Cours1_Rappels

January 19, 2024

1 Rappels (connaissances pré-requises pour cette UE)

1.0.1 Statistique : Données → connaissances

1.1 Catégories de statistiques :

Descriptives *décrire, résumer ou représenter un ensemble de données (un échantillon).*	**Inférentielles** *généraliser les observations d'un échantillon (les données) à l'ensemble de la population dont l'échantillon est issu*	
moyenne, médiane… $\bar{x} = \mu$	intervalle de confiance $\hat{\mu} \pm IC$	
Variance de l'échantillon $\frac{\sum_{i=1}^{n}(x_i-\hat{\mu})^2}{\mathbf{n}}$	Variance de la population $\frac{\sum_{i=1}^{n}(x_i-\hat{\mu})^2}{\mathbf{n-1}}$	
	paramétrique Basé sur • des paramètres ***estimés*** : $\widehat{paramètre}$ ↳ degrés de libertés • des hypothèses sur la distribution des données Bonne puissance statistique Grande flexibilité des analyses	**non-paramétrique** Basé soit sur des permutations soit sur des rangs (ordonnancement des données) Peu d'hypothèses Puissance statistique amoindrie pour les tests sur les rangs
	Quantification de la confiance dans l'inférence (IC, vraisemblance, AIC, *p*.value, Probabilité a posteriori, ...)	

1.2 Définitions importantes

1.2.1 Hypothèses nulle et alternative d'un test statistique

Classiquement, un test statistique met en jeu deux hypothèses mutuellement incompatibles (si l'une est vraie, l'autre est fausse et vice-versa). Ces deux hypothèses, nulle et alternative (H_0 et H_1) sont l'objet de l'analyse statistique qui vise à tester si les données suggèrent que H_0 est fausse, et donc que H_1 est vrai.

Typiquement, H_0 est plus simple que H_1. *Par exemple*, une variable X (ex, la taille) est étudiée dans 2 populations A et B à l'aide des échantillons x_A et x_B. On souhaite tester si la moyenne de la variable diffère entre les populations A et B. Nos deux hypothèses sont donc $\mu_A \neq \mu_B$, et $\mu_A = \mu_B$. Si les moyennes sont égales, elles doivent être *exactement* égales, or si elles sont différentes, il y a tout un continuum de différences possibles. Ici, c'est donc $\mu_A = \mu_B$ qui correspond à H_0, et $\mu_A \neq \mu_B$ à H_1.

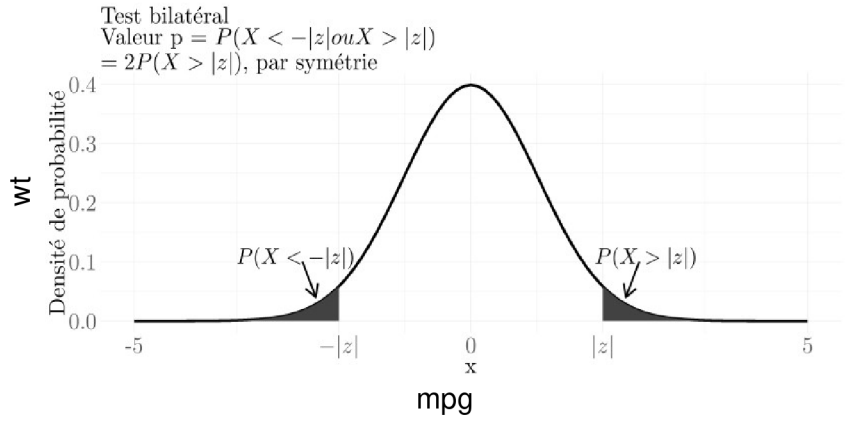
Ces hypothèses H_0 et H_1 portent sur une ou plusieurs populations ; et leur plausibilité est étudiée via des échantillons représentatifs de ces populations.

1.2.2 *p*.value

La *p*.value quantifie l'évidence contre l'hypothèse nulle (H_0) ; = en faveur de l'hypothèse alternative (H_1). Spécifiquement, pour réaliser un test statistique, une fois que H_0 et H_1 ont été établies, on utilise **toujours** les 3 étapes ci-dessous.

1. **Établir une *statistique observée*** (S_{obs}) qui i) utilise l'ensemble du jeu de donnée concerné par le test et ii) décrit la tendance du jeu de donnée à dévier de H_0 pour se rapprocher de H_1 (c.-à-d. à être incompatible avec H_0). *Par exemple*, pour $H_0 : \mu_A = \mu_B$, on peut utiliser $S_{obs} = |\widehat{\mu_A} - \widehat{\mu_B}|$. Si H_0 est vrai, ce S_{obs} sera probablement 'proche' de 0 et improbablement 'loin' de 0. La notion de 'proche' ou 'loin' de 0 est définie par la *distribution nulle*.
2. **Établir la *distribution nulle*** de S_{obs} qui correspond à la densité de probabilité décrivant, quelles sont les valeurs de S_{obs} qui sont probables ou improbables si H_0 est vraie ?
3. Extraire de la *distribution nulle*, la ***p*-value** : la probabilité d'obtenir une valeur de S_{obs} au moins aussi extrême que celle observée. Mathématiquement, on a donc *p*.value $= P\left(|S| \geq |S_{obs}| \middle| H_0\right)$

[9]:
```
options(repr.plot.width = 4.5, repr.plot.height = 2.3, repr.plot.res = 300)
File = "./Figures/p-value.png"
source('./PlotFile.R') ; Plot
```



Si une *p*.value est proche de 0 (typiquement < 0.05), cela signifie qu'il serait improbable d'observer les données que l'on a si H_0 était vrai. Or nos données sont vraies. C'est donc que H_0 est probablement fausse, on la rejette donc, ce qui revient de facto à accepter H_1. On peut souligner ici qu'une *p*.value décrit donc la probabilité des données si H_0 est vrai, ou "sachant H_0", pas la probabilité de H_0 sachant les données. Pour obtenir ce type de probabilité, il vous faudra utiliser les statistiques 'Bayésiennes', une branche des statistiques inférentielles, alternative aux statistiques 'fréquentistes', basées sur la *p*.value.

Si une *p*.value est loin de 0 (typiquement > 0.05), cela signifie qu'il est tout à fait *plausible* d'observer les données que l'on a si H_0 est vrai (H_1 est fausse). **Mais cela ne prouve en rien qu'il soit improbable d'observer nos données si H_0 est fausse.** On n'accepte donc **pas** H_0. Simplement le test est **inconclusif**.

Cependant, *en pratique*, dans certains cas de figure, où l'on souhaiterait accepter H_0 (test sur des conditions d'application d'un autre test) lorsque la p.value est loin de 0, si H_0 est en fait fausse, elle ne l'est que faiblement, et donc on considère qu'elle est vrai. *En d'autre terme*, si on a suffisamment de données et qu'une p.value est non-significative (loin de 0), alors H_0 est soit vrai, soit presque vrai. Lorsque l'on est dans ce cas de figure et que la p.value porte sur un test sur des conditions d'application d'un autre test, par abus de langage, on dit alors que l'on accepte H_0.

1.2.3 degrés de liberté

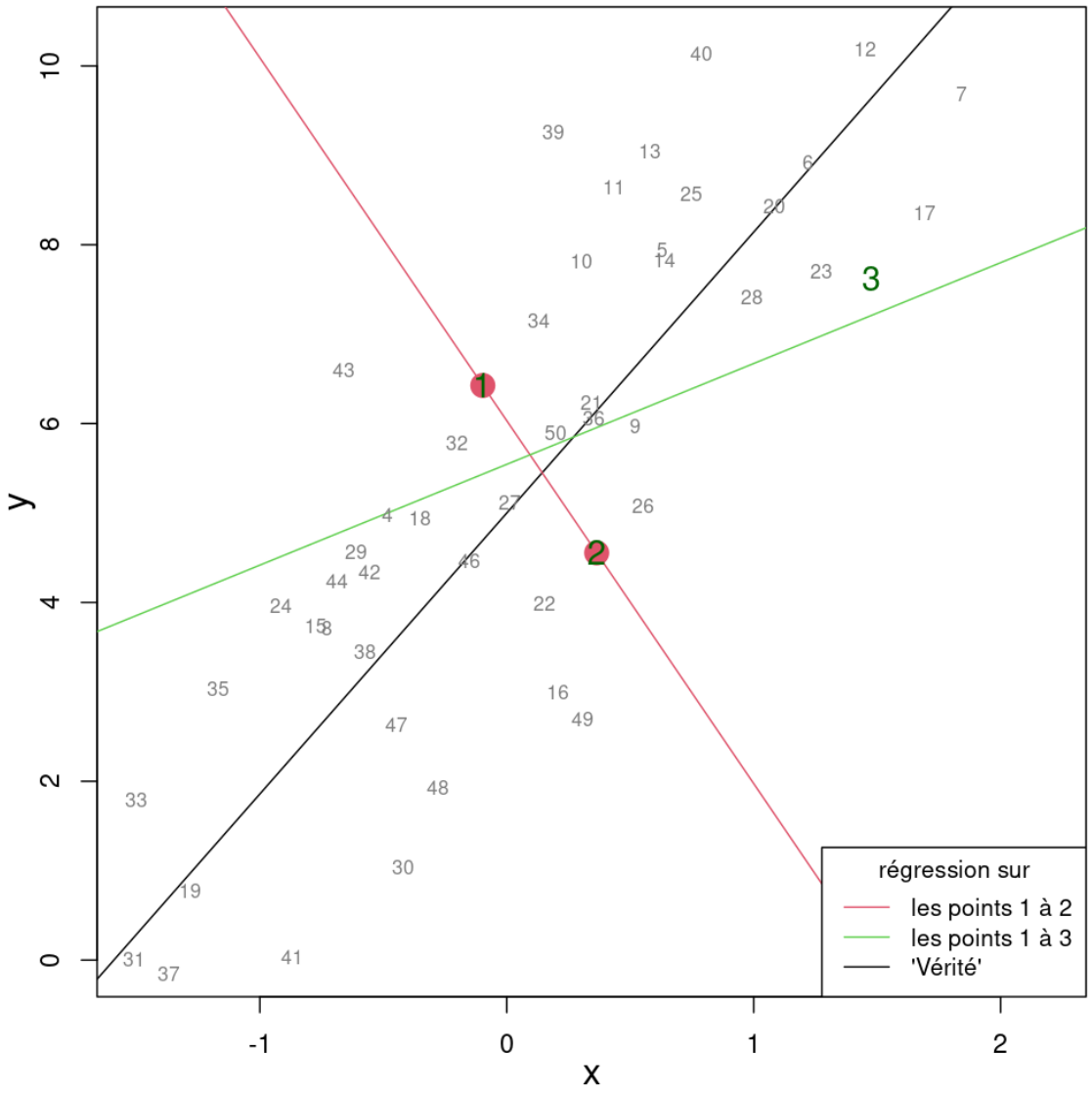
Le nombre de valeurs qui sont libres de varier. Supposez que vous vouliez choisir 5 nombres dont la moyenne fasse 3. Vous êtes libre de choisir les 4 premiers, mais le cinquième est automatiquement déterminé par la valeur des 4 premiers et la *contrainte* que la moyenne fasse 3.

En statistique, les degrés de liberté permettent de mesurer de la précision que pourra avoir une analyse statistique lorsqu'elle est appliquée à une certaine quantité de données.

Ils représentent la quantité d'information utilisée pour calculer des paramètres impliqués dans une analyse. information utilisée pour calculer les paramètres n'est plus disponible pour le reste de l'analyse.

Spécifiquement, le nombre de degrés de liberté (d'un modèle, ou d'un test) est égal au nombre d'observations moins le nombre de 'relations' entre ces observations dans le modèle ou le test. On entend par 'relations', les paramètres utilisés par le test (moyenne, variance, coefficients, etc) et qui sont autant de contraintes impliquées dans le calcul de la p.value.

[8]:
```
options(repr.plot.width = 7, repr.plot.height = 7, repr.plot.res = 140)
source("Figures/Fig_ddl.R")
```



1.3 Précision par rapport au cours précédent

1. "Si $n \geq 30$ le théorème central limite s'applique, et donc on peut s'affranchir de vérifier les conditions d'applications dans des tests de students et les modèles linéaires (ANOVA, régressions linéaires, etc.)." **C'est vrai, mais seulement lorsque les effectifs sont à peu près 'balancés'.** On parle d'effectifs 'balancés' lorsque les effectifs sont similaires dans tous les groupes comparés. C'est généralement le cas pour des données expérimentales, où les individus sont répartis aléatoirement entre les différents traitement et contrôle. Mais c'est bien moins souvent le cas pour des données observationnelles (épidémiologie, sociologie, écologie, etc.). Lorsque les effectifs sont fortement non 'balancés' (ex 3 fois plus d'individus dans un groupe que dans l'autre), une forte in-homogénéité des variances (hétéroscédasticité, liens 1 et 2) peut conduire ces tests à donner des conclusions déconnectées de la réalité des données.

2. Après une sélection de variables basée sur l'AIC (fonction R `step`), effectuer une 'validation' du modèle en testant les variables sélectionnées sur la base de l'AIC conduit à une inflation du taux d'erreur de type 1 : les *p*.values ainsi obtenues sont sous-estimées. En effet, cela revient i) à faire des tests multiples implicites, lors de l'utilisation de la fonction `step` et ii) à ne pas corriger les *p*.values obtenues sur le modèle final pour les tests multiples. C'est d'autant plus problématique que le nombre de variables éliminées par la fonction `step`. > Harrison et al., 2018 : Stepwise deletion procedures have come under heavy criticism; they can overestimate the effect size of significant predictors (Whittingham et al., 2006; Forstmeier & Schielzeth, 2011; Burnham, Anderson & Huyvaert, 2011) and force the researcher to focus on a single best model as if it were the only combination of predictors with support in the data. > Lorsque vous avez une incertitude sur le modèle à utiliser, il est préférable d'utiliser du model averaging (liens : 1, 2, 3 et 4). Cependant, pour le stepAIC, comme pour le modèle averaging, gardez à l'esprit que : > "Let the computer find out" is a poor strategy and ***usually*** reflects the fact that the researcher did not bother to think clearly about the problem of interest and its scientific setting (Burnham and Anderson, 2002).

1.4 Rappel sur les modèles linéaires

Un modèle linéaire vise à décrire l'ajustement entre une[1] variable 'expliquée' (ou 'dépendante') et une ou plusieurs variables explicatives (ou 'indépendantes'), catégorielles (ANOVA) ou continues (régressions) ou les deux à la fois.

On peut décrire un modèle linéaire comme système d'équation, avec une équation par individu statistique. Par exemple, pour une régression, on a :

$$y_1 = \mathbf{a} \times x_1 + \mathbf{b} + e_1 \; y_2 = \mathbf{a} \times x_2 + \mathbf{b} + e_2 \; y_{...} = \mathbf{a} \times x_{...} + \mathbf{b} + e_{...} \; y_n = \mathbf{a} \times x_n + \mathbf{b} + e_n$$

que l'on note plus classiquement : $y_i = \mathbf{a} \times x_i + \mathbf{b} + e_i$

Pour estimer l'ajustement entre la variable expliquée (y) et la ou les variables explicatives (ici x), le modèle choisit les valeurs des coefficients a et b qui minimisent la variance des erreurs de prédictions (e), ou plus spécifiquement, la somme du carré des écarts, $\sum_{i=1}^{n} e_i^2$ (méthode dite des moindres carrées = Ordinary Least Squares [OLS]).

Les coefficients a et b correspondent respectivement à la pente et à l'ordonnée à l'origine (souvent appelé 'intercept' par anglicisme).

Pour une ANOVA avec une variable à trois catégories (I, II et III), on a

$$y_i = \{a_I,\ a_{II},\ a_{III}\}_i + e_i$$

Où le coefficient a_I ou a_{II} ou a_{III} est choisi en fonction de la catégorie à laquelle appartient l'observation i.

Pour un modèle linéaire où l'on aurait deux variables explicatives, une continue (x) et une variable discrète à trois catégories (I, II et III), et que l'on veuille tester l'interaction[2] entre ces deux variables, si l'effet de la variable x sur la variable y dépend de la catégorie de la catégorie I, II ou III, on a mathématiquement :

$$y_i = a \times x_i + b + \{c_I,\ c_{II},\ c_{III}\}_i + \{d_I,\ d_{II},\ d_{III}\}_i \times x + e_i$$

Régression | ANOVA | Interaction | erreurs de prédiction

Les coefficients a et b correspondent à la régression (pente et intercept) les coefficients $c_{I,IIetIII}$ à

l'ANOVA, et les coefficients $d_{I,IIetIII}$, à l'interaction. Les coefficients de l'interaction $d_{I,IIetIII}$ sont choisis en fonction de la catégorie de l'observation i (comme pour l'ANOVA), mais ils sont aussi multipliés à la variable x.

[1]: en général, il n'y a qu'une seule variable expliquée par modèle, mais dans le cas des MANOVAs, il y en a plusieurs à la fois. [2]: pour une interaction, mathématiquement, > l'effet de la variable x sur la variable y dépend de la catégorie de la catégorie I, II ou III > est strictement identique à > l'effet de la catégorie I, II ou III sur la variable y dépend de la valeur de la variable x >

1.5 Observations *indépendantes et identiquement distribuées*

Une hypothèse cruciale et souvent négligée que font **tous** les tests et les modèles statistiques est que, au sein de chaque groupe, les observations sont 1. **indépendantes**, et 2. **identiquement distribuées**.

La non-prise en compte de ce point est une des principales sources d'erreur dans les analyses statistiques : 1. Observations **indépendantes** : Dans un groupe, les individus sont indépendants si le fait de connaitre la valeur de la variable pour un individu du groupe ne nous informe en rien sur les autres individus. Exemple de données non indépendantes : on étudie le nombre de grains de beauté sur les habitant e s d'un quartier. Le fait d'observer un nombre de grains de beauté très élevé sur une personne nous informe sur le fait que les membres de sa famille risquent d'en avoir, eux-mêmes, plus que la moyenne du quartier. Les membres d'une même famille ne sont pas indépendants les uns des autres, car ils partagent le même fond génétique. Dans une analyse statistique, **la non-indépendance entre observations est liée au fait que des individus statistiques ont une valeur similaire pour la variable d'intérêt qui est étudiée, parce qu'ils ont aussi une valeur proche pour une ou plusieurs autres variables confondantes qui n'ont pas été mesurées, mais qui ont un effet sur la variable d'intérêt.** Ce problème est typiquement présent pour des données spatiales ou temporelles. Par exemple, si on cherche à tester l'effet du régime alimentaire sur la santé des personnes âgées de 60 ans en France. Si on ne prend pas en compte la spatialité des données, le lieu ou ont vécu les personnes étudiées, cela revient à négliger toutes les variables confondantes spatialisées comme le niveau de pollution, l'altitude, le climat, etc. Dans le cadre d'un modèle statistique, on ne fait pas l'hypothèse que les observations sont indépendantes, mais que les erreurs (a.k.a résidus, e_i) sont indépendants. C.-à-d., que toutes les variables induisant de la non-indépendance entre les observations dans le modèle ont été prises en compte dans l'analyse.

3. observations **identiquement distribuées** : Les observations d'un échantillon proviennent toutes de la même population. Cette population peut être décrite par une distribution particulière. En d'autre terme, les valeurs observées sur différents individus résultent des mêmes phénomènes sous-jacents.

En résumé, des données (ou les résidus dans le cas des moèles) doivent être **i**ndépendantes et **i**dentiquement **d**istribuées (en abrégé `i.i.d.`) pour être analysées statiquement.

Même si ce n'est pas toujours mentionné, tous les intervalles de confiances et les tests statistiques font l'hypothèse que les données sont i.i.d.

Lorsque les données ne sont pas indépendantes, il existe différentes approches pour prendre en compte la structure de dépendance entre les données. L'analyse de **données appariées** pour des tests 'simples', ou des **effets aléatoires** dans le cas de modèles statistiques permettent de prendre en compte la structure de dépendance entre les données.

2 Programme *prévisionnel* de l'UE

Pour cette UE, nous avons choisi de vous donner un aperçu relativement large de la diversité des méthodes que vous pourrez être amenés à utiliser. La conséquence est que nous ne pourrons approfondir aucune des méthodes que nous vous présenterons. Si vous travaillez correctement cette UE, et que nous sommes bons, vous devriez à la fin être en statistique des 'Jack of $_{\text{all}}$ *many* trades, master of none'.

Pour utiliser ces méthodes suite à cette UE, c'est à vous qu'il reviendra d'approfondir vos connaissances sur celles que vous souhaiterez utiliser.

Ci-après, vous trouverez le programme *prévisionnel* de l'UE : celui-ci sera adapté selon le rythme auquel nous avançons.

Dans le tableau ci-dessous, n'hésitez pas au cours de l'UE à compléter la colonne "Références" pour pouvoir approfondir vos connaissances sur ces méthodes.

Intervenant·e	**Descriptive** *décrire, résumer ou représenter un ensemble de données (un échantillon).*	**Inférentielle** *généraliser les observations d'un échantillon (les données) à l'ensemble de la population dont l'échantillon est issu*		**Références**
		Fréquentiste *p*.value	**Bayésienne** Probabilité a posteriori	
Hugo Mathé-Hubert		Bootstraps, vraisemblance et Modèles linéaires généralisés (GLM)		Bootstraps : 1
Hugo Mathé-Hubert		Effets aléatoires et Modèle mixtes (généralisés ou non) (G·LMM)		G·LMM : 1, 2, 3 et 4
Christelle Gonindard	Analyses multivariées : analyses en composantes principales (ACP) et méthodes de classification			
Christelle Gonindard	modèles de forets aléatoires (RF : Random Forest)	modèles de forets aléatoires (RF : Random Forest)		RF : généralités, *p*.value, high order interactions, sur-apprentissage et Random survival forests
Matthias Grenié			Introduction aux statistiques Bayésiennes et MCMC	1