{p}_1$, $\hat{q}_2 = 1 - \hat{p}_2$, $\hat{p} = (n_{11} + n_{21})/(n_{1.} + n_{2.})$, $\hat{q} = 1 - \hat{p}$ et

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}\hat{q}(1/n_{1.} + 1/n_{2.})}} \quad \text{ou} \quad z^2 = \frac{n(n_{11}n_{22} - n_{12}n_{21})^2}{n_{1.}n_{2.}n_{.1}n_{.2}}.$$

Ici, $n_{.1} = n_{11} + n_{21}$ et $n_{.2} = n_{12} + n_{22}$, $n = n_{11} + n_{12} + n_{21} + n_{22}$.

Le terme "étude prospective" est utilisé en épidémiologie lorsque A indique la présence d'un "facteur antécédant" ou un "facteur de risque" (par exemple, fumer) et B la présence d'une certaine maladie (par exemple, le cancer du poumon). Dans ce type d'étude, un échantillon d'individus avec A et un échantillon d'individus sans A sont suivis pendant un certain temps et on détermine combien parmi ces deux groupes auront développé la maladie. Les tailles des deux échantillons sont fixées de façon à obtenir une puissance suffisante.

Etude rétrospective. L'hypothèse d'indépendance peut aussi être exprimée de la façon suivante:

$$H_0 : P(A|B) = P(A|\bar{B}).$$

Pour la tester on prend deux échantillons aléatoires de tailles fixées $n_{.1}$ et $n_{.2}$; le premier est pris dans la sous-population avec B , le deuxième dans la sous-populations sans B . Dans chaque échantillon on compte les individus avec et sans A . On peut alors présenter les résultats comme indiqué dans le tableau 2×2 suivant.

Nombre d'individus avec et sans A dans les deux échantillons

	B	\bar{B}	
A	n_{11}	n_{12}	
\bar{A}	n_{21}	n_{22}	
Total	$n_{.1}$	$n_{.2}$	

Les probabilités $P(A|B)$ et $P(A|\bar{B})$ peuvent être estimées par $\hat{p}_1 = n_{11}/n_{.1}$ et $\hat{p}_2 = n_{12}/n_{.2}$ et le test bilatéral décrit dans la Section 11.2 est appliqué avec $\hat{q}_1 = 1 - \hat{p}_1$, $\hat{q}_2 = 1 - \hat{p}_2$, $\hat{p} = (n_{11} + n_{12})/(n_{.1} + n_{.2})$, $\hat{q} = 1 - \hat{p}$ et

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}\hat{q}(1/n_{.1} + 1/n_{.2})}} \quad \text{ou} \quad z^2 = \frac{n(n_{11}n_{22} - n_{12}n_{21})^2}{n_{.1}n_{.2}n_{.1}n_{.2}}.$$

Ici, $n_{.1} = n_{11} + n_{12}$, $n_{.2} = n_{21} + n_{22}$, $n = n_{11} + n_{12} + n_{21} + n_{22}$.

Le terme "étude rétrospective" est utilisée en épidémiologie lorsque deux échantillons aléatoires sont pris, le premier dans une population d'individus ("cas") ayant contracté la maladie et le deuxième parmi des individus ("contrôles") ne l'ayant pas eue. On détermine alors combien de ceux-ci avaient le facteur A suspect par le passé. Parfois il est nécessaire d'utiliser les dossiers médicaux d'individus décédés à cause de la maladie B ou sans la maladie B et de rechercher la présence de A dans les informations disponibles. La qualité de ces informations n'est pas toujours bonne.

Etude transversale. Enfin, l'hypothèse d'indépendance peut être exprimée de la façon suivante:

$$\begin{aligned} H_0 : P(A \cap B) &= P(A)P(B), \\ P(A \cap \bar{B}) &= P(A)P(\bar{B}), \\ P(\bar{A} \cap B) &= P(\bar{A})P(B), \\ P(\bar{A} \cap \bar{B}) &= P(\bar{A})P(\bar{B}). \end{aligned}$$

Pour la tester on prend un seul échantillon de taille n tiré de façon aléatoire de la population entière et on classe les individus en fonction de la présence ou l'absence de A et B . (Les totaux marginaux sont donc aléatoires.) On obtient la table de comptage suivante.

Nombre d'individus selon leur caractères A et B

	B	\bar{B}	Total
A	n_{11}	n_{12}	$n_{.1}$
\bar{A}	n_{21}	n_{22}	$n_{.2}$
Total	$n_{.1}$	$n_{.2}$	$n_{..} = n$

Les probabilités conjointes et les probabilités marginales sont estimées par des proportions:

proportion	estimation de
$\hat{p}_{11} = n_{11}/n_{..}$	$P(A \cap B)$
$\hat{p}_{12} = n_{12}/n_{..}$	$P(A \cap \bar{B})$
$\hat{p}_{21} = n_{21}/n_{..}$	$P(\bar{A} \cap B)$
$\hat{p}_{22} = n_{22}/n_{..}$	$P(\bar{A} \cap \bar{B})$
$\hat{p}_{1.} = n_{1.}/n_{..}$	$P(A)$
$\hat{p}_{2.} = n_{2.}/n_{..}$	$P(\bar{A})$
$\hat{p}_{.1} = n_{.1}/n_{..}$	$P(B)$
$\hat{p}_{.2} = n_{.2}/n_{..}$	$P(\bar{B})$

Sous H_0 , les probabilités conjointes peuvent aussi être estimées par les produits des proportions marginales; il faut donc s'attendre à des petits écarts

$$\hat{p}_{11} - \hat{p}_{1.}\hat{p}_{.1}, \quad \hat{p}_{21} - \hat{p}_{2.}\hat{p}_{.1}, \quad \hat{p}_{12} - \hat{p}_{1.}\hat{p}_{.2}, \quad \hat{p}_{22} - \hat{p}_{2.}\hat{p}_{.2}.$$

Plus précisément, on considère la statistique de test

$$s = \frac{(\hat{p}_{11} - \hat{p}_{1.}\hat{p}_{.1})^2}{(\hat{p}_{1.}\hat{p}_{.1}/n_{..})} + \frac{(\hat{p}_{12} - \hat{p}_{1.}\hat{p}_{.2})^2}{(\hat{p}_{1.}\hat{p}_{.2}/n_{..})} + \frac{(\hat{p}_{21} - \hat{p}_{2.}\hat{p}_{.1})^2}{(\hat{p}_{2.}\hat{p}_{.1}/n_{..})} + \frac{(\hat{p}_{22} - \hat{p}_{2.}\hat{p}_{.2})^2}{(\hat{p}_{2.}\hat{p}_{.2}/n_{..})}.$$

On démontre alors que

$$s = \frac{n(n_{11}n_{22} - n_{12}n_{21})^2}{n_{1.}n_{2.}n_{.1}n_{.2}} = z^2.$$

En outre, sous H_0 , z^2 est la valeur observée d'une variable aléatoire Z^2 qui suit une distribution χ^2 à 1 degré de liberté. Au niveau α , on pourra donc rejeter H_0 si $z^2 > \chi_{1,1-\alpha}^2$.

11.4 Corrections de continuité

Lorsque $K \sim \mathcal{B}(n, p)$ le théorème limite centrale fournit l'approximation

$$P(K \leq k) \approx \Phi \left((k - np) / \sqrt{np(1-p)} \right).$$

Cette approximation peut être améliorée en utilisant une correction appelée *correction de continuité* qui consiste à ajouter 0.5 à k , c'est-à-dire:

$$P(K \leq k) \approx \Phi \left((k + 0.5 - np) / \sqrt{np(1-p)} \right).$$

Les formules (1)–(2) peuvent être modifiées pour tenir compte de cette correction (Blyth, 1986):

$$p_i = \frac{1}{1+c} \left(\hat{p}_+ + c/2 - \sqrt{c^2/4 + c\hat{p}_+(1-\hat{p}_+)} \right),$$

$$p_s = \frac{1}{1+c} \left(\hat{p}_- + c/2 + \sqrt{c^2/4 + c\hat{p}_-(1-\hat{p}_-)} \right),$$

où $c = z_{1-\alpha/2}^2/n$, $\hat{p}_+ = (k + 0.5)/n$ et $\hat{p}_- = (k - 0.5)/n$.

La formule de z^2 sera modifiée de la façon suivante:

$$z^2 = \frac{n(|n_{11}n_{22} - n_{12}n_{21}| - \frac{1}{2}n)^2}{n_{1.}n_{2.}n_{.1}n_{.2}}.$$

11.5 Exemples

Nous allons illustrer, à l'aide d'un même exemple, l'étude transversale, l'étude prospective et l'étude rétrospective. L'exemple est fictif: on désire analyser l'association éventuelle entre l'âge maternel et le poids à la naissance de l'enfant. Soit:

$A = \text{âge maternel} \leq 20 \text{ ans}$

$B = \text{poids à la naissance} \leq 2500g.$

Afin de limiter l'effet d'autres facteurs tels que l'origine, la classe socio-économique, etc, on suppose que l'échantillonnage se fait dans une tranche bien déterminée de la population. (Toutes les valeurs de z^2 sont calculées avec la correction de continuité.)

Etude transversale

Dans ce cas on fixe la taille totale de l'échantillon, 200 dans notre exemple. Supposons que les données obtenues soient celles présentées dans la table suivante.

Age maternel	Poids à la naissance		Total
	$\leq 2500g$	$> 2500g$	
≤ 20	10	40	50
> 20	15	135	150
Total	25	175	200

Les données seront de préférence montrées sous forme de proportions conjointes (en divisant chaque valeur de la table par la taille totale).

Age maternel	Poids à la naissance		Total
	$\leq 2500g$	$> 2500g$	
≤ 20	.050 (= \hat{p}_{11})	.200 (= \hat{p}_{12})	.25 (= $\hat{p}_{1.}$)
> 20	.075 (= \hat{p}_{21})	.675 (= \hat{p}_{22})	.75 (= $\hat{p}_{2.}$)
Total	.125 (= $\hat{p}_{.1}$)	.875 (= $\hat{p}_{.2}$)	1.

On trouve

$$z^2 = \frac{200(|10 \times 135 - 15 \times 40| - 200/2)^2}{50 \times 150 \times 25 \times 175} = 2.58,$$

qui n'atteint pas la valeur critique 2.71 (au niveau $\alpha = 0.10$).

Etude prospective

Dans une étude prospective, les populations sont définies par le facteur antécédent: on connaît l'âge de la mère avant la naissance de l'enfant. On fixe la taille des échantillons à prendre dans chacune des populations: par exemple, on prend 100 mères de moins de 20 ans et 100 mères de plus de 20 ans, puis on observe les poids à la naissance de leurs enfants.

Imaginons que les résultats soient parfaitement cohérents avec ceux de l'étude transversale, c'est-à-dire que le taux de bébés de très petit poids parmi les jeunes mères est

$$P(B|A) = p_{11}/p_{1.} = .05/.25 = .20 (= 20\%)$$

et de façon semblable, pour les mères de plus de 20 ans,

$$P(B|\bar{A}) = p_{21}/p_{2.} = .075/.75 = .10 (= 10\%) .$$

Il en résulte donc la table ci-dessous.

Age maternel	Poids à la naissance		Total	Proportion de faibles poids à la naissance
	$\leq 2500g$	$> 2500g$		
≤ 20	20	80	100	.20 [= $\hat{p}(B A)$]
> 20	10	90	100	.10 [= $\hat{p}(B \bar{A})$]
Total	30	170	200	

Pour cette table on obtient

$$z^2 = 3.18,$$

qui se situe entre la valeur critique 2.71 (au niveau $\alpha = 0.10$) et la valeur critique 3.84 (au niveau $\alpha = 0.05$).

Etude rétrospective

Ici les populations sont définies par l'événement d'intérêt: poids inférieur à 2500g ou pas. On fixe les tailles des échantillons à prendre dans chacune des populations: par exemple, 100 bébés de moins de 2500g et 100 bébés de plus de 2500g, puis on détermine les âges de leurs mères.

Supposons que les résultats soient une fois encore parfaitement cohérents avec ceux de l'étude transversale, c'est-à-dire que la proportion de jeunes mères parmi les enfants de très petit poids est

$$P(A|B) = p_{11}/p_{.1} = .05/.125 = .40 (= 40\%),$$

et de même, pour les enfants de plus de 2500g,

$$P(A|\bar{B}) = p_{12}/p_{.2} = .20/.875 = .23 (= 23\%).$$

Donc on aurait trouvé la table ci-dessous.

Poids à la naissance	Age maternel		Total	Proportion de très jeunes mères
	≤ 20 ans	> 20 ans		
$\leq 2500g$	40	60	100 (= N_B)	.40 [= $\hat{p}(A B)$]
$> 2500g$	23	77	100 (= $N_{\bar{B}}$)	.23 [= $\hat{p}(A \bar{B})$]
Total	63	137	200	

Cette fois,

$$z^2 = 5.93,$$

indiquant que l'on peut rejeter l'hypothèse de non association au niveau $\alpha = 0.025$.

Cette "évolution" des valeurs des statistiques de test est remarquable car les effectifs totaux étaient les mêmes (200 dans les 3 cas) et les tableaux ont été construits en utilisant les mêmes probabilités sous-jacentes. En fait, il est généralement vrai qu'une étude rétrospective avec 2 échantillons de même taille est plus puissante qu'une étude transversale avec le même effectif total. De plus, si la fréquence de l'événement d'intérêt (B) est plus extrême (plus loin de 50%) que la fréquence du facteur antécédent (A), alors l'étude rétrospective avec 2 échantillons de même taille est plus puissante qu'une étude prospective avec 2 échantillons de même taille.

Chapitre 12

Tests et intervalles de confiance pour moyennes

Dans ce chapitre nous considérons les techniques d'inférence usuelles pour la moyenne d'une variable aléatoire quantitative et le test de Student pour comparer les moyennes de deux variables quantitatives.

12.1 Inférences relatives à une seule moyenne

Soient X_1, \dots, X_n les variables aléatoires qui représentent un échantillon. Supposons que X_1, \dots, X_n soient i.i.d. et que $E(X_i) = \mu$, $\sigma^2(X_i) = \sigma^2$. Soient x_1, \dots, x_n les valeurs observées de X_1, \dots, X_n . Nous nous intéressons aux problèmes suivants:

- estimer μ ;
- tester l'hypothèse

$$H_0 : \mu = \mu_0,$$

où μ_0 est une valeur donnée (par exemple, 0), contre une des alternatives suivantes:

$$H_1 : \mu \neq \mu_0 \quad \text{ou bien} \quad H_1 : \mu > \mu_0 \quad \text{ou bien} \quad H_1 : \mu < \mu_0;$$

- construire un intervalle de confiance pour μ .

Les procédés paramétriques usuels se basent sur la condition d'application suivante:

$$C : X_i \sim \mathcal{N}(\mu, \sigma^2), \quad i = 1, \dots, n.$$

Ils utilisent les statistiques

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

comme estimateurs de μ et de σ .

Résultat théorique

Sous C , la variable aléatoire

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

a une distribution t à $(n-1)$ degrés de liberté.

La distribution t a une densité symétrique qui dépend d'un paramètre à valeurs entières appelé degrés de liberté (Chapitre 7). Lorsque la valeur de ce paramètre est fixée, la distribution est complètement spécifiée. Soit $t_{1-\alpha, n-1}$ le percentile $1-\alpha$ de cette distribution. D'après la table on a par exemple:

$$t_{95\%, 9} = 1.833, \quad t_{97.5\%, 9} = 2.262, \quad t_{5\%, 9} = -1.833, \quad t_{2.5\%, 9} = -2.262.$$

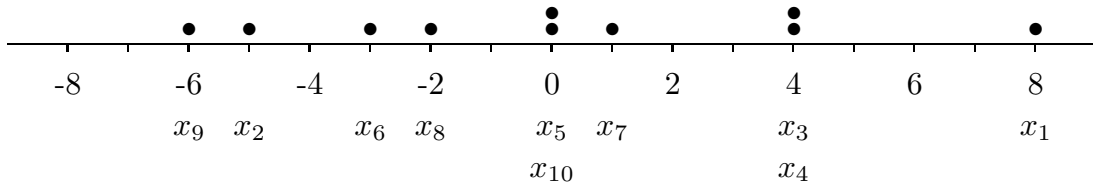
Dans les applications, les pas suivants sont nécessaires.

Premier pas: analyse graphique

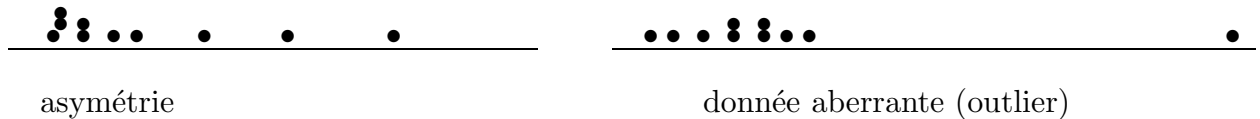
Il convient d'abord de représenter les données graphiquement dans le but de s'assurer que la condition d'application ne soit pas clairement violée.

Exemple 1. Les valeurs observées x_i sont:

$$x_1 = 8, x_2 = -5, x_3 = 4, x_4 = 4, x_5 = 0, x_6 = -3, x_7 = 1, x_8 = -2, x_9 = -6, x_{10} = 0.$$



L'analyse graphique de ces données ne suggère pas de violations de la condition C . Par contre, il ne serait pas approprié d'appliquer les procédés paramétrique basés sur la condition C aux données suivantes:



Dans le cas d'une distribution asymétrique, un autre modèle paramétrique pourrait être pris en considération; dans le cas d'une distribution symétrique "contaminée" par des outliers un procédé non-paramétrique (Chapitre 13) pourrait être approprié.

Deuxième pas: calcul

Calculer

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i, \quad \hat{\sigma} = \sqrt{\frac{1}{n-1} \sum (x_i - \bar{x})^2}, \quad \hat{\sigma}_{\bar{x}} = \frac{\hat{\sigma}}{\sqrt{n}}$$

et

$$t = \frac{\bar{x} - \mu_0}{\hat{\sigma}_{\bar{x}}}.$$

Troisième pas: règle de décision pour le test au niveau α

- Rejeter H_0 en faveur de $H_1 : \mu \neq \mu_0$ si $t > t_{1-\alpha/2, n-1}$ ou $t < -t_{1-\alpha/2, n-1}$.
- Rejeter H_0 en faveur de $H_1 : \mu > \mu_0$ si $t > t_{1-\alpha, n-1}$.
- Rejeter H_0 en faveur de $H_1 : \mu < \mu_0$ si $t < -t_{\alpha, n-1}$.

Ce procédé est connu comme le *test de Student* ou *t-test pour un seul échantillon*.

Exemple 1 (continuation). Pour $H_0 : \mu_0 = 0$, on obtient: $\bar{x} = 0.1$, $\hat{\sigma} = 4.35$, $\hat{\sigma}_{\bar{x}} = 1.38$ et $t = 0.072$. Donc, $-2.262 < t < 2.262$ et on ne peut pas rejeter H_0 au niveau 10%.

Intervalle de confiance pour μ

Le résultat théorique mentionné implique que

$$P(\bar{X} - t_{1-\alpha/2, n-1} \hat{S} / \sqrt{n} \leq \mu \leq \bar{X} + t_{1-\alpha/2, n-1} \hat{S} / \sqrt{n}) = 1 - \alpha.$$

L' intervalle

$$(\bar{x} - t_{1-\alpha/2, n-1} \hat{\sigma}_{\bar{x}}, \quad \bar{x} + t_{1-\alpha/2, n-1} \hat{\sigma}_{\bar{x}})$$

est donc un intervalle de confiance avec coefficient de couverture $1 - \alpha$ pour μ .

Exemple 1 (continuation). On a $t_{97.5\%, 9} \hat{\sigma}_{\bar{x}} = 3.12$. L'intervalle $(-3.02, 3.22)$ est un intervalle de confiance avec coefficient de couverture 95% pour μ .

12.2 Données appariées et non appariées

Par la suite, nous considérons le problème de comparer deux ensembles de données. Selon leur structure, il faut distinguer deux cas.

Données appariées

Deux mesures sont effectuées pour chaque unité d'observation.

Exemple 2. Est-ce que la concentration de cholestérol, triglycérides et d'autres substances se modifie si des échantillons de sang sont conservés pendant un certain temps ? Evidemment, la réponse à cette question est une information importante pour l'organisation du travail de laboratoire. Dans une étude publiée, les échantillons de sang de 10 sujets d'une certaine population ont été analysés immédiatement après la prise de sang et 8 mois après. Les mesures sont donc appariées.

Avant:	74	80	75	136	104	102	90	100	95	84
Après:	66	85	71	132	104	105	89	102	101	84

On se demande si les deux mesures de chaque échantillon sont suffisamment éloignées pour qu'on puisse décider qu'il y a un effet de la conservation.

Autre exemple: dans l'étude de l'effet d'un traitement il convient souvent d'apparier des sujets similaires. Chaque sujet traité est comparé à un sujet non-traité: le sexe, l'âge, la gravité de la maladie et tout autre facteur contrôlable pouvant avoir une influence sur la réponse sont identiques chez les deux sujets. Les deux sujets appariés constituent l'unité d'observation.

Données non-appariées

On considère deux ensembles de mesures non-appariées.

Exemple 3. Concentration lipidique chez des sujets avec un problème circulatoire (insuffisance artérielle périphérique des membres inférieurs) et chez des sujets sains:

Sains:	4.90	5.40	5.60	5.90	6.20	6.75		
Malades:	5.40	6.00	6.25	6.50	6.60	6.75	7.40	7.90

12.3 Le t-test pour données appariées

On calcule les différences entre les données appariées et on réduit l'analyse des données originales à l'analyse des différences. Le t-test se base sur la condition d'application

$$C : D_i \sim \mathcal{N}(\mu, \sigma^2), \quad i = 1, \dots, n,$$

où D_1, \dots, D_n sont les variables aléatoires qui représentent les différences. On teste $H_0 : \mu = \mu_0$ où $\mu_0 = 0$ en utilisant les techniques décrites dans la Section 12.1.

Exemple 2 (continuation). Les différences d_i entre les paires de mesures sont $d_1 = 8, d_2 = -5, d_3 = 4, d_4 = 4, d_5 = 0, d_6 = -3, d_7 = 1, d_8 = -2, d_9 = -6, d_{10} = 0$. Ces différences coïncident avec les données analysées dans l'Exemple 1.

12.4 Le t-test pour données non-appariées

Soient X_1, \dots, X_m les variables aléatoires qui représentent les observations du premier échantillon, Y_1, \dots, Y_n celles qui représentent le deuxième. Soient $x_1, \dots, x_m, y_1, \dots, y_n$ les données.

Condition d'application

Les deux échantillons proviennent de deux populations Gaussiennes avec la même variance:

$$C : X_i \sim \mathcal{N}(\mu_1, \sigma^2) \quad \text{et} \quad Y_i \sim \mathcal{N}(\mu_2, \sigma^2).$$

En outre, les X_i et les Y_i sont indépendantes.

Si cette condition est admise, la comparaison est réduite au test de l'hypothèse

$$H_0 : \mu_1 = \mu_2$$

contre l'une des alternatives suivantes:

$$H_1 : \mu_1 \neq \mu_2 \quad \text{ou bien} \quad H_1 : \mu_1 < \mu_2 \quad \text{ou bien} \quad H_1 : \mu_1 > \mu_2.$$

Le t-test pour données non-appariées utilise les statistiques

$$\bar{X} = \sum_{i=1}^m X_i/m, \quad \bar{Y} = \sum_{j=1}^n X_j/n,$$

$$S_x = \left[\sum_{i=1}^m (X_i - \bar{X})^2 / (m-1) \right]^{1/2}, \quad S_y = \left[\sum_{i=1}^n (Y_i - \bar{Y})^2 / (n-1) \right]^{1/2}.$$

Résultat théorique

Si la condition d'application C est satisfaite, et si l'hypothèse nulle H_0 est vraie, alors la statistique

$$T = \frac{D}{S_D}$$

où

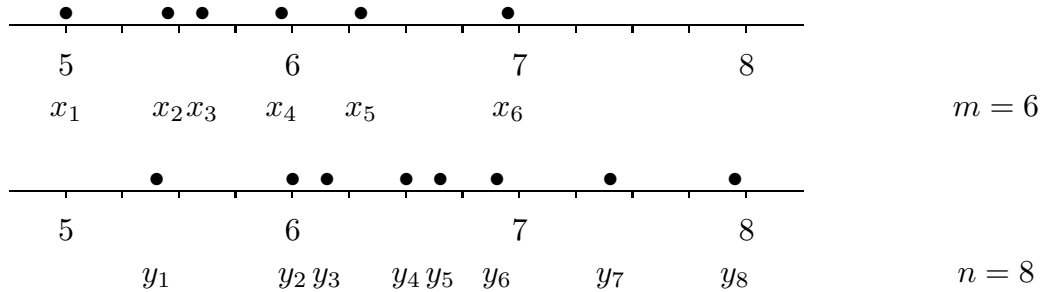
$$D = \bar{X} - \bar{Y}, \quad S_D = \sqrt{\frac{1}{m} + \frac{1}{n}} \sqrt{\frac{(m-1)S_x^2 + (n-1)S_y^2}{(m-1) + (n-1)}},$$

suit une distribution t à $(m+n-2)$ degrés de liberté. (S_D est un estimateur de $\sigma(D)$.)

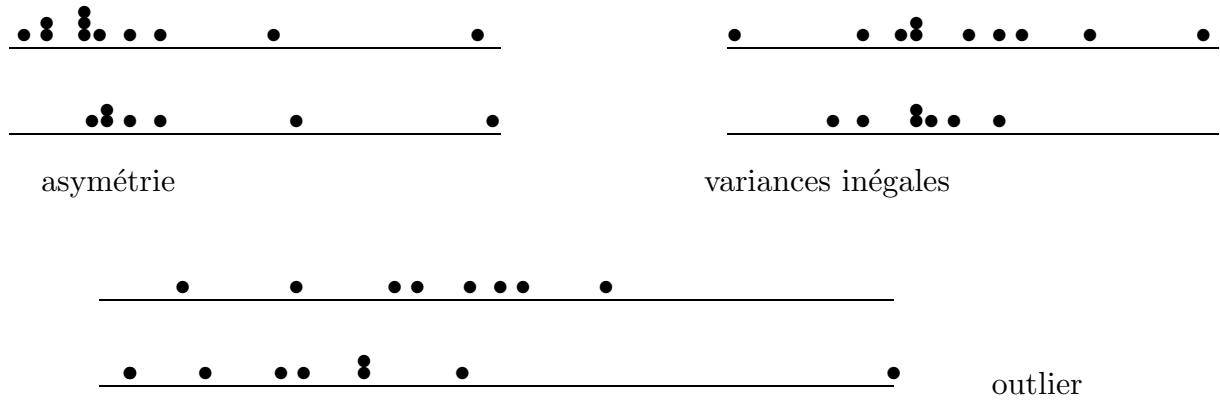
Premier pas: analyse graphique

Il convient d'abord de représenter les données graphiquement dans le but de déterminer si la condition d'application est approximativement satisfaite. Si oui, le t-test pourra être utilisé. Si non, un autre procédé sera nécessaire.

Exemple 3 (continuation).



L'analyse graphique de ces données ne suggère pas de violations de la condition C . Par contre, il ne serait pas approprié d'appliquer le t-test aux données suivantes:

*Deuxième pas: calcul*

Calculer

$$\bar{x} = \sum_{i=1}^m x_i/m, \quad \bar{y} = \sum_{i=1}^n y_i/n, \quad d = \bar{x} - \bar{y},$$

$$\hat{\sigma}_x = \sqrt{\frac{1}{m-1} \sum (x_i - \bar{x})^2}, \quad \hat{\sigma}_y = \sqrt{\frac{1}{n-1} \sum (y_i - \bar{y})^2},$$

$$\hat{\sigma}_d = \sqrt{\frac{1}{m} + \frac{1}{n}} \sqrt{\frac{(m-1)\hat{\sigma}_x^2 + (n-1)\hat{\sigma}_y^2}{(m-1) + (n-1)}},$$

$$t = \frac{d}{\hat{\sigma}_d}.$$

Troisième pas: règle de décision pour le test au niveau α

- Rejeter H_0 en faveur de $H_1 : \mu_1 \neq \mu_2$ si $t > t_{1-\alpha/2, m+n-2}$ ou $t < -t_{1-\alpha/2, m+n-2}$.
- Rejeter H_0 en faveur de $H_1 : \mu_1 < \mu_2$ si $t < -t_{1-\alpha, m+n-2}$.
- Rejeter H_0 en faveur de $H_1 : \mu_1 > \mu_2$ si $t > t_{1-\alpha, m+n-2}$.

Exemple 3 (continuation). On obtient:

$$\bar{x} = 5.79; \quad \bar{y} = 6.60; \quad d = -0.81; \quad \hat{\sigma}_x = 0.782; \quad \hat{\sigma}_y = 0.645; \quad \hat{\sigma}_d = 0.393; \quad t = -2.06.$$

En outre, $t_{95\%,12} = 1.782$, $t_{97.5\%,12} = 2.179$ et $-2.179 < t < -1.782$. On peut donc rejeter H_0 en faveur de $H_1 : \mu_1 < \mu_2$ au niveau 5%, mais il n'est pas possible de rejeter H_0 au niveau 2.5%.

Intervalle de confiance pour $\mu_1 - \mu_2$

L'intervalle

$$(d - t_{1-\alpha/2, m+n-2} \hat{\sigma}_d, \quad d + t_{1-\alpha/2, m+n-2} \hat{\sigma}_d)$$

est un intervalle de confiance avec coefficient de confiance $1 - \alpha$ pour la différence des moyennes $\mu_1 - \mu_2$.

Chapitre 13

Tests nonparamétriques pour un et deux échantillons

Ce chapitre décrit brièvement deux tests qui peuvent être utilisés dans des situations similaires à celles traitées dans le Chapitre 12. Toutefois, les conditions d'application sont moins exigeantes: en particulier, il n'est pas requis que les données proviennent de populations Gaussiennes. Les statistiques de test n'utilisent pas directement les valeurs observées mais seulement leur rang, c'est-à-dire leur position dans l'échantillon ordonné. Ces statistiques sont donc peu influencées par les outliers. Le livre de Lehmann (1975) est recommandé à tous ceux qui souhaitent approfondir le thème des tests non-paramétriques.

13.1 Le test de Wilcoxon pour données appariées

Comme pour le test paramétrique (Chapitre 12, Section 12.3), l'analyse de deux ensembles de données appariées se réduit à l'analyse d'un seul échantillon, celui des différences. Notons par D_1, \dots, D_n les variables aléatoires qui représentent ces différences et par d_1, \dots, d_n leurs valeurs observées.

Condition d'application

Les différences D_1, \dots, D_n sont i.i.d. selon une distribution symétrique centrée autour d'une certaine valeur Δ .

La forme de la distribution n'est pas spécifiée à l'aide d'un modèle. Le paramètre d'intérêt est la différence moyenne $\Delta = E(D_i)$ et l'hypothèse nulle est

$$H_0 : \Delta = 0.$$

Si H_0 est rejetée une des hypothèses alternatives suivantes est acceptée:

$$H_1 : \Delta \neq 0 \quad \text{ou bien} \quad H_1 : \Delta > 0 \quad \text{ou bien} \quad H_1 : \Delta < 0.$$

Une analyse graphique suffit généralement pour s'assurer que la condition d'application ne soit pas violée. Le test de Wilcoxon (connu aussi comme *test des rangs signés*) ne devra pas être utilisé si la distribution des différences est clairement asymétrique; toutefois, la présence de quelques outliers n'aura pas de conséquences importantes sur la décision.

Les statistiques de test

Les pas suivants sont nécessaires.

1. Eliminer les d_i qui sont nuls et réduire n conformément.
2. Ranger en ordre croissant les valeurs absolues des d_i (ignorer leur signe).
3. Attribuer le rang 1 à la plus petite valeur, le rang 2 à la suivante, etc.
Si deux ou plusieurs valeurs sont identiques, on leur donne un *rang moyen*, défini comme la moyenne des rangs que ces valeurs auraient si elles étaient différentes.
4. Signer les rangs, par exemple en plaçant un signe “-” derrière les rangs négatifs.
5. Calculer les statistiques de test v_r et v_s : si r_1, \dots, r_n désignent les rangs (moyens),

$$v_s = \sum_{d_i > 0}^n r_i \quad \text{et} \quad v_r = \sum_{d_i < 0}^n r_i.$$

v_s est la somme des “rangs positifs” et v_r la somme des “rangs négatifs”.

Soient V_r et V_s les variables aléatoires dont les valeurs observées sont v_s et v_r . Notons que $V_r + V_s = n(n+1)/2$.

Règles de décision au niveau α

Pour un test unilatéral, déterminer à l'aide des résultats théoriques mentionnés ci-dessous, la valeur a telle que $P(V_r \geq a) = P(V_s \geq a) \approx \alpha$.

- Rejeter H_0 en faveur de $H_1 : \Delta > 0$ si $v_s > a$.
- Rejeter H_0 en faveur de $H_1 : \Delta < 0$ si $v_r > a$.

Pour un test bilatéral, déterminer a tel que $P(V_r \geq a) = P(V_s \geq a) \approx \alpha/2$.

- Rejeter H_0 en faveur de $H_1 : \Delta \neq 0$ si $v_s > a$ ou $v_r > a$.

Théorie

Si H_0 et la condition d'application sont satisfaites on peut calculer la distribution de V_r et de V_s (Lehmann, 1975) et, pour $v = 0, \dots, n(n+1)/2$,

$$P(V_s \geq v) = P(V_r \geq v).$$

Ces probabilités sont partiellement tabulées (Tables) pour divers choix de $n \leq 15$. Si $n > 15$, l'approximation suivante est utilisable:

$$P(V_s \geq v) = P(V_r \geq v) \approx 1 - \Phi \left(\frac{v - n(n+1)/4}{\sqrt{n(n+1)(2n+1)/24}} \right).$$

Exemple. Nous reprenons l'exemple 3 du Chapitre 12. Les différences d_i sont:

$$d_1 = 8, d_2 = -5, d_3 = 4, d_4 = 4, d_5 = 0, d_6 = -3, d_7 = 1, d_8 = -2, d_9 = -6, d_{10} = 0.$$

L'analyse graphique présentée au Chapitre 12 ne suggère pas de violations de la condition d'application. On obtient:

valeurs absolues des d_i :	8	5	4	4	0	3	1	2	6	0
rangs signés:	8	6-	4.5	4.5	*	3-	1	2-	7-	*

Les deux valeurs absolues "4" et "4" devraient recevoir les rangs 4 et 5. Comme elles sont identiques, on leur donne le "rang moyen" 4.5 (= moyenne entre 4 et 5). La valeur de n est réduite à 8 car deux différences sont nulles.

Donc, $v_r = 18$, $v_s = 18$ et dans la table on trouve $P(V_r \geq 18) = 0.527$ (pour $n = 8$); il n'y a donc pas de raisons de croire que la conservation a modifié les taux de triglycéridémie.

13.2. Le test de Wilcoxon pour données non-appariées

Soient X_1, \dots, X_m les variables aléatoires qui représentent les observations du premier échantillon, Y_1, \dots, Y_n celles qui représentent le deuxième. Soient $x_1, \dots, x_m, y_1, \dots, y_n$ les données.

Condition d'application

On suppose que X_1, \dots, X_m i.i.d. $\sim F(x)$ et que Y_1, \dots, Y_n i.i.d. $\sim F(x - \Delta)$.

Les deux distributions sont donc de forme identique mais la forme n'est pas spécifiée de façon paramétrique. Δ est un *paramètre de déplacement* qui mesure le décalage entre les deux distributions. L'hypothèse nulle est

$$H_0 : \Delta = 0.$$

Si l'hypothèse nulle est rejetée une des hypothèses alternatives suivantes est acceptée:

$$H_1 : \Delta \neq 0 \quad \text{ou bien} \quad H_1 : \Delta > 0 \quad \text{ou bien} \quad H_1 : \Delta < 0.$$

Une analyse graphique suffit généralement pour s'assurer que la condition d'application n'est pas violée. Si la condition n'est pas satisfaite, le test de Wilcoxon pour deux échantillons non appariés (connu aussi comme *test de la somme des rangs* ou *test de Mann-Whitney*) ne pourra pas être utilisé. A fortiori, le t-test ne sera pas approprié.

Les statistiques de test

Les pas suivants sont nécessaires.

1. Représenter les 2 échantillons par $x_1, \dots, x_m, y_1, \dots, y_n$ avec $m \leq n$. Ranger en ordre croissant les observations des deux échantillons réunis. La condition $m \leq n$ est nécessaire pour l'utilisation des tables.
2. Attribuer le rang 1 à la plus petite observation, le rang 2 à la suivante, etc. Si certaines observations sont identiques, leur attribuer un rang moyen, défini comme la moyenne de rangs que ces observations auraient si elles étaient différentes.
3. Calculer la somme w des rangs des y_j . Si r_1, \dots, r_n sont les rangs des y_j :

$$w = \sum_{j=1}^n r_j.$$

Soit $w_{XY} = w - n \cdot (n + 1)/2$.

La statistique w est connue comme la *statistique de la somme des rangs* proposée par Wilcoxon. La statistique w_{XY} est connue comme la *statistique de Mann-Whitney*. On peut démontrer que $0 \leq w_{XY} \leq m \cdot n$ et que

$$\begin{aligned} w_{XY} = & \text{nombre de paires } (x_i, y_j) \text{ telles que } x_i < y_j \\ & + \frac{1}{2} \text{nombre de paires } (x_i, y_j) \text{ telles que } x_i = y_j. \end{aligned}$$

On peut fonder le test sur l'une ou l'autre de ces statistiques. Toutefois, la statistique w_{XY} offre certains avantages dans le calcul de la distribution nulle. Soit W_{XY} la variable aléatoire dont la valeur observée est w_{XY} .

Règles de décision au niveau α

Pour un test unilatéral, déterminer, à l'aide des résultats théoriques mentionnés ci-dessous, la valeur a telle que $P(W_{XY} \leq a) = P(W_{XY} \geq m \cdot n - a) \approx \alpha$.

- Rejeter H_0 en faveur de $H_1 : \Delta < 0$ si $w_{XY} \leq a$.
- Rejeter H_0 en faveur de $H_1 : \Delta > 0$ si $w_{XY} \geq m \cdot n - a$.

Pour un test bilatéral, déterminer a tel que $P(W_{XY} \leq a) = P(W_{XY} \geq m \cdot n - a) \approx \alpha/2$.

- Rejeter H_0 en faveur de $H_1 : \Delta \neq 0$ si $w_{XY} \leq a$ ou $w_{XY} \geq m \cdot n - a$.

Théorie

Supposons que toutes les observations soient différentes. Si H_0 et la condition d'application sont satisfaites, on peut calculer la distribution de W_{XY} (Lehmann, 1975). Elle est symétrique et, pour $w = 0, \dots, mn$,

$$P(W_{XY} \leq w) = P(W_{XY} \geq m \cdot n - w).$$

Cette distribution est partiellement tabulée (Tables) pour divers choix de $m \leq 10$, $n \leq 10$. Pour l'utilisation de la table il faut que $m \leq n$. Si $n > 10$ et $m > 10$, l'approximation suivante peut être utilisée:

$$P(W_{XY} \leq w) \approx \Phi \left(\frac{w - mn/2}{\sqrt{mn(m+n+1)/12}} \right).$$

Ces résultats sont utilisables si le nombre d'observations identiques est faible. Au cas contraire, la distribution nulle doit être calculée d'une autre façon (Lehmann, 1975).

Exemple. Nous considérons les données de l'Exemple 3, Chapitre 12. L'analyse graphique suggère que les deux populations peuvent être considérées comme identiques de forme avec, peut-être, un décalage $\Delta > 0$ entre la deuxième et la première. Les données rangées sont:

4.90, 5.40, 5.40, 5.60, 5.90, 6.00, 6.20, 6.25, 6.50, 6.60, 6.75, 6.75, 7.40, 7.90.

Ici, les y_j ont été soulignés. Les rangs moyens sont

1, 2.5, 2.5, 4, 5, 6, 7, 8, 9, 10, 11.5, 11.5, 13, 14.

Donc, $w = 2.5 + 6 + 8 + 9 + 10 + 11.5 + 13 + 14 = 74$, d'où $w_{XY} = 74 - 8 \cdot 9/2 = 38$. Pour tester l'alternative $\Delta > 0$ on peut évaluer $P(W_{XY} \geq 38)$, que l'on transforme par symétrie en $P(W_{XY} \leq 10)$; on trouve

$$P(W_{XY} \geq 38) = P(W_{XY} \leq 10) = 0.0406.$$

Comme $P(W_{XY} \leq 10) < 0.05$, on peut rejeter H_0 en faveur de $H_1 : \Delta > 0$ au niveau 5%.

Chapitre 14

Tests d'adéquation et d'indépendance par la méthode du chi-carré

Un procédé graphique pour vérifier l'adéquation d'un modèle de distribution avait été introduit au Chapitre 8. Ce chapitre présente un procédé général de test de l'adéquation d'un modèle mathématique de distribution. Ce procédé est adapté au test d'indépendance entre deux caractères A et B avec h , respectivement k niveaux, lorsque $h \geq 2$ et $k \geq 2$.

14.1 Test d'adéquation

Supposons que dans une expérience on observe k résultats avec des fréquences respectives n_1, n_2, \dots, n_k appelées ici *fréquences observées*, alors que les fréquences auxquelles on peut s'attendre selon un certain modèle sont a_1, a_2, \dots, a_k . Les a_j sont appelés *fréquences attendues ou espérées*. On souhaite tester s'il y a une différence significative entre les fréquences attendues et les fréquences observées: plus précisément, on souhaite tester

$$H_0 : E(n_j) = a_j, \quad j = 1, \dots, k.$$

Une mesure de l'écart entre ces deux distributions de fréquences est donnée par

$$s = \sum_{j=1}^k \frac{(n_j - a_j)^2}{a_j} \quad \text{avec} \quad \sum_{j=1}^k n_j = \sum_{j=1}^k a_j = n,$$

n étant le nombre d'observations individuelles. Soit S la variable aléatoire dont la valeur observée est s . On démontre que, sous H_0 , S suit approximativement une distribution χ^2 avec ν degrés de liberté où

- $\nu = k - 1$ si les a_i peuvent être calculés grâce au modèle sans avoir à estimer des paramètres inconnus;
- $\nu = k - \ell - 1$ si les a_i sont calculés après avoir estimé ℓ paramètres.

Un test de niveau α rejettera donc H_0 si s est supérieure au quantile $1 - \alpha$ de la distribution χ^2 à ν degrés de liberté.

Exemple 1. On jette 4 dés et on s'intéresse au "nombre de 6". Selon le modèle binomial,

$$P(\text{nombre de 6} = j) = \binom{4}{j} (1/6)^j (5/6)^{4-j}, \quad j = 0, 1, 2, 3, 4.$$

Supposons de répéter $n = 10000$ fois cette expérience et d'obtenir les fréquences n_j indiquées dans le tableau ci-dessous. Selon le modèle, les fréquences attendues sont

$$a_j = np_j, \quad \text{où} \quad p_j = P(\text{nombre de 6} = j).$$

Table 1

j	n_j	a_j	$(n_j - a_j)^2$	$\frac{(n_j - a_j)^2}{a_j}$
0	5023	$a_0 = p_0 n = 4822.5$	$4.020 \cdot 10^4$	8.3360
1	3687	$a_1 = p_1 n = 3858.1$	$2.928 \cdot 10^4$	7.5880
2	1132	$a_2 = p_2 n = 1157.4$	$6.452 \cdot 10^2$	0.5574
3	148	$a_3 = p_3 n = 154.3$	$3.969 \cdot 10^1$	0.2572
4	10	$a_4 = p_4 n = 7.7$	5.290	0.6870
	10000	10000		$s = 17.43$

On obtient $s = 17.43$. Comme $s > 13.3 = \chi_{4,99\%}^2$ le modèle binomial doit être rejeté au niveau $\alpha = 1\%$.

En général, supposons que l'on souhaite tester l'hypothèse qu'une certaine variable aléatoire X suit une distribution F donnée. Dans ce but, on observe n valeurs de X , on partage le domaine des valeurs en k classes disjointes C_1, C_2, \dots, C_k (le plus souvent des intervalles) et on détermine les fréquences n_j d'observations dans chaque classe.



On aura:

$$a_j = n \cdot p_j \quad j = 1, \dots, k$$

où p_j est la probabilité d'observer une valeur de X dans la classe C_j calculée à l'aide de la distribution F . Alors

$$s = \sum_{j=1}^k \frac{(n_j - np_j)^2}{np_j}.$$

Souvent, la distribution F est choisie dans une "famille paramétrique" de distributions, par exemple, la famille de toutes les distributions $\mathcal{N}(\mu, \sigma^2)$, μ et σ^2 inconnus. Il faut alors estimer des paramètres (par exemple, μ et σ) pour la fixer. Si le nombre de paramètres est ℓ et si le modèle F est correct, la distribution de S sera alors approximativement une distribution χ^2 avec $\nu = k - \ell - 1$ degrés de liberté. En d'autres termes, ce modèle pourra être rejeté dès que la valeur observée s de S dépasse le percentile $1 - \alpha$ de la distribution χ^2 à ν degrés de liberté.

Remarque. Pour assurer une bonne approximation de la distribution de S avec la distribution χ^2 il faut que $a_j > 5$ pour presque tous (par exemple, 4 sur 5) les j .

Exemple 2. Les données de la Table 2 représentent les tailles (en cm) de $n = 216$ filles d'une certaine école. L'histogramme de ces tailles (Figure 1) suggère que la distribution de Gauss peut être un modèle adéquat. L'hypothèse nulle est donc

$$\text{Taille} \sim \mathcal{N}(\mu, \sigma^2).$$

Dans ce cas μ et σ doivent être estimés à l'aide des données. On a

$$\hat{\mu} = \text{estimation de } \mu = \frac{1}{n} \sum_{i=1}^{216} \text{Taille}_i = 153,$$

$$\hat{\sigma} = \text{estimation de } \sigma = \sqrt{\frac{1}{n-1} \sum_{i=1}^{216} (\text{Taille}_i - \hat{\mu})^2} = 13.5.$$

Avec les classes de l'histogramme on a:

$$p_j = \int_{c_{j-1}}^{c_j} f(x) dx$$

où f est la densité de la distribution $\mathcal{N}(153, 13.5^2)$ et les c_j sont les limites des classes (voir Table 3). Pour satisfaire la règle $a_j > 5$ pour tous les j , certaines classes doivent être regroupées. On obtient $s = 3.785$ et, comme $\chi_{10;95\%}^2 = 18.31$, l'hypothèse peut être acceptée.

Table 2

144	143	139	161	134	135	157	158	144	138	155	162	151	134
141	121	137	162	133	147	125	154	154	150	145	157	165	157
152	164	145	187	153	144	175	150	162	158	174	150	161	165
163	124	142	152	183	141	146	172	148	133	151	164	153	176
160	148	136	160	154	148	161	149	147	161	158	155	175	137
147	132	129	154	161	132	168	147	159	118	179	162	143	151
158	157	156	146	146	157	136	161	166	152	167	160	137	156
145	184	162	153	146	144	155	139	173	151	145	166	141	148
161	155	142	150	159	166	173	171	127	164	163	159	148	143
148	149	155	131	139	142	156	149	154	163	149	164	142	136
146	153	154	144	172	150	143	150	169	157	164	152	165	116
177	169	168	138	143	160	137	140	164	142	132	176	138	165
167	158	153	162	134	153	156	137	169	161	172	143	171	158
125	170	153	170	152	145	154	134	156	153	146	138	173	179
156	154	147	180	145	165	170	127	135	156	157	174	167	169
174	150	152	137	166	142								

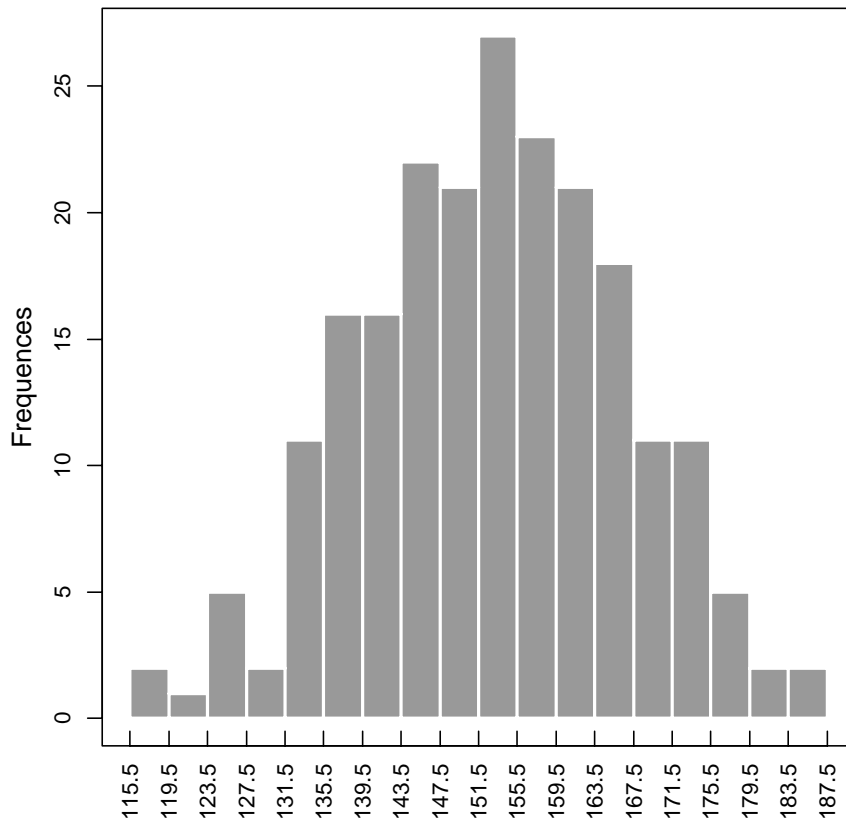


Figure 1

Table 3

Limite	n_j	p_j	np_j
115.5	2	0.05592	12.1
119.9			
123.5			
127.5			
131.5			
135.5	11	0.04088	8.8
139.5	16	0.0619	13.4
143.5	16	0.0833	18.0
147.5	22	0.0989	21.4
151.5	21	0.1153	24.9
155.5	27	0.1191	25.7
159.5	23	0.1091	23.6
163.5	21	0.0979	21.1
167.5	18	0.0754	16.3
171.5	11	0.0570	12.3
175.5	11	0.0378	8.2
179.5	5	0.0475	10.3
183.5	2		
187.5	2		
Table	216	1.000	216.1

14.2 Test d'indépendance dans un tableau de contingence

On veut tester l'indépendance de deux caractères A et B lorsque A présente h niveaux et B en présente k . Le cas $h = k = 2$ a été discuté au Chapitre 11, Section 11.3. Supposons que n observations aient été classées dans un tableau de comptage selon deux caractères A et B . Le *tableau de contingence* (ou de comptage) a la forme de la Table 4.

Table 4

		B						Σ
		1	2	...	j	...	k	
A	1	n_{11}	n_{12}		n_{1j}		n_{1k}	$n_{1.}$
	2	n_{21}	n_{22}		n_{2j}		n_{2k}	$n_{2.}$
	\vdots							\vdots
	i	n_{i1}	n_{i2}		n_{ij}		n_{ik}	$n_{i.}$
	\vdots							\vdots
	h	n_{h1}	n_{h2}		n_{hj}		n_{hk}	$n_{h.}$
Σ	$n_{.1}$	$n_{.2}$		$n_{.j}$		$n_{.k}$	n	

Dans ce tableau, n_{ij} est le nombre d'observations (individus) ayant obtenu le i -ème niveau du caractère A et le j -ème niveau du caractère B . En outre,

$$n_{i.} = \sum_{j=1}^k n_{ij}, \quad n_{.j} = \sum_{i=1}^h n_{ij}, \quad n = \sum_{i=1}^h n_{i.} = \sum_{j=1}^k n_{.j}.$$

Une estimation de la proportion du i -ème niveau du caractère A dans la population est fournie par le rapport $n_{i.}/n$. De même, la proportion du j -ème niveau du caractère B est estimée par $n_{.j}/n$. Supposant que les deux caractères sont indépendants dans la population, une estimation naturelle de la proportion d'individus dans la population avec le i -ème niveau du caractère A et le j -ème niveau du caractère B est

$$\frac{n_{i.}}{n} \cdot \frac{n_{.j}}{n}.$$

Ainsi, la fréquence théorique (c'est-à-dire, sous l'hypothèse d'indépendance) a_{ij} pour un échantillon de taille n est

$$n \cdot \frac{n_{i.}}{n} \cdot \frac{n_{.j}}{n}.$$

Le test d'indépendance est basé sur la statistique

$$s = \sum_{i=1}^h \sum_{j=1}^k \frac{(n_{ij} - a_{ij})^2}{a_{ij}}.$$

Soit S la variable aléatoire correspondante. On démontre que, sous l'hypothèse d'indépendance entre les caractères A et B , S suit la distribution χ^2 avec $\nu = (h - 1)(k - 1)$ degrés de liberté.

Par un développement algébrique on obtient

$$s = n \cdot \left(\sum_{i=1}^h \sum_{j=1}^k \frac{n_{ij}^2}{n_i \cdot n_j} - 1 \right).$$

Justification théorique

Pour chaque individu (observation) soit X le niveau de A et Y le niveau de B . Soit

$$p_{ij} = P(X = i \text{ et } Y = j), \quad i = 1, \dots, h \text{ et } j = 1, \dots, k$$

la distribution conjointe de X et Y . Les distributions marginales sont:

$$\begin{aligned} p_{i.} &= P(X = i), & i &= 1, \dots, h, \\ p_{.j} &= P(Y = j), & j &= 1, \dots, k. \end{aligned}$$

L'hypothèse (modèle théorique) à tester est:

$$H_0 : p_{ij} = p_{i.} \cdot p_{.j}, \quad i = 1, \dots, h \text{ et } j = 1, \dots, k.$$

Les $p_{i.}$ et les $p_{.j}$ représentent les paramètres inconnus du modèle. Ils seront estimés par $n_{i.}/n$ et $n_{.j}/n$ respectivement. Notons que seuls $h - 1 + k - 1$ paramètres doivent être estimés, car $\sum p_{i.} = \sum p_{.j} = 1$. La statistique S suivra donc une distribution χ^2 avec $\nu = hk - [h - 1 + k - 1] - 1 = (h - 1)(k - 1)$ degrés de liberté.

Remarque

Ce même test peut être utilisé dans la situation suivante. Soient $X^{(1)}, \dots, X^{(h)}$ h variables aléatoires discrètes avec les mêmes modalités (par exemple $X^{(i)} = \hat{\text{âge}}$ d'un habitant pris au hasard dans la ville i). Soit

$$p_j^{(i)} = P(X^{(i)} = j) \quad j = 1, \dots, k$$

la distribution de $X^{(i)}$. On dispose d'un échantillon de taille n_i pour la variable $X^{(i)}$ tel que

$$n_{ij} \text{ valeurs sont égales à } j \text{ et } \sum_j n_{ij} = n_i,$$

et ceci pour $i = 1, \dots, h$. On pourra compter les données dans un tableau comme celui utilisé auparavant et tester l'hypothèse

$$H_0 : p_j^{(1)} = p_j^{(2)} = \dots = p_j^{(h)} \quad \text{pour } j = 1, \dots, k,$$

en utilisant la même statistique s et la même distribution de référence.

Exemple 3. Comparaison de h distributions continues. Il faudra partager leurs domaines des valeurs en k classes, etc.

Exemple 4. Dans une étude sur l'efficacité de médicaments contre la nausée post-opératoire, 30 patients (pris au hasard parmi 167) ont été traités avec le médicament P, 67 avec le médicament C et 70 ont reçu un placebo. (L'allocation des traitements aux patients a été effectuée à l'aide d'une randomisation; voir Chapitre 15). Les nombres de patients souffrant d'une nausée grave, modérée, légère ou nulle sont donnés dans la table suivante.

Traitement	Nausée				Total
	grave	modérée	légère	absente	
Placebo	8	8	19	35	70
Médicament P	2	3	5	20	30
Médicament C	3	4	15	45	67
Total	13	15	39	100	167

On obtient $s = 6.358$. Comme le quantile 95% de la distribution χ^2 avec $(3-1) \times (4-1) = 6$ degrés de liberté est 12.59, on ne peut pas rejeter l'hypothèse que les 3 niveaux de traitement ont des effets identiques.

Chapitre 15

Etudes expérimentales, randomisation et causalité

Ce chapitre introduit le concept d'étude expérimentale et met en évidence les différences fondamentales entre étude expérimentale et étude d'observation. La possibilité de manipuler la cause hypothétique (traitement, intervention) d'un certain effet est pour l'étude expérimentale un avantage sans équivalent dans l'étude d'observation. La randomisation du traitement est à la fois le fondement rigoureux du test de l'hypothèse d'absence d'effet, et le moyen d'éliminer statistiquement les éventuelles différences entre les groupes comparés. Elle est donc un excellent support de l'inférence causale. Ce chapitre se limite à la présentation de ses fondements logiques.

15.1 Relation de causalité de nature déterministe

En général, on parle d'une *relation de causalité de nature déterministe* lorsque la présence de la cause implique l'effet et, réciproquement, si on observe l'effet, la cause est présente au départ.

Exemple. Selon l'état des connaissances actuelles, et en supposant qu'une seule particule virale viable soit suffisante pour causer la maladie, un humain non vacciné, infecté par la morsure d'un animal enragé, développe nécessairement la rage. D'autre part, si on observe la rage chez un humain, celui-ci a été infecté par le virus de la rage. Le lien entre l'infection par suite d'une morsure et la rage chez l'humain apparaît comme un lien de causalité de nature déterministe.

15.2 Relation de causalité de nature probabiliste

En général, dans un groupe de personnes exposées à un certain *facteur de risque*, la fréquence d'une certaine maladie est plus élevée que dans un groupe non exposé. On parle, dans ce cas, d'une *relation de causalité de nature probabiliste*: si la cause est présente, l'effet suit avec une certaine probabilité. Réciproquement, si on observe l'effet, la cause est présente au départ avec une certaine probabilité.

Exemple. Le fait de fumer, pour une personne, n'entraîne pas nécessairement un cancer du poumon. Un non-fumeur peut, par ailleurs, développer un cancer du poumon. Le fait de fumer augmente le risque d'être affecté par un cancer du poumon.

La liaison entre le facteur et son effet est souvent exprimée par des mesures statistiques d'association, comme la différence de fréquences de cancer entre deux groupes ou l'écart moyen entre les mesures d'une certaine variable dans les deux groupes. Toutefois, ces mesures n'indiquent pas nécessairement une relation de cause à effet. Elles peuvent seulement témoigner d'une relation statistique.

Avant qu'une association observée entre un facteur et une maladie ne soit déclarée causale, certaines précautions doivent être prises pour établir un tel jugement. En particulier, il faut s'assurer que les groupes soient comparables par rapport à toute caractéristique des sujets (âge, sexe, etc.) qui peut influencer l'association. Seul le facteur doit faire la différence.

15.3 Etudes expérimentales

L'étude expérimentale est caractérisée par le fait que *le chercheur peut manipuler le facteur étudié* (traitement ou intervention).

Exemple. L'*essai clinique* est le cas le plus connu d'étude expérimentale en médecine. Il s'agit d'évaluer l'efficacité d'un certain traitement, souvent nouveau. Dans ce but le chercheur peut appliquer le traitement à des sujets de son choix et comparer les réponses avec celles produites par des sujets non traités.

Exemple. Un investigateur veut évaluer l'efficacité d'une intervention en santé publique. Le facteur, ici l'intervention, sera manipulé si l'investigateur décide d'appliquer cette intervention à une population de son choix. En la comparant, à partir de certains critères, à une autre population non soumise à l'intervention, il pourra se prononcer sur l'efficacité de la prévention.

Pour "éliminer" les différences entre les groupes et passer de l'association statistique à la causalité, le chercheur peut faire appel à deux procédés:

- (a) Constitution de *blocs*. Le chercheur applique l'intervention à des groupes d'individus les plus homogènes possible par rapport à tout facteur *connu* (âge, sexe etc.) qui peut influencer le résultat de la comparaison.
- (b) *Randomisation*. Dans l'ensemble des sujets à disposition, le chercheur choisit au hasard les individus auxquels il donnera le traitement (l'intervention).

Nous verrons que la randomisation joue un rôle clé dans la détermination d'une relation de causalité. Elle permet d'éliminer avec une haute probabilité *toute différence* entre les groupes.

15.4 Etudes d'observation ou non expérimentales

Dans les études d'observation l'investigateur observe la réalité telle qu'elle se présente: *Le chercheur ne manipule pas le facteur étudié.*

Exemple. Si le but de l'étude est de mettre en évidence une relation entre un facteur et une maladie, l'investigateur comparera la fréquence de la maladie dans un groupe de personnes "naturellement" exposées au facteur et dans un groupe de personnes non exposées.

Pour éliminer des éventuelles sources de distorsion l'investigateur peut faire appel à la constitution de blocs. Le chercheur effectue les comparaisons dans des groupes de sujets les plus homogènes possible par rapport à tout facteur *connu*. Dans le meilleur des cas, les comparaisons sont effectuées à l'intérieur de paires d'individus "jumeaux" (même âge, même sexe, etc.) où "appariés" (Chapitres 12 et 13).

Toutefois *le chercheur ne dispose pas de moyens, comme la randomisation, pour éliminer l'effet éventuel de facteurs inconnus.*

15.5 Randomisation et test de randomisation

Nous nous limitons à un exemple typique d'étude expérimentale: un chercheur veut évaluer l'efficacité d'un nouveau traitement T pour soigner une certaine maladie en le comparant à un standard S. Pour simplifier nous allons supposer que S est un placebo.

Planification. Dans la phase de planification de l'étude, le chercheur choisit le type de test qu'il va utiliser en fonction de la nature attendue des données (quantitatives, en catégories ordonnées, binaires) et détermine le nombre de sujets en fonction de la puissance souhaitée (ce thème n'est pas traité dans ce cours). Supposons que pour des raisons pratiques (ici didactiques) il ne dispose que de 5 sujets malades A, B, C, D, E. Il décide alors d'en traiter 3 avec T et 2 avec S.

Randomisation. Le chercheur choisit les k sujets à traiter avec T au hasard parmi n sujets donnés. Dans ce but il peut procéder de deux façons équivalentes. Par exemple, pour $n = 5$ et $k = 3$:

- Il écrit les lettres A, B, C, D, E sur 5 étiquettes identiques qu'il dépose dans une urne; ensuite il en tire 3 au hasard. On parle dans ce cas d'une *allocation aléatoire des sujets aux traitements*. Supposons que les sujets choisis soient A, B et D.
- Il écrit la lettre T sur 3 étiquettes identiques et la lettre S sur 2 étiquettes identiques qu'il dépose dans une urne; ensuite il tire les étiquettes de façon aléatoire l'une après l'autre et il associe la lettre observée aux sujets A, B, C, D, E dans l'ordre. Dans ce cas on parle d'une *allocation aléatoire des traitements aux sujets*. Supposons que la suite des lettres observées soit T, T, S, T, S; les sujets traités sont donc A, B et D.

En pratique le tirage aléatoire est "simulé" par ordinateur ou par l'utilisation d'une table de nombres aléatoires. (Le procédé de randomisation a été proposé en 1935 par le biologiste et statisticien Sir Ronald Fisher.)

Réalisation de l'expérience et récolte des données. Supposons qu'après avoir traité les 5 sujets, le chercheur obtienne les résultats suivants:

Sujet	A	B	D	C	E
Traitement	T	T	T	S	S
Evolution	1	2	1	3	2

Ici, la réponse "1" signifie "le patient va mieux", "2" signifie "l'état du patient est resté stable" et "3" signifie "l'état du patient a empiré". Dans cet exemple, la réponses est donc en catégories ordonnées.

Test de randomisation. Il faut déterminer si T est un traitement supérieur à S (placebo sans effet), en d'autres termes, si T est la cause probable d'une évolution favorable de la maladie. Dans ce but, le checheur souhaite rejeter l'hypothèse nulle:

$$H_0 : \text{ T est équivalent à S.}$$

Il faut donc utiliser une statistique de test sensible à un éventuel effet de T. Une statistique bien adaptée à des données en catégories ordonnées est la somme des rangs signés de Wilcoxon (Chapitre 13). Dans notre exemple, le calcul est effectué de la façon suivante.

1. Ranger en ordre croissant les observations des deux échantillons réunis:

1 1 2 2 3

2. Souligner les données obtenues avec T:

1 1 2 2 3

3. Remplacer les observations par leur rang:

1 2 3 4 5

4. Remplacer les rangs des observations identiques par leur rang moyen:

1.5 1.5 3.5 3.5 5

5. Calculer la somme des rangs moyens soulignés: $w = \underline{1.5} + \underline{1.5} + \underline{3.5} = 6.5$.

Remarquablement, le calcul de la distribution de W (variable aléatoire dont la valeur observée est w) sous H_0 peut être effectué uniquement sur la base des deux points suivants:

- sous H_0 , les réponses sont des caractéristiques des sujets (et non du traitement);
- toutes les allocations de T à trois sujets parmi 5 sont équiprobables.

On peut donc imaginer que les réponses soient déjà inscrites sur les étiquettes avant le tirage. Pour chaque allocation il y a une valeur précise de W , comme indiqué dans la Table 1, et la distribution de W (*distribution de randomisation*) s'obtient en tenant compte du fait que les différents choix possibles (10 dans l'exemple) sont équiprobables. Cette distribution est représentée dans la Figure 1.

Table 1. Allocations possibles et valeurs de W .

Traitement	A B C	A B D	A B E	A C D	A C E	A D E	B C D	B C E	B D E	C D E
Evolution	1 2 3	1 2 1	1 2 2	1 2 1	1 3 2	1 1 2	2 3 1	2 3 2	2 1 2	3 1 2
Contrôle	D E	C E	C D	B E	B D	B C	A E	A D	A C	A B
Evolution	1 2	3 2	3 1	2 2	2 1	2 3	1 2	1 1	1 3	1 2
Statistique W	10	6.5	8.5	8.0	10	6.5	10	12	8.5	10

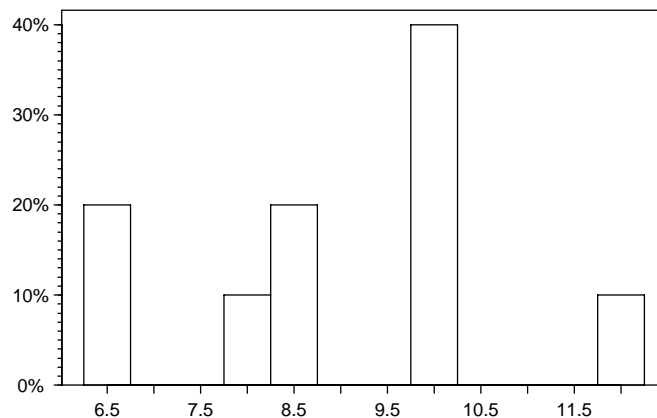


Figure 1. Distribution de randomisation de la statistique W .

Calcul du p-value. Dans l'exemple, la valeur effectivement observée de W est 6.5. Bien que 6.5 soit la valeur minimale de W le chercheur ne rejettera pas l'hypothèse H_0 , car

$$p = \text{Prob}(W \leq 6.5) = 20\%,$$

et cette valeur n'est pas suffisamment petite selon les règles habituelles.

Remarque. Dans cet exemple il n'est pas possible de rejeter H_0 car $p = 20\%$. En effet, avec 5 sujets seulement, il faut s'attendre à une faible puissance du test.

15.6 Choix de la statistique de test

Nous avons utilisé la statistique W de Wilcoxon car elle se prête bien au traitement de données en catégories ordonnées comme celles de l'exemple. Pour l'analyse d'une réponse quantitative (par exemple le poids, la taille, la pression artérielle) nous avons le choix entre la statistique W et la statistique T du t-test basée sur la différence des moyennes des réponses à T et à S. Si T et S étaient, par exemple, deux régimes diététiques et les résultats des réductions de poids en unités de 100g:

Sujet	A	B	D	C	E
Traitement	T	T	T	S	S
Réponse	12	7	10	-2	8

on obtiendrait la Table 2 et la Table 3 ainsi que les distributions indiquées dans la Figure 2 et la Figure 3.

Toutefois, la statistique T est très sensible à la présence d'outliers (cas atypiques, erreurs grossières). De ce fait, il peut devenir difficile de rejeter une fausse hypothèse H_0 si des observations aberrantes se trouvent parmi les données (perte de puissance). La statistique W et sa distribution de randomisation sont stables ("robustes") par rapport à la présence d'outliers. Son utilisation est donc avantageuse, avec des données quantitatives aussi.

Si enfin la réponse n'était que "succès" (1) ou "échec" (0), par exemple:

Sujet	A	B	D	C	E
Traitement	T	T	T	S	S
Réponse	1	0	1	0	0

on pourrait utiliser la statistique z^2 (Chapitre 11) pour les 10 tableaux 2×2 correspondant aux 10 choix. On obtiendrait ainsi la Table 4 et la distribution de la Figure 4. (Ce test est une forme du *test exact de Fisher*.)

15.7 Calcul de la distribution de randomisation

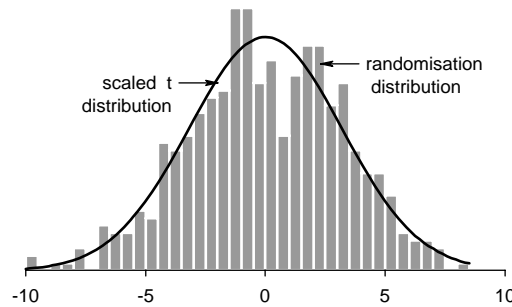
Pour certaines statistiques, l'analyse combinatoire permet de calculer exactement la distribution de randomisation pour tout n et k . Par exemple, la distribution de randomisation de W est calculable de cette façon (Lehmann, 1975) et tabulée (Tables). En principe, la distribution de randomisation de toute statistique est calculable à l'aide d'un programme informatique qui calcule tous les choix de k sujets parmi n . Toutefois, le temps de calcul peut être excessif car le nombre de choix peut être très élevé. Par exemple, il y a $\binom{5}{3} = 10$ choix possibles de 3 sujets parmi 5, mais il y a $\binom{11}{6} = 462$ choix possibles de 6 sujets parmi 11. Si ce nombre devient prohibitif, des approximations doivent être utilisées.

Une première possibilité consiste à calculer un nombre limité d'allocations prises au hasard, par exemple 5000. Chaque allocation est obtenue par *échantillonnage sans remise* de k sujets parmi n . La statistique de test est ensuite calculée pour chacun de ces choix et la distribution empirique des 5000 valeurs obtenus est utilisée comme approximation.

Parfois, des modèles paramétriques peuvent être utilisés. Par exemple, si les réponses sont quantitatives et si les nombres k et $n - k$ des sujets attribués à T, respectivement S, sont assez grands, la distribution t à $n - 2$ degrés de liberté est une bonne approximation de la distribution de randomisation de T . La figure ci-après donne une idée de cette approximation pour les données:

T	T	T	T	T	T	S	S	S	S	S
26.6	23.7	28.5	25.3	17.9	24.3	29.9	11.4	14.2	16.5	21.1

Le lecteur pourra calculer le p-value.



15.8 Conséquences pour l'étude d'une relation de causalité

Supposons que l'état de santé initial des sujets A, B, et D est nettement meilleur que celui des sujets C et E et que l'évolution de la maladie est influencée par cet état. Grâce à la randomisation il y a seulement une chance sur dix pour que les sujets A, B et D soient tous traités par T et qu'il y ait donc confusion (*confounding*) totale entre le traitement et le facteur. Il est clairement possible de réduire (en dessous d'une limite préfixée) la probabilité d'une telle confusion en augmentant le nombre de sujets. Avec 6 sujets traités par T et 5 par S, les chances de confusion totale sont de $1/462 \approx 2\text{‰}$. *De ce fait, on peut rendre improbable la détermination erronée d'une relation de causalité*: l'inférence causale est bien fondée. Bien évidemment, la randomisation n'assure pas l'équilibre de façon certaine. La Table 5 montre comment les caractéristiques principales des 79 sujets participant à un essai clinique randomisé ont été réparties entre les deux groupes. On remarquera une certaine différence pour la caractéristique "severity of admission".

Remarque. Les conclusions tirées d'un test de randomisation sont strictement valables seulement pour les sujets utilisés dans l'expérience (*validité interne*). Ces conclusions peuvent être inférées à une population mère si les sujets sont un échantillon aléatoire de la population mère. Malheureusement ceci est souvent impossible en pratique.

Sans randomisation il n'est pas possible d'assurer la comparabilité des groupes. Dans les études d'observations, pour passer de l'association statistique à la causalité, le chercheur doit faire appel à un certain nombre de critères auxiliaires. Voici les plus connus:

- (a) *Constance de l'association observée.* Des études conduites en des moments et des lieux différents sur d'autres populations produisant des résultats semblables accréditent l'hypothèse de causalité.
- (b) *Intensité de l'association.* L'action des biais sur l'association se fait vraisemblablement moins sentir lorsque l'intensité ou la force de l'association est grande. En ce sens, une plus forte intensité favorise le jugement de causalité.

- (c) *Spécificité de l'association.* Plus un facteur est exclusif par rapport à une maladie, plus l'interprétation causale est plausible.
- (d) *Cohérence chronologique.* La cause doit précéder l'effet.
- (e) *Présence d'une relation dose-effet.* L'effet augmente lorsque la dose augmente.
- (f) *Cohérence avec les connaissances bio-médicales.* Harmonie avec l'histoire naturelle connue de la maladie, plausibilité de l'association observée, cohérence avec les résultats chez l'animal dans le cadre de l'expérimentation.

Toutefois, déclarer à partir de ces critères qu'un facteur est causal *ne signifie pas qu'il y a preuve irréfutable de causalité*, mais seulement qu'il y a forte présomption en faveur de celle-ci. En outre, s'il n'y a pas de randomisation on ne peut pas calculer la probabilité de la configuration de données observées et l'inférence statistique rencontre des difficultés.

Cependant, des test statistiques sont parfois utilisés dans des études non randomisées expérimentales et d'observation. Pour donner un sens à cette inférence, il faut imaginer que les sujets traités (ou exposés) et non traités (non exposés) soient des échantillons aléatoires de deux populations: l'une traitée (exposée) et l'autre non traitée (non exposée). On parle dans ce cas d'*inférence descriptive*. Par exemple le p-value peut se référer à l'hypothèse que les fréquences de "succès" (cancer) dans les deux populations (fumeurs, non fumeurs) sont les mêmes. Clairement, le p-value peut être très petit sans impliquer que le traitement (l'exposition) a un effet.

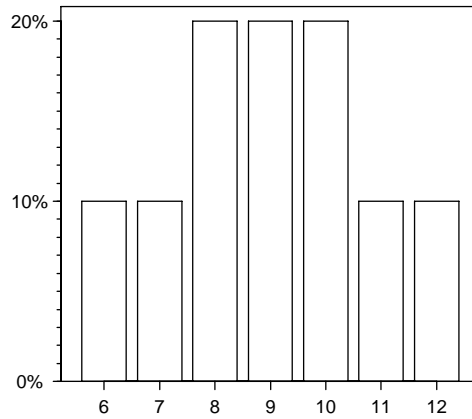
Enfin, il est souvent difficile de démontrer que les sujets sont un échantillon aléatoire d'une population. Lorsqu'il n'y a ni échantillonnage aléatoire ni randomisation, il est recommandé de ne pas utiliser des méthodes d'inférence statistique. Les statistiques descriptives et exploratoires sont le seul moyen approprié pour présenter, analyser et interpréter les résultats. Ces moyens comprennent les analyses multifactorielles et notamment les mesures quantitatives des effets des facteurs multiples (voir par exemple, régression multiple). Pour réaliser ces analyses il faut en général ventiler les données sur de nombreuses strates; de grandes quantités de données sont donc nécessaires.

Conclusions

- L'acte physique de la randomisation permet de calculer exactement le niveau de signification (p-value) du test choisi *sans aucune hypothèse concernant la forme de la distribution des données.*
- Grâce à la randomisation, on peut rendre petite la probabilité qu'un biais quelconque fausse la comparaison. De ce fait, *la probabilité que la détermination d'une relation de causalité soit erronée est aussi petite.*
- *En l'absence de randomisation* (études expérimentales non randomisées et études d'observation) *d'éventuels biais inconnus ne peuvent pas être exclus.*
- Dans les études d'observation *sans échantillonnage aléatoire seules les statistiques descriptives sont appropriées.*

Table 2. Allocations possibles et valeurs de W pour réponse quantitative.

Traitement	A B C	A B D	A B E	A C D	A C E	A D E	B C D	B C E	B D E	C D E
Poids	12 7 -2	12 7 10	12 7 8	12 -2 10	12 -2 8	12 10 8	7 -2 10	7 -2 8	7 10 8	-2 10 8
Contrôle	D E	C E	C D	B E	B D	B C	A E	A D	A C	A B
Poids	10 8	-2 8	-2 10	7 8	7 10	7 -2	12 8	12 10	12 -2	12 7
Statistique W	8	11	10	10	9	12	7	6	9	8

Figure 2. Distribution de randomisation de W pour réponse quantitative.Table 3. Choix possibles et valeurs de T pour réponse quantitative.

Traitement	A B C	A B D	A B E	A C D	A C E	A D E	B C D	B C E	B D E	C D E
Poids	12 7 -2	12 7 10	12 7 8	12 -2 10	12 -2 8	12 10 8	7 -2 10	7 -2 8	7 10 8	-2 10 8
Contrôle	D E	C E	C D	B E	B D	B C	A E	A D	A C	A B
Poids	10 8	-2 8	-2 10	7 8	7 10	7 -2	12 8	12 10	12 -2	12 7
Statistique T	-0.624	1.598	1.023	-0.147	-0.455	2.043	-1.023	-1.598	0.624	-0.810

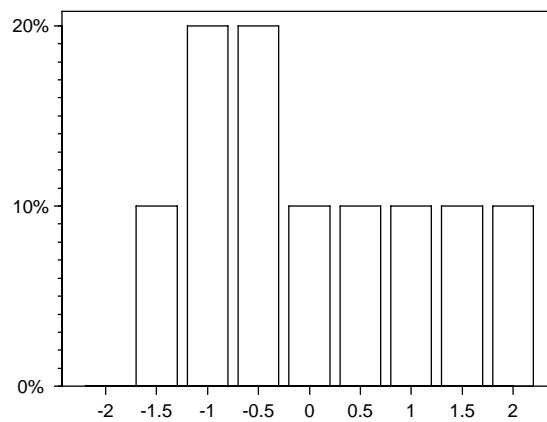
Figure 3. Distribution de randomisation de T pour réponse quantitative.

Table 4. Allocations possibles et valeurs de Z^2 pour réponse binaire.

Traitement	A B C	A B D	A B E	A C D	A C E	A D E	B C D	B C E	B D E	C D E
Résultat	1 0 0	1 0 1	1 0 0	1 0 1	1 0 0	1 1 0	0 0 1	0 0 0	0 1 0	0 1 0
Contrôle	D E	C E	C D	B E	B D	B C	A E	A D	A C	A B
Résultat	1 0	0 0	0 1	0 0	0 1	0 0	1 0	1 1	1 0	1 0
Statistique Z^2	0.139	2.222	0.139	2.222	0.139	2.222	0.139	5.000	0.139	0.139

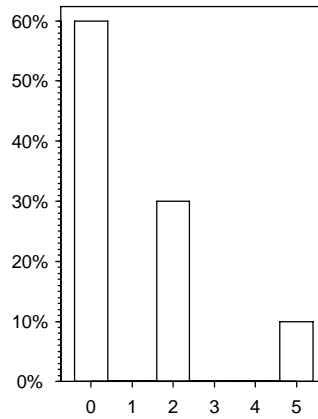
Figure 4. Distribution de randomisation de Z^2 pour réponse binaire.

Table 5. Caractéristiques de 79 sujets participant à un essai clinique randomisé (Colton 1974).

Characteristic	Antioxin group	No antitoxin group
Total cases	41	38
Locality and race:		
Nigeria; Black	35	32
West Indies		
Black	5	3
Others	1	3
Sex:		
Male	23	21
Female	18	17
Mean values:		
Age (yr)	24.8	27.1
Incubation period (days)	10.1 (26 cases)	9.7 (29 cases)
Period of onset (days)	3.0 (33 cases)	3.2 (34 cases)
Intercurrent disease or complications:		
None	18	15
Pulmonary infection	5	10
Hyperpyrexia	6	7
Laryngospasm	1	2
Postpartum	6	5
Miscellaneous	12	6
Severity on admission:		
Local	0	2
General		
Mild	9	4
Moderate	12	9
Severe	10	23
Site of Wound:		
Limbs	22	22
Head and neck	1	1
Trunk	1	1
Genital tract	6	5
Not known	11	9
Sepsis (excluding postpartum cases):		
Present	10	8
Absent	25	25

Chapitre 16

Une introduction au bootstrap

Ce chapitre a pour but de présenter le principe général du bootstrap et quelques-unes de ses applications. Nous présentons également un intervalle de confiance bootstrap ainsi que quelques tests bootstrap, pour un et deux échantillons. Pour une introduction plus approfondie, voir le livre de Efron (1992).

16.1 Introduction

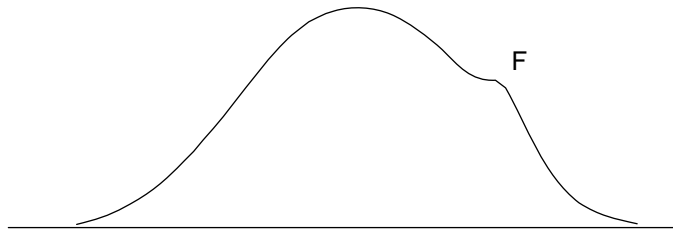
Basons-nous sur un cas bien connu: soit une population \mathcal{P} de taille N dont on veut estimer un paramètre, par exemple une moyenne μ . Pour cela, on tire un échantillon aléatoire de n individus de cette population et on mesure la caractéristique qui nous intéresse sur chaque individu de l'échantillon. Nous notons x_i ($i = 1, \dots, n$) la valeur de cette caractéristique sur l'individu i . Une estimation classique de μ est la moyenne arithmétique

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n x_i.$$

La question générale qui va nous amener au bootstrap est la suivante: est-ce que cette estimation est "bonne" ? Une manière d'y répondre est de calculer la variance $\sigma^2(\hat{\mu})$ de l'estimateur $\hat{\mu}$ vu comme variable aléatoire. (Remarquons que l'on pourrait aussi étudier son biais $E(\hat{\mu}) - \mu$, mais on ne le fera pas ici.)

Pour calculer la variance de l'estimateur $\hat{\mu}$, il faut étudier sa distribution d'échantillonnage, dont une approximation pourrait se construire de la manière suivante.

1. Les individus de \mathcal{P} , étudiés selon la caractéristique précitée, suivent une certaine distribution F :



2. De cette population, on tire un nombre k d'échantillons de taille n et on calcule alors k estimations $\hat{\mu}_1, \dots, \hat{\mu}_k$ de la moyenne inconnue μ .
3. La distribution empirique de ces estimations $\hat{\mu}_1, \dots, \hat{\mu}_k$ est une approximation de la distribution d'échantillonnage de l'estimateur $\hat{\mu}$ (Chapitre 9).

Une estimation de la variance cherchée $\sigma^2(\hat{\mu})$ serait alors

$$\hat{\sigma}^2(\hat{\mu}) = \frac{1}{k-1} \sum_{i=1}^k (\hat{\mu}_i - \bar{\hat{\mu}})^2$$

où $\bar{\hat{\mu}} = (1/k) \sum_{i=1}^k \hat{\mu}_i$ est la moyenne des k estimations de μ trouvées auparavant.

Notons que pour avoir la véritable variance $\sigma^2(\hat{\mu})$, il faudrait avoir tous les échantillons possibles de n éléments pris dans la population \mathcal{P} , ce qui serait un grand travail ! On aurait alors $\hat{\mu} = \mu$ et $\sigma^2(\hat{\mu}) = (1/k) \sum_1^k (\hat{\mu}_i - \mu)^2$ où $k = \binom{N}{n}$ est le nombre d'échantillons de \mathcal{P} . Mais le procédé décrit ci-dessus est utopique; en effet, il exige d'avoir plusieurs échantillons alors que la réalité ne nous en donne qu'un seul. Le bootstrap permet de remédier à ce problème grâce à la *simulation d'échantillons*.

16.2 Le bootstrap

L'estimateur considéré (plus haut $\hat{\mu}$), ou plus généralement la statistique S d'intérêt, a une distribution d'échantillonnage notée F_S . Cette distribution dépend de la distribution F de la variable aléatoire X dont les valeurs observées sont x_1, \dots, x_n . On écrit $F_S(s, F_X)$ au Chapitre 9, Section 1, où F_X (ici F) est la distribution de X . Comme F est inconnue, on travaille avec une estimation de F , que l'on note \hat{F} et qui est:

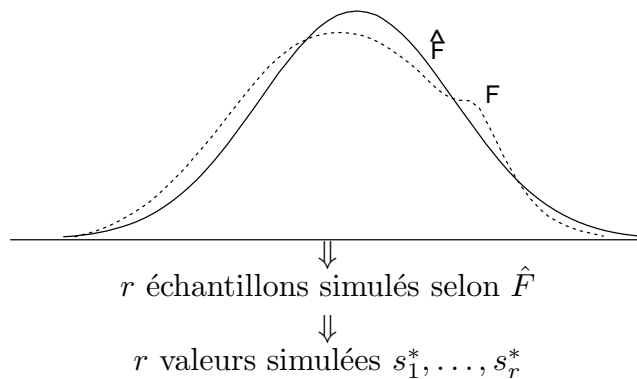
- soit un modèle paramétrique connu (normal, exponentiel, ou autre) qui ajuste assez bien les données,
- soit la distribution empirique F_n des données.

Approximation de la distribution d'échantillonnage

Le fait de remplacer F par l'un des \hat{F} vus ci-dessus va donner une distribution d'échantillonnage F_S également modifiée puisque F_S dépend de F . On écrit dans ce cas $F_S(s, \hat{F})$ au lieu de $F_S(s, F)$.

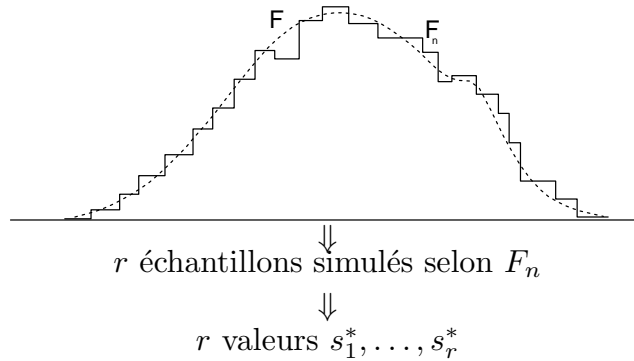
Il y a deux approches pour calculer F_S , utilisant donc les deux types d'estimation de F mentionnés ci-dessus: *l'approche mathématique* qui consiste à déterminer $F_S(s, \hat{F})$ grâce à la logique et à l'analyse mathématique (par exemple, le théorème centrale limite, Chapitre 9); *l'approche computationnelle* ou *bootstrap* qui consiste à déterminer $F_S(s, \hat{F})$ par la simulation de pseudo-données.

Il y a donc deux formes de bootstrap: le bootstrap paramétrique et le bootstrap non-paramétrique. Dans le *bootstrap paramétrique*, on remplace d'abord F par un modèle paramétrique \hat{F} qui semble bien ajuster les données. On simule ensuite r échantillons ($r = 1000$ par exemple) de taille n , indépendamment les uns des autres, qui proviennent de la distribution \hat{F} (Chapitre 9, Complément 1). Enfin, on calcule la statistique S pour chacun des r échantillons simulés et l'on obtient les valeurs simulées s_1^*, \dots, s_r^* .



Si nous ne pouvons pas attribuer un modèle paramétrique aux données, nous utilisons $\hat{F} = F_n$ comme approximation de F , où F_n est la distribution empirique; c'est le *bootstrap*

non-paramétrique. On génère alors r échantillons de taille n provenant de F_n . De nouveau, on calcule les r valeurs s_1^*, \dots, s_r^* de S pour ces r échantillons simulés.



Dans les deux cas, la distribution empirique des valeurs simulées s_1^*, \dots, s_r^* fournit une approximation de la distribution d'échantillonnage de S . On appelle cette approximation la *distribution bootstrap* de S . À l'aide de cette distribution on peut alors calculer, par exemple, une approximation de $\sigma^2(S)$:

$$\sigma^2(S)^* = \frac{1}{r-1} \sum_{i=1}^r (s_i^* - \bar{s}^*)^2 \quad \text{où} \quad \bar{s}^* = \frac{1}{r} \sum_{i=1}^r s_i^*.$$

On peut démontrer (Chapitre 9, Complément 1) que remplacer F par $\hat{F} = F_n$ et générer un échantillon de taille n selon cette distribution F_n revient au même que de tirer avec remise n éléments de l'ensemble de données originales $\{x_1, \dots, x_n\}$.

Exemple

Soit un échantillon de taille 10 tiré aléatoirement parmi les 49 plus grandes villes des États-Unis d'Amérique en 1920. On donne leur population de 1920 (en milliers d'habitants) et leur population de 1930 et l'on s'intéresse au "ratio" (rapport) "population de 1930 divisée par la population de 1920".

en 1920,	x_i :	138	93	61	179	48	37	29	23	30	2
en 1930,	y_i :	143	104	69	260	75	63	50	48	111	50

On obtient les moyennes $\bar{x} = 64.0$ et $\bar{y} = 97.3$ et le ratio

$$s = \frac{\text{moyenne des populations des villes en 1930}}{\text{moyenne des populations des villes en 1920}} = \frac{\bar{y}}{\bar{x}} = 1.52.$$

Mais quelle est la variance de $S = \bar{Y}/\bar{X}$? Pour répondre à cette question en utilisant le bootstrap non-paramétrique, on effectue un tirage aléatoire avec remise dans cet échantillon de paires de taille 10. On obtient, par exemple, l'échantillon simulé:

x_i^* :	37	29	93	93	61	61	2	29	93	30
y_i^* :	63	50	104	104	69	69	50	50	104	111

Dans cet échantillon, la 1ère paire de l'échantillon initial n'apparaît aucune fois, la 2ème apparaît 3 fois, la 3ème paire 2 fois, etc. On observe la valeur $s_1^* = 1.466$ de S . De la même manière, on construit r échantillons bootstrap, et on calcule la statistique s pour chacun de ces échantillons. Prenons ici $r = 9$:

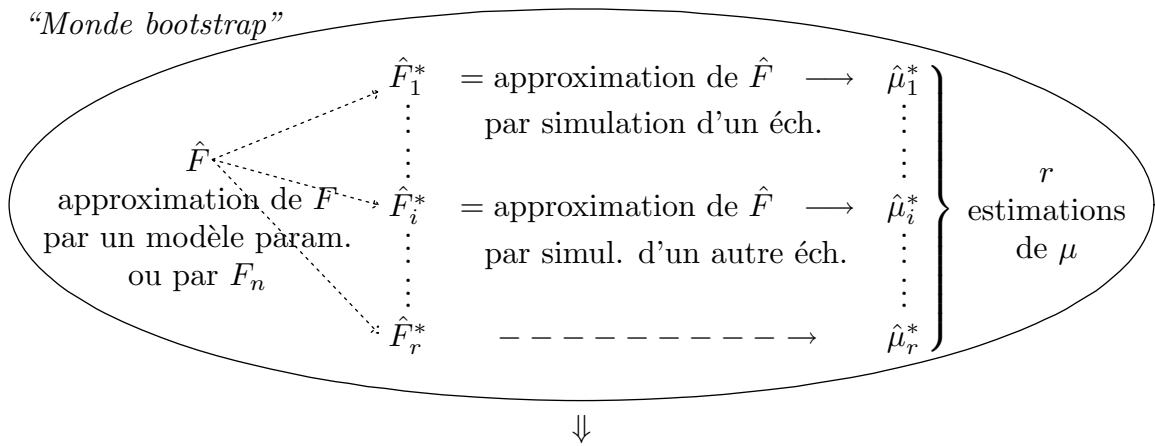
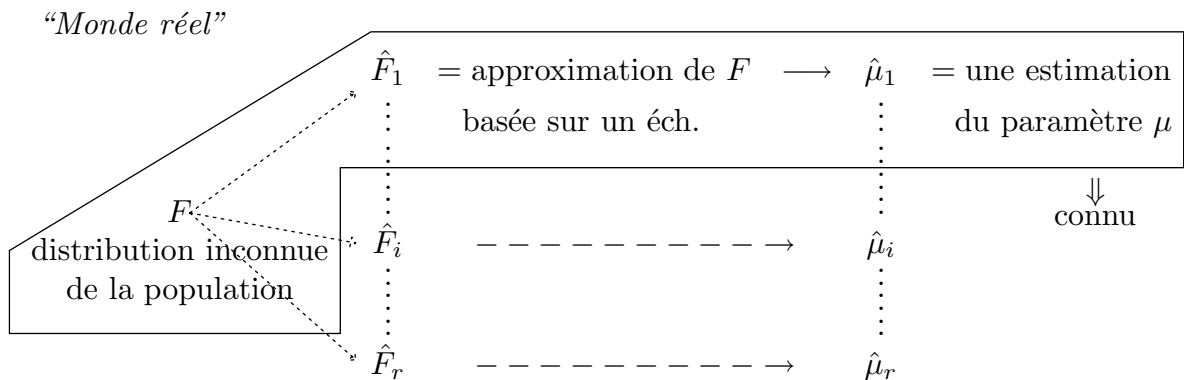
	Nbre de fois que chaque paire est tirée										Statistique	
échantillon observé	1	1	1	1	1	1	1	1	1	1	$s = 1.520$	
1		3	2			1	2		1	1	$s_1^* = 1.466$	
2		1		1		2	2	1		2	1	$s_2^* = 1.761$
3		1	1		1		1			4	2	$s_3^* = 1.951$
4			1	2		1	1	2	2		1	$s_4^* = 1.542$
5		3			1	3		1	1	1		$s_5^* = 1.371$
6		1	1	2			1		1	1	3	$s_6^* = 1.686$
7		1	1	2	2	2		1			1	$s_7^* = 1.378$
8		2		1		3	1	1	1	1		$s_8^* = 1.420$
9			1	1	1	2	1		2	1	1	$s_9^* = 1.660$

La moyenne des s_j^* ($j = 1, \dots, r$) est $\bar{s}^* = 1.582$. La variance estimée de S est alors $\hat{\sigma}^2(S)^* = \sum_{j=1}^9 (s_j^* - 1.582)^2 / 8 = 0.03907$. Notons qu'on a estimé la variance d'un rapport !

Remarques

1. On parle souvent de *rééchantillonnage* dans le cas du bootstrap non-paramétrique car on reconstruit un ensemble d'échantillons en partant de l'échantillon de départ.
2. Le bootstrap non-paramétrique est approprié lorsqu'il est difficile de trouver un bon modèle pour F_X .
3. Plus le nombre d'échantillons générés r est grand, meilleures sont les approximations bootstrap, mais plus l'ordinateur prend du temps à calculer aussi. Un r de 100 ou 200 peut suffire à avoir une bonne estimation de $\sigma^2(\hat{\mu})$ mais un $r=500$ ou 1000 donne une meilleure précision. Pour calculer des intervalles de confiance bootstrap (Section 16.4), on choisit en général $1000 \leq r \leq 5000$.

4. Nous présentons ci-dessous un schéma mettant en valeur le principal avantage du bootstrap:



Ici, tout est connu et on peut générer tous les échantillons nécessaires.

16.3 Exemple

Appliquons le bootstrap non-paramétrique au tableau de données présenté ci-dessous (Table 1). Ces données ont été récoltées lors d'une campagne de baguement des oiseaux migrateurs dans les Préalpes Vaudoises en automne 1994. Cet échantillon est constitué de 40 observations tirées au hasard parmi les 244 oiseaux qui étaient à disposition. La variable d'intérêt est la longueur de la 3ème rémige, donnée en dixièmes de millimètres.

sexe	age	date	h	rem	pds	adp	sexe	age	date	h	rem	pds	adp
1	4	10/10/94	8	720	235	1	2	3	14/10/94	14	620	215	2
1	3	21/10/94	7	680	235	2	2	3	13/10/94	13	640	185	1
2	3	10/10/94	7	705	200	2	1	4	20/10/94	12	675	225	2
2	3	10/10/94	17	600	205	2	2	3	13/10/94	12	665	220	1
2	3	10/10/94	8	625	195	2	2	3	11/10/94	10	620	210	1
1	4	13/10/94	16	680	215	1	2	4	14/10/94	16	630	200	2
2	4	18/10/94	12	630	205	2	1	4	10/10/94	17	680	220	1
2	3	14/10/94	17	650	205	2	1	3	20/10/94	11	660	220	2
2	3	18/10/94	13	635	180	1	1	4	14/10/94	16	690	205	1
2	3	10/10/94	13	640	205	1	1	4	14/10/94	15	720	225	2
1	3	11/10/94	10	675	210	1	1	4	11/10/94	12	650	210	1
2	3	21/10/94	11	620	215	3	1	3	14/10/94	7	680	225	1
2	3	11/10/94	13	625	195	1	1	4	10/10/94	17	700	235	1
1	4	10/10/94	8	700	230	1	1	4	11/10/94	10	685	220	1
2	3	14/10/94	14	620	215	2	2	4	20/10/94	11	625	175	1
2	4	11/10/94	8	600	205	1	1	3	14/10/94	11	690	210	1
1	4	11/10/94	16	690	225	2	2	3	14/10/94	17	650	200	2
2	3	20/10/94	15	615	200	1	2	3	10/10/94	17	645	215	1
1	4	11/10/94	16	720	215	2	1	3	11/10/94	10	690	225	1
1	3	10/10/94	17	655	225	1	1	3	14/10/94	14	670	220	2

Table 1. Données de capture de 40 pinsons des arbres (*Fringilla coelebs*). Le codage des variables est le suivant: sexe: 1=mâle, 2=femelle; âge: 3=jeune de l'année, 4=adulte; h: heure de capture; rem: longueur de la 3ème rémige en 1/10 de mm; pds: poids en 1/10 de gramme; adp: code de 1 à 5 estimant la réserve de graisse.

L'estimateur choisi est la moyenne arithmétique des longueurs de rémiges: $\hat{\mu} = \sum_1^{40} x_i/40$. Pour l'échantillon d'origine, on obtient une moyenne $\hat{\mu} = 659.25$ dixièmes de millimètres et un écart-type estimé de la moyenne $\hat{\sigma}(\hat{\mu}) = 5.39$. En simulant $r = 500$ échantillons, on obtient 500 nouvelles estimations du paramètre $\hat{\mu}$, notées $\hat{\mu}_1^*, \dots, \hat{\mu}_{500}^*$, et un écart-type approximé de la moyenne

$$\hat{\sigma}(\hat{\mu})^* = \sqrt{\sum_1^{500} (\hat{\mu}_i^* - \bar{\mu}^*)^2 / 499} = 5.35.$$

Dans la Figure 1 (en haut), on trouve l'histogramme des $r = 500$ estimations simulées $\hat{\mu}_1^*, \dots, \hat{\mu}_{500}^*$ de $\hat{\mu}$. La distribution est approximativement normale; ceci se confirme sur le qq-plot (en bas).

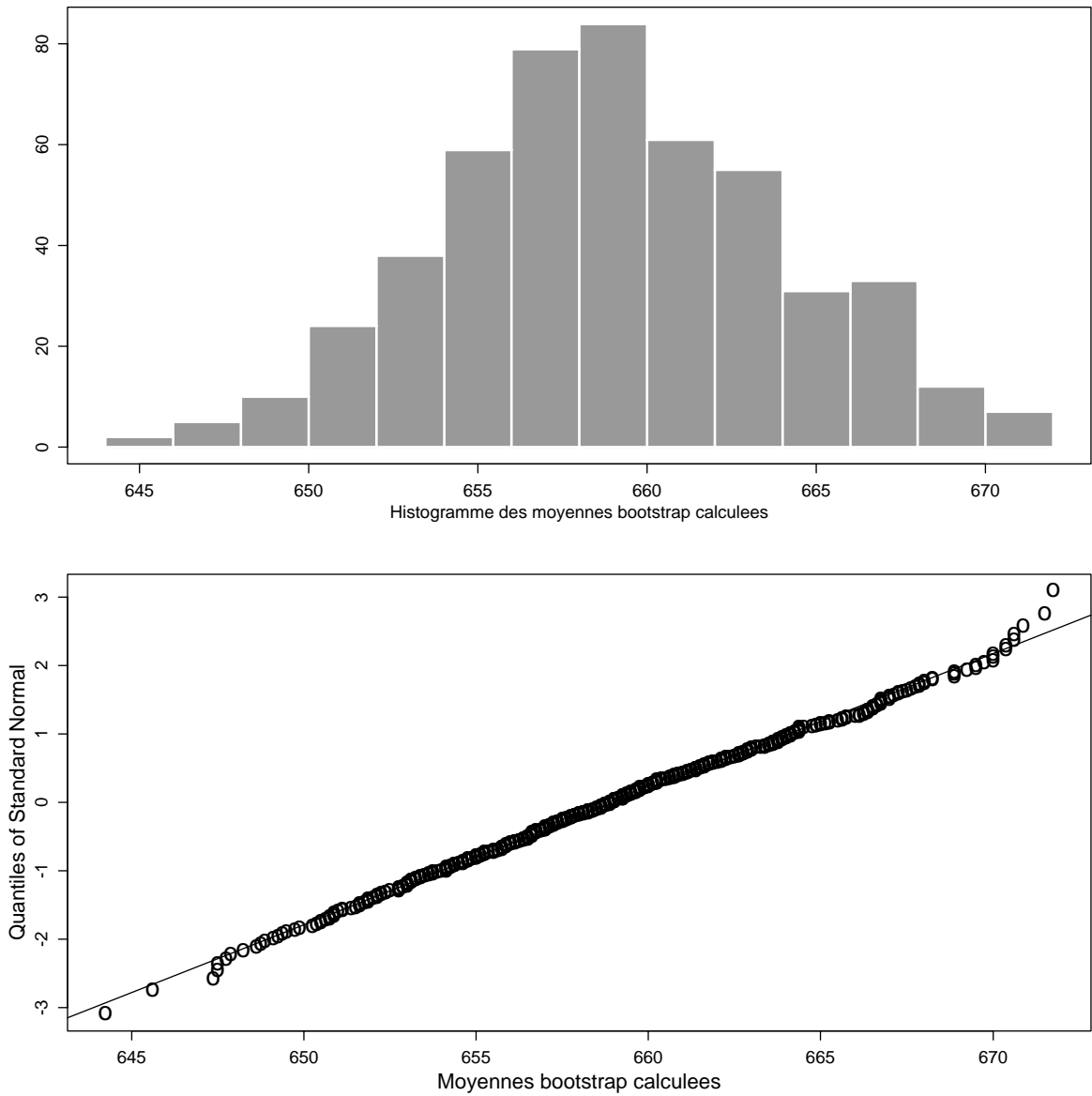


Figure 1. Histogramme et qq-plot de 500 valeurs simulées $\hat{\mu}^*$ de $\hat{\mu}$.

16.4 Intervalles de confiance bootstrap

Grâce au bootstrap, on peut déterminer des intervalles de confiance pour le paramètre inconnu μ . Il y a plusieurs types d'intervalles de confiance utilisant la simulation bootstrap mais un seul d'entre eux, d'ailleurs simple à déterminer, est présenté ici. Précisons qu'il est utilisable si la distribution bootstrap de l'estimateur est approximativement normale. Si ce n'est pas le cas, d'autres types d'intervalles de confiance sont à utiliser (Efron, 1992).

On appelle *intervalle percentile 5% – 5%* pour μ l'intervalle $(\hat{\mu}_{[0.05 \cdot r]}^*, \hat{\mu}_{[0.95 \cdot r]}^*)$ dont les bornes sont simplement les percentiles 5% et 95% de la distribution bootstrap des $\hat{\mu}_i^*$ (Figure 2). On démontre qu'il s'agit d'un intervalle de confiance dans le sens habituel: si on prenait un grand nombre d'échantillons de même taille que l'échantillon original, dans les mêmes conditions, et que l'on construisait pour chacun d'entre eux un intervalle percentile 5%-5%, le 90% de ces intervalles contiendrait la valeur inconnue μ . Ce 90% est le coefficient de couverture de l'intervalle.

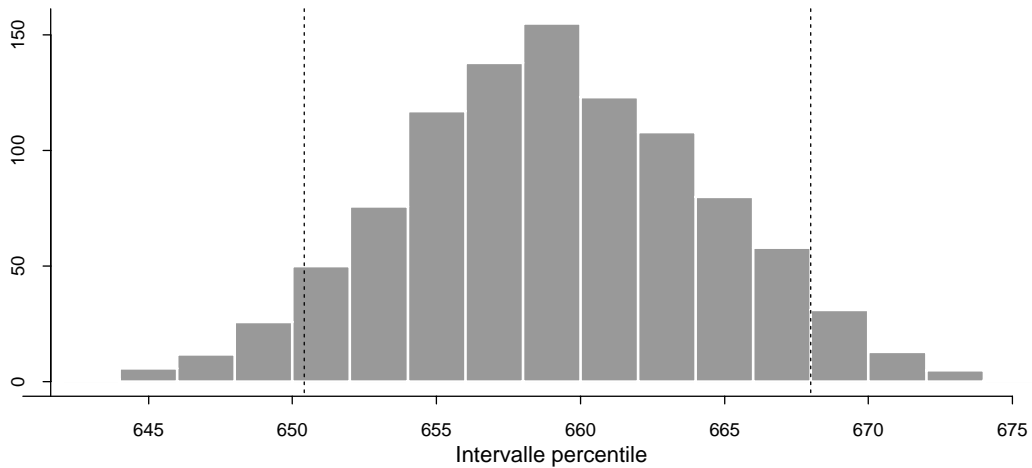


Figure 2. Intervalle percentile 5%-5% calculé à l'aide d'une distribution bootstrap de la moyenne des longueurs de rémiges, avec $r = 1000$.

Remarque. On peut, bien entendu, travailler sur plusieurs populations de distributions F_1, F_2, \dots et devoir estimer plusieurs paramètres μ_1, μ_2, \dots . De la même manière que ci-dessus, on peut alors calculer des intervalles de confiance pour $\mu_1, \mu_2, \mu_2 - \mu_1, \mu_1/\mu_2$, etc.

16.5 Tests bootstrap

Dans les cas où il est difficile de déterminer mathématiquement la distribution d'une statistique de test, nous pouvons utiliser les tests bootstrap. Nous présentons ci-dessous les tests bootstrap pour les cas suivants:

- μ est la moyenne d'une variable $X \sim F$. L'hypothèse nulle est $H_0 : \mu = \mu_0$, où μ_0 est une valeur spécifiée. On dispose d'un échantillon $\mathcal{E} = \{x_1, \dots, x_n\}$ de X .
- μ_1 et μ_2 sont les moyennes des variables $X \sim F_1$ et $Y \sim F_2$. L'hypothèse nulle est $H_0 : \mu_1 = \mu_2$. On dispose de deux échantillons indépendants $\mathcal{E}_1 = \{x_1, \dots, x_m\}$ et $\mathcal{E}_2 = \{y_1, \dots, y_n\}$ de X et de Y .
- $X \sim F_1$ et $Y \sim F_2$. L'hypothèse nulle est $H_0 : F_1 \equiv F_2$ (identité de deux distributions). On dispose de deux échantillons indépendants $\mathcal{E}_1 = \{x_1, \dots, x_m\}$ et $\mathcal{E}_2 = \{y_1, \dots, y_n\}$ de X et de Y .

Il y a une approche paramétrique et une approche non-paramétrique aux tests bootstrap. Nous décrivons uniquement une approche "semi-paramétrique" qui assume que μ , μ_1 et μ_2 sont des paramètres de positions de F , F_1 et F_2 , respectivement. On dit que μ est un *paramètre de position* de X si $X \sim G(x - \mu)$ pour une certaine distribution G , non nécessairement spécifiée par un modèle.

Les tests bootstrap procèdent en deux étapes:

- Estimation des distributions des variables aléatoires X et Y sous H_0 .
- Simulation d'échantillons selon les distributions estimées en (1) et calcul de valeurs simulées de la statistique de test.

Test bilatéral de $H_0 : \mu = \mu_0$ au niveau α

Comme dans le t-test pour un échantillon il convient d'utiliser la statistique

$$S = \frac{\hat{\mu} - \mu_0}{\hat{\sigma}(\hat{\mu})},$$

où $\hat{\sigma}(\hat{\mu})$ est l'écart type empirique de $\hat{\mu}$. On estime F par la distribution empirique des données centrées en μ_0 , c'est-à-dire la distribution empirique de $\tilde{x}_i = x_i - \hat{\mu} + \mu_0$ ($i = 1, \dots, n$) et on note par \hat{F}_0 cette estimation. Ensuite, les pas suivants sont accomplis:

- Calculer la valeur de S pour l'échantillon \mathcal{E} donné: soit s_0 la valeur observée.
- Simuler r échantillons selon \hat{F}_0 , et obtenir ainsi r valeurs simulées s_i^* de S :

$$s_i^* = \frac{\hat{\mu}_i^* - \mu_0}{\hat{\sigma}(\hat{\mu}_i^*)}, \quad i = 1, \dots, r.$$

- Calculer la *p-value bootstrap*

$$p^* = \frac{1}{r} (\text{Nombre de } s_i^* > s_0)$$

et rejeter H_0 si $p^* < \alpha/2$ ou si $p^* > 1 - \alpha/2$.

Test bilatéral de $H_0 : \mu_1 = \mu_2$ au niveau α

Il convient d'utiliser une statistique similaire à celle du t-test pour deux échantillons:

$$S = \frac{\hat{\mu}_1 - \hat{\mu}_2}{\sqrt{\hat{\sigma}^2(\hat{\mu}_1) + \hat{\sigma}^2(\hat{\mu}_2)}},$$

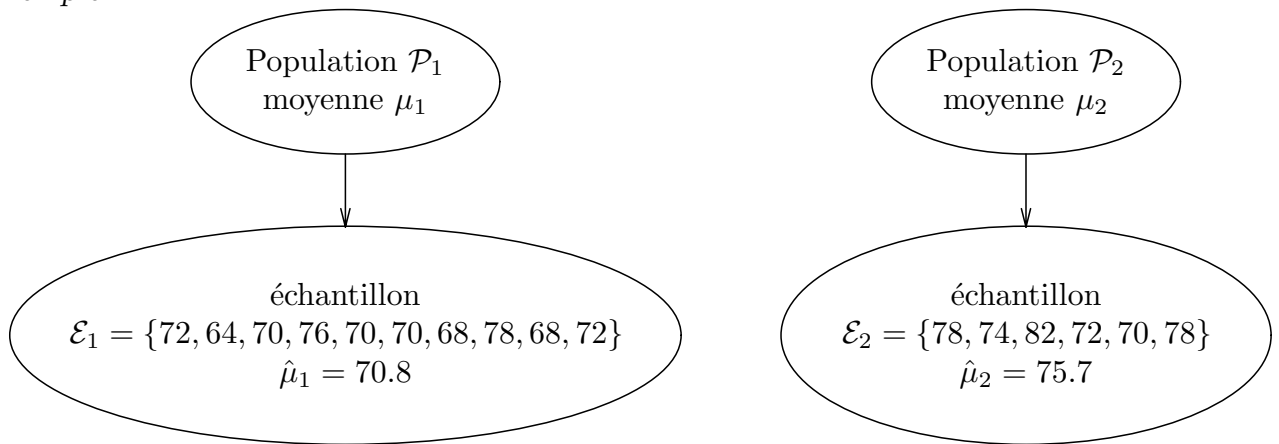
où $\hat{\mu}_1 = \sum x_i/m$, $\hat{\mu}_2 = \sum y_j/n$; $\hat{\sigma}(\hat{\mu}_1)$ et $\hat{\sigma}(\hat{\mu}_2)$ sont les écarts type empiriques de $\hat{\mu}_1$ et $\hat{\mu}_2$. On définit les estimations \hat{F}_{1_0} de F_1 et \hat{F}_{2_0} de F_2 comme les distributions empiriques des échantillons déplacés $\tilde{\mathcal{E}}_1 = \{\tilde{x}_1, \dots, \tilde{x}_m\}$ et $\tilde{\mathcal{E}}_2 = \{\tilde{y}_1, \dots, \tilde{y}_n\}$ où $\tilde{x}_i = x_i - \hat{\mu}_1$ ($i = 1, \dots, m$) et $\tilde{y}_j = y_j - \hat{\mu}_2$ ($j = 1, \dots, n$). Ensuite, les pas suivants sont accomplis:

1. Calculer la valeur s_0 de S pour les données originales:
2. Simuler r paires échantillons tirés de \hat{F}_{1_0} et \hat{F}_{2_0} et calculer r valeurs simulées s_1^*, \dots, s_r^* de S .
3. Calculer

$$p^* = \frac{1}{r} (\text{Nombre de } s_i^* > s_0)$$

et rejeter H_0 si $p^* < \alpha/2$ ou si $p^* > 1 - \alpha/2$.

Exemple



On obtient

$$\hat{\sigma}^2(\hat{\mu}_1) = \frac{1}{10} \cdot \frac{1}{9} \sum_{i=1}^{10} (x_i - 70.8)^2 = 1.6,$$

$$\hat{\sigma}^2(\hat{\mu}_2) = \frac{1}{6} \cdot \frac{1}{5} \sum_{j=1}^6 (y_j - 75.7)^2 = 2.3,$$

et $s_0 = 2.48$. On simule ensuite r paires d'échantillons selon \hat{F}_{1_0} et \hat{F}_{2_0} par tirage avec remise des éléments des échantillons déplacés

$$\tilde{\mathcal{E}}_1 = \{1.2, -6.8, -0.8, 5.2, -0.8, -0.8, -2.8, 7.2, -2.8, 1.2\},$$

$$\tilde{\mathcal{E}}_2 = \{2.3, -1.7, 6.3, -3.7, -5.7, 2.3\},$$

etc.

Test de $H_0 : F \equiv G$ au niveau α

Plusieurs statistiques peuvent être utilisées, par exemple

$$S = \frac{\hat{\mu}_1 - \hat{\mu}_2}{\sqrt{\hat{\sigma}^2(\hat{\mu}_1) + \hat{\sigma}^2(\hat{\mu}_2)}}$$

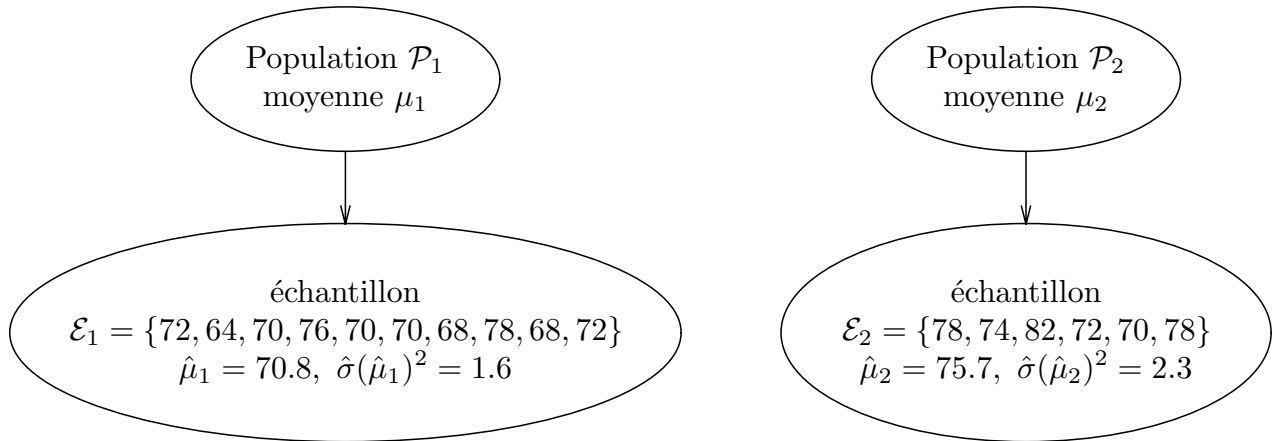
comme dans le cas précédent. Sous H_0 , les distributions de X et de Y sont les mêmes. On peut donc considérer l'échantillon combiné $\mathcal{E} = \{x_1, \dots, x_n, y_1, \dots, y_m\}$. et estimer F et G par la distribution empirique F_{n+m} de \mathcal{E} . Ensuite, les pas suivants sont accomplis:

1. Calculer la valeur s_0 de S pour les données originales.
2. Simuler r échantillons de taille n et r échantillons de taille m par tirages aléatoires avec remise des éléments de l'échantillon combiné et obtenir r valeurs simulées s_1^*, \dots, s_r^* de S .
3. Calculer

$$p^* = \frac{1}{r} (\text{Nombre de } s_i^* > s_0)$$

et rejeter H_0 si $p^* < \alpha/2$ ou si $p^* > 1 - \alpha/2$.

Exemple



On obtient $s_0 = 2.48$ et l'échantillon combiné est

$$\mathcal{E} = \{72, 64, 70, 76, 70, 70, 68, 78, 68, 72, 78, 74, 82, 72, 70, 78\}$$

On construit r échantillons combinés simulés \mathcal{E}^* de taille $n + m = 16$ par tirage avec remise de \mathcal{E} et on décompose chaque \mathcal{E}^* en deux parties \mathcal{E}_1^* et \mathcal{E}_2^* de tailles 10 et 6 respectivement. On obtient par exemple:

$$\mathcal{E}^* = \underbrace{\{64, 64, 72, 72, 68, 70, 68, 78, 74, 82\}}_{\mathcal{E}_1^*} \quad \underbrace{\{64, 74, 70, 72, 78, 78\}}_{\mathcal{E}_2^*}$$

$$\hat{\mu}_1^* = 71.2, \hat{\sigma}(\hat{\mu}_1^*)^2 = 3.31 \quad \hat{\mu}_2^* = 72.7, \hat{\sigma}(\hat{\mu}_2^*)^2 = 3.74$$

$$s_1^* = (\hat{\mu}_2^* - \hat{\mu}_1^*) / \sqrt{\hat{\sigma}(\hat{\mu}_1^*)^2 + \hat{\sigma}(\hat{\mu}_2^*)^2} = 0.57$$

etc.

Remarque. On procède de la même manière pour tester $H_0 : p_1 = p_2$, l'égalité de 2 proportions. Les échantillons tirés seront formés de 1 et de 0 où 1 indique la présence du caractère considéré et 0 son absence. Evidemment, la statistique de test sera adaptée.

16.6 Tests de permutation

Nous présentons ci-dessous une famille de tests basés sur des procédés de simulation similaires aux tests bootstrap: les tests de permutation. Un test de permutation permet de tester par exemple:

- (c) $H_0 : F \equiv G$,
- (d) $H_0 : 2$ caractères X et Y sont indépendants,
- (e) H_0 : un certain traitement n'a pas d'effet.

La marche-à-suivre présentée ci-dessous est en gros la même pour tous les tests de permutation. Nous la présentons dans le cadre dans le cas d'une étude randomisée (Chapitre 15) et de l'hypothèse (e).

Test de permutation pour une étude randomisée

Soient k individus d'une population. Donnons de manière aléatoire à m d'entre eux un certain traitement et aux $k - m$ restants un placebo. Notons $\mathcal{E}_1 = \{x_1, \dots, x_m\}$ et $\mathcal{E}_2 = \{y_1, \dots, y_{k-m}\}$ les ensembles des réponses obtenues sous traitement, respectivement sous placebo et remarquons que, sous H_0 , toutes les permutations aléatoires de ces réponses sont équiprobables. Les pas suivants seront accomplis:

1. Calculer les moyennes $\hat{\mu}_1$ et $\hat{\mu}_2$ de \mathcal{E}_1 et \mathcal{E}_2 .
Construire l'ensemble combiné des réponses $\mathcal{E} = \{x_1, \dots, x_m, y_1, \dots, y_{k-m}\}$
2. Simuler $r \leq \binom{k}{m}$ ensembles \mathcal{E}^* de réponses de taille k par permutation aléatoire (ou tirage *sans* remise) de \mathcal{E} . Partager chaque \mathcal{E}^* en deux parties \mathcal{E}_1^* et \mathcal{E}_2^* de tailles m et $k - m$ respectivement. Calculer les moyennes de \mathcal{E}_1^* et \mathcal{E}_2^* et obtenir ainsi r paires de moyennes simulées $\hat{\mu}_{1i}^*$ et $\hat{\mu}_{2i}^*$ ($i = 1, \dots, r$).
3. Calculer

$$p^* = \frac{1}{r}(\text{Nombre de } (\hat{\mu}_{2i}^* - \hat{\mu}_{1i}^*) > (\hat{\mu}_2 - \hat{\mu}_1))$$

et rejeter H_0 si $p^* < \alpha/2$ ou si $p^* > 1 - \alpha/2$.

Exemple

$\mathcal{E}_1 = \{0, 1, 0, 0, 0, 1, 1, 0, 0, 0\}$, $\mathcal{E}_2 = \{0, 1, 1, 0, 1, 1, 0, 1, 1, 1, 0\}$, $\hat{\mu}_1 = 0.33$, $\hat{\mu}_2 = 0.67$, $\hat{\mu}_2 - \hat{\mu}_1 = 0.34$. On a $\mathcal{E} = \{0, 1, 0, 0, 0, 1, 1, 0, 0, 0, 0, 1, 1, 0, 1, 1, 0, 1, 1, 1, 0\}$ et, par exemple,

$$\mathcal{E}^* = \left\{ \underbrace{1, 1, 0, 1, 0, 0, 1, 0, 0, 1}_{\substack{\mathcal{E}_1^* \\ \hat{\mu}_1^* = \frac{5}{10}}}, \underbrace{1, 0, 1, 1, 1, 1, 0, 0, 0, 1, 0, 0}_{\substack{\mathcal{E}_2^* \\ \hat{\mu}_2^* = \frac{6}{12}}} \right\}$$

$$\hat{\mu}_2^* - \hat{\mu}_1^* = 0.$$

Remarques

1. Les tests de randomisation peuvent être généralisés à la comparaison de plusieurs groupes.
2. Les tests de permutation sont dûs à R.A.Fisher (années 1930). L'idée générale est simple; la réalisation nécessite peu d'hypothèses mathématiques.