Tests statistiques par permutation

Master parcours SSD - UE Statistique Computationnelle

Pierre Mahé - bioMérieux & Université de Grenoble-Alpes

Définition

- Approche par permutation pour réaliser des tests statistiques
- Approche non-paramétrique : pas d'hypothèse sur la loi de la variable aléatoire sous-jacente.
- Principe : construire la distribution de la statistique de test sous H_0 par ré-échantillonnage
- Versions exactes ou approximées ("randomisées")
- Particulièrement intéressant si peu d'observations

Introduction

Exemple introductif

On a mesuré la pression sanguine dans deux populations de souris soumises ou non un médicament et obtenu les valeurs suivantes :

$$x_A = \{89, 90, 92, 93, 93, 96, 99, 99, 99, 102, 103, 104, 105, 106, 106, 107, 108, 108, 110, 110, 112, 114, 116, 116\}$$
$$x_B = \{86, 88, 89, 89, 92, 93, 94, 94, 94, 95, 95, 96, 96, 97, 97, 98, 98, 99, 99, 101, 106, 107, 110, 113, 116, 118\}$$

Exemple introductif

On a mesuré la pression sanguine dans deux populations de souris soumises ou non un médicament et obtenu les valeurs suivantes :

$$x_A = \{89, 90, 92, 93, 93, 96, 99, 99, 99, 102, 103, 104, 105, 106, 106, 107, 108, 108, 110, 110, 112, 114, 116, 116\}$$

$$x_B = \{86, 88, 89, 89, 92, 93, 94, 94, 94, 95, 95, 96, 96, 97, 97, 98, 98, 99, 99, 101, 106, 107, 110, 113, 116, 118\}$$

On veut tester si le médicament a un effet :

$$H_0 : \mu_A = \mu_B \quad \text{contre} \quad H_1 : \mu_A \neq \mu_B$$

Exemple introductif

On a mesuré la pression sanguine dans deux populations de souris soumises ou non un médicament et obtenu les valeurs suivantes :

$$x_A = \{89, 90, 92, 93, 93, 96, 99, 99, 99, 102, 103, 104, \\ 105, 106, 106, 107, 108, 108, 110, 110, 112, 114, 116, 116\}$$
$$x_B = \{86, 88, 89, 89, 92, 93, 94, 94, 94, 95, 95, 96, 96, \\ 97, 97, 98, 98, 99, 99, 101, 106, 107, 110, 113, 116, 118\}$$

On veut tester si le médicament a un effet :

$$H_0 : \mu_A = \mu_B \quad \text{contre} \quad H_1 : \mu_A \neq \mu_B$$

$\Rightarrow$ le test de Student nous donne une p-valeur de 0.048...

Exemple introductif

On a mesuré la pression sanguine dans deux populations de souris soumises ou non un médicament et obtenu les valeurs suivantes :

$$x_A = \{89, 90, 92, 93, 93, 96, 99, 99, 99, 102, 103, 104, \\ 105, 106, 106, 107, 108, 108, 110, 110, 112, 114, 116, 116\}$$
$$x_B = \{86, 88, 89, 89, 92, 93, 94, 94, 94, 95, 95, 96, 96, \\ 97, 97, 98, 98, 99, 99, 101, 106, 107, 110, 113, 116, 118\}$$

On veut tester si le médicament a un effet :

$$H_0 : \mu_A = \mu_B \quad \text{contre} \quad H_1 : \mu_A \neq \mu_B$$

$\Rightarrow$ le test de Student nous donne une p-valeur de 0.048...

$\Rightarrow$ est-on confiant pour rejeter H_0 ?

Exemple introductif

Quelles sont les hypothèses importantes du test de Student ?

Exemple introductif

Quelles sont les hypothèses importantes du test de Student ?

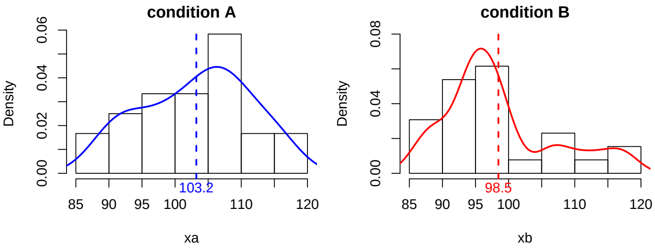
1. échantillons indépendants
2. populations (approximativement) normales
   - robuste si +/- normales (symétriques et unimodales)

Exemple introductif

Quelles sont les hypothèses importantes du test de Student ?

1. échantillons indépendants
2. populations (approximativement) normales
   - robuste si +/- normales (symétriques et unimodales)

Distributions observées ($n = 25$) :

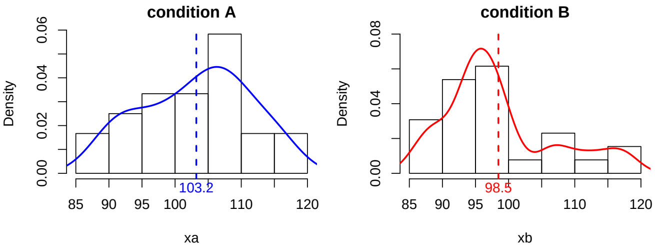


Exemple introductif

Quelles sont les hypothèses importantes du test de Student ?

1. échantillons indépendants
2. populations (approximativement) normales
   - robuste si +/- normales (symétriques et unimodales)

Distributions observées ($n = 25$) :



$\Rightarrow$ dur de vérifier / de se convaincre qu'elles sont normales

- (sans autre information a priori)

Exemple introductif

Vers une approche non paramétrique ?

Exemple introductif

Vers une approche non paramétrique ?

Non paramétrique : pas d'hypothèse sur la distribution de la statistique de test sous H_0

- vs ici une loi de Student à $(2n-2)$ degrés de liberté

Exemple introductif

Vers une approche non paramétrique ?

Non paramétrique : pas d'hypothèse sur la distribution de la statistique de test sous H_0

- vs ici une loi de Student à $(2n-2)$ degrés de liberté

Revenons à notre l'hypothèse nulle : $H_0 : \mu_A = \mu_B$

Exemple introductif

Vers une approche non paramétrique ?

Non paramétrique : pas d'hypothèse sur la distribution de la statistique de test sous H_0

- vs ici une loi de Student à $(2n-2)$ degrés de liberté

Revenons à notre l'hypothèse nulle : $H_0 : \mu_A = \mu_B$

- si la moyenne est la même dans les deux groupes,

Exemple introductif

Vers une approche non paramétrique ?

Non paramétrique : pas d'hypothèse sur la distribution de la statistique de test sous H_0

- vs ici une loi de Student à $(2n-2)$ degrés de liberté

Revenons à notre l'hypothèse nulle : $H_0 : \mu_A = \mu_B$

- si la moyenne est la même dans les deux groupes,
- alors l'affectation de chaque observation importe peu,

Exemple introductif

Vers une approche non paramétrique ?

Non paramétrique : pas d'hypothèse sur la distribution de la statistique de test sous H_0

- vs ici une loi de Student à $(2n-2)$ degrés de liberté

Revenons à notre l'hypothèse nulle : $H_0 : \mu_A = \mu_B$

- si la moyenne est la même dans les deux groupes,
- alors l'affectation de chaque observation importe peu,
- donc si on les ré-affecte on reste sous H_0.

Exemple introductif

Vers une approche non paramétrique ?

Non paramétrique : pas d'hypothèse sur la distribution de la statistique de test sous H_0

- vs ici une loi de Student à $(2n-2)$ degrés de liberté

Revenons à notre l'hypothèse nulle : $H_0 : \mu_A = \mu_B$

- si la moyenne est la même dans les deux groupes,
- alors l'affectation de chaque observation importe peu,
- donc si on les ré-affecte on reste sous H_0.

$\Rightarrow$ en affectant aléatoirement les observations aux groupes, on peut **construire** la distribution de la statistique sous H_0.

Exemple introductif

Procédure de ré-échantillonage :

1. ré-affecter les observations aux groupes par permutations
2. calculer les valeurs de la statistique de test
3. utiliser cette distribution comme distribution sous H_0

Exemple introductif

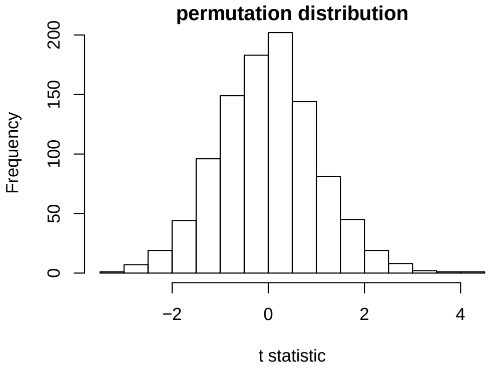
Procédure de ré-échantillonage :

1. ré-affecter les observations aux groupes par permutations
2. calculer les valeurs de la statistique de test
3. utiliser cette distribution comme distribution sous H_0

$\Rightarrow$ on estime alors la p-valeur en comparant la valeur observée aux quantiles de cette distribution.

Exemple introductif

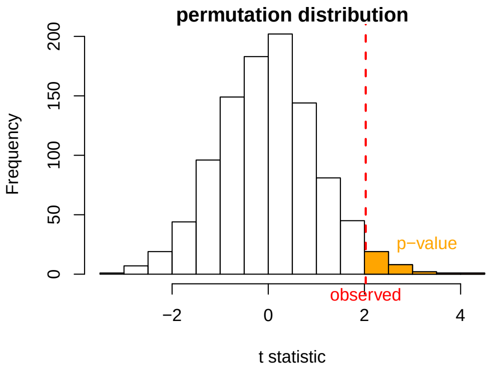
Illustration :



- distribution de la statistique du test de Student pour 1000 permutations.

Exemple introductif

Illustration :



- p-valeur= proportion de valeurs supérieures à l'observée
  - estimation empirique de la définition théorique

Définition

Plus formellement

On considère les tests de comparaison d'échantillons :

$$H_0 : P_X = P_Y \quad \text{contre} \quad H_1 : P_X \neq P_Y.$$

P_X et P_Y = distributions générant les échantillons X et Y :

- $X = \{X_1, ..., X_n\}$, de taille n
- $Y = \{Y_1, ..., Y_m\}$, de taille m

On considère une statistique de test $\hat{\theta}(X, Y)$.

Plus formellement (sur notre exemple)

On considère les tests de comparaison d'échantillons :

$$H_0 : P_X = P_Y \quad \text{contre} \quad H_1 : P_X \neq P_Y.$$

$\Rightarrow H_0 : \mu_0 = \mu_1 \quad \text{contre} \quad H_1 : \mu_0 \neq \mu_1.$

P_X et P_Y : distributions générant les échantillons X et Y :

- $X = \{X_1, ..., X_n\}$, de taille n
- $Y = \{Y_1, ..., Y_m\}$, de taille m

$\Rightarrow P_X = \mathcal{N}(\mu_0, \sigma)$ et $P_Y = \mathcal{N}(\mu_1, \sigma)$

On considère une statistique de test $\hat{\theta}(X, Y)$.

$\Rightarrow t = (\bar{X} - \bar{Y})/s_p\sqrt{\frac{1}{n} + \frac{1}{m}}, \;\; s_p = \sqrt{\frac{(n-1)s_X^2 + (m-1)s_Y^2}{n+m-2}}$

Plus formellement

Sous H_0 : X, Y et $Z = (X, Y)$ sont générés par P_X.

Plus formellement

Sous H_0 : X, Y et $Z = (X, Y)$ sont générés par P_X.

$\Rightarrow$ La distribution de permutation de $\hat{\theta}$ est donnée par :

$$\{\hat{\theta}^*\} = \big\{\hat{\theta}(X^*, Y^*)\big\},$$

où $\big\{(X^*, Y^*)\big\}$ = **toutes** les permutations de Z.

Plus formellement

Sous H_0 : X, Y et $Z = (X, Y)$ sont générés par P_X.

$\Rightarrow$ La distribution de permutation de $\hat{\theta}$ est donnée par :

$$\{\hat{\theta}^*\} = \{\hat{\theta}(X^*, Y^*)\},$$

où $\{(X^*, Y^*)\}$ = **toutes** les permutations de Z.

$\Rightarrow$ chaque permutation $Z_\pi = Z^* = (X^*, Y^*)$ a la probabilité :

$$\frac{1}{C_N^n} = \frac{n!(N-n)!}{N!} = \frac{n!m!}{N!} = \frac{n!m!}{(n+m)!}$$

d'être observée.

- c'est le "**permutation lemma**"
- π = permutation du vecteur $\{1, ..., N\}$

Plus formellement

On peut donc **construire** la distribution de $\hat{\theta}^*$ sous H_0 :

$$F_{\hat{\theta}^*}(t) = P(\hat{\theta}^* \leq t) = \frac{1}{C_N^n} \sum_{i=1}^{C_N^n} \mathbf{1}(\hat{\theta}^*(i) \leq t).$$

$\Rightarrow$ on compte les permutations pour lesquelles $\hat{\theta}^* \leq t$.

Plus formellement

On peut donc **construire** la distribution de $\hat{\theta}^*$ sous H_0 :

$$F_{\hat{\theta}^*}(t) = P(\hat{\theta}^* \leq t) = \frac{1}{C_N^n} \sum_{i=1}^{C_N^n} \mathbf{1}(\hat{\theta}^*(i) \leq t).$$

$\Rightarrow$ on compte les permutations pour lesquelles $\hat{\theta}^* \leq t$.

On en déduit de la même manière la p-valeur de la statistique observée $\hat{\theta} = \hat{\theta}(X, Y)$:

$$P(\hat{\theta}^* \geq \hat{\theta}) = \frac{1}{C_N^n} \sum_{i=1}^{C_N^n} \mathbf{1}(\hat{\theta}^*(i) \geq \hat{\theta})$$

- pour une hypothèse alternative unilatérale supérieure
- on procède de même si inférieure ou bilatérale

En pratique

En pratique, difficile d'évaluer **toutes** les permutations.

En pratique

En pratique, difficile d'évaluer **toutes** les permutations.

On procède par Monte Carlo :

1. calculer la statistique sur les données : $\hat{\theta} = \hat{\theta}(X, Y)$
2. Pour $b = 1, ..., B$:
   - générer une permutation $\pi^{(b)}$
   - calculer $\hat{\theta}^{(b)} = \hat{\theta}^*(X^{*(b)}, Y^{*(b)})$
3. Calculer la p-valeur (empirique) comme :

$$\hat{p} = \frac{1 + \#\{\hat{\theta}^{(b)} \geq \hat{\theta}\}}{B+1} = \frac{1 + \sum_{b=1}^{B} \mathbf{1}(\hat{\theta}^{(b)} \geq \hat{\theta})}{B+1}$$

- là aussi, pour une alternative unilatérale supérieure.
- on ajoute 1 au numérateur et au dénominateur car sous H_0 on peut inclure $\hat{\theta}$ dans la distribution de permutation.

En pratique

Exemple de mise en oeuvre en R :

```
> B = 1000       # nombre de permutations
> z = c(x,y)
> t0 = t.test(x,y)$statistic
> t.perm = numeric(B)
> for(b in 1:B){
ind = sample(length(z),n) # n = length(x)
x1 = z[ind]
y1 = z[-ind]
t.perm[b] = t.test(x1,y1)$statistic }
> t.perm = c(t.perm, t0)
> pval = mean(t.perm >= t0)
```

Hypothèses, avantages & inconvénients

Hypothèse : 1 seule mais critique = échangeabilité sous H_0

- exemple précédent : $\Rightarrow$ même variance dans les groupes.

Hypothèses, avantages & inconvénients

Hypothèse : 1 seule mais critique = échangeabilité sous H_0

- exemple précédent : $\Rightarrow$ même variance dans les groupes.

Avantages :

- approche non paramétrique
  - hypothèses jamais parfaitement vérifiées/vérifiables
- applicable quand peu d'observations
- permet de considérer d'autres statistiques de test
  - e.g., pour lesquelles on n'a pas de forme paramétrique
  - exemple précédent : différence de médianes

Hypothèses, avantages & inconvénients

Hypothèse : 1 seule mais critique = échangeabilité sous H_0

- exemple précédent : $\Rightarrow$ même variance dans les groupes.

Avantages :

- approche non paramétrique
  - hypothèses jamais parfaitement vérifiées/vérifiables
- applicable quand peu d'observations
- permet de considérer d'autres statistiques de test
  - e.g., pour lesquelles on n'a pas de forme paramétrique
  - exemple précédent : différence de médianes

Inconvénient :

- coût calculatoire
- – pertinent/applicable quand beaucoup d'observations
  - mais les tests paramétriques deviennent alors valides

Lien avec les épisodes précédents...

Lien avec "méthodes MC pour l'inférence"

Lien avec les épisodes précédents...

Lien avec "méthodes MC pour l'inférence" .... c'est une méthode MC pour l'inférence !

Lien avec les épisodes précédents...

Lien avec "méthodes MC pour l'inférence" .... c'est une méthode MC pour l'inférence !

Néanmoins avant :

1. on simulait des données selon un modèle
2. on évaluait les performances d'un test paramétrique

$\Rightarrow$ Intérêt = évaluer effet de la taille de l'échantillon et/ou des écarts aux hypothèses.

Lien avec les épisodes précédents...

Lien avec "méthodes MC pour l'inférence" .... c'est une méthode MC pour l'inférence !

Néanmoins avant :

1. on simulait des données selon un modèle
2. on évaluait les performances d'un test paramétrique

$\Rightarrow$ Intérêt = évaluer effet de la taille de l'échantillon et/ou des écarts aux hypothèses.

Ici :

1. approche totalement non-paramétrique
2. on ne simule pas de nouvelles données
3. on est rigoureusement orienté "test"
    - approche moins applicable pour estimer des IC

Lien avec les épisodes précédents...

Lien avec le bootstrap :

- les deux approches sont non paramétriques
- le bootstrap est davantage orienté estimation et IC
  - on peut également faire des tests par bootstrap mais c'est moins courant
- avec le bootstrap on tire avec remise, ici on permute
  - ici chaque observation est utilisée 1 seule fois par tirage
  - avec le bootstrap, certaines apparaitront plusieurs fois

Tests exacts et approximés

Si on considère **toutes** les permutations : le test est **exact**.

- on obtient la distribution de toutes les valeurs possibles sous H_0

$\Rightarrow$ il contrôle le risque de 1ère espèce au niveau attendu.

Tests exacts et approximés

Si on considère **toutes** les permutations : le test est **exact**.

- on obtient la distribution de toutes les valeurs possibles sous H_0

$\Rightarrow$ il contrôle le risque de 1ère espèce au niveau attendu.

Les tests paramétriques usuels sont également exacts...dès lors que leurs hypothèses sont **parfaitement** vérifiées.

- pas forcément le cas en pratique

$\Rightarrow$ terminologie : tests exacts $=$ tests par permutation

Tests exacts et approximés

Si on considère **toutes** les permutations : le test est **exact**.

- on obtient la distribution de toutes les valeurs possibles sous H_0

$\Rightarrow$ il contrôle le risque de 1ère espèce au niveau attendu.

Les tests paramétriques usuels sont également exacts...dès lors que leurs hypothèses sont **parfaitement** vérifiées.

- pas forcément le cas en pratique

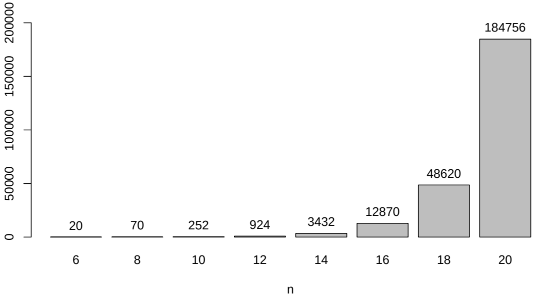
$\Rightarrow$ terminologie : tests exacts = tests par permutation

En pratique, on peut difficilement considérer toutes les permutations : on procède par Monte Carlo.
$\Rightarrow$ on parle de tests approximés ou "randomisés"

Tests exacts et approximés

Illustration du nombre de permutations $C_n^{n/2}$:

- i.e., choisir $n/2$ points parmi n



- $n = 30$ (15 observations par groupe) : $> 155 \times 10^6$
- $n = 40$ (20 observations par groupe) : $> 137 \times 10^9$

Applications

Applications

Quelques applications typiques :

- test d'égalité de moyennes
- test de corrélation
- test d'adéquation
- test d'indépendance de deux variables aléatoires

Applications

Quelques applications typiques :

- test d'égalité de moyennes
- test de corrélation
- test d'adéquation
- test d'indépendance de deux variables aléatoires

Démarche générale :

1. bien poser l'hypothèse nulle
2. en déduire une stratégie de permutation
3. implémenter la procédure (exacte ou par MC)
   - permutations + calcul p-valeur

Test d'égalité de moyennes

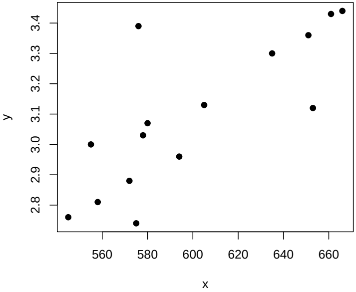
Voir l'exemple introductif.

Remarques :

- permet de faire une version non paramétrique du t-test
  - pour petits échantillons, pas très Gaussiens...
- permet de le généraliser à d'autres statistiques de test
  - simple différence des moyennes, de médianes, ...
  - ... ou des choses arbitrairement plus compliquées !

Test de corrélation

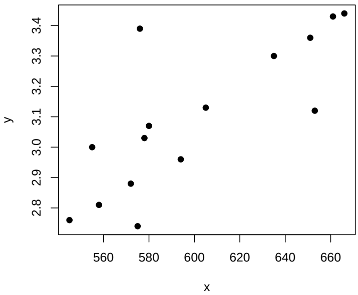
On observe deux variables aléatoires sur un échantillon :



Leur corrélation ρ est-elle significativement différente de 0 ?

Test de corrélation

On observe deux variables aléatoires sur un échantillon :



Leur corrélation ρ est-elle significativement différente de 0 ?

$\Rightarrow$ TP : proposer une procédure de permutation pour tester

$$H_0 : \rho = 0 \quad \text{contre} \quad H_1 : \rho \neq 0.$$

Test d'adéquation

Exemple : test d'adéquation à une loi multinomiale

- soit une variable aléatoire X pouvant prendre K valeurs
- on observe un échantillon $(X_1, ..., X_n)$ iid $\sim$ la loi de X
- on veut tester si X suit une loi multinomiale $(p_1, ..., p_K)$
  - $p_i = P(X = i)$, $p_i \geq 0 \; \forall i$, $\sum_{i=1}^{K} p_i = 1$

Test d'adéquation

Exemple : test d'adéquation à une loi multinomiale

- soit une variable aléatoire X pouvant prendre K valeurs
- on observe un échantillon $(X_1, ..., X_n)$ iid $\sim$ la loi de X
- on veut tester si X suit une loi multinomiale $(p_1, ..., p_K)$
  - $p_i = P(X = i)$, $p_i \geq 0 \; \forall i$, $\sum_{i=1}^{K} p_i = 1$

Cadre paramétrique : test du χ^2 (d'adéquation).

- statistique de test : $\displaystyle\sum_{i=1}^{K} \frac{(np_i - n\hat{p}_i)^2}{n} \rightarrow \chi^2(K-1)$

Test d'adéquation

Exemple : test d'adéquation à une loi multinomiale

- soit une variable aléatoire X pouvant prendre K valeurs
- on observe un échantillon $(X_1, ..., X_n)$ iid $\sim$ la loi de X
- on veut tester si X suit une loi multinomiale $(p_1, ..., p_K)$
  - $p_i = P(X = i)$, $p_i \geq 0 \; \forall i$, $\sum_{i=1}^{K} p_i = 1$

Cadre paramétrique : test du χ^2 (d'adéquation).

- statistique de test : $\displaystyle\sum_{i=1}^{K} \frac{(np_i - n\hat{p}_i)^2}{n} \rightarrow \chi^2(K-1)$

Limite du test : imprécis si n est petit.

- au moins 5 représentants par catégories

Test d'adéquation

Exemple : test d'adéquation à une loi multinomiale

- soit une variable aléatoire X pouvant prendre K valeurs
- on observe un échantillon $(X_1, ..., X_n)$ iid $\sim$ la loi de X
- on veut tester si X suit une loi multinomiale $(p_1, ..., p_K)$
  - $p_i = P(X = i)$, $p_i \geq 0 \; \forall i$, $\sum_{i=1}^{K} p_i = 1$

Cadre paramétrique : test du χ^2 (d'adéquation).

- statistique de test : $\displaystyle\sum_{i=1}^{K} \frac{(np_i - n\hat{p}_i)^2}{n} \to \chi^2(K-1)$

Limite du test : imprécis si n est petit.

- au moins 5 représentants par catégories

$\Rightarrow$ TP : proposer une procédure de permutation pour ce test.

Test d'indépendance entre variables aléatoires

Exemple : indépendance entre deux variables binaires

	B = 0	B = 1	total
A = 0	a	b	a + b
A = 1	c	d	c + d
total	a + c	b + d	a+b+c+d = n

$\Rightarrow$ test d'indépendance : $H_0 : P(AB) = P(A)P(B)$.

Test d'indépendance entre variables aléatoires

Exemple : indépendance entre deux variables binaires

	B = 0	B = 1	total
A = 0	a	b	a + b
A = 1	c	d	c + d
total	a + c	b + d	a+b+c+d = n

$\Rightarrow$ test d'indépendance : $H_0 : P(AB) = P(A)P(B)$.

Illustration avec un échantillon de taille $n = 20$:

	B = 0	B = 1	total
A = 0	5	5	10
A = 1	5	5	10
total	10	10	20

$\Rightarrow$ très probablement indépendant

	B = 0	B = 1	total
A = 0	1	9	10
A = 1	9	1	10
total	10	10	20

$\Rightarrow$ probablement pas indépendant

Test d'indépendance entre variables aléatoires

Exemple : indépendance entre deux variables binaires

	B = 0	B = 1	total
A = 0	a	b	a + b
A = 1	c	d	c + d
total	a + c	b + d	a+b+c+d = n

$\Rightarrow$ test d'indépendance : $H_0 : P(AB) = P(A)P(B)$.

Test d'indépendance entre variables aléatoires

Exemple : indépendance entre deux variables binaires

	B = 0	B = 1	total
A = 0	a	b	a + b
A = 1	c	d	c + d
total	a + c	b + d	a+b+c+d = n

$\Rightarrow$ test d'indépendance : $H_0 : P(AB) = P(A)P(B)$.

Cadre paramétrique : test du χ^2 (d'indépendance).

Test d'indépendance entre variables aléatoires

Exemple : indépendance entre deux variables binaires

	B = 0	B = 1	total
A = 0	a	b	a + b
A = 1	c	d	c + d
total	a + c	b + d	a+b+c+d = n

$\Rightarrow$ test d'indépendance : $H_0 : P(AB) = P(A)P(B)$.

Cadre paramétrique : test du χ^2 (d'indépendance).

Limite du test : imprécis si n est petit.

- au moins 5 représentants par catégories

Test d'indépendance entre variables aléatoires

Exemple : indépendance entre deux variables binaires

	B = 0	B = 1	total
A = 0	a	b	a + b
A = 1	c	d	c + d
total	a + c	b + d	a+b+c+d = n

$\Rightarrow$ test d'indépendance : $H_0 : P(AB) = P(A)P(B)$.

Cadre paramétrique : test du χ^2 (d'indépendance).

Limite du test : imprécis si n est petit.

- au moins 5 représentants par catégories

Alternative par permutations = test exact de Fisher
$\Rightarrow$ **exemple fondateur** des tests par permutations

- mais pas stricto sensu un test par permutations

Test exact de Fisher

Test exact de Fisher : analyse de tables de contingences

Test exact de Fisher

Test exact de Fisher : analyse de tables de contingences

Exemple historique : the lady testing tea experiment

- une collègue prétendait reconnaître si le lait avait été mis avant ou après le thè dans sa tasse.

Test exact de Fisher

Test exact de Fisher : analyse de tables de contingences

Exemple historique : the lady testing tea experiment

- une collègue prétendait reconnaître si le lait avait été mis avant ou après le thè dans sa tasse.

Formalisation :

- évènement A : lait mis avant ou non
- evènement B : lait prétendu mis avant ou non

	$B = 0$	$B = 1$	total
$A = 0$	a	b	$a + b$
$A = 1$	c	d	$c + d$
total	$a + c$	$b + d$	$a{+}b{+}c{+}d = n$

Test exact de Fisher

Test exact de Fisher : analyse de tables de contingences

Exemple historique : the lady testing tea experiment

- une collègue prétendait reconnaître si le lait avait été mis avant ou après le thè dans sa tasse.

Formalisation :

- évènement A : lait mis avant ou non
- evènement B : lait prétendu mis avant ou non

	B = 0	B = 1	total
A = 0	a	b	a + b
A = 1	c	d	c + d
total	a + c	b + d	a+b+c+d = n

$\Rightarrow P(AB) = P(A)P(B)$ si la décision est prise au hasard.

Test exact de Fisher - principe

Principe du test exact de Fisher :

Test exact de Fisher - principe

Principe du test exact de Fisher :

1. calculer la probabilité d'observer une table de contingence donnée.

Test exact de Fisher - principe

Principe du test exact de Fisher :

1. calculer la probabilité d'observer une table de contingence donnée.
2. considérer toutes les configurations possibles, pour les **mêmes valeurs marginales**.

Test exact de Fisher - principe

Principe du test exact de Fisher :

1. calculer la probabilité d'observer une table de contingence donnée.
2. considérer toutes les configurations possibles, pour les **mêmes valeurs marginales**.

$\Rightarrow$ la p-valeur du test est égale à la somme des probabilités d'observer une table "plus extrême" que la table observée.

Test exact de Fisher - principe

Principe du test exact de Fisher :

1. calculer la probabilité d'observer une table de contingence donnée.
2. considérer toutes les configurations possibles, pour les **mêmes valeurs marginales**.

$\Rightarrow$ la p-valeur du test est égale à la somme des probabilités d'observer une table "plus extrême" que la table observée.

Table plus "extrême" = plus déséquilibrée que l'observée

- (dans la même direction pour un test unilatéral)

Test exact de Fisher - principe

Principe du test exact de Fisher :

1. calculer la probabilité d'observer une table de contingence donnée.
2. considérer toutes les configurations possibles, pour les **mêmes valeurs marginales**.

$\Rightarrow$ la p-valeur du test est égale à la somme des probabilités d'observer une table "plus extrême" que la table observée.

Table plus "extrême" = plus déséquilibrée que l'observée

- (dans la même direction pour un test unilatéral)

$\Rightarrow$ il faut les énumérer

- NB : pas toutes les tables, seulement les plus extrêmes

Test exact de Fisher - principe

Principe du test exact de Fisher :

1. calculer la probabilité d'observer une table de contingence donnée.
2. considérer toutes les configurations possibles, pour les **mêmes valeurs marginales**.

$\Rightarrow$ la p-valeur du test est égale à la somme des probabilités d'observer une table "plus extrême" que la table observée.

Table plus "extrême" = plus déséquilibrée que l'observée

- (dans la même direction pour un test unilatéral)

$\Rightarrow$ il faut les énumérer

- NB : pas toutes les tables, seulement les plus extrêmes

$\Rightarrow$ la table observée entre dans le calcul de la p-valeur

Test exact de Fisher - probabilité d'une table

Soit la table de contingence :

	B = 0	B = 1	total
A = 0	a	b	a + b
A = 1	c	d	c + d
total	a + c	b + d	a+b+c+d = n

Sa probabilité = la probabilité d'observer (a, b, c, d) est[1] :

$$p = \frac{C^a_{a+b} C^c_{c+d}}{C^{a+c}_n}$$

⇒ loi hypergéometrique

- intuitivement : combinatoire en contraignant 3 marges

1. NB : on peut vérifier qu'elle vaut aussi $C^a_{a+c} C^b_{b+d} / C^{a+b}_n$

Test exact de Fisher - probabilité d'une table

Soit cette première table de contingence :

	B = 0	B = 1	total
A = 0	1	9	10
A = 1	11	3	14
total	12	12	24

$\Rightarrow$ sa probabilité est 0.001346076.

Soit cette seconde table de contingence :

	B = 0	B = 1	total
A = 0	4	6	10
A = 1	8	6	14
total	12	12	24

$\Rightarrow$ sa probabilité est 0.2332077.

$\Rightarrow$ + de possibilité de l'obtenir avec $n = 24$ échantillons.

Test exact de Fisher - tables "extrêmes"

Soit la table de contingence :

	$B = 0$	$B = 1$	total
$A = 0$	1	9	10
$A = 1$	11	3	14
total	12	12	24

Test exact de Fisher - tables "extrêmes"

Soit la table de contingence :

	$B = 0$	$B = 1$	total
$A = 0$	1	9	10
$A = 1$	11	3	14
total	12	12	24

$\Rightarrow$ avec les mêmes marges, on peut obtenir 11 tables

Test exact de Fisher - tables "extrêmes"

Soit la table de contingence :

	B = 0	B = 1	total
A = 0	1	9	10
A = 1	11	3	14
total	12	12	24

$\Rightarrow$ avec les mêmes marges, on peut obtenir 11 tables

$\Rightarrow$ 1 seule est plus extrême :

- quand on met 0 dans la case $(A = 0; B = 0)$

Test exact de Fisher - tables "extrêmes"

Soit la table de contingence :

	B = 0	B = 1	total
A = 0	1	9	10
A = 1	11	3	14
total	12	12	24

⇒ avec les mêmes marges, on peut obtenir 11 tables

⇒ 1 seule est plus extrême :

- quand on met 0 dans la case $(A = 0; B = 0)$

⇒ on obtient la p-valeur en sommant les deux probabilités.

- celle de l'observée + celle de la plus extrême

Test exact de Fisher et permutations ?

Test exact de Fisher : **pas** une procédure de permutations

Analogue par permutations :

- associer une statistique de test à une table de contingence
  - e.g., la statistique du test du χ^2 "classique"
- "tirer" des tables respectant les valeurs marginales
  - $\sim$ permuter les observations
- calculer la p-valeur empririque
  - % de statistiques de test plus elevées que l'observée

$\Rightarrow$ remplacé par l'utilisation conjointe de la loi géométrique
$+$ l'énumération explicite des tables extrêmes

- pas les mêmes moyens informatiques à l'époque...

Test exact de Fisher - mise en oeuvre

Mise en oeuvre en R avec la fonction `fisher.test` :

```
> M = matrix(c(1,9,11,3),nrow=2,byrow=T)
> p1 = fisher.test(M, alternative="less")$p.value
> # NB : consider one-sided alternative
> print(p1)
[1] 0.001379728
> p2 = chisq.test(M)$p.value
> print(p2)
[1] 0.00375221
```

⇒ Exercice : retrouver p_1 en appliquant la formule de la loi géométrique (sur la table et la plus extrême).

Echangeabilité

Echangeabilité

Hypothèse forte d'échangeabilité sous H_0.

- le principe même des permutations

Echangeabilité

Hypothèse forte d'échangeabilité sous H_0.

- le principe même des permutations

Conséquences :

1. hypothèses implicites sur les lois des variables aléatoires
2. traitement spécifique des plans d'expérience structurés

Echangeabilité

Hypothèse forte d'échangeabilité sous H_0.

- le principe même des permutations

Conséquences :

1. hypothèses implicites sur les lois des variables aléatoires
2. traitement spécifique des plans d'expérience structurés

Solutions :

1. en avoir conscience...et mettre en oeuvre des transformations pour préserver/garantir l'échangeabilité
   - e.g., si on connaît les variances par groupe...
2. mettre en oeuvre des stratégies de permutation par blocs pour prendre en compte des covariables

$\Rightarrow$ ne sera pas couvert dans ce cours.

Echangeabilité et différence de moyenne

On considère deux échantillons :

- $(X_1, , ..., X_n)$ distribués selon $\mathcal{N}(\mu_0, \sigma_0)$
- $(Y_1, , ..., Y_n)$ distribués selon $\mathcal{N}(\mu_1, \sigma_0)$

$\Rightarrow$ on veut tester $H_0 : \mu_0 = \mu_1$ vs $H_1 : \mu_0 \neq \mu_1$

Echangeabilité et différence de moyenne

On considère deux échantillons :

- $(X_1, , ..., X_n)$ distribués selon $\mathcal{N}(\mu_0, \sigma_0)$
- $(Y_1, , ..., Y_n)$ distribués selon $\mathcal{N}(\mu_1, \sigma_0)$

$\Rightarrow$ on veut tester $H_0 : \mu_0 = \mu_1$ vs $H_1 : \mu_0 \neq \mu_1$

Sous H_0, $\{X_i\}$ et $\{Y_i\}$ sont iid et donc échangeables.
$\Rightarrow$ le test par permutation est valide.

Echangeabilité et différence de moyenne

On considère deux échantillons :

- $(X_1, , ..., X_n)$ distribués selon $\mathcal{N}(\mu_0, \sigma_0)$
- $(Y_1, , ..., Y_n)$ distribués selon $\mathcal{N}(\mu_1, \sigma_0)$

$\Rightarrow$ on veut tester $H_0 : \mu_0 = \mu_1$ vs $H_1 : \mu_0 \neq \mu_1$

Sous H_0, $\{X_i\}$ et $\{Y_i\}$ sont iid et donc échangeables.
$\Rightarrow$ le test par permutation est valide.

En revanche, si les $\{Y_i\} \to \mathcal{N}(\mu_1, \sigma_1)$, avec $\sigma_0 \neq \sigma_1$:

- $\{X_i\}$ et $\{Y_i\}$ ne sont pas iid sous H_0
- les observations ne sont pas échangeables

$\Rightarrow$ le test par permutation n'est plus valide

- il ne contrôle pas le risque α au niveau attendu

Echangeabilité et différence de moyenne

On considère deux échantillons :

- $(X_1, , ..., X_n)$ distribués selon $\mathcal{N}(\mu_0, \sigma_0)$
- $(Y_1, , ..., Y_n)$ distribués selon $\mathcal{N}(\mu_1, \sigma_0)$

$\Rightarrow$ on veut tester $H_0 : \mu_0 = \mu_1$ vs $H_1 : \mu_0 \neq \mu_1$

Sous H_0, $\{X_i\}$ et $\{Y_i\}$ sont iid et donc échangeables.
$\Rightarrow$ le test par permutation est valide.

En revanche, si les $\{Y_i\} \rightarrow \mathcal{N}(\mu_1, \sigma_1)$, avec $\sigma_0 \neq \sigma_1$:

- $\{X_i\}$ et $\{Y_i\}$ ne sont pas iid sous H_0
- les observations ne sont pas échangeables

$\Rightarrow$ le test par permutation n'est plus valide

- il ne contrôle pas le risque α au niveau attendu

$\Rightarrow$ **ce test fait implicitement l'hypothèse que $\sigma_0 = \sigma_1$.**

Echangeabilité et différence de moyenne

Illustration :

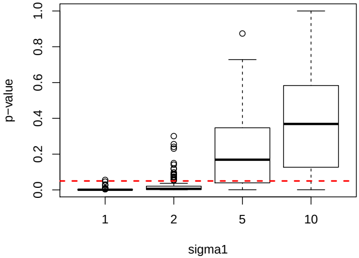
- on génère 2 échantillons à $n = 10$ sous $\mathcal{N}(0,1)$ et $\mathcal{N}(2,\sigma)$ pour $\sigma \in \{1,2,5,10\}$.
- on teste la différence de moyenne par permutation.

Echangeabilité et différence de moyenne

Illustration :

- on génère 2 échantillons à $n = 10$ sous $\mathcal{N}(0,1)$ et $\mathcal{N}(2,\sigma)$ pour $\sigma \in \{1,2,5,10\}$.
- on teste la différence de moyenne par permutation.

Distribution des p-valeurs obtenues pour 1000 répétitions :



$\Rightarrow$ la puissance chute fortement quand σ augmente, i.e., quand les données sont de - en - échangeables

Conclusion

Remarques et conclusion

- Approche non paramétrique des tests d'hypothèses
  - $\neq$ des approches MC précédentes : pas de simulations

Remarques et conclusion

- Approche non paramétrique des tests d'hypothèses
  - $\neq$ des approches MC précédentes : pas de simulations
- Intérêt principal : petits échantillons
  - $+$ généralement : quand hypothèses non vérifiées

Remarques et conclusion

- Approche non paramétrique des tests d'hypothèses
  - $\neq$ des approches MC précédentes : pas de simulations
- Intérêt principal : petits échantillons
  - + généralement : quand hypothèses non vérifiées
- Version par permutation de tests paramétriques
  - utilise les mêmes statistiques de test
  - relâche hypothèse(s) sur leurs distributions sous H_0

Remarques et conclusion

- Approche non paramétrique des tests d'hypothèses
  - $\neq$ des approches MC précédentes : pas de simulations
- Intérêt principal : petits échantillons
  - $+$ généralement : quand hypothèses non vérifiées
- Version par permutation de tests paramétriques
  - utilise les mêmes statistiques de test
  - relâche hypothèse(s) sur leurs distributions sous H_0
- Ouvre la porte à d'autres statistiques de test
  - pour lesquelles on ne connaît pas la distribution sous H_0
  - e.g., plus complexes ou propres à l'application

Remarques et conclusion

- Procédure simple à mettre en oeuvre
  - même type de procédure que cours précédents

Remarques et conclusion

- Procédure simple à mettre en oeuvre
  - même type de procédure que cours précédents
- Versions exactes ou approximées
  - exacte : **toutes** les permutations
  - souvent impossible $\rightarrow$ approximation par MC

Remarques et conclusion

- Procédure simple à mettre en oeuvre
  - même type de procédure que cours précédents
- Versions exactes ou approximées
  - exacte : **toutes** les permutations
  - souvent impossible $\rightarrow$ approximation par MC
- Hypothèse importante d'échangeabilité
  - seule hypothèse mais limitation importante
  - hypothèses implicites sur la loi des variables
  - extensions par transformation ou stratification