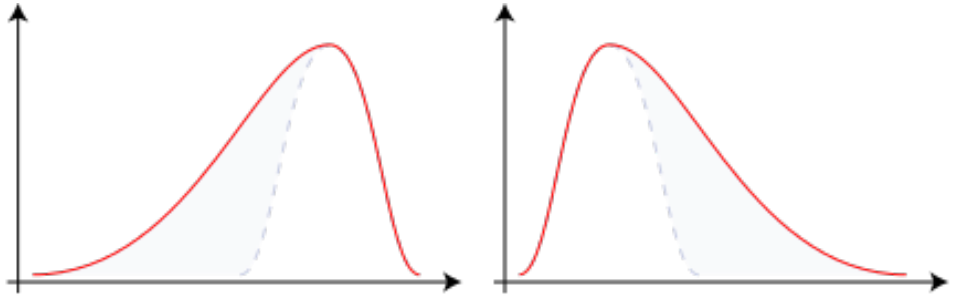
Statistiques non-paramétriques
Ch.1. Tests non-paramétriques 2018-19
M2 CEE

Pr. Philippe Polomé, Université Lumière Lyon 2

Motivation : Quand utiliser ?

- Hyp. de normalité pas facilement acceptable
- Données nominales ou ordinales [+loin]
  - Ou des outliers qu'on ne veut pas enlever
- Tests sur des fréquences, médianes ou quantiles
  - plutôt que des moyennes ou des variances
  - Souvent les mesures associées au revenu sont asymétriques
    - dépenses, dispositions à payer...



  - Alors, la médiane représente mieux les obs. que la moyenne,
  - qui est tirée dans la queue asymétrique de la distribution

Quand utiliser ?

- Quand on a peu de données
  - Les tests np ont peu de puissance (ci-dessous)
  - mais les tests param. peuvent donner des résultats aberrants
    - si l'hyp. de normalité est fausse
- Inconvénients
  - Est plus une collection de recettes
    - Qu'une méthodo bien claire
  - Donc : difficile de s'y retrouver

Sommaire

Rappel sur les tests

Tests

Corrélation non-paramétrique

Rappel sur les tests

- Les tests ont pour objet une hyp. nulle H_0 d'égalité
  - p.e. la moyenne dans une certaine population = celle d'un autre groupe
  - Si elle est rejetée, on peut accepter l'alternative H_1
    - H_1 peut être une $\neq$, ou une inégalité si on sait qu'un des côtés de l'= est impossible
- Les tests sont construits sur plusieurs autres hyp. que H_0
  - Si l'une de ces hyp. vient à faillir, les résultats du test sont aberrants
    - Mais c'est difficile à avérer
    - Moins un test présuppose, plus il est général

Procédure du test

- On calcule une statistique du test
  - Cette stat suit une distribution connue **si H_0 est vraie**
    - p.e. χ^2 , t, F,... – c'est la partie difficile à démontrer
    - dépend aussi de la taille n de l'échantillon
- On compare la stat avec une valeur tabulée de cette distribution connue
  - Cette valeur tabulée est choisie arbitrairement
    - Ce choix est le niveau de significativité α du test
    - On prend généralement 5%
  - Si la stat est + grande, en valeur absolue, que cette valeur tabulée,
    - on RH_0

Rappel des principaux tests paramétriques

- test Z
  - Tout test dont la distribution de la stat sous H_0 est approximée par la normale
  - À cause du thm limite centrale, beaucoup de stat de test convergent à la normale dans de grands échantillons
    - p.e. la t-stat converge à une normale pour $n > 30$ dans certaines conditions
  - Approprié pour comparer des moyennes
- test t : somme de normales
  - Approprié pour comparer des moyennes, avec moins de données, mais normales
- test F : ratio de t
  - Sert à comparer des variances
- test χ^2 : somme de carrés de normales
  - Nombreux usages en np, mais aussi en paramétrique

Les erreurs du test

- Erreur de type I : RH_0 alors que H_0 est vraie
  - $\Pr\{RH_0 \mid H_0 \textit{ vraie}\} = \alpha$
    - Parfois appelé taille du test
    - $RH_0 \mid H_0$ *vraie* est un **faux positif** : détecter une différence, alors qu'il n'y en a pas
    - Rem. dans une régression, un régresseur non pertinent a 5% de passer pour significatif
  - Plus α est grand, plus on RH_0 faussement
    - p.e. un coef. β d'une régression est + facilement significatif quand $\alpha = .1$ que lorsque $\alpha = .05$
- Erreur de type II : $\neg RH_0$ alors que H_0 est fausse
  - $\Pr\{\neg RH_0 \mid H_0 \textit{ fausse}\} = \beta$
    - Cette Pr décroit à mesure que n croit
  - $1 - \beta = \Pr\{RH_0 \mid H_0 \textit{ fausse}\}$ est appelé **puissance du test**
    - Mais pas connu en pratique car dépend de la "fausseté" de H_0

Sommaire

Rappel sur les tests
Tests & corrélation : mesures
Tests
Nominal : Test du χ^2
Un échantillon
Plusieurs échantillons
2 échantillons reliés
2 échantillons indépendants
k échantillons
Procédure de randomisation
Corrélation non-paramétrique

Tests & mesures de corrélation np

- $\exists$ une vingtaine de tests np connus
  - D'autres moins connus
  - et une demi-douzaine de mesure de corrélation np
  - Selon les situations : mesure & échantillon
  - On va voir les mesures ci-dessous
- L'échantillon
  - 1 échantillon
  - 2 ou +
    - indépendants ou reliés
- Au sein d'un échantillon
  - Pour les tests, les obs doivent être indépendantes
  - Pour les mesures de corrélations, non

La mesure

- Les données sont toutes codées par des chiffres
  - Mais ces chiffres n'ont pas tous la même signification
  - Nominal - Ordinal - Cardinal

Mesure nominale / "qualitative"

- = classification
  - p.e. couleurs
  - Les classes sont mutuellement exclusives
- La seule relation entre classes est l'équivalence (=)
  - Pas de $\geq, \leq$ ou de $+$,-,$\times, \div$
- Seul le mode et les fréquences sont définis
  - Classe la + fréquemment observée
  - Médiane, moyenne, variance ne sont pas définies
- On teste sur des fréquences

Mesure ordinale (classement, rangs, hiérarchie)

- Il y a un ordre entre les classes
  - + grands, difficiles, ...
  - Cet ordre ne pourrait prévaloir qu'entre certaines classes mais ici il faut qu'il prévale partout
- Relations d'équivalence $=$ et d'ordre $\succ$
  - ordre : irréflexif, asymétrique et transitif
- Pas d'opérations $+$, $-$, $\times$, $\div$
  - La distance entre 2 classes n'a pas de sens
  - Même si une classe est codée 1 et une autre 3, cette dernière n'est pas "deux de plus" que la 1º
  - La moyenne et la variance ne sont donc pas définies
  - Mais la médiane et les quantiles (percentiles, déciles...) bien
  - On teste à partie de rangs / classements
- Les tests applicables au nominal sont applicables à l'ordinal

Mesure d'intervalle / cardinale / quantitative

- La distance entre les classes est définie
  - p.e. la température
- La même relation d'ordre prévaut
  - mais en plus les opérations sont définies
- Tous les moments ont du sens
  - moyenne, variance, ...
  - On teste sur ces moments dit des paramètres
    - De localisation ou autre
    - Les $\neq$ paramètres d'une distribution
- Les tests applicables à l'ordinal et/ou au nominal sont applicables à l'intervalle
  - Les tests param. ne sont applicables qu'à la mesure d'intervalle
    - p.e. les tests t & F des régressions

Sommaire

Rappel sur les tests

Tests

Corrélation non-paramétrique

Récapitulatif sur les tests np (réf. Siegel)

Mesure	1 éch.	2 échantillons		k éch.	
		Reliés	Indépendants	Reliés	Indép.
Nomin.	Binomial χ^2	McNemar	Fisher χ^2	Cochran Q	χ^2
Ordin.	Runs Kolmogorov -Smirnov (K-S)	Sign Wilcoxon	Médiane Mann-Whitney U K-S Wald-Wolfowitz	Friedman 2-way ANOVA	Médiane Kruskal-Walis 1-way ANOVA

- Seuls les tests colorés sont présentés
  - Info abondante, souvent sur Wikipedia et aide de R

Sommaire

Rappel sur les tests
Tests & corrélation : mesures
Tests
Nominal : Test du χ^2
Un échantillon
Plusieurs échantillons
2 échantillons reliés
2 échantillons indépendants
k échantillons
Procédure de randomisation
Corrélation non-paramétrique

Les tests du χ^2

- Un test χ^2 est tout test d'hyp
  - dont la stat suit une dist. χ^2 quand H_0 est vraie
    - Sans autre précision, il s'agit svt du test χ^2 de Pearson
    - qui est celui qui nous intéresse
  - Même si χ^2 est une dist. paramétrique, le test ne porte pas sur un paramètre
- 2 types de comparaison

  Ajustement tester si une distribution observée diffère d'une distribution théorique
  - Pour un éch.
  - "Goodness-of-fit"

  Indépendance tester si des obs non reliées sur 2 variables sont indépendantes l'une de l'autre
  - Pour 2+ éch.
  - s'exprime en *table de contingence*

χ^2 1 échantillon – Nominal

- On cherche à comparer
  - un groupe de fréquences absolues observées
    - k catégories i sont observées
  - avec un groupe de fréquences théoriques
- La stat de test est

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

O_i = nombre de cas observés dans la cat. i
E_i = nombre de cas espérés dans la cat. i sous H_0

- Il faut donc spécifier une dist espérée
- S'il y a peu de $\neq$ entre O et E
  - donc si les fréq. obs. sont proches des espérées
  - alors, χ^2 sera petit, donc pas dans les extrêmes de la dist.
- Le nombre de degrés de liberté est $k - p$
  - p est le nombre de paramètres de la dist. espérée
    - p.e. 2 si c'est une normale (espérance, variance)

χ^2 1 échantillon – Nominal : exemple

- On lance un dé à 6 faces 60 fois
  - Le dé est-il pipé (biaisé) selon le test de χ^2 de Pearson ?

- Dé régulier
  - $\Pr\{face\} = 1/6$

i	O_i	E_i	$\frac{(O_i - E_i)^2}{E_i}$
1	5	10	2.5
2	8	10	.4
3	9	10	.1
4	8	10	.4
5	10	10	0
6	20	10	10

- $k = 6$
- Degrés de liberté 6-1=5
  - Je pense que c'est plutôt 4 car U(a,b)
- $\chi^2 = 13.4$
  - $\chi^2_{.95;5} = ?$
  - R qchisq(.95, df=5) = 11.07
  - RH_0

χ^2 2 ou + échantillons indépendants – Nominal

Figure – Table de contingence

Groupe i (échantillon)	Résultat j				$\sum$ ligne
	1	2	...	C	
1	O_{11}				$O_{1.}$
2					
⋮			O_{ij}		$O_{i.}$
R				O_{RC}	
$\sum$ col	$O_{.1}$		$O_{.j}$		N

- Table de contingence : fréq. absolues observées O_{ij}
- H_0 : La distribution des résultats est indépendante des groupes
  - Peut-on dire que les lignes sont indépendantes entre elles ?
  - On ne précise pas de quelle distribution viendraient ces lignes

χ^2 2 ou + échantillons indépendants – Nominal

- Il y a N observations au total
- Somme par ligne (résultats) : $O_{i.} = \sum_j O_{ij}$
  - Soit $p_{i.} = O_{i.}/N$
- Somme par colonne (groupes) : $O_{.j} = \sum_i O_{ij}$
  - Soit $p_{.j} = O_{.j}/N$
- On définit les fréquences absolues espérées comme

$$E_{ij} = N p_{i.} p_{.j}$$

- La stat de test est

$$\chi^2 = \sum_{i=1}^{R} \sum_{j=1}^{C} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

χ^2 2 ou + échantillons indépendants – Nominal

- Même logique que à un échantillon
- Le nombre de degrés de liberté est $RC - (R + C - 1)$
  - RC est bien le nbr de catégories comme à 1 éch.
  - La réduction vient de ce qu'il faut calculer $R + C - 1$ quantités
    - pour calculer les fréquences espérées
- Limitations
  - Il faut que les $O_{ij} \geq 5$
    - ≥ 10 lorsqu'il y a un seul degré de liberté
    - Sinon, la stat du test ne converge pas à une χ^2
  - Ne peut déterminer la causalité
    - Groupe $\leftrightarrow$ résultat

χ^2 2 ou + échantillons indépendants – Nominal : exemple

- Y-a-t-il indépendance entre la pratique du sport et le choix de logement ?
  - Échantillon d'étudiants d'une université USA

Groupe i	Sport j			$\sum$ ligne
	Aucun	Peu	Régulier	
Résidence U	32	30	28	90
Appart. sur campus	74	64	42	180
Appart. hors campus	110	25	15	150
Chez ses parents	39	6	5	50
$\sum$ col	255	125	90	470

χ^2 2 ou + échantillons indépendants – Nominal : exemple

- Questions posée dans ce genre d'exercice :
  - Le lieu de vie affecte-il la pratique du sport ?
- Mais le test ne peut déterminer si
  - les étudiants ont choisi leur logement pour pratiquer le sport
  - le logement induit la pratique du sport
  - Pas de un test de causalité
    - Sauf à conclure à l'indépendance (pas de causalité)
- Intéressant pour chercher un régresseur
  - qu'on penserait significatif
  - mais pas linéaire

χ^2 2 ou + échantillons indépendants – Nominal : exemple

- Calcul des fréquences espérées
  - p.e. $E_{Aucun,Resid} = Np_{Aucun.}p_{.Resid} = 48.82$
  - $p_{Aucun.} = O_{Aucun.}/N = 255/470$ $p_{.Resid} = O_{.Resid}/N = 90/470$

Groupe i	Sport j			$\sum$ ligne
	Aucun	Peu	Régulier	
Résidence U	32	30	28	90
	48.8	23.9	17.2	
Appart. sur campus	74	64	42	180
	97.7	47.9	34.5	
Appart. hors campus	110	25	15	150
	81.4	39.9	28.7	
Chez ses parents	39	6	5	50
	27.1	13.3	9.6	
$\sum$ col	255	125	90	470

χ^2 2 ou + échantillons indépendants – Nominal : exemple

- Il faut que chaque $E_{ij} \geq 5$
  - c'est le cas
- La χ^2 est calculée en sommant tous les $\frac{(O_{ij}-E_{ij})^2}{E_{ij}}$
  - $\frac{(32-48.8)^2}{48.8} + \ldots = 60.5$
  - Les degrés de liberté sont

$$RC - (R + C - 1) = 12 - (4 + 3 - 1) = 6$$

  - $\chi^2_{.95;6} = 12.59 \implies RH_0$
    - Il n'y a pas de quoi conclure à l'indépendance du logement et sport

χ^2 2 ou + échantillons indépendants – Nominal : exemple

- Si on ramène la table de contingence en fréquences relatives
  - On voit bien qu'il y a un clivage dans/hors campus

Groupe i	Sport j			$\sum$ ligne
	Aucun	Peu	Régulier	
Résidence U	36%	33%	31%	100%
Appart. sur campus	41%	36%	23%	100%
Appart. hors campus	73%	17%	10%	100%
Chez ses parents	78%	12%	10%	100%
$\sum$ col	54%	27%	19%	100%

χ^2 dans R "Pearson's Chi-squared Test for Count Data"

chisq.test(x, y = NULL, correct = TRUE, p = rep(1/length(x), length(x)), rescale.p = FALSE, simulate.p.value = FALSE, B = 2000)

- x vecteur ou matrice numérique
- y vecteur numérique ; ignoré si x est une matrice
  - x & y peuvent aussi être des factors
  - Si x est un factor, y devrait être un factor de même longueur
- correct logique, indique s'il faut appliquer une correction de continuité pour 1 table 2 par 2
- p vecteur de proba de même longueur que x
  - permet de spécifier des proba ad hoc
- simulate.p.value logique indicant si calculer les p-valeurs par simulation
- B entier = nombre de réplications pour la simulation

χ^2 dans R – exemple

M <- as.table(rbind(c(762, 327, 468), c(484, 239, 477)))

- les données

dimnames(M) <- list(gender = c("F", "M"), party = c("Democrat","Independent", "Republican"))

- nommer les variables

(Xsq <- chisq.test(M))

- Calcul du test, on met dans Xsq & imprime le résumé

Xsq$observed

- Comptes observés ($\equiv$ M)

Xsq$expected

- Comptes espérés sous H_0
- Conclusion
  - on rejette l'indépendance des "échantillons"
  - les femmes votent différemment des hommes

χ^2 dans R – exercices

- Refaites les 2 tests précédents dans R
  - Dé et sport
- Un échantillon de 44 hommes et 56 femmes
  - A-t-il été tiré d'une population où les hommes et les femmes sont en égales proportions ?
- Une certaine procédure médicale est associée à une morbidité ± importante
  - On voudrait savoir si cette dernière dépendrait d'un score dit Apgar

	Morbidité		
Apgar	Nulle	Mineure	Majeure
0-4	21	20	16
5-6	135	71	35
7-10	158	62	35

Sommaire

Rappel sur les tests
Tests & corrélation : mesures
Tests
Nominal : Test du χ^2
Un échantillon
Plusieurs échantillons
2 échantillons reliés
2 échantillons indépendants
k échantillons
Procédure de randomisation
Corrélation non-paramétrique

Un échantillon – Nominal : Binomial – Intro

- Population partagée en 2 classes
  - Nombreux exemples
- Proportion dans une classe est P
  - Dans l'autre classe $1 - P$
  - Proportion fixe dans une pop.
    - mais v.a. dans un éch. à cause de l'échantillonnage
- Distribution binomiale
  - Distribution d'échantillonnage de proportions calculées sur des échantillons d'une population à 2 classes
- Test H_0 : la proportion est P dans la population
  - P est un certain chiffre
  - Le test indique s'il est raisonnable de croire que la proportion observée dans l'éch. puisse provenir d'une pop. avec proportion P

Un échantillon – Nominal : Binomial – méthode

- Dans un éch. de taille N
  - extrait d'une pop à 2 classes
  - la proba. d'obtenir x objets d'une classe
    - et donc $N - x$ de l'autre classe
  - est

$$p(x) = \left(\begin{array}{c} N \\ x \end{array} \right) P^x (1 - P)^{N-x}$$

avec $\left(\begin{array}{c} N \\ x \end{array} \right) = \frac{N!}{x!\,(N-x)!}$

  - Ce n'est que de la combinatoire, on compte les évènements
- Proba d'obtenir *au plus* x objets d'une classe $= \sum_{i=0}^{x} p(i)$

Un échantillon – Nominal : Binomial – méthode

- p.e. un dé à 6 faces tiré 5 fois, on obtient 2 "6"
  - $N = 5$, $x = 2$
    - "succès" est le 6 si c'est ça qui nous intéresse
  - P proportion supposée de $6 = \frac{1}{6} = H_0$
    - H_0 est essentiellement : la face 6 a-t-elle une proba "régulière" ?
    - On ne sait rien sur les autres faces
  - Proportion observée $\hat{P} = \frac{2}{5} = .4$
    - Un estimateur "naturel"

Test binomial : Procédure "exacte"

- Sous H_0, quelle est la proba d'observer un évènement aussi extrême ou plus extrême que celui observé ?
  - $\Pr\{\textit{deux}\ 6\ \textit{ou}\ +\}$
- $\Pr\{\textit{deux}\ 6\}$
  - $p(2) = \binom{5}{2} \left(\frac{1}{6}\right)^2 \left(1 - \frac{1}{6}\right)^{5-2} = .161$
- $\Pr\{\textit{moins de deux}\ 6\} = p(0) + p(1) = .402 + .402 = .804$
  - donc $\Pr\{\textit{deux}\ 6\ \textit{ou plus}\} = 1 - .804 = .196$
  - Alternative :
    - $p(3) = \binom{5}{3} \left(\frac{1}{6}\right)^3 \left(1 - \frac{1}{6}\right)^{5-3} = .032$
    - $p(4) = \binom{5}{4} \left(\frac{1}{6}\right)^4 \left(1 - \frac{1}{6}\right)^{5-4} = .003$
    - $p(5) = \binom{5}{5} \left(\frac{1}{6}\right)^5 \left(1 - \frac{1}{6}\right)^{5-5} = .0001$
  - $\Pr\{\textit{deux}\ 6\ \textit{ou}\ +\} = p(2) + p(3) + p(4) + p(5) = .196$
    - Sous $H_0 : P = \frac{1}{6}$, on a $\Pr\{x \geq 2\} = .196$
    - probabilité dite "exacte" car on n'a pas recourt à une approx p.e. normale
  - La proba obtenue est $\geq .05 \implies \neg RH_0$
    - deux 6 sur 5 jets de dés n'est pas incompatible avec $p = 1/6$

Test binomial : Procédure "exacte" – Remarques

- On n'utilise pas le $\hat{p} = 2/5 = .4$
- One-sided ou two-sided ?
  - Ici, on a regardé $x \geq 2$
    - les évènements "plus extrêmes", donc plus petite proba que pour $x \leq 2$
    - Donc : One-sided "greater"
  - Pour un "two-sided", $x \neq 2$, doubler la proba
- C'est de plus en plus long à mesure que N augmente
  - Mais on peut montrer que la $\sum$ de binomiales $\underset{N \to \infty}{\to}$ normale
  - Cela amène au test "approximatif" suivant
    - En fait, asymptotique - on aura une autre significaiton pour "approximatif" avec le test de randomisation

Un échantillon – Nominal : Binomial – Score

- Le test du score
  - Stat $\dfrac{\hat{p}-p}{\sqrt{p\left(1-p\right)/n}} = \dfrac{.4-.2}{\sqrt{.2\left(1-.2\right)/5}} \simeq 1.118$
  - Sous H_0, le score suit asymptotiquement une distribution normale
    - Ici, avec 5 obs, on ne peut pas l'employer
  - On compare avec
    - 1.96 en valeur absolue pour le 2-tail
    - 1.644 pour le 1-tail "greater"
    - On est donc loin de R aussi

Un échantillon – Nominal : Binomial – R

- binom.test(x, n, p = 0.5, alternative = c("two.sided", "less", "greater"), conf.level = 0.95)
  - x nombre de "1" (succès, vrai...)
  - n taille de l'échantillon
  - p Proba de succès que l'on teste $=H_0$
  - alternative indique l'hyp. alternative
    - "two.sided", "greater" ou "less" (on n'indique que la 1º lettre)
    - par défaut two.sided
  - conf.level le niveau de confiance choisi pour calculer l'intervalle de confiance
- Test exact, quel que soit n

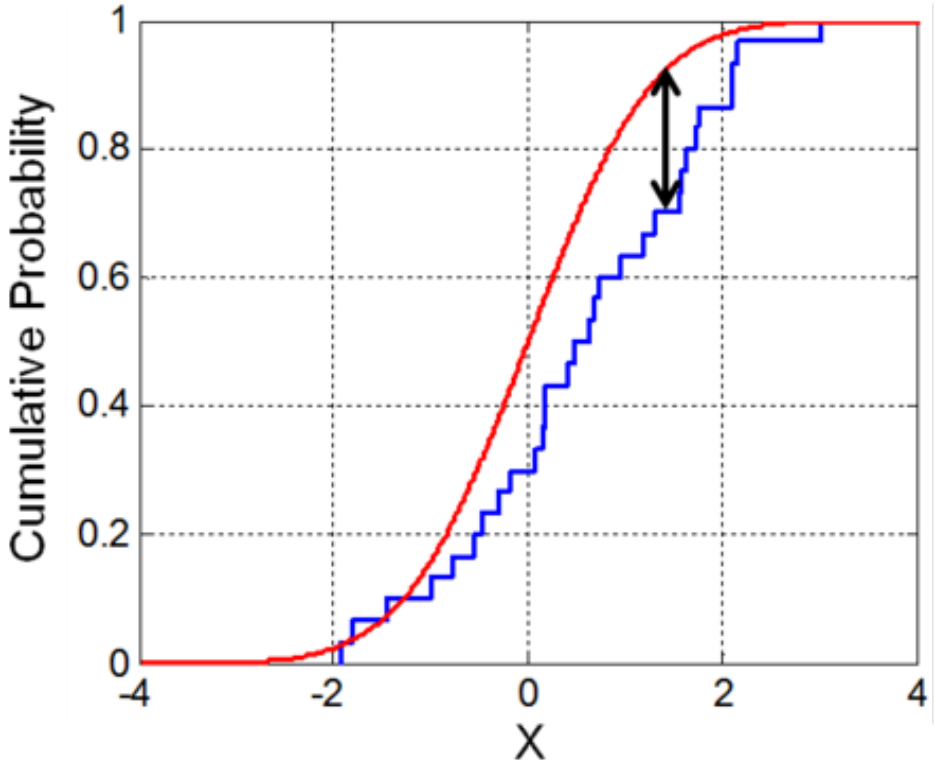
Un échantillon – Nominal : Binomial – Exemple

- En supposant l'hérédité mendélienne
  - un croisement de 2 variétés d'une certaine plante produirait 1/4 de "naines" et 3/4 de "géantes"
  - Dans une expérience pour déterminer si cette hyp. est raisonable
    - on obtient 243 naines et 682 géantes.
- On teste $H_0 : p = 3/4$, en prenant "géante" comme le succès
  - binom.test(682,682+243,p=3/4)
  - résultat : p-valeur>5% $\neg RH_0$
- Autre exemple
  - Sur un échantillon de 24 pièces manufacturées, 13 passent le test de qualité
  - Peut-on dire que la fiabilité est meilleure que 35% ?

Un échantillon – Ordinal : Kolmogorov-Smirnov

- KS teste l'égalité de 2 dist. uni-dimensionnelles continues
  - À 1 éch. : compare la dist. empirique de celui-ci avec une distribution connue
    - spécifiée dans le test
    - H_0 la dist. empirique est compatible avec la théorique
    - H_1 2 tailed : pas compatible, sans dominance claire
    - H_1 1-tailed : pas compatible, l'une domine l'autre
  - À 2 éch. : compare les deux dist. empiriques
- La stat du test est la plus grande distance possible entre les deux distributions
  - Cette stat suit une dist. de Kolmogorov sous H_0

Un échantillon – Ordinal : Kolmogorov-Smirnov



En rouge, une dist. théorique, en bleu une empirique

La flèche noire est la stat du test KS

Un échantillon – Ordinal : Kolmogorov-Smirnov – remarques

- Comme les dist. sont continues, la proba de 2 points égaux est nulle
  - En pratique, ks ne devrait pas s'appliquer à 2 éch. avec bcp de points identiques
- Pour tester la normalité, $\exists$ des test spécialisés + puissants
  - p.e. Shapiro–Wilk ou Anderson–Darling
- Si les paramètres de la dist. connue sont estimés
  - dans le sens où ils résultent d'une estimation
  - alors le test n'est pas valable
- Une généralisation à un cas multivarié existe
- La fonction ks.test de R implémente le test – prochaine dia
  - Mais permet aussi que la distribution théorique soit discrète

Un échantillon – Ordinal : Kolmogorov-Smirnov dans R

ks.test(x, y, ..., alternative = c("two.sided", "less", "greater"), exact = NULL)

- Arguments
  - x vecteur numérique de données
  - y soit un vecteur numérique de données pour le test à 2 échantillons
    - ou une chaine de caractère qui nomme une fonction de distribution continue p.e. pnorm
  - ... d'éventuels paramètres de la distribution indiquée dans y
  - alternative – voir test binomial
  - exact : NULL ou bien TRUE ou 1 pour indiquer si une p-valeur exacte doit être calculée (pas calculable dans tous les cas)

Un échantillon – Ordinal : Kolmogorov-Smirnov dans R – détails

- Si y est numérique, c'est le test à 2 éch. H_0 : x et y sont tirés de la même distribution
  - Sinon, y est une chaine de caractères qui teste si x est généré par la distribution nommée
    - avec les paramètres indiqués par ...
- Des égalités causent un avertissement
  - car l'hyp. sous-jacente de distribution continue ne les permet pas
  - Les arrondis peuvent générer des erreurs importantes ici.
- Les NA sont omises silencieusement

Un échantillon – Ordinal : Kolmogorov-Smirnov – exemple

- x <- rnorm(50)
  - Est-ce que x+1.2 vient d'une distribution gamma avec paramètres de forme 3 et taux 2 ?
  - ks.test(x+1.2, "pgamma", 3, 2) # deux côtés (two-sided)
  - ks.test(x+1.2, "pgamma", 3, 2, exact = FALSE)
  - ks.test(x+1.2, "pgamma", 3, 2, alternative = "gr")

Un échantillon – Ordinal : Runs

- Wald-Wolfowitz “Runs” runs.test
  - Installer le package randtests
  - Skip

Sommaire

Rappel sur les tests
Tests & corrélation : mesures
Tests
Nominal : Test du χ^2
Un échantillon
Plusieurs échantillons
2 échantillons reliés
2 échantillons indépendants
k échantillons
Procédure de randomisation
Corrélation non-paramétrique

2 ou plus échantillons

- Échantillons **reliés**
  - Il y a un lien entre les obs des $\neq$ éch.
    - p.e. pls obs de la même personne : poids, taille, revenu..
    - dans le temps : poids en t & t+1 pour des personnes
    - n juges classent k vins. Un vin est-il classé systématiquement plus haut ou plus bas que les autres ?
- Échantillons **indépendants**
  - Pas de lien dans le processus d'échantillonnage
  - On veut alors souvent comparer les distributions
    - Et tester l'indépendance

Sommaire

Rappel sur les tests
Tests & corrélation : mesures
Tests
Nominal : Test du χ^2
Un échantillon
Plusieurs échantillons
2 échantillons reliés
2 échantillons indépendants
k échantillons
Procédure de randomisation
Corrélation non-paramétrique

2 échantillons reliés – Nominal : McNemar

- Test de significativité des changements
  - mcnemar.test
- Skip

2 échantillons reliés – Ordinal : Signe

- Teste pour une différence systématique entre paires d'obs. (x, y)
  - p.e. le poids de sujets avant & après traitement
- Le test est plus utile si les comparaisons ne peuvent être exprimées que par $x > y, x = y, x < y$
  - Si (x, y) peuvent être exprimés cardinalement
    - p.e. $x = 17, y = 12 \rightarrow$ t-test
    - Mais on peut utiliser le test du signe pour vérifier si la **médiane** de $x - y$ est $\neq 0$
  - Si (x, y) ne peuvent être exprimés que ordinalement
    - p.e. x est 1º rang de tous les x, y est 7º $\rightarrow$ test des rangs signés de Wilcoxon (prochain test)

2 échantillons reliés – Ordinal : Signe – Principe

- S'il n'y a pas de $\neq$ entre x & y, $p = \Pr\{x > y\} = .5$
  - $\Pr\{x = y\} = 0$ pour des variables continues
    - les paires t.q. $x = y$ sont omises du test
  - Donc $H_0 : p = .5$
- Soit W le nombre de paires positives
  - "paire positive" peut-être $x > y$ qualitatif, ou $x - y$ quanti
  - Sous H_0, W suit une distribution binomiale
  - Il faut supposer l'indépendance entre paires
- On utilise donc binom.test comme dans le test binomial avec $p = .5$

2 échantillons reliés – Ordinal : Signe – Exemple 1

- Longueur des pattes antérieures et postérieures G de cerfs

i	Avt	Arr.	$\neq$
1	142	138	+
2	140	136	+
3	144	147	−
4	144	139	+
5	142	143	−
6	146	141	+
7	149	143	+
8	150	145	+
9	142	136	+
10	148	146	+

- H_0 : pas de $\neq$ de longueur
  - two-tailed, pas d'à priori
  - Si on disait H_0 Avt>Arr. : one-tailed
- Écrivez le test dans R
  - p-value = .109 > .05
  - $\neg RH_0$
  - À retenir : 8/10 pas signif $\neq$ 5/10
  - Comparer si c'était 80/100

2 échantillons reliés – Ordinal : Signe – Exemple 2

- Un fabriquant a 2 produits, A & B
  - Il veut savoir si les consommateurs préfèrent B à A
- Éch de 10 consommateurs
  - Chacun reçoit A & B et indique ce qu'il préfère
- H_0 B n'est pas préféré à A
  - H_1 B est préféré à A
  - $\Longrightarrow$ C'est un test à un côté
- Sur l'éch. 8 consommateurs préfèrent B, un préfère A et un est indifférent
  - Tester
    - binom.test(8, 9, p = 0.5, alternative = c("greater"))
    - p-valeur=.0195 RH_0

2 échantillons reliés – Ordinal : Signe – Exemple 3 médiane

- Survie en semaines pour 10 sujets :
  - 49, 58, 75, 110, 112, 132, 151, 276, 281, 362+
    - 362+ veut dire >362
- La médiane est-elle supérieure à 200 ?
  - Dans cet éch., la médiane est toute valeur entre 112 & 132
  - Si H_0 est vraie, médiane =200,
    - alors on doit s'attendre à la moitié des sujets en vie après 200 semaines
    - on assigne + aux obs >200 et - aux <
    - Donc 7 - et 3 +
  - vs H_1 Médiane $\neq$ 200 : 2-tailed
    - Tester – p-val=.34 : $\neg RH_0$
  - vs H_1 Médiane $>$ 200 : 1-tailed
    - Tester – p-val=. ? ? : ? ?

2 échantillons reliés – Ordinal : Wilcoxon

- Test des rangs signés ("Wilcoxon signed-ranks test")
  - H_0 : la dist. des paires est symmétrique autour de zéro
    - lorsqu'on ne peut pas supposer la normalité
    - Si on peut supp. la norm. : test t
- Hyp
  - Les données sont en paires et viennent de la même pop.
  - Chaque paire est choisie au hasard, indépendamment
- Soit N le nombre de paires
  - éventuellement en omettant les paires égales
  - Calculer $|x_{2i} - x_{1i}|$ et $signe\,(x_{2i} - x_{1i}) = s_i$
  - Assigner un rang R_i aux $|x_{2i} - x_{1i}|$
    - Le rang 1 va à la plus petite
  - Stat du test : $W = \sum_i s_i R_i$ la somme de rangs signés
    - Si $|W|$ est grande, un groupe domine l'autre
    - RH_0 si $|W| > W_{critical,N}$

2 échantillons reliés – Ordinal : Wilcoxon – R

wilcox.test(x, y, alternative = c("t"), mu = 0, paired = 0/1, exact = NULL, ...)

- Arguments
  - x : vecteur de données numériques, NA omises
  - y : vecteur optionnel de données numériques, comme x
  - alternative : comme binomial
  - mu : un nombre optionnel – 'Détails'
    - Permet de tester si les $\neq$ sont sym. autour de mu, au lieu de zéro
  - paired : 0/1 indique si les obs. sont en paires (2 éch.) – Détails
  - exact : 0/1 indique une p-valeur exacte doit être calculée

2 échantillons reliés – Ordinal : Wilcoxon

- Détails “paired”
  - Si seul x est donné, ou x & y, mais paired=1,
    - un test de Wilcoxon rangs signés est calculé
    - H_0 : la distribution de x (1 éch.) ou de x-y (2 éch. paired) est symmétrique autour de mu
  - Si x et y sont donnés, mais paired=0,
    - un test équivalent au test de Mann-Whitney est calculé
    - H_0 : les distributions de x et de y diffère de mu (localisition) et l'alternative est qu'ils diffèrent par une autre valeur
  - Avec moins de 50 obs : la dist. exacte du test est calculée
    - sinon approx normale

2 échantillons reliés – Ordinal : Wilcoxon – R exemple

- Mesures d'importance de dépression sur 9 patients
  - Avant & après traitements
  - Ordinal car on ne peut comparer entre patients
    - Donc : pas normalité, pas de t-test
  - On se demande si le traitement a un effet
    - H_0 : dist. sym autour de zéro
    - Alternative : dist sym autour de mu>0 "greater" – comme avec binomial
    - $\Longrightarrow$ One-tailed test

```
avt <- c(1.83, 0.50, 1.62, 2.48, 1.68, 1.88, 1.55, 3.06, 1.30)
apr <- c(0.878, 0.647, 0.598, 2.05, 1.06, 1.29, 1.06, 3.14, 1.29)
wilcox.test(avt, apr, paired = TRUE, alternative = "greater")
```

- Quelle conclusion ?

Sommaire

Rappel sur les tests
Tests & corrélation : mesures
Tests
Nominal : Test du χ^2
Un échantillon
Plusieurs échantillons
2 échantillons reliés
2 échantillons indépendants
k échantillons
Procédure de randomisation
Corrélation non-paramétrique

2 échantillons indépendants – Nominal

- Test de Fisher
  - Skip
- Test du χ^2
  - Voir la section χ^2

2 échantillons indépendants – Ordinal

- Test de la Médiane
  - Skip
- Test "Mann-Whitney U"
  - Même test que Wilcoxon à 2 éch.
  - R wilcox.test( )

2 échantillons indépendants – Ordinal : Kolmogorov-Smirnov

- Voir Kolmogorov-Smirnov 1 échantillon
- ks.test(x, y, ..., alternative = c("two.sided", "less", "greater"), exact = NULL)
  - x & y vecteurs numériques de données
- Exemple
  - x <- rnorm(50)
  - y <- runif(30)
  - H_0 x et y viennent de la même distribution
  - ks.test(x, y)
  - H_0 x est (stochastiquement) plus grand que x2
  - x2 <- rnorm(50, -1)
  - plot(ecdf(x), xlim = range(c(x, x2)))
  - plot(ecdf(x2), add = TRUE, lty = "dashed")
  - ks.test(x, x2, alternative = "l")
  - Comparer avec
    - t.test(x, x2, alternative = "g")
    - wilcox.test(x, x2, alternative = "g")

Sommaire

Rappel sur les tests
Tests & corrélation : mesures
Tests
Nominal : Test du χ^2
Un échantillon
Plusieurs échantillons
2 échantillons reliés
2 échantillons indépendants
k échantillons
Procédure de randomisation
Corrélation non-paramétrique

k échantillons indépendants

- Je laisse tomber les éch. reliés
- Éch. indép.
  - Nominal
    - Voir section χ^2
  - Ordinal : Extension du test de la médiane
    - Skip

k échantillons indépendants – Ordinal : Kruskal-Wallis

- Teste si des éch. proviennent de la même dist.
  - Assez bien une extension de Wilcoxon
- La significativité (RH_0)
  - indique qu'au moins un éch. domine stochastiquement au moins un autre éch.
    - mais n'identifie pas où cette dominance pourrait se placer
    - D'autres tests pourraient aider
  - Si on peut faire l'hyp. d'une même distribution pour tous les groupes à l'exception de la médiane
    - En rappelant que la moyenne & la variance ne sont pas définies
    - alors H_0 : toutes les médianes sont égales
    - H_1 : au moins une médiane est $\neq$ de celle d'une autre groupe
  - Plus généralement, K-W teste l'égalité des paramètres de localisation (=position) entre les groupes
    - Moyenne (si définie), médiane et mode

k éch. indép. – Ordinal : Kruskal-Wallis – méthode

- Classer toutes les N obs. des k éch. ensemble
  - Sans tenir compte de l'appartenance de groupes - donc de 1 à N
- La stat de test est

$$H = (N-1)\frac{\sum_{i=1}^{g} n_i \left(\bar{r}_{i.} - \bar{r}\right)^2}{\sum_{i=1}^{g}\sum_{j=1}^{n_i}\left(r_{ij} - \bar{r}\right)^2}$$

n_i est le nombre d'obs. du groupe i
r_{ij} est le classement de l'obs. j du groupe i
$\bar{r}_{i.} = \dfrac{\sum_{j=1}^{n_i} r_{ij}}{n_i}$ est le classement moyen des obs. du groupe i
$\bar{r} = \frac{1}{2}(N+1)$ est la moyenne des r_{ij}

k éch. indép. – Ordinal : Kruskal-Wallis – méthode

- Je ne discute pas les gestions des égalités
  - mais s'il n'y a pas d'égalité, alors

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{g} n_i \bar{r}_{i.}^2 - 3(N+1)$$

    qui donc ne comprend que des moyennes de rangs (par groupe)
- On voit que si les groupes sont "mélangés"
  - donc les rangs dans un groupe sont répartis dans tout l'éch.
  - alors H sera relativement petit
    - plus petit que si un groupe concentre des rangs élevés
    - Cela, à cause du carré
- Sous H_0, $H \sim \chi^2_{g-1}$
  - Donc si $H < \chi^2_{g-1;0.95}$ alors aucun ne domine un autre stochastiquement
  - Si c'est le cas contraire, il faut chercher entre paires de groupes

k éch. indép. – Ordinal : Kruskal-Wallis – Remarques

- H ne suit pas exactement une χ^2_{g-1}
  - La différence devient importante lorsque certains groupes ont moins de 5 membres
  - Il est possible de trouver une dist. exacte pour H
- Le test des rangs signés de Wilcoxon est essentiellement un cas particulier de celui-ci
- Kruskal-Walis est parfois appelé “ANOVA 1-way for ranks”
  - ANOVA = ANalysis Of VAriance
  - KW n’est pas du tout un test sur les variances
  - Mais la structure en tableau évoque ANOVA
  - ANOVA teste les moyennes
    - KW la dominance stoch
    - qui devient la médiane en cas de symmétrie

k éch. indép. – Ordinal : Kruskal-Wallis – R

- Kruskal-Wallis Rank Sum Test kruskal.test(x, ...)
  - x vec. numérique des données ou liste de vec. num de données
    - Les éventuels éléments non-num. seront forcés (coerced), avec avertissement
  - g vec ou factor qui indique le groupe pour chaque élément de x
    - Ignoré si x est une liste (car les groupes sont alors définis)
  - formula une formule de type "response ~ group" comme pour les régressions
    - voir exercice 3
  - data matrice ou frame optionnel qui contient les données de la formula
  - subset perment éventuellement de prendre un sous-ensemble des données
  - na.action fonction qui indique ce qu'il faut faire des NAs.

k éch. indép. – Ordinal : Kruskal-Wallis – Exemple

- Mesure d'efficience d'un certain médicament
  - pour retirer des poussières de sujets
    - normaux x <- c(2.9, 3.0, 2.5, 2.6, 3.2)
    - avec une maladie des voies respiratoires y <- c(3.8, 2.7, 4.0, 2.4)
    - avec asbestose z <- c(2.8, 3.4, 3.7, 2.2, 2.0)
  - kruskal.test(list(x, y, z))
    - p-valeur ? Conclusion : pas de dominance stoch

k éch. indép. – Ordinal : Kruskal-Wallis – Exercice 1

- 24 sujets (N)
  - 3 groupes A, B, C de 8, pour une interview
  - mais 1 sujet en B et 2 en C ne viennent pas
    - $n_A = 8$, $n_B = 7$, $n_C = 6$
    - Éch. de $\neq$ tailles : argument pour np ?
- Doivent classer 3 vins
  - En fait : le même vin, mais
    - l'interview est faite pour que les A s'attendent à de bons vins, les C de mauvais

A	B	C
6.4	2.5	1.3
6.8	3.7	4.1
7.2	4.9	4.9
8.3	5.4	5.2
8.4	5.9	5.5
9.1	8.1	8.2
9.4	8.2	
9.7		

k éch. indép. – Ordinal : Kruskal-Wallis – Exercice 1

Niveaux			Rangs			
A	B	C	A	B	C	
6.4	2.5	1.3	11	2	1	
6.8	3.7	4.1	12	2	4	
7.2	4.9	4.9	13	5.5	5.5	
8.3	5.4	5.2	17	8	7	
8.4	5.9	5.5	18	10	9	
9.1	8.1	8.2	19	14	15.5	$\bar{r} =$
9.4	8.2	NA	20	15.5		11
9.7	NA	NA	21			$\sum r_{ij}$
Somme rangs			131	58	42	231
Moyenne rangs			16.4	8.3	7.0	$\bar{r}_{i.}$

Entrez les données à la main et réalisez le test

k éch. indép. – Ordinal : Kruskal-Wallis – Exercice 2 & 3

- Ex. 2. Données simulées SimKW.csv sur site diplôme
  - 3 éch avec la même moyenne (43.5), la même médiane (27.5),
    - mais un résultat Kruskal-Wallis tranché, avec p-valeur=0.025
- Ex. 3. Illustration de la version "formule" de la commande
  - Airquality : Daily air quality measurements in New York, May to September 1973.
    - Ozone = niv d'ozone
    - Month = 5,6...9
  - Core distribution R
    - utiliser data(airquality)
    - boxplot(Ozone ~ Month, data = airquality)
    - kruskal.test(Ozone ~ Month, data = airquality)

k échantillons reliés – Nominal

- Cochran Q
  - Skip
- Friedman analyse de variance 2 voies
  - "ANOVA 2-way for ranks"
  - Sur R friedman.test
  - Skip

Résumé tests np dans R

$m0.09| > m0.18| > p0.18| > p0.18| > p0.18| > p0.18|| > m0.09| > m0.18| > p0.18| > p0.18| > p0.18| > p0.18|01234567891011$

	2 échantillons	k éch.
–		

1. Installer package *randtests*

Sommaire

Rappel sur les tests
Tests & corrélation : mesures
Tests
Nominal : Test du χ^2
Un échantillon
Plusieurs échantillons
2 échantillons reliés
2 échantillons indépendants
k échantillons
Procédure de randomisation
Corrélation non-paramétrique

Tests de randomisation

- Divers noms
  - randomisation, permutation, exact
- R package **coin**
  - Voir vignette coin_implementation.pdf
- Randomisation/permutation = une **classe** de tests
  - Qui ont pour objectif commun de tester des hyp. d'indépendance
  - Beaucoup de tests y sont dispo, dont ceux qu'on a vu
    - χ^2
    - Wilcoxon
    - Kruskal-Walis

Idée de la randomisation

- Pour un certain test
  - np ou autre, sur 2 ou plus éch.
- La distribution de la stat du test est obtenue
  - en calculant la stat du test dans tous les réarrangements possibles des étiquettes des points de donnée
  - l'étiquette est un groupe auquel appartient le point de donnée
    - La ligne d'un tableau de contingence
- H_0 : les étiquettes sont interchangeables
  - Équivalent à "même distribution entre 2 groupes de données"
  - Un rejet de H_0 conduit à l'indépendance des 2 éch.

Exemple. Randomisation sur des moyennes

- Soit 2 groupes A & B
  - de tailles n_A et n_B
  - avec moyennes $\bar{x}_A$ et $\bar{x}_B$
    - L'appartenance à un groupe est **l'étiquette** de chaque point de donnée
- On veut tester si les 2 groupes ont la même distribution (H_0)
  - La **stat du test de permutation** est la $\neq$ entre les moyennes $\bar{x}_A - \bar{x}_B$
    - Appelée $T(obs)$
    - $T(obs)$ est-elle assez grande pour rejeter H_0 ?
  - On calcule $T(obs)$
    - Puis on **mélange** les obs. des 2 échantillons

Exemple. Randomisation sur des moyennes

- On calcule la $\neq$ des moyennes pour chacune des possibilités de faire 2 groupes de tailles n_A et n_B dans ces données
  - *L'ensemble de ces* $\neq$ *est l'exacte dist. de* $T\,(obs)$ *sous* H_0
  - H_0 *: les 2 groupes ont les mêmes moyennes / étiquettes*
  - Si n_A et/ou n_B est très grand, il faudra en calculer moins
- p-valeur one-tailed
  - proportion de ces permutations pour lesquelles la $\neq$ de moyennes est $\geq T\,(obs)$
    - si <5% : RH_0
- p-valeur two-tailed
  - proportion de ces permutations pour lesquelles la valeur absolue de la $\neq$ de moyennes est $\geq |T\,(obs)|$
    - Remarque : ici on ne teste qu'une égalité de moyennes, pas l'indépendance
    - parce que la stat de test est une $\neq$ de moyenne
- On peut calculer un Intervalle de Confiance pour $T\,(obs)$:
  - Trier toutes les $\neq$ calculées de la + petite à la + grande
  - La borne inf de l'IC est p.e. la 25º ou 250º valeur, selon nbr d'obs

Tests de randomisation – Package coin

- On ne part pas sur une stat de test à priori
  - Ds l'ex., une moyenne
  - Mais on peut refaire la stat de test de n'importe quel test
    - p.e. rangs signés de Wilcoxon ou somme de carrés de χ^2
    - En général, il s'agit de test d'indépendance sur base d'une table de contingence
- Au lieu de faire une hypothèse sur la distribution asymptotique de cette stat
  - p.e. le test χ^2 approche la dist. de la stat de test par celle de la χ^2
  - le test calcule la stat de test sur base de permutations de l'éch.
    - Si on peut faire toutes les permutations : "exact"
    - Sinon : "approximée" – on en prend "beaucoup"
    - par opposition à "asymptotique" qui est le cas classique

Tests de randomisation – Package coin

- En général, les permutations reproduisent les tables de contingence
  - Avec chaque table, on calcule la stat de test
  - Ça donne une distribution "empirique" de la stat de test
  - Si la stat de test est dans la queue de la distribution, on RH_0
- C'est la commande qui "choisit" le test par défaut
  - Donc, il faudrait connaitre + de tests np pour comprendre ce test
    - En particulier, évaluer si le choix est approprié
  - Mais quelques exemples pour comprendre la base

Tests de randomisation – Package coin

- Charger le package coin
  - La commande principale est IndependenceTest( )
- En premier on met une formule
  - Comme dans le dernier exemple de Kruskal-Walis
    - independence_test(Ozone ~ Month, data = airquality)
    - Les résultats sont bien différents !
  - Mais on peut introduire des groupes (blocks)
    - y ~ x | block
    - p.e. y = satisfaction de son emploi, x = groupes de revenus, block = ho & fe
- C'est la commande qui "choisit" le test
  - Par défaut – on verra + loin une façon de changer
- On peut choisir un test "exact", "approximate", "asymptotic"
- + des options diverses

Tests de randomisation – Package coin : exemple simple

- Données asat de toxicologie
  - mesures avec traitement "compound" et sans (témoin)
  - boxplot(asat ~ group, data = asat)
  - independence_test(asat ~ group, data = asat)
- Que concluez-vous ?

Tests de randomisation – Package coin : exemple

- Données “job satisfaction”
  - en 4 cat.
  - selon 4 niveaux de revenu
  - pour les hommes et les femmes
  - data("jobsatisfaction", package = "coin")
- Attention, les données sont en “Table”
  - pas en Frame
  - Une Table dans R est une table de contingence
    - présente le comptes (fréq. abs.) dans chque cellule / combinaison de factor
  - On l’utilise directement dans le test

Tests de randomisation – Package coin : exemple

- Visualisation
  - ftable "flat contingency tables"
    - ftable(js)
    - Tableau de contingence à pls dim
  - Installer & charger le package vcd
    - Mosaic plot pour des donnés cat pluridim
    - mosaic(Income ~ Gender + Job.Satisfaction, data = js, split_vertical=c(TRUE, FALSE, TRUE))
    - Peut-être pas le + joli – je vous laisse expérimenter

Tests de randomisation – Package coin : exemple

- independence_test(js)
  - Le test décide seul
    - Il emploit une stat de Max : peu d'info, peu de puissance
- independence_test(js, teststat = "quadratic", distribution = asymptotic())
  - On impose un test classique
    - Ici Cochran-Mantel-Haenszel – mais c'est là qu'il faut connaitre bcp de tests
    - p-valeur $\backsimeq .33$: H_0 pas rejetée
    - H_0 était "pas de différence entre les éch. / groupes" : donc même distribution

Tests de randomisation – Package coin : exemple

- independence_test(js, distribution = approximate(B = 10000), scores = list(Job.Satisfaction = 1 :4, Income = 1 :4))
  - scores, ici, force Job.Satisfaction et Income, de factor qu'ils étaient à "facteur ordonné"
    - Donc : exploite le fait que la satisfaction et le revenu sont des ordres (ordinal, pas nominal)
    - Le test a donc un peu plus d'info
  - p-valeur $\backsimeq .01$: H_0 rejetée
    - Donc : indépendance – les groupes ne viennent pas tous de la même dist.
    - Donc, une différence significative entre hommes et femmes
  - Réaliser qu'on teste avec 2 ordinales et une nominale

Conclusion

- Ceci conclut les tests np
- On en a vu un échantillon
  - Plutôt à titre didactique que pour être exhaustif
- On a vu qu'on pouvait tester dans des situations
  - où on ne pourrait utiliser des régressions
    - cfr dernier exemple
  - mais où la puissance est sans doute faible
    - p.e. les différents résultats des tests

Sommaire

Rappel sur les tests

Tests

Corrélation non-paramétrique

Corrélation non-paramétrique

Équivalent du coef de corrélation pour données "quali."

\+ précisément : "coef de corrélation de Pearson" r ou R – paramétrique

Mesure	Corrélation
Nomin.	Coefficient de contingence, ϕ,...
Ordin.	Coef. de corrélation de rangs de Spearman r_s
	Coef. de corrélation de rangs de Kendall τ

Il y en a quelques autres

Nominal : Coefficient de contingence, ϕ,...

- Plusieurs mesures ont été définies
  - pour des données nominales
  - à partir des tables de contingence
  - pour essayer de décrire la force de l'association entre variables
  - principalement à partir des valeurs de la stat du test χ^2
- Il est vrai qu'il est souvent utile de décrire la relation
  - mais je crois que la table de contingence y arrive mieux
    - qu'un coefficient résumé

Ordinal : Corrélation de rang de Spearman ρ

- Parfois noté ρ, parfois r_s
- Les obs. sont en paires $\langle x, y \rangle$
  - La formule est simplement celle de la corrélation
    - appliquée au rangs

$$\rho = \frac{cov\left(r_x, r_y\right)}{\sqrt{var\left(x\right) var\left(y\right)}}$$

  - Si les rangs de x sont proches de ceux de y
    - $\rho \to 1$
  - S'ils sont complètement à l'opposé : $\rho \to -1$

cov(x, y = NULL, use = "everything", method = c("pearson", "kendall", "spearman"))

Ordinal : Corrélation de Kendall τ

- Dans les 2 éch. appareillés
  - une paire de paires $\langle x_i, y_i \rangle$ et $\langle x_j, y_j \rangle$ est dite concordante si
    - soit $x_i > x_j$ et $y_i > y_j$
    - soit $x_i < x_j$ et $y_i < y_j$
    - sinon elle est discordante
    - Une double égalité sort du calcul du τ
- Soit n_c et n_d les nombres de paires concordantes et discordantes, respect.,

$$\tau = \frac{n_c - n_d}{n\left(n-1\right)/2}$$

  est le τ de Kendall
- Le dénominateur est le total de combinaisons de paires
  - s'il y a parfaite concordance, $\tau = 1$
  - s'il y a parfaite disconcordance, $\tau = -1$
  - s'il y a indépendance, $\tau = 0$

Corrélation de Kendall τ : exemple

- Données "survey" préchargées dans R
  - 237 variables
  - smoke <- factor(survey$Smoke, levels=c("Never","Occas","Regul","Heavy"))
  - exer <- factor(survey$Exer, levels=c("None","Some","Freq"))
  - m <- cbind(exer, smoke)
    - on refait une petite matrice
- cor(m, method="kendall", use="pairwise")
  - On utilise pairwise sinon les NA causent un résultat NA
  - La corrélation est donc le chiffre hors de la diag princ.

Devoir # 1 : Tests np

1. Données simulées SimKW.csv sur site diplôme
   - 3 éch avec la même moyenne (43.5), la même médiane (27.5)
   - Montrez que ces 3 éch. sont indép. avec Kruskal-Wallis
     - Résultat tranché, avec p-valeur=0.025
2. Kurskal-Walis version “formule”
   - Airquality : Daily air quality measurements in New-York, May to Sept. 1973. Ozone = niv d'ozone ; Month = 5,6...9
   - Utilisez data(airquality) ; boxplot(Ozone ~ Month, data = airquality) ; kruskal.test(Ozone ~ Month, data = airquality)
     - Interprétez

Devoir #1 : test np

3. Survie en semaines pour 10 sujets : 49, 58, 75, 110, 112, 132, 151, 276, 281, 362+
   - 362+ veut dire >362
   - La médiane est-elle supérieure à 200 ? Dans cet éch., la médiane est toute valeur entre 112 & 132
4. Une certaine procédure médicale est associée à une morbidité ± importante. On voudrait savoir si cette dernière dépend d'un score dit Apgar.

	Morbidité		
Apgar	Nulle	Mineure	Majeure
0-4	21	20	16
5-6	135	71	35
7-10	158	62	35

5. Répliquez chacun de ces tests en utilisant coin ; essayez de spécifier des options pour améliorer la puissance.