INTERPRÉTATION D'UN TEST D'HYPOTHÈSE
Comparaison des approches fréquentiste et Bayésienne

François PIRAUX, Emmanuelle GOURDAIN, Florent DUYME, Emmanuelle HERITIER
ARVALIS Institut du végétal – 11 décembre 2015

ARVALIS
Institut du végétal

Pourquoi utiliser des approches Bayésiennes ?

Quelques situations où les méthodes Bayésiennes pourraient être appropriées:

- vous avez de l'information **a priori** concernant les paramètres de votre modèle et vous voulez la prendre en compte dans l'analyse de vos données
- vous voulez faire de l'inférence sur des **combinaisons non-linéaires** des paramètres de votre modèle
- vous voulez prendre en compte l'**incertitude** des estimations des composantes de la variance dans un modèle mixte
- **les p-values sont souvent mal interprétées, et vous voulez utiliser la distribution a posteriori relative à l'hypothèse que vous testez**

L'inférence fréquentiste est largement répandue

Schéma classique :

Expérience

↓

anova

→

```
Response: rdt
            Sum Sq Df    F value    Pr(>F)
(Intercept)  83404  1 19419.2179 < 2.2e-16
variete        740  9    19.1535 1.567e-09
bloc            64  3     5.0048  0.006892
Residuals      116 27
```

↓

comparaisons multiples

→

```
contrast          estimate       SE df t.ratio p.value
BASTION - ARKAS     8.2300 1.465425 27   5.616  <.0001
ECHO - ARKAS       -2.3675 1.465425 27  -1.616  0.4987
FLAMBARD - ARKAS    2.9375 1.465425 27   2.005  0.2871
HERMES - ARKAS      4.2075 1.465425 27   2.871  0.0532
JERICHO - ARKAS     8.6425 1.465425 27   5.898  <.0001
KOLIBRI - ARKAS    -3.1225 1.465425 27  -2.131  0.2326
ROCK - ARKAS        4.6425 1.465425 27   3.168  0.0270
SIROCCO - ARKAS     6.4600 1.465425 27   4.408  0.0012
WIM - ARKAS         9.5500 1.465425 27   6.517  <.0001
```

↓

IC (éventuellement)

L'inférence fréquentiste est largement répandue

The *p*-value is arguably the most used and most misused quantity in all of statistical practice (Littell et al, 2006)

La p-value associée à un test d'hypothèse est sans doute la quantité la plus utilisée (et souvent mal utilisée) pour interpréter des données.

QUIZZ - 1

Test réalisé par Haller et Krauss (2002) sur des professeurs d'université enseignant la statistique, des chercheurs en psychologie et des étudiants en psychologie

On a un traitement médical que l'on soupçonne d'altérer la performance d'une certaine tâche. On compare la moyenne d'un groupe ayant reçu le traitement à celle d'un groupe ayant reçu un placebo.

Le résultat du test t est significatif : t=2.7, ddl=18, p_value=0.01

Affirmations :

1. Nous connaissons la probabilité que l'hypothèse nulle soit vraie (le traitement n'a pas d'effet)
2. Nous pouvons déduire du résultat la probabilité qu'a l'hypothèse alternative d'être vraie (le traitement altère la performance)
3. Si, de façon hypothétique, l'expérience était répétée un grand nombre de fois, nous obtiendrions un résultat significatif dans 99% des cas

QUIZZ - 1

Réponses possibles au quizz 1 :

A : les 3 affirmations sont vraies

B : 2 affirmations sont vraies

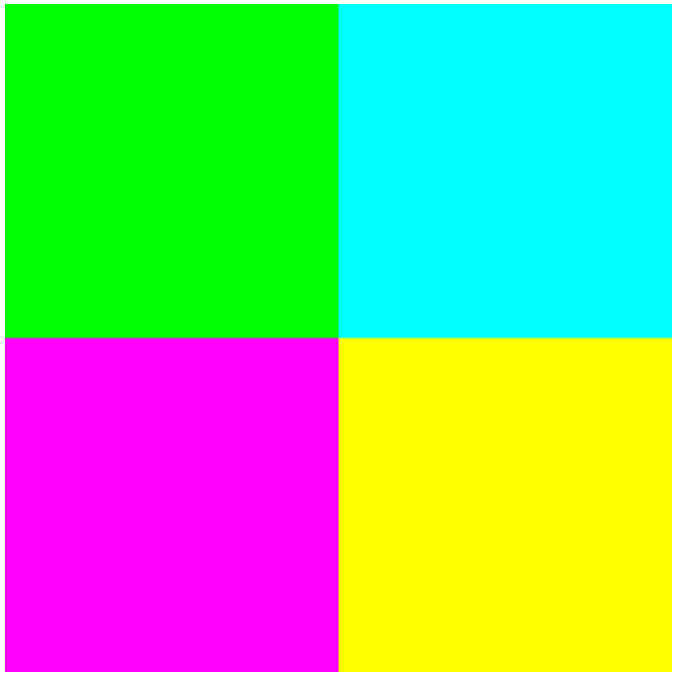
C : 1 affirmation est vraie

D : aucune affirmation n'est vraie

QUIZZ - 1

Dépouillement des résultats avec VotAR

1. Orientez la face à 4 couleurs vers l'animateur



2. Tournez la feuille de manière à lire correctement la lettre correspondant à votre réponse (ici, on choisit la réponse B)

D

C A

B

QUIZZ - 2

Test réalisé par Hoekstra et al (2014) sur 120 chercheurs et 476 étudiants, tous dans le domaine de la psychologie

Le professeur tournesol mène une expérience, analyse les données et commente :

« L'intervalle de confiance à 95% de la moyenne est compris entre 0.1 et 0.4 »

Affirmations :

1. L'hypothèse nulle Ho : « la moyenne est égale à 0 » est probablement fausse
2. Nous sommes sûrs à 95% que la vraie moyenne se situe entre 0.1 et 0.4
3. Si nous répétions l'expérience un grand nombre de fois, alors dans 95% des cas la vraie moyenne se situerait entre 0.1 et 0.4.

QUIZZ - 2

Réponses possibles au quizz 2 :

A : les 3 affirmations sont vraies

B : 2 affirmations sont vraies

C : 1 affirmation est vraie

D : aucune affirmation n'est vraie

Dépouillement des résultats avec VotAR

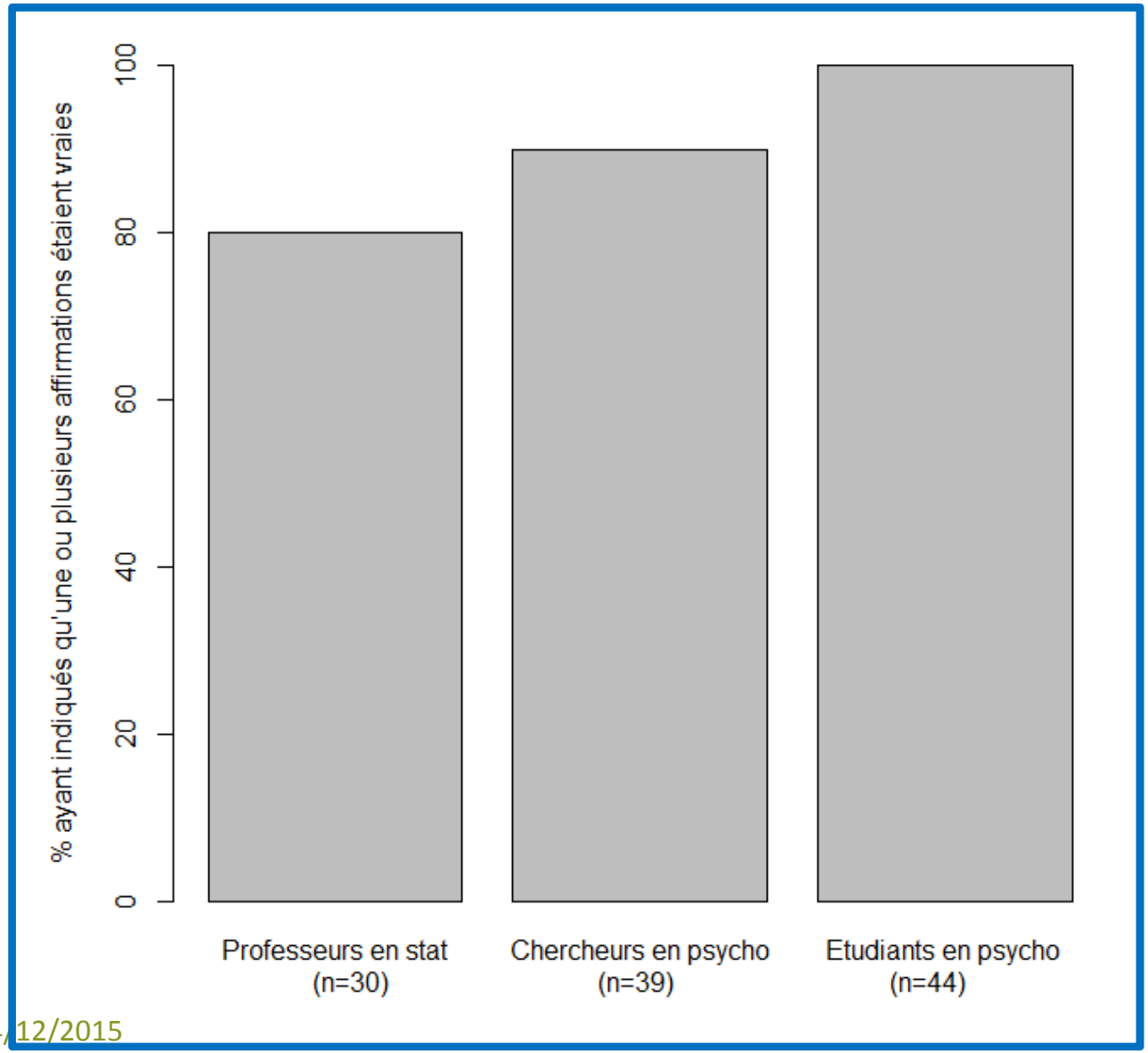
QUIZZs - résultats

TOUTES LES AFFIRMATIONS SONT FAUSSES

Il fallait répondre la réponse D

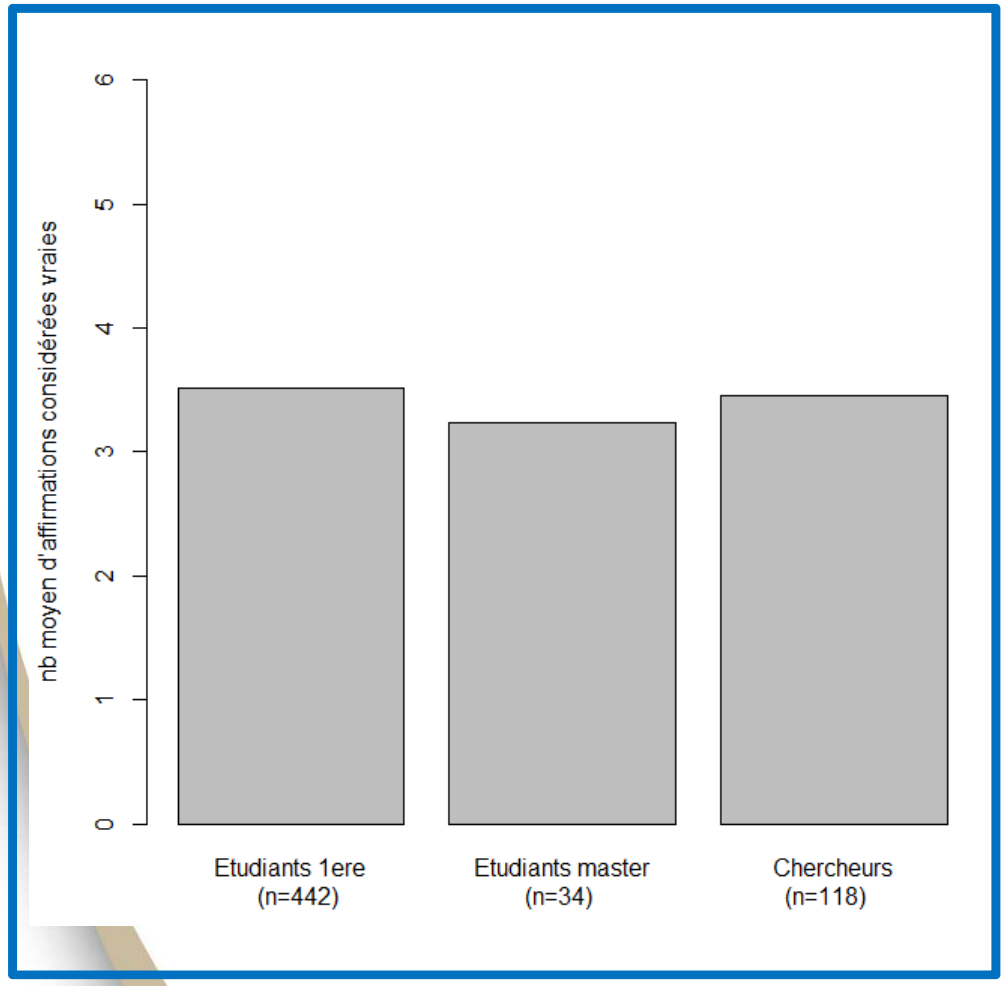
QUIZZs – résultats de la biblio

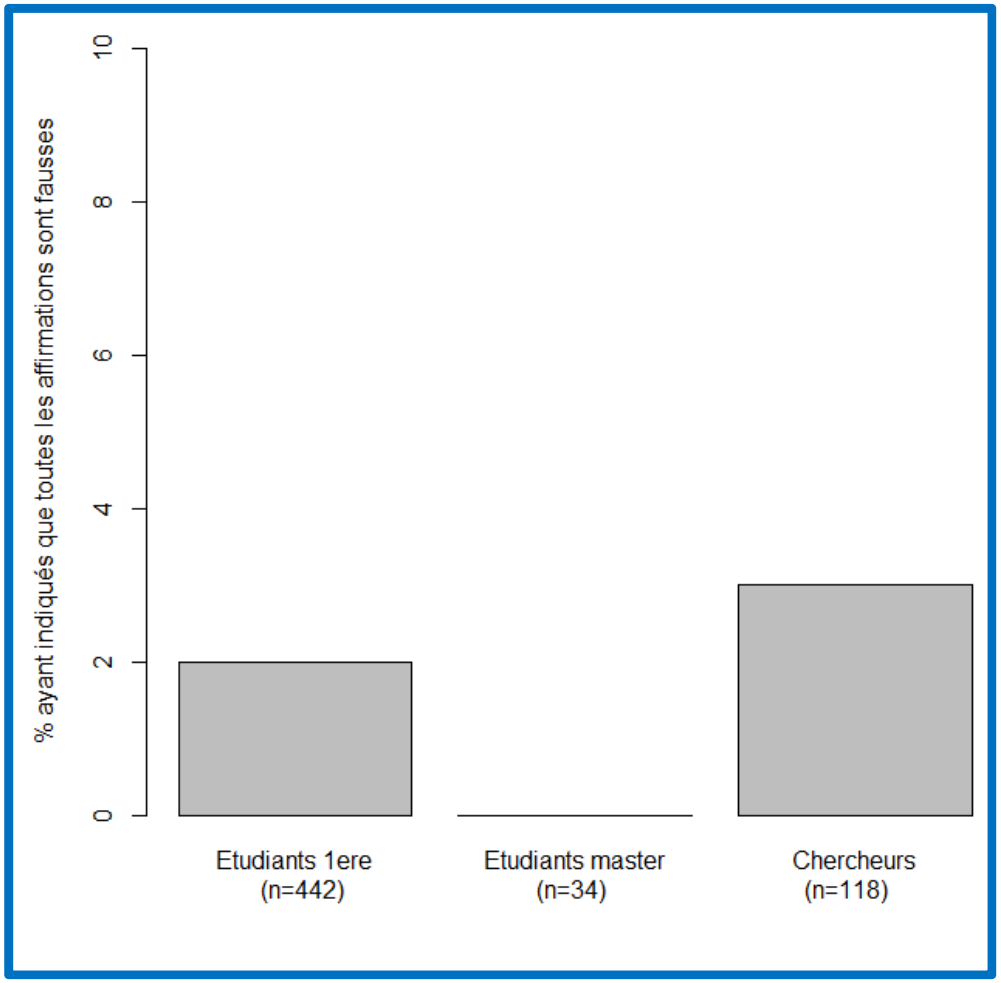
p-value



QUIZZs – résultats de la biblio

Intervalle de confiance





Définition

Probabilité critique (p_value) : probabilité d'observer une statistique de test T au moins aussi élevée que celle calculée à partir des données observées, quand Ho est vraie

ou

Probabilité d'observer des données au moins aussi extrêmes que celles observées lorsque Ho est vraie

Définition

Intervalle de confiance à 95% d'un paramètre : C'est un intervalle calculé de façon à ce que si on répète l'expérience un grand nombre de fois, 95% des fois il contiendra la vraie valeur du paramètre.

Réponses QUIZZ - 1

On a un traitement médical que l'on soupçonne d'altérer la performance d'une certaine tâche. On compare la moyenne d'un groupe ayant reçu le traitement à celle d'un groupe ayant reçu un placebo.

*Le résultat du test t est **significatif** : t=2.7, ddl=18, **p_value=0.01***

TOUT EST FAUX

1. Nous connaissons la probabilité que l'hypothèse nulle soit vraie (le traitement n'a pas d'effet) : la différence **théorique** n'est pas une variable aléatoire, on ne peut pas lui associer de probabilité, elle est égale ou pas à 0

2. Nous pouvons déduire du résultat la probabilité qu'a l'hypothèse alternative d'être vraie (le traitement altère la performance) : la différence **théorique** n'est pas une variable aléatoire, on ne peut pas lui associer de probabilité, elle est égale ou pas à 0

3. Si, de façon hypothétique, l'expérience était répétée un grand nombre de fois, nous obtiendrions un résultat significatif dans 99% des cas : on ne peut pas dire cela à partir de la p-value. Si l'hypothèse nulle est vraie, la distribution des p-values du test est une distribution uniforme et la probabilité d'avoir un résultat significatif est égale à 5%

Réponses QUIZZ - 2

Le professeur tournesol mène une expérience, analyse les données et commente :

« L'intervalle de confiance à 95% de la moyenne est compris entre 0.1 et 0.4 »

TOUT EST FAUX

1. L'hypothèse nulle Ho : « la moyenne est égale à 0 » est probablement fausse : la moyenne **théorique** n'est pas une variable aléatoire, on ne peut pas lui associer de probabilité, elle est **égale** ou **pas** à 0

2. Nous sommes sûrs à 95% que la vraie moyenne se situe entre 0.1 et 0.4 : la moyenne **théorique** n'est pas une variable aléatoire, on ne peut pas lui associer de probabilité, elle est **comprise** ou **pas** entre 0.1 et 0.4

3. Si nous répétions l'expérience un grand nombre de fois, alors dans 95% des cas la vraie moyenne se situerait entre 0.1 et 0.4 : la valeur de la vraie moyenne est **comprise** ou **pas** entre 0.1 et 0.4. Ce n'est pas une variable aléatoire, contrairement aux bornes de l'IC : il y a 95% des IC calculés qui contiennent la vraie moyenne

Fréquentiste vs Bayésien

La raison principale pour laquelle la p-value d'un test d'hypothèse est souvent mal interprétée, est sans doute qu'**elle ne répond pas à la question** que se pose le chercheur

- p-value = probabilité d'observer des données au moins aussi extrêmes que celles observées si H_0 est vraie : $P(data|H_0)$
- Ce que le chercheur veut connaître, c'est la probabilité que son hypothèse soit vraie, sachant les données observées : $P(H_0|data)$
- $P(H_0|data)$ est la probabilité a posteriori de H_0
- Dans le cadre Bayésien, on associe une probabilité à H_0

Fréquentiste vs Bayésien

		Vous allez chez le médecin parce que vous toussez
FRÉQUENTISTE	p-value = $P(data\|H_0)$	probabilité de tousser sous l'hypothèse d'avoir un cancer de la gorge
BAYÉSIEN	probabilité a posteriori de H_0 = $P(H_0\|data)$	probabilité d'avoir un cancer de la gorge sachant qu'on tousse

Le raisonnement Bayésien

Exemple de raisonnement bayésien : **un médecin reçoit un patient qui tousse**

$P(H)$ = probabilité a priori (prévalence)

$$P(cancer) = 0.05\%$$
$$P(gastro) = 25\%$$
$$P(angine) = 28\%$$

$P(data|H)$ = vraisemblance

$$P(tousser|cancer) = 95\%$$
$$P(tousser|gastro) = 0.3\%$$
$$P(tousser|angine) = 89\%$$

$$P(H|data) = \frac{P(data|H)\,P(H)}{P(data)} = proba\ a\ posteriori$$

$$P(cancer|tousser) \cong 0.05 * 95 = 4.75$$
$$P(gastro|tousser) \cong 25 * 0.3 = 7.5$$
$$P(angine|tousser) \cong 28 * 89 = 2492$$

$2492/4.75=525$
$2492/7.5=332$

Pourquoi l'interprétation Bayésienne est-elle plus intuitive ?

Le cerveau Bayésien

Cours dispensés par Stanislas Dehaene, psychologue cognitif et neuroscientifique

http://www.college-de-france.fr/site/stanislas-dehaene/

Le cerveau statisticien

Importance de l'a priori

Le cerveau Bayésien

Le test d'hypothèse Bayésien

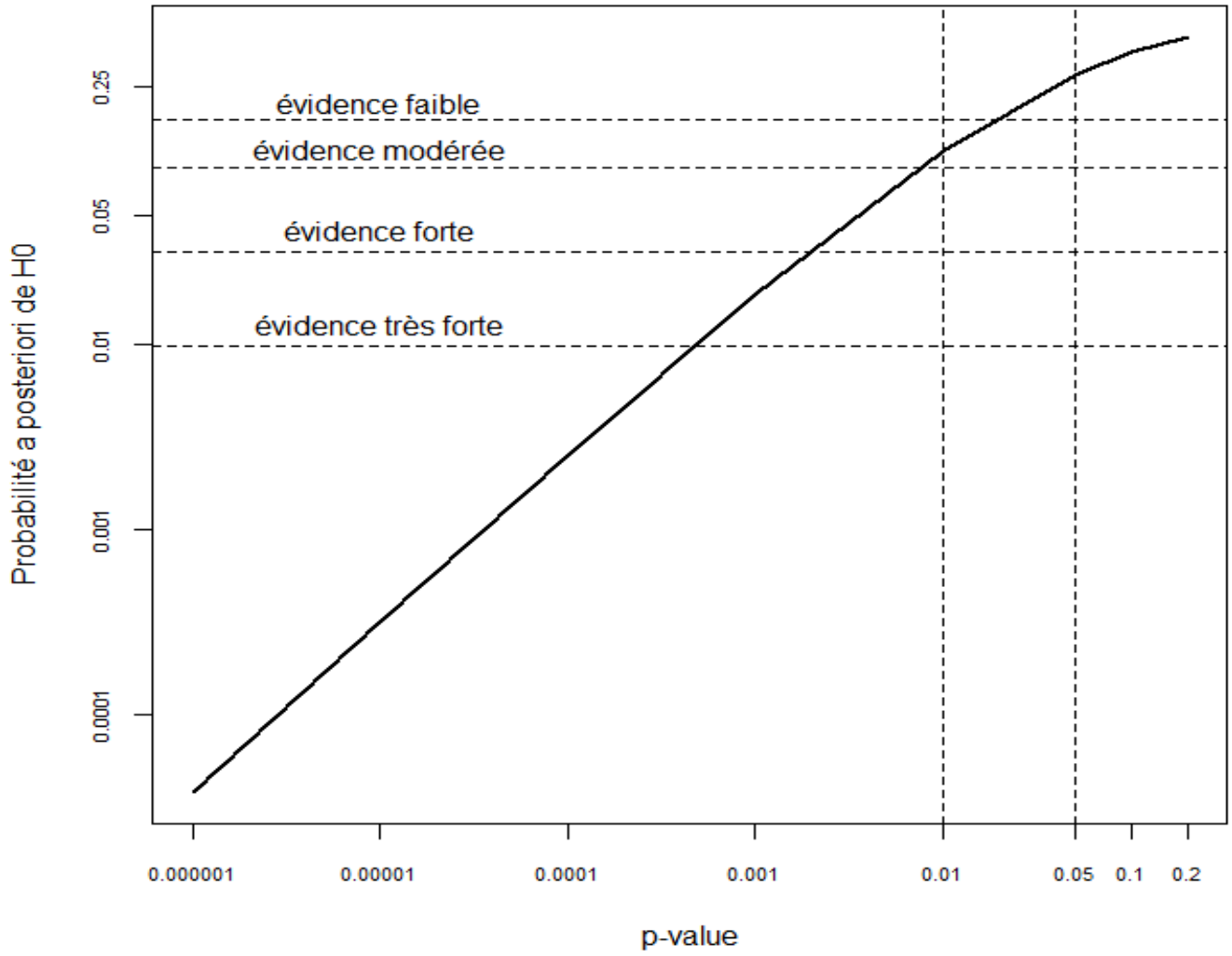
Le test d'hypothèse Bayésien repose sur le calcul de la probabilité a posteriori de H_0 : $P(H_0|data)$

On rejettera l'hypothèse nulle lorsque la probabilité a posteriori de H_0 est inférieure à un seuil d'évidence donné :

Probabilité a posteriori	Force de l'évidence
0.17	Faible
0.09	Modérée
0.03	Forte
0.01	Très forte

Relation entre p-value et probabilité *a posteriori*

Il existe une relation approximative entre la p-value et la probabilité a posteriori de H_0



Relation entre p-value et probabilité *a posteriori*

- La décision de rejeter l'hypothèse nulle en fonction d'un seuil de significativité donné est **équivalent** à rejeter l'hypothèse nulle lorsque la probabilité a posteriori de H_0 est inférieure à un seuil d'évidence donné.

- Cependant, le seuil classique de 5% utilisé en statistique fréquentiste correspond souvent à un seuil d'évidence plutôt faible dans le cas bayésien

L'information a priori dans le test d'hypothèse Bayésien

Pour calculer $P(H_0|data)$, il est nécessaire de spécifier la **probabilité a priori** de H_0, $P(H_0)$

- quand on compare un traitement dont le mode d'action est connu, à un témoin non traité, on peut s'attendre à ce que le traitement étudié soit très efficace et on considérera une probabilité a priori pour H_0 relativement faible, par exemple $\boldsymbol{P(H_0) = 0.1}$
- quand on compare deux traitements dont le mode d'action est similaire, on peut s'attendre à ce que les traitements ne soient pas différents et on considérera une probabilité a priori pour H_0 relativement élevée, par exemple $\boldsymbol{P(H_0) > 0.75}$
- souvent notre a priori est plutôt vague et incertain et dans ce cas on considérera un a priori non informatif, c'est-à-dire, $\boldsymbol{P(H_0) = 0.5}$

L'information a priori dans le test d'hypothèse Bayésien

Exemple chiffré : comparaison d'un nouveau produit fertilisant A à un produit de référence B

$H_0 : \mu_A = \mu_B$

moy_A - moy_B = 0.9 q/ha, n=41

Résultats avec différents a priori :

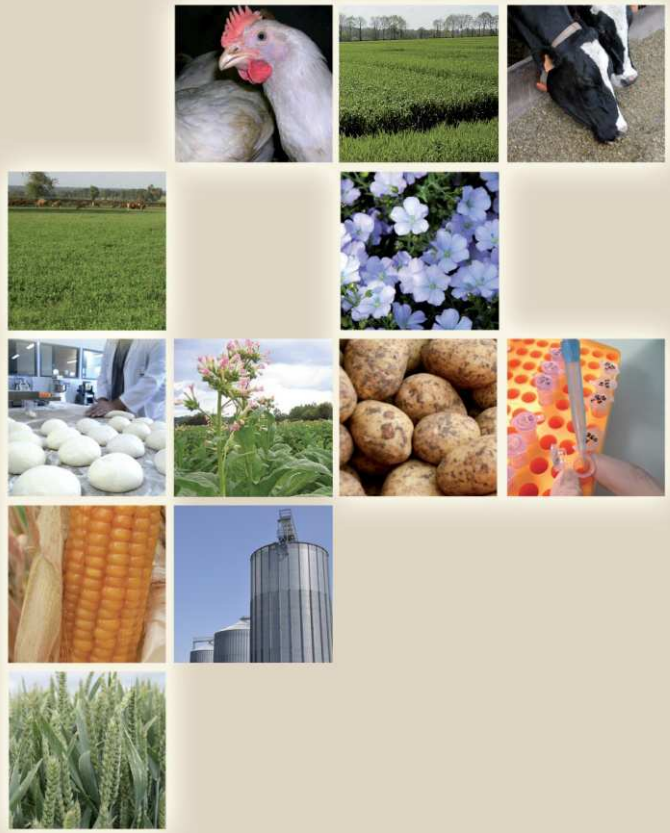
Probabilité a priori	Probabilité a posteriori	Évidence
$P(H_0) = 0.1$	$P(H_0 \mid data) = 0.0082$	Très forte
$P(H_0) = 0.5$	$P(H_0 \mid data) = 0.069$	Modérée
$P(H_0) = 0.9$	$P(H_0 \mid data) = 0.40$	Faible

Conclusions

- De nombreux utilisateurs réalisent leurs tests d'hypothèses dans un cadre fréquentiste en adoptant une interprétation bayésienne de la p-value, interprétation erronée
- Cependant, les conclusions obtenues avec un test d'hypothèse fréquentiste sont équivalentes à celles obtenues avec une approche Bayésienne (c'est juste une question de seuil)
- Les tests d'hypothèses Bayésiens apparaissent plus intuitifs à interpréter, mais surtout permettent de prendre en compte une information a priori

Références

- Berger J. O. (2003). Could Fisher, Jeffreys and Neyman Have Agreed on Testing? Statistical Science 2003, Vol. 18, No. 1, 1–32.
- Cohen, J. (1994). The earth is round (p <.05). American Psychologist, 49, 997–1003.
- Dagnelie P. (1992). Statistique Théorique et Appliquée, Tome 1. Gembloux, Presses Agronomiques, 492 p.
- Fisher, R. A. (1955). Statistical methods and scientific induction. *Journal of the Royal Statistical Society B*, 17, 69-78.
- Haller, H., Krauss, S., 2002. Misinterpretations of significance: a problem students share with their teachers? Methods of Psychological Research—Online [On-line serial], 7, 1–20.
- Hoekstra R., Morey R.D., Rouder J.N. & Wagenmakers E.-J. (2014). Robust misinterpretation of confidence intervals. Psychon Bull Rev, 8p.
- Johnson V. E. (2013). Revised standards for statistical evidence. Proc Natl Acad Sci U S A. 2013; 110(48):19313–19317. doi: 10.1073/pnas.1313476110 www.pnas.org/cgi/doi/10.1073/pnas.1313476110
- Lecoutre B. (2005). Et si vous étiez un bayésien qui s'ignore ? Revue MODULAD, numéro 32, 92-105.
- Littell, R. C., Milliken, G. A., Stroup, W. W., Wolfinger, R. D., and Schabenberger, O. 2006. SAS for Mixed Models, 2nd edition. SAS Press, Cary, NC (USA). http://ebooks.cawok.pro/SAS.Publishing.SAS.for.Mixed.Models.2nd.Edition.Mar.2006.pdf
- Madden L. V., Shah D. A. and Esker P. D. (2015). Does the *P* Value Have a Future in Plant Pathology? http://dx.doi.org/10.1094/PHYTO-07-15-0165-LE
- Morey, R., Hoekstra, R., Rouder, J., Lee, M., & Wagenmakers, E.-J. (2015). The fallacy of placing confidence in confidence intervals. Psychonomic Bulletin & Review, 1–21. http://doi.org/10.3758/s13423-015-0947-8
- Sellke, T., Bayarri, M. J., and Berger, J. O. 2001. Calibration of p values for testing precise null hypotheses. Am. Stat. 55: 62-71.
- Tukey J. W. (1991). *The Philosophy of Multiple* Comparisons. Statistical Science, Vol. 6, No. 1, pp. 100-116.



ARVALIS

Institut du végétal