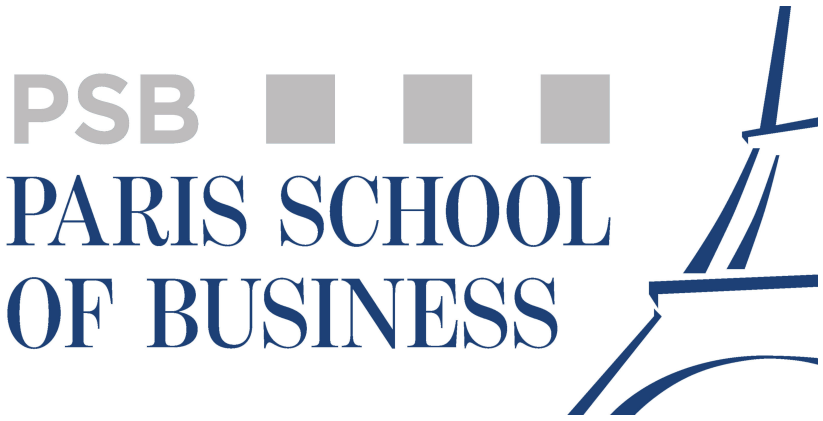
PSB Paris School of Business



Probability Value

Hakim DAIF

Zakaria RIDADARAJAT

15 novembre 2020

*

Table des matières

Table des figures

Liste des tableaux

Première partie

Concepts historiques

1 Introduction

Lorsque des chercheurs lisent ou publient des résultats de tests d'hypothèses, ils portent généralement une attention particulière aux coefficients de signification (p). Dans le second cas, ils espèrent que leur valeur sera inférieure ou égale à 0,05. De plus, la plupart des rédacteurs et évaluateurs de revues savantes ont les mêmes préoccupations, d'où, peut-être, une partie de leur intérêt pour la *p-value*. Le fait d'obtenir des résultats statistiquement significatifs accroît leurs chances d'être publiés (Maddock[1] et Rossi, 2001 ; Nakagawa, 2004 ; Poitevineau, 2004).

Or, selon plusieurs auteurs, la valeur *p-value* ne devrait constituer qu'une étape de l'interprétation des tests d'hypothèses. En fait, il faudrait, généralement, que son rôle soit appuyé par d'autres informations comme la taille de l'effet ; mais, dans tous les cas, il n'aurait qu'une importance pratique limitée (Cohen[2], 1962 ; Gigerenzer, 1993 ; Kline, 2004 ; Thompson, 1989).

2 Problématique

Depuis le début des années 1960, des études continuent d'identifier un recours presque exclusif à la valeur de la *p-value* pour l'interprétation des tests d'hypothèses et affichent, de plus, une puissance expérimentale insuffisante (Bezeau[3] et Graves, 2001). Cette situation persiste même si de nombreux auteurs ont tenté de provoquer un changement dans les pratiques en sciences humaines et sociales (Carver 1978).

Les préoccupations quant aux répercussions de cette situation sur la rigueur de la recherche sont assez importantes pour que soit créé, aux États-Unis, un groupe de travail sur l'inférence statistique (Wilkinson[4] and the Task Force on Statistical Inference, 1999). Des réactions similaires proviennent également du monde francophone.

Au Québec, en 1991, Blais proposait une réflexion sur la pratique de la statistique en éducation. Ce texte s'avère d'ailleurs un précurseur du présent article mais, comme il a été publié dans une revue avec tirage limité et maintenant disparue, il peut être difficilement accessible à la communauté scientifique. Par conséquent, il paraît approprié de réactualiser et de compléter le propos puisque, outre ce texte, la préoccupation envers l'adéquation de l'usage des tests d'hypothèses dans la francophonie ne semble pas déborder le cadre de la psychologie et de la médecine.

3 Objectif de recherche

Dans l'étude qu'on va mener, nous souhaitons faire le point sur l'utilisation des tests d'hypothèses sur les aléas du recours exclusif à la valeur de la *p-value*. Plus précisément, après avoir présenté un bref historique des tests d'hypothèses, nous aborderons les concepts de signification statistique, de taille de l'effet et de puissance des tests statistiques ; ensuite, nous consacrons toute une partie à l'étude de la *p-value* en donnant quelques exemples illustratifs.

4 Cadre théorique : les tests d'hypothèses

Le test d'hypothèses ne trouve son utilité que lorsque l'étude de la population entière est impossible et que le chercheur doit plutôt analyser un échantillon de cette population . Dans ce cas, comme l'échantillonnage comporte inévitablement une marge d'erreur, le test d'hypothèses vise à indiquer la probabilité d'obtenir les statistiques observées sur la base d'une hypothèse quant à la valeur d'un paramètre de la population. Si le premier test d'hypothèses connu, le test du khi-carré, peut être attribué à Karl Pearson , c'est Ronald Fisher qui a d'abord esquissé la logique méthodologique des tests d'hypothèses.

4.1 L'approche de Fisher

Selon Fisher, dans le cadre d'un plan expérimental, le test d'hypothèses vise à réfuter une hypothèse donnée, sans lui adjoindre d'hypothèse concurrente. La logique de Fisher débute donc avec la formulation d'une hypothèse *H*, selon laquelle la statistique (la moyenne, par exemple) d'un échantillon aléatoire, tiré d'une population hypothétique infinie, est égale à une valeur donnée. Ensuite, on teste la différence entre le paramètre de la distribution d'échantillonnage théorique et la statistique observée dans l'échantillon. L'hypothèse sera rejetée si les valeurs comparées diffèrent de plus d'un écart convenu d'avance .

Au départ, une certaine confusion peut provenir de l'évolution dans le temps du discours de Fisher : dans ses premiers écrits, il prône l'adoption de critères de signification fixes, alors que, dans les années 1950, il change de position et propose que ces critères puissent être variables. Le chercheur devrait alors publier la probabilité exacte obtenue, et non la valeur retenue du critère de signification .

Par ailleurs, en l'absence de résultats significatifs, c'est-à-dire de résultats permettant de rejeter *H*, l'hypothèse n'est pas acceptée : le chercheur suspend alors son jugement. Selon Fisher, la finalité des tests était l'inférence inductive, bien que la probabilité obtenue soit d'obtenir les données observées en postulant la véracité de l'hypothèse nulle, donc **P(D/H)**.

4.2 L'approche de Neyman et Pearson

La contribution de Jerzy Neyman et d'Egon Pearson se voulait une tentative de consolider les travaux de Fisher de la transformer en une approche plus cohérente et rigoureuse . Par conséquent, Neyman et Pearson délaissent l'inférence inductive pour mettre les tests d'hypothèses au service de la prise de décision dans des contextes pragmatiques. Ainsi, ils ajoutent notamment à l'approche de Fisher une analyse dans une logique de coûts et de bénéfices .

D'abord, là où Fisher ne posait qu'une seule hypothèse, Neyman et Pearson formulent une hypothèse testée **(H0)** et une contre-hypothèse ou hypothèse alternative **(H1)**. Ces deux hypothèses se doivent d'être exhaustives et mutuellement exclusives, de sorte que le rejet de l'une implique l'acceptation de l'autre, et vice-versa. Il s'ensuit l'introduction de deux types d'erreurs et de leur probabilité associée : l'erreur de type I (α), soit rejeter H0 à tort, et l'erreur de type II (β), soit conserver H0 à tort. Dans cette logique, les valeurs acceptables de α et β sont fixées a priori : il appartient au chercheur de déterminer les risques d'erreurs qu'il est prêt à assumer, et ce, en tenant compte des coûts relatifs à chaque type d'erreurs.

Soulignons qu'avec le concept d'erreur de type II apparaît aussi le concept de puissance statistique, qui est son complément en termes de probabilités ($1 - \beta$). De plus, c'est le caractère pragmatique de l'approche de Neyman et Pearson qui confère son utilité au calcul de la taille des effets observés. Précisons finalement que, contrairement à Fisher, qui posait l'hypothèse d'une population infinie pour pouvoir utiliser le concept de distribution d'échantillonnage, Neyman et Pearson entrevoient leur interprétation des tests d'hypothèses dans un contexte de répétition ; ainsi, la probabilité α devient le pourcentage d'erreurs de type I, commises par le chercheur sur une grande série de répétitions de la même expérience et provenant du tirage successif d'échantillons aléatoires d'une même population .

4.3 L'approche de Bayes

Les deux approches discutées précédemment appartiennent toutes deux à l'allégeance fréquentiste, pour emprunter le terme anglo-saxon, par opposition à l'approche bayésienne. Cette dernière découle des travaux de Thomas Bayes et se distingue des approches précédentes en ce qu'elle vise à établir le degré de certitude par rapport à une hypothèse sur la base des données obtenues. Il s'agit donc de $\mathbf{P(H/D)}$, soit la probabilité de l'hypothèse H conditionnelle à l'observation des données **D**, le Saint Graal de l'inférence statistique, à une nuance près, c'est-à-dire le degré de certitude envers la vraisemblance de l'hypothèse et non de la probabilité que l'hypothèse soit effectivement vraie. Or, le calcul de cette vraisemblance (évaluée comme une probabilité) requiert notamment du chercheur qu'il attribue une probabilité initiale à la véracité de l'hypothèse, ce qui devient subjectif dans la mesure où cette probabilité théorique est généralement inconnue.

Ce procédé devient d'autant plus subjectif que l'ignorance du chercheur quant au phénomène observé est grande. De plus, comme plusieurs chercheurs pourraient attribuer des probabilités initiales différentes à une même hypothèse, le procédé risque de refléter davantage les opinions du chercheur que la réalité.

5 Critiques de la p-value

Les tests de significativité statistique ne sont pas parfaits, et ont dernièrement fait l'objet de nombreuses critiques. Plusieurs limites connues existent. En voici quelques-unes.

5.1 Mauvaise interprétation de la p-value

La p-value est souvent mal interprétée, y compris dans la littérature scientifique. Une erreur courante consiste à l'interpréter comme la probabilité que l'hypothèse nulle («il n'y a pas d'effet») soit vraie.

5.2 Oubli de la taille de l'effet

Au-delà de la significativité statistique d'un résultat, il est malheureusement courant d'oublier de commenter la taille de l'effet mesuré : si l'effet existe mais qu'il est d'une taille très petite, est-ce vraiment nécessaire d'y consacrer de l'attention ?

Une version encore plus problématique consiste à interpréter la p-value comme la taille de l'effet : ça n'est pas parce que la p-valueest très petite (donc que l'effet mesuré est statistiquement très significatif) que l'effet en question est nécessairement important en taille.

5.3 Seuil de la p-value

Le choix du seuil de lap-valueva nécessairement influencer les résultats qui seront considérés comme fiables des autres : plus le seuil est haut, plus on est «tolérant à l'imprécision».

Cela étant, dans certains domaines les données empiriques sont par nature très imprécises – typiquement, les sciences humaines et sociales. Cela «oblige» les chercheurs de ces disciplines à accepter une «tolérance à l'imprécision» plus grande. Cette tolérance est, en comparaison, nettement plus restreinte en physique des particules.

5.4 Sur-interprétation

Un test unique de significativité ne permet pas en lui-même de conclure à l'existence (ou à l'inexistence) d'un effet. En général, d'autres études sont nécessaires pour corroborer le résultat du test.

5.5 Taille de l'échantillon

Lorsque l'échantillon utilisé pour mesurer l'effet est très grand, n'importe quel effet extravagant pourra être mesuré. Cette limite est décrite parla loi des très grands nombres.

Deuxième partie

L'inférence statistique

Traditionnellement, lorsque les on essaye d'apprendre pour la première fois à analyser des expériences, on met fortement l'accent sur la vérification des hypothèses et la prise de décisions basées sur des *p-values*. La vérification des hypothèses est importante pour déterminer s'il y a des effets statistiquement significatifs. Cependant, il ne faut absolument pas négliger l'importance des *p-values*.il est aussi très importante de réaliser que les *p-values* sont affectées par la taille de l'échantillon et qu'une faible *p-values* ne suggère pas nécessairement un effet important ou un effet pratiquement significatif.

Avant d'entamer l'étude sur la *p-values*,il indispensable de comprendre ce qu'est l'inférence statistique

6 Qu'est-ce que l'inférence statistique ?

— Considérons une population (par exemple, l'ensemble des êtres humains).
— Considérons une variable d'intérêt $\mathbf{Y}$ (par exemple, la taille des êtres humains).
— Considérons par ailleurs une variable explicative $\mathbf{X}$ (par exemple, le sexe(homme ou femme) des individus considérés).

Pour étudier et expliquer la taille des êtres humains "en général", on est obligé d'échantillonner la variable $\mathbf{Y}$, puis de tenter d'extrapoler le résultat obtenu sur l'échantillon à l'ensemble de la population. De même, si l'on cherche à savoir si $\mathbf{X}$ à un effet "en général" sur $\mathbf{Y}$, il faudra déterminer dans quelle mesure on peut extrapoler l'éventuelle différence observée entre les groupes à la population.

C'est cette généralisation (ou extrapolation) de l'échantillon à la population que l'on appelle inférence statistique.

6.1 Un premier exemple d'inférence statistique : l'estimation d'un paramètre

Considérons par exemple, que l'on ait mesuré 5 êtres humains "au hasard" sur terre, et qu'ils mesurent respectivement 159, 179, 183, 166 et 167cm. On s'intéresse à la taille moyenne (μ) de la population humaine.

```
Console R:
  > quelques_tailles=c(159,179,183,166,167)
  > mean(quelques_tailles)
   [1] 170.8
```

Avec une taille d'échantillon aussi petite (n=5), il est évidemment difficile d'extrapoler les résultats pour affirmer **La taille moyenne des êtres humains est 170.8cm**.

En effet, la capacité à réaliser l'inférence statistique dépend de la taille d'échantillon.

Considérons maintenant un échantillon plus grand :

```
Console R:
> representations=read.table("../../datasets/representations.csv",
                                                    sep=";",header=T)
> attach(representations)
> length(taille) # taille d'echantillon
  [1] 199
> mean(taille)
  [2] 170.7638
```

Avec cet échantillon, on est déjà un peu plus confiant lorsque l'on décrit la taille des êtres humains "en général". Malgré cela, il reste impossible d'affirmer " La taille moyenne des êtres humains est 170.7638 ". En effet, la moyenne observée sur l'échantillon est une estimation $\widehat{\mu}$, a priori différente de la valeur réelle de taille moyenne μ du fait du hasard d'échantillonnage.

Il convient donc d'être "prudent" dans nos affirmations et d'assortir, par exemple, notre estimation de moyenne d'un intervalle de confiance.L'intervalle de confiance (IC) à 95% est un intervalle de valeurs qui ont 95% de chance de contenir la vraie valeur du paramètre μ.Le calcul de l'intervalle de confiance repose sur un modèle.

Ici, supposons que les mesures Y_i soient indépendantes les unes des autres et qu'elles

soient toutes distribuées selon la même loi gaussienne (ou normale) autour de la moyenne μ.

On peut ajuster ce modèle comme suit :

```
Console R:
> modele=lm(taille~1) # modele lineaire: correspond aux hypotheses
                                                susmentionnees

> modele # renvoie l'estimation du parametre du modele

Call:
lm(formula = taille ~ 1)

Coefficients:

(Intercept)

      170.8
```

Partant de ce modèle, un intervalle de confiance pour μ est :

$$\widehat{\mu} \pm 1.96\sqrt{\frac{Sd(Y))}{n}}$$

Dans ce cas, au vu des données, l'estimation de μ est $\widehat{\mu}$=170.8 et il y a 95% de chances que la valeur de μ soit comprise entre 169.177cm et 172.3507cm :

```
Console R:
> confint(modele)

                2.5 %    97.5 %

      (Intercept) 169.177 172.3507
```

Notez que l'incertitude sur cette estimation était évidemment beaucoup plus grande lorsque la taille d'échantillon était n=5 :

```
Console R:
> modele_bis=lm(quelques_tailles~1)

    confint(modele_bis)

                  2.5 %    97.5 %

(Intercept) 158.4956 183.1044
```

6.2 Deuxième exemple d'inférence statistique : détection de l'existence d'un lien entre deux variables

Considérons maintenant le lien entre la variable **X** et la variable **Y**.

```
Console R:
> apply(taille, sexe, "mean")

   femme     homme

   164.7708  176.3495
```

Il semblerait, au vu de l'échantillon, que la taille moyenne diffère selon que l'on s'intéresse aux hommes ou aux femmes...

On peut, comme on l'a fait auparavant, essayer d'estimer la taille moyenne, cette fois-ci en fonction du sexe.

```
Console R:
> modele=lm(taille~sexe+0)

    confint(modele)

                  2.5 %    97.5 %

     sexefemme  162.8019  166.7398

     sexehomme  174.4486  178.2504
```

Ici, le fait que les intervalles de confiance soient disjoints tend à prouver que les tailles moyennes sont effectivement différentes en fonction du sexe. Néanmoins, si les intervalles de confiance se recouvraient en partie, ils ne nous permettraient pas d'affirmer si, oui ou non, le facteur de groupementXXavraimentun effet, en moyenne, sur la variableYY.

Une autre forme d'inférence statistique vise précisément à extrapoler à l'ensemble de la population le fait qu'il existe un effet de **X** sur **Y**. Il s'agit des tests d'hypothèse.

Les tests d'hypothèse s'appuient sur un modèle statistique qui décrit Y en fonction de X. Un tel modèle s'accompagne typiquement des questions suivantes :

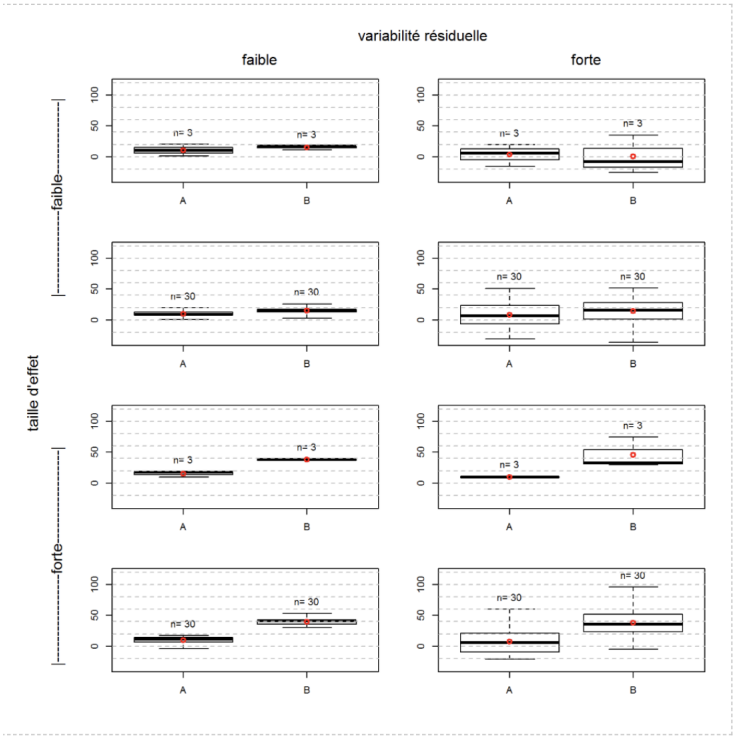
- En termes effet : quel est l'effet de X sur Y ?
- En termes de significativités : Est-ce que l'effet observé de X sur Y est significatif, ou au contraire pourrait-il simplement s'expliquer par le hasard d'échantillonnage ? C'est à cette question que le test d'hypothèse est censé apporter une réponse.
- En termes de prédiction et de qualité d'ajustement : Serait-on en mesure de

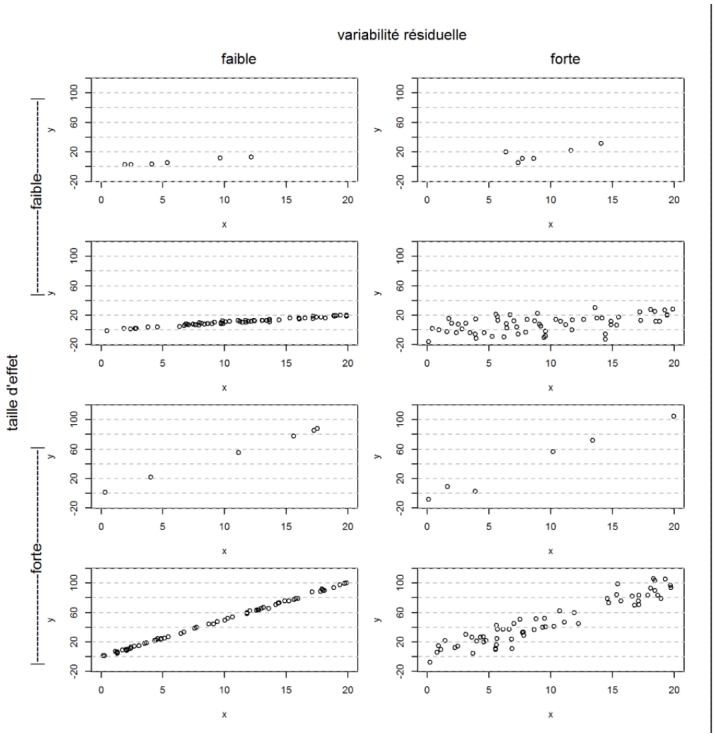
prédire la valeur de Y, connaissant celle de X, avec précision ? Yest-elle "étroitement" liée à X ou juste "vaguement" ?). Autrement dit, la variabilité résiduelle des observations par rapport au modèle proposé est-elle forte ?

Les figures 1 et 2 vous permettent d'apprécier "intuitivement" l'influence de :

— la taille d'échantillon
— la taille d'effet
— la variabilité résiduelle

sur la capacité à extrapoler l'effet observé de X sur Y à la population.





Troisième partie

Chapitre 3 : P-value en pratique

7 Package utilisé dans ce chapitre

Les packages utilisés dans ce chapitre comprennent :

— lsr

Les commandes suivantes permettent d'installer ce package s'il n'est pas déjà installé : if(!require(lsr))install.packages("lsr")

8 Test d'hypothèses

8.1 Erreurs d´décisionnelles et risques

		Etat de la nature	
		H_0	H_1
Décision	Rejet de H_0	α	$1-\beta$
	Non rejet de H_0	$1-\alpha$	β

Table 1: Risques décisionnels conditionnels à l'état de la nature

Le principe de base d'un test de signification est de considérer une hypothèse privilégiée H0 et une alternative H1, puis de bâtir une règle permettant de décider de rejeter ou pas H0.

Le tableau 1 résume les 4 situations possibles. L'erreur de première espèce est de rejeter l'hypothèse privilégiée H0 alors qu'elle est vraie. L'erreur de seconde espèce est de ne pas rejeter H0 alors qu'elle est fausse. α est la probabilité de rejeter 'a tort l'hypothèse H0 ; α est aussi appelé risque de première espèce, ou niveau du test β est la probabilité de ne pas rejeter H0 alors que l'hypothèse alternative H1 est vraie ; β est appelé risque de seconde espèce. La valeur 1 β est la puissance du test, et traduit la faculté de rejeter H0 quand l'alternative H1 est vraie. Dans la pratique, α est fixé par l'expérimentateur (les valeurs les plus courantes sont 0,05 ou 0,01. On dit qu'on contrôle le risque de première espèce. Par contre, β peut être difficile à calculer. Heureusement, ce calcul n'est pas n´nécessaire sauf si l'on veut comparer plusieurs procédures de tests. Dans la littérature, H0 est aussi appelée hypothèse nulle ou encore hypothèse principale. Elle joue un rôle prédominant par rapport 'à l'hypothèse H1 qui est souvent l'hypothèse alternative contraire. On cherche à contrôler le risque α de rejeter 'a tort H0 en lui imposant une valeur relativement faible (au plus 0,05). Le fait d'imposer une valeur faible à α conduit à n'abandonner l'hypothèse H0 que dans des cas qui semblent sortir nettement de l'ordinaire si H0 était vraie.

9 Définition de la p-value

Dans un test statistique, la p-value (en anglais probability value), parfois aussi appelée p-valeur, est la probabilité pour un modèle statistique donné sous l'hypothèse nulle d'obtenir la même valeur ou une valeur encore plus extrême que celle observée.1

Notons bien que plus α est choisi petit, plus la règle de d´décision est stricte (ou conservative) dans la mesure où elle aboutit à rejeter H0 que dans des cas rarissimes et donc à conserver cette hypothèse quelque fois à tort. Une vision moderne, liée 'a l'explosion de la puissance des ordinateurs et de processus numérique d'approximation rapides et précis, est d'afficher la p-value ou probabilité critique pc.

Par d´définition, la p-value est la plus petite des valeurs de risque de première espèce pour lesquelles la d´décision serait de rejeter H0. La valeur pc est calculée à

partir des observations et de leurs propriétés distributionnelles sous H0. Comme Pc est le plus petit niveau de signification auquel on rejette l'hypothèse H0, il est aussi appelé niveau de signification observée. L'amélioration fulgurante des capacités de calcul permet maintenant de baser les règles de d´décision sur les probabilités critiques sans forcément comparer la statistique de test avec une valeur seuil, comme cela se faisait classiquement. La définition formelle de la p-value donnée ci-dessus est difficile 0 ingurgiter et peut conduire 'a une mauvaise utilisation et/ou une mauvaise interprétation de l'inférence statistique .

Une définition littérale et plus parlante aux non-initiés peut être la suivante : la p-value est une mesure de la compatibilité des données avec l'hypothèse privilégiée. Plus cette p-value est proche de zéro, plus la compatibilité est faible et donc conduit 0 rejeter cette hypothèse. La proximité à zéro dépend de la sévérité que l'on s'impose à travers le risque α.2

9.1 La règle de décision

Les tests statistiques reposent en général sur la vérification d'une hypothèse nulle, qui a une formulation spécifique pour chaque test. L'hypothèse nulle décrit toujours le cas où, par exemple, deux groupes ne sont pas différents ou il n'y a pas de corrélation entre deux variables, etc.

L'hypothèse alternative est le contraire de l'hypothèse nulle, et décrit donc les cas où il y a une différence entre les groupes ou une corrélation entre deux variables, etc.

Notez que les définitions d'hypothèse nulle et d'hypothèse alternative n'ont rien à voir avec ce que vous voulez ou ne voulez pas trouver, ou ce qui est intéressant ou non, ou ce que vous vous attendez à trouver ou ce que vous ne vous attendez pas à trouver. Si vous compariez la taille des hommes et des femmes, l'hypothèse nulle serait que la taille des hommes et la taille des femmes ne sont pas différentes. Pourtant, vous pourriez être surpris si vous trouviez cette hypothèse vraie pour une population que vous étudiez. De même, si vous étudiez les revenus des hommes et des femmes, l'hypothèse nulle serait que les revenus des hommes et des femmes ne sont pas différents, dans la population que vous étudiez. Dans ce cas, vous pouvez espérer que l'hypothèse nulle est vraie, mais vous ne serez pas surpris si l'autre hypothèse est vraie. Dans tous les cas, l'hypothèse nulle prendra la forme qu'il n'y a pas de différence entre les groupes, qu'il n'y a pas de corrélation entre deux variables, ou qu'il n'y a pas d'effet de cette variable dans notre modèle.

La p-value sera déterminée en effectuant le test statistique.

Cette p-value est ensuite comparée à une valeur alpha prédéterminée. Le plus souvent, une valeur alpha de 0,05 est utilisée, mais cette valeur n'a rien de magique.

— Si la p-value du test est inférieure à la valeur alpha, nous rejetons l'hypothèse nulle.

— Si la p-value est supérieure ou égale à alpha, nous ne rejetons pas l'hypothèse nulle.

9.2 En pratique. . .

En pratique, que ce soit dans les manuels de statistique, dans le cadre de la formation des chercheurs ou dans les textes que publient les revues savantes en sciences humaines et sociales , l'interprétation des tests d'hypothèses fait appel à une logique mixte combinant des éléments des trois approches présentées. D'abord, en accord avec Fisher, seule l'hypothèse nulle est généralement formulée (lorsqu'elle l'est explicitement), la visée est surtout l'inférence inductive où, conformément à sa position la plus récente, plusieurs seuils de signification sont utilisés (0,05 ; 0,01 et 0,001) et, souvent, les probabilités exactes sont publiées. Puis, dans la veine des travaux de Neyman et Pearson, apparaissent les concepts d'erreurs de types I et II et, plus rarement, des considérations sur la taille de l'effet (maintenant largement utilisée en psychologie) et la puissance statistique. Enfin, l'interprétation des résultats est souvent bayésienne, dans le sens où l'on généralise régulièrement les résultats des tests d'hypothèses en les associant à la population de référence (et ce, bien que la probabilité conditionnelle obtenue en réalité aille dans l'autre sens).

9.3 La signification statistique

Essayons maintenant d'illustrer, par un exemple, l'application de cette logique hybride des tests d'hypothèses. Supposons que nous voulons tester la différence entre garçons et filles d'un cours de sciences quant à la fréquence à laquelle ils ont à recopier les notes que leur enseignant écrit au tableau (Conseil des ministres de l'Éducation du Canada, 2007). Notons que, comme garçons et filles fréquentent les mêmes classes, il n'y a pas lieu de croire en la présence d'une différence significative. L'approche fréquentiste attribue une probabilité à la valeur d'une statistique obtenue par calcul à partir des données de recherche et testée à partir d'une distribution de probabilités connue . C'est le cas, par exemple, du test du khi-carré (^{2}) : la différence entre les fréquences attendues et les fréquences observées permet de calculer la valeur de la statistique ^{2}, à laquelle est attribuée une probabilité p sur la base de la distribution de probabilités du khi-carré. Dans notre exemple, nous cherchons à explorer l'association entre le sexe de l'élève et la pratique de l'enseignant, exprimée comme une fréquence parmi quatre catégories possibles. Notre échantillon compte 21 961 élèves et nous obtenons un résultat statistiquement significatif (^{2}[6] = 29,00,p= 0,00) au seuil de 0,05. Ici, comme il est illogique de croire que garçons et filles ne sont pas exposés à la même quantité de cours magistraux, que signifie ce résultatsignificatif ? C'est que le résultat du test d'hypothèses sera déclaréstatistiquement significatifou non sur la base de la valeur dep.Ainsi, sips'avère inférieur à un seuil de signification arbitraire, généralement 0,05en sciences sociales, le résultat est dit statistiquement significatif. Or, sous l'influence de Fisher et

de ses derniers écrits, la valeur depest souvent mise en parallèle avec plusieurs valeurs critères, chaque seuil semblant correspondre à une certitude plus grande quant au rejet de l'hypothèse nulle, ou alors, la valeur exacte depsera communiquée (p= 0,00 dans notre exemple) et considérée significative ou non à partir des mêmes critères. Ici,p(0,00) désigne la probabilité d'obtenir une valeur de ^{2} supérieure ou égale à 29,00 si cette valeur est de 0 dans la population (l'hypothèse nulle postule un effet d'une ampleur donnée, généralement nulle, dans la population hypothétique d'où serait tiré l'échantillon). Le rejet de cette hypothèse nulle (et de sa conséquence statistique, $^{2} = 0$) constitue d'ailleurs un abus de langage puisqu'elle est, pratiquement, toujours fausse . Il faut ici comprendre que, comme l'explique Chow (1996), il est théoriquement possible que l'hypothèse nulle soit vraie : étant donné des échantillons aléatoires tirés d'une population théorique, poser l'hypothèse nulle revient à soutenir qu'il n'y a, dans les données, aucune autre variation que celle attribuable aux fluctuations aléatoires d'échantillonnage. Or, cette condition correspond ausystème isolé sans frottementdes physiciens : il s'agit d'une condition idéale, théorique, rarement rencontrée dans la pratique, puisqu'un nombre important de variables confondues, dont l'effet peut être plus ou moins important, viennent souvent menacer le postulat selon lequel les groupes comparés sont équivalents au départ. De plus, le chercheur ne dispose pratiquement jamais d'échantillons aléatoires , ne serait-ce qu'à cause de l'obligation éthique d'obtenir le consentement des sujets qui donne un caractère volontaire à presque tout échantillon, qu'il soit originellement aléatoire ou non. Justement pour ces raisons, l'ouvrage de Chow a d'ailleurs suscité un vif débat, dans la revueBehavioral and Brain Sciences, quant à l'applicabilité pratique de son raisonnement théorique qui, lui, n'est pas contesté. Dans ce numéro, dirigé par Paul Bloom et Barbara L. Finlay , trente-sept auteurs ont réagi à l'ouvrage de Chow en publiant de courts articles. Nous avons repris les arguments énumérés dans ces articles, mais à partir d'autres textes, plus étoffés, dont certains provenant des auteurs ci-haut mentionnés.

Notons ici que les tests d'hypothèses postulent que l'hypothèse nulle est exactement vraie dans la population, pour ensuite calculer une probabilité d'obtenir les résultats observés à partir d'un échantillon aléatoire tiré de cette population. La probabilité conditionnelle évaluée est donc P(D—H0) et non P(H0—D) . Dans ce cas-ci, il s'agit de la probabilité d'observerDconditionnellement à la véracité de l'hypothèse nulle. Contrairement à une conception erronée, mais néanmoins largement répandue, les résultats de tests d'hypothèses ne donnent pas d'informations sur la population dans la mesure où la probabilité que l'hypothèse nulle y soit exactement vraie est, pratiquement, infinitésimale. Par contre,une valeur de p [statistiquement] non significative ne signifie pas qu'il n'y a pas de différence mais plutôt qu'aucune preuve de l'existence d'une différence n'a pu être trouvée.

9.4 La taille de l'effet

Dans une optique pragmatique, l'une des conceptions associées à la signification statistique est que l'on confond signification et importance. Ainsi, un résultat

statistiquement significatif ne se traduit pas nécessairement par une importance au niveau pratique. Il faut ici considérer qu'un échantillon suffisamment vaste mènera généralement au rejet de l'hypothèse nulle ou, en d'autres mots, à un résultat statistiquement significatif et ce, peu importe la taille réelle de l'effet (pourvu qu'elle ne soit pas nulle).

La taille de l'effet désigne à quel degré un phénomène donné est présent dans la population. Ainsi, une autre façon de concevoir l'hypothèse nulle est de dire que la taille de l'effet est nulle. Dans cette optique, la taille de l'effet décrit dans quelle mesure l'hypothèse nulle est fausse : plus la taille de l'effet est grande, plus il est justifié de rejeter l'hypothèse nulle. Or, pour un grand échantillon, un effet même minime suffira à la faire rejeter. Concrètement, ce problème peut être envisagé à travers l'exemple présenté plus haut : notre résultat significatif au test du khi-carré s'accompagne d'un effet (calculé à l'aide duVde Cramer) considéré comme significatif, mais néanmoins négligeable, selon les critères de Cohen (V= 0,02 ;p= 0,00). La portée pratique d'une telle conclusion peut difficilement être vue comme importante. Néanmoins, comme lors de l'interprétation dep, il faut éviter de porter un jugement mécanique sur la taille de l'effet . Dans la mesure où les exigences du test d'hypothèses sont respectées, un résultat statistiquement non significatif malgré une taille d'effet importante signifie que cet effet, aussi grand soit-il, pourrait être obtenu avec une probabilité supérieure à la valeur critère . Pour éviter une telle situation, le chercheur doit veiller, notamment, à minimiser l'erreur de mesure. Cela dit, afin de fournir au lecteur toute l'information requise pour juger des résultats obtenus, le chercheur devrait toujours fournir la taille de l'effet au terme d'un test d'hypothèse . En fait, la taille de l'effet peut s'avérer plus utile à l'interprétation des résultats que la valeur dep, surtout si l'ampleur de cet effet peut influencer les décisions pratiques subséquentes.

9.5 La puissance statistique

Supposons maintenant qu'une biologiste place une lamelle sous son microscope pour vérifier la présence de bactéries et qu'elle constate que la lamelle semble vierge. Peut-elle en déduire que c'est obligatoirement le cas ? En fait, deux explications sont possibles : 1) la lamelle est effectivement exempte de bactéries ou 2) l'objectif n'est pas assez puissant pour distinguer des corps aussi petits que des bactéries. Notons que, dès qu'une bactérie est visible, la question ne se pose pas.

Dans la logique de Neyman et Pearson (le concept de puissance statistique ne s'applique pas à l'approche de Fisher), le même raisonnement peut être appliqué aux tests d'hypothèses. Lorsque le chercheur obtient un résultat non significatif, est-ce dû au fait que cet échantillon pourrait provenir d'une population où l'hypothèse nulle est vraie ou est-ce dû à un manque de puissance statistique ? Encore ici, la question n'a de sens qu'en l'absence de résultats statistiquement significatifs.

Selon Cohen (1988), la puissance d'un test statistique peut se définir commela probabilité qu'il produise des résultats statistiquement significatifs quand H0 est fausse.

Au regard d'une hypothèse nulle donnée, la puissance peut aussi être comprise comme la probabilité qu'un test donné mène à son rejet. La puissance d'un test statistique dépend principalement de trois paramètres : le seuil de signification, la taille de l'échantillon et la taille de l'effet.

Ainsi, quand le seuil de signification est augmenté (par exemple, passe de 0,05à 0,10), la puissance augmente aussi. La direction d'un test a également une influence sur la puissance. Un test statistique peut être bilatéral (two-tailed) ou unilatéral (one-tailed). Dans un test bilatéral, l'hypothèse nulle peut être rejetée sur la base d'un écart dans chaque direction par rapport à l'hypothèse nulle. Au contraire, dans un test unilatéral, un écart dans une seule direction permet de rejeter l'hypothèse nulle. Les tests unilatéraux sont plus puissants que les tests bilatéraux, mais leur puissance est nulle dans la direction autre que celle spécifiée .

Par ailleurs, la fiabilité d'un estimateur provenant d'un échantillon désigne la précision avec laquelle il parvient à identifier approximativement la valeur du paramètre correspondant relié à la population de référence. En d'autres termes, un estimateur fiable est celui qui minimise l'erreur due à l'échantillonnage. La taille de cette erreur varie de façon inverse au nombre d'unités statistiques incluses dans l'échantillon : plus l'échantillon est grand, plus l'erreur sera faible et plus l'estimateur sera fiable. Conséquemment, quand la taille de l'échantillon augmente, la puissance statistique augmente aussi. À la limite, comme la probabilité de retrouver un effet exactement nul dans la population est pratiquement inexistante, tout échantillon suffisamment vaste finira par produire des résultats significatifs (Vacha-Haase et Nilsson, 1998). Enfin, plus l'effet est grand, plus il sera facile à détecter. Par conséquent, si la taille de l'effet augmente, la puissance statistique augmente aussi.

En outre, comme le fait remarquer Maxwell (2004), de nombreux devis de recherche impliquent des tests d'hypothèses multiples, par exemple, lors de l'utilisation de régressions multiples ou d'analyses de variance à plan factoriel. Ainsi, une ANOVA donnée peut mener au calcul de trois puissances statistiques distinctes : la probabilité de détecter au moins un effet significatif, la probabilité de détecter un effet significatif donné et la probabilité de détecter tous les effets significatifs. Ces trois puissances statistiques seront généralement d'ampleurs différentes : la probabilité de détecter au moins un effet sera plus élevée que celle de détecter un effet donné, qui à son tour sera plus élevée que celle de détecter tous les effets. Conséquemment, une étude peut démontrer une puissance statistique suffisante pour détecter au moins un effet, mais rendre la détection de tous les effets peu probables. Autrement dit, il se peut que la probabilité de commettre une erreur de type II pour au moins un effet soit très élevé malgré une puissance statistique suffisante pour détecter au moins un effet (la puissance peut aussi être considérée comme la probabilité complémentaire à la probabilité d'erreur de type II : $1 - \beta$). Certains auteurs, à la suite de Neyman et Pearson, questionnent d'ailleurs la priorité accordée systématiquement à l'erreur de type I dans les écrits : selon la situation étudiée, l'erreur de type II peut avoir des conséquences pratiques plus graves que l'erreur de type I. Alors, pourquoi cette préoccupation à se prémunir uniquement contre cette dernière ? Les seuils acceptables pour α et β ne devraient-ils pas dépendre

des coûts relatifs associés à chaque type d'erreurs dans la situation étudiée ?

10 Exemples simples de calcul de la p-value (dans R)

10.1 Exemple de pièces de monnaie

Afin d'expliquer l'utilisation de la p-value pour tester une hypothèse, imaginez que vous avez une pièce de monnaie que vous lancerez 100 fois. L'hypothèse nulle est que la pièce est non truquée, c'est-à-dire que la probabilité d'obtenir pile ou face est la même (0,5). L'hypothèse alternative est que la pièce est truquée.

Supposons que pour cette expérience, vous lanciez la pièce 100 fois et on obtient 95 fois pile (sur les 100 fois). La p-value dans ce cas serait la probabilité d'obtenir 95, 96, 97, 98, 99 ou 100 fois pile, ou uniquement 0, 1, 2, 3, 4 ou 5 fois pile, en supposant que l'hypothèse nulle est vraie.

C'est ce que nous appelons un test bilatéral, puisque nous testons les deux extrêmes suggérés par nos données : obtenir 95 fois pile ou plus ou obtenir 95 fois face ou plus. Dans la plupart des cas, nous utiliserons des tests bilatéraux.

Vous pouvez imaginer que la p-value de ces données sera assez faible. Si l'hypothèse nulle est vraie, et que la pièce est non truquée, il y aurait une faible probabilité d'obtenir 95 fois pile ou plus ou 95 fois face ou plus.

En utilisant un test binomial, la p-value est ¡ 0,0001.

(En fait, R la rapporte comme ¡ 2,2e-16 (notation scientifique = 0,00000000000000022, avec 15 zéros après la virgule).

```
Console R:
> binom.test(5, 100, 0.5)

 Exact binomial test

number of successes = 5, number of trials = 100,
                            p-value < 2.2e-16

alternative hypothesis: true probability of success
                              is not equal to 0.5
```

10.2 Exemple de réussite et d'échec

Autre exemple : imaginez que nous ayons deux salles de classe et que nous comptions les étudiants qui ont réussi un certain examen. Nous voulons savoir si une classe a statistiquement plus de réussites ou d'échecs que l'autre.

Dans notre exemple, chaque classe comptera 10 élèves. Les données sont classées dans un tableau de contingence.

groupe de classe	réussite	échec
A	8	2
B	3	7

Nous utiliserons le test de Fisher pour vérifier s'il existe un lien entre la classe et le nombre d'étudiants ayant réussi et échoué. L'hypothèse nulle : il n'y a pas de lien entre le groupe et réussite/échec , sur la base des comptes relatifs dans chaque cellule du tableau de contingence.

```
Console R:
> Input =("
groupe reussite  echec
A         8          2
B         3          7
")
> Matrix = as.matrix(read.table(textConnection(Input),
                                header=TRUE, row.names=1))

> Matrix

        reussite echec
    A      8       2
    B      3       7


> fisher.test(Matrix)

Fisher's Exact Test for Count Data

p-value = 0.06978
```

La p-value est de 0,070. Si nous utilisons un alpha de 0,05, alors la p-value est supérieure à l'alpha, nous ne pouvons donc pas rejeter l'hypothèse nulle. C'est-à-dire que nous n'avions pas suffisamment de preuves pour dire qu'il y a une association entre le groupe et réussite/échec.

Dans ce cas, les données seraient plus extrêmes si les chiffres en haut à gauche ou en bas à droite (ou les deux) étaient plus élevés.

Références

[1] Maddock, J. E. et Rossi, J. S. (2001) (2001). Statistical power of articles published in three health psychology-related journals. Health Psychology, 20(1), 76-78.

[2] Cohen, J. (1962). The statistical power of abnormal-social psychological research : a review. Journal of abnormal and social psychology, 65(3), 145-153.

[3] Bezeau, S. et Graves, R. (2001). Statistical power and effect sizes of clinical neuropsychology research. Journal of clinical and experimental neuropsychology, 23(3), 399-406.

[4] Wilkinson, L. and the Task Force on Statistical Inference (1999). Statistical methods in psychology journals : guidelines and explanations. American psychologist, 54(8), 594-604.

[5] Blais, J.-G. (1991). Statistique, méthodes quantitatives et analyse des données. Repères, essais en éducation, (13), 63-90.