Les tests statistiques

Tests de signification et tests d'hypothèse : utilisation quelque peu délicate de deux notions proches mais différentes.

M. L. Delignette-Muller
VetAgro Sup

21 septembre 2020

Objectifs pédagogiques

- Savoir définir les notions suivantes : test de signification, test d'hypothèse, différence significative, risques d'erreur de première et deuxième espèces, p-value (valeur de p ou degré de signification), puissance.
- Savoir réaliser à la main un test à partir de sa fiche technique.*
- Savoir interpréter le résultat d'un test et notamment avoir les idées claires sur les conclusions qu'on peut tirer d'un test.

* *savoir faire évalué uniquement en S5*

Exemple introductif

On tire au sort aléatoirement n étudiants vétérinaires sur lesquels on estime la fréquence de filles. **A partir des données observées** (sans utiliser de connaissance *a priori*) **peut-on conclure que la fréquence de filles parmi les étudiants vétérinaires est différente de** 50% ?

Imaginons 5 cas :

1. 2 filles sur $n = 2$: 100%
2. 6 filles sur $n = 10$: 60%
3. 15 filles sur $n = 20$: 75%
4. 37 filles sur $n = 50$: 74%
5. 68 filles sur $n = 100$: 68%

Tentez de répondre à la question dans chaque cas et notez votre réponse.

Plan

Le test de signification

Un concept original proposé par Karl Pearson en 1900 puis popularisé dans les années 1920 par Ronald Aylmer Fisher.

R A. Fisher

Définition de l'hypothèse nulle H_0

H_0 **: hypothèse de différence nulle**
Dans notre exemple on compare une fréquence observée
(f fréquence de filles)
à une valeur de référence
($p_0 = 0.5$).
H_0 : hypothèse selon laquelle la proportion de filles parmi les étudiants vétérinaires est de 50%.
Notre objectif va être de voir si les données nous permettent de réfuter cette hypothèse.

Définition de la "p-value" : résultat de la confrontation des données à H_0

Les données sont-elles probables sous H_0 ?

Calcul de la "p-value" (p) (appelé aussi degré de signification ou valeur de p) : **probabilité, si on est sous H_0, d'observer une différence au moins aussi grande que celle observée sur les données.**

Si p **est faible** (en général si $p < 5\%$) **on rejette** H_0 et on en conclut qu'il existe bien une différence, que la **différence est significative**, sous entendu que la différence observée n'est pas uniquement due aux fluctuations d'échantillonnage mais est le reflet d'une différence réelle dans la population.

Utilisation d'une variable de décision pour le calcul de la p-value

Dans notre exemple il s'agit de comparer f la fréquence observée à p_0 la fréquence de référence (ici 50%)
Le **théorème de l'approximation normale**, s'il est applicable (n assez grand) nous dit :

$F \sim N(p_0, \sqrt{\frac{p_0(1-p_0)}{N}})$.

Donc la variable centrée réduite $u = \frac{F-p_0}{\sqrt{\frac{p_0(1-p_0)}{N}}} \sim N(0,1)$

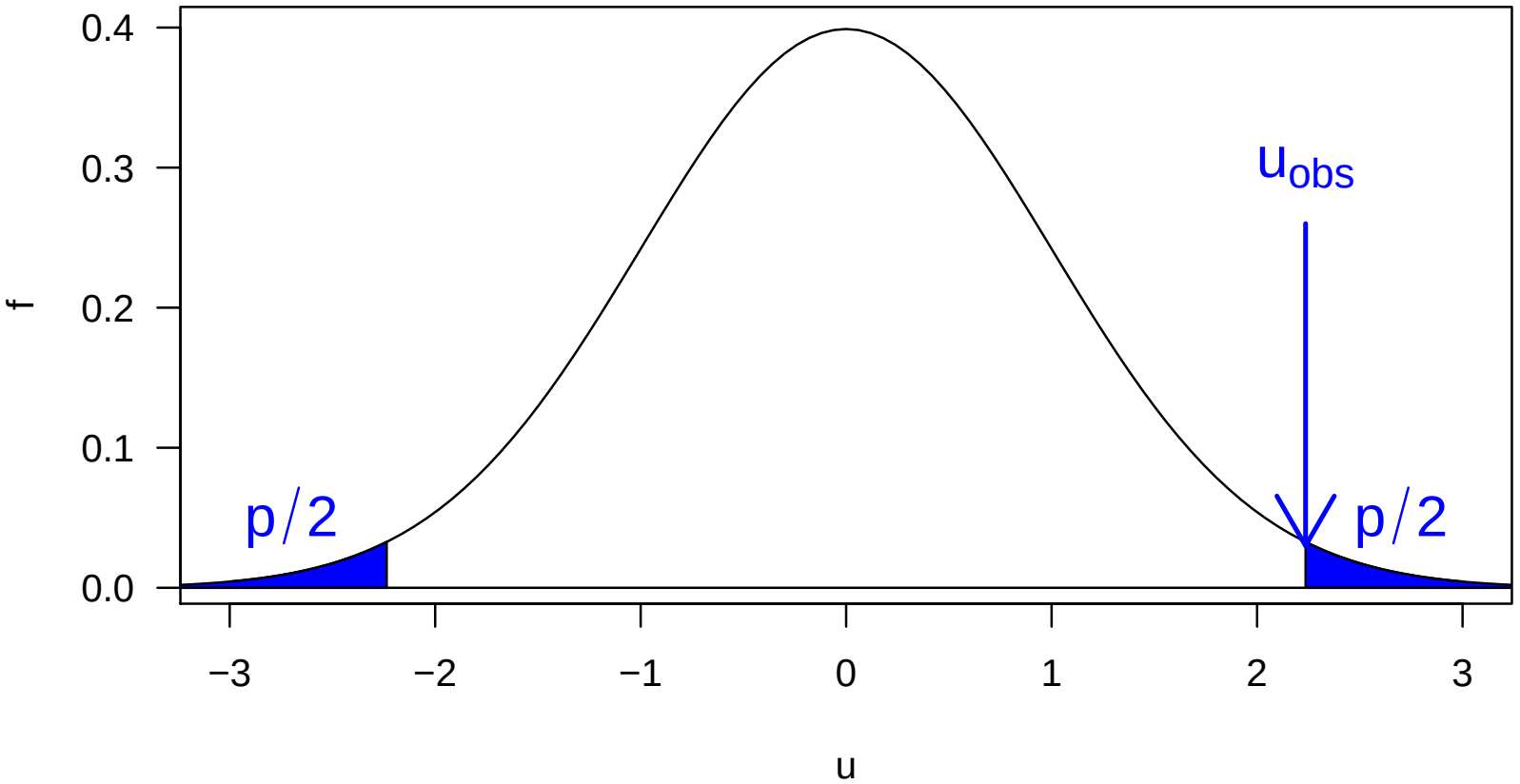
On va utiliser u comme **variable de décision** :

- On **calcule la variable de décision u sur les données observées** : $u_{obs} = \frac{f-p_0}{\sqrt{\frac{p_0(1-p_0)}{N}}}$
- on **confronte u_{obs} à la loi qu'elle est censée suivre sous H_0 pour quantifier la p-value**.

Application sur notre ex. (cas 3 - visualisation de p)

Cas 3 : 15 filles sur $n = 20$ **:** $75\% \rightarrow u_{obs} = 2.24$

p < 0.05 ?

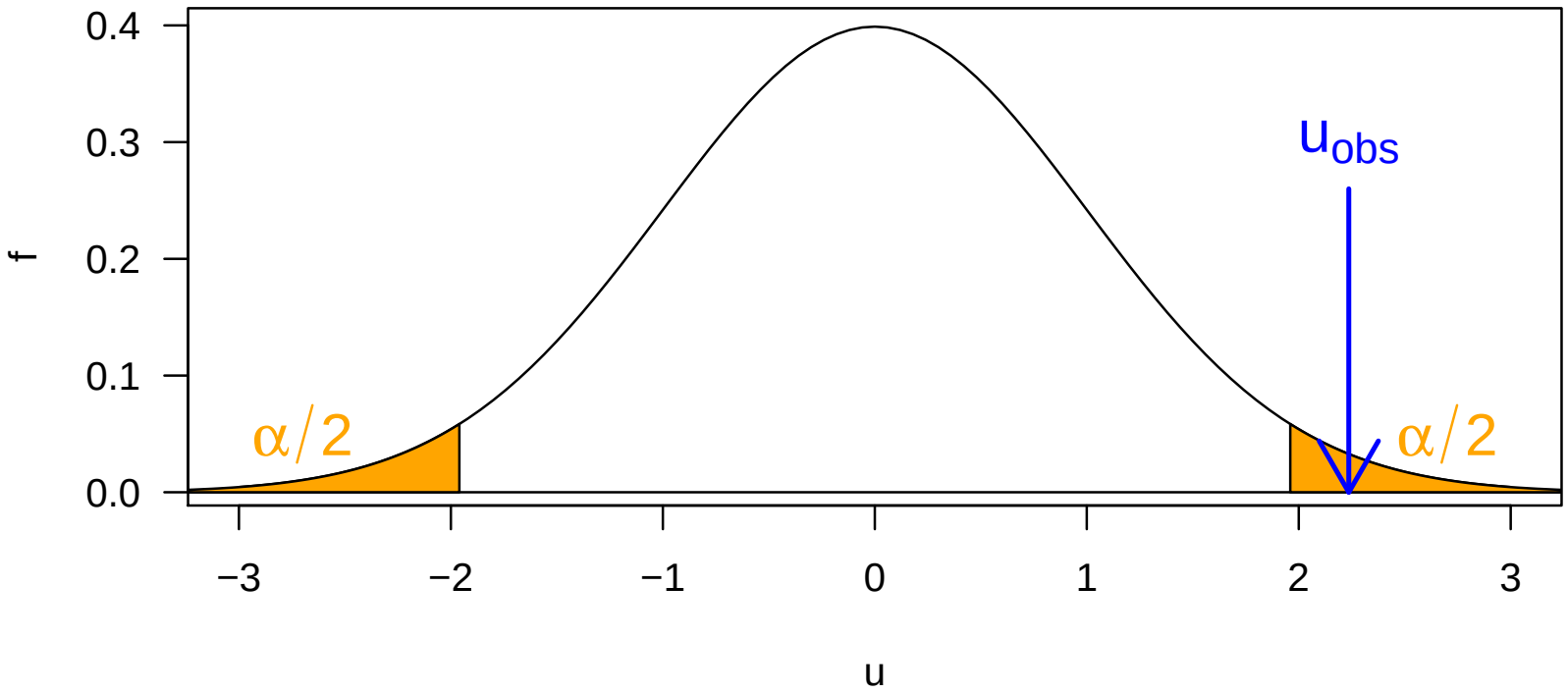


Application sur notre ex. (cas 3 - $p < 0.05$?)

Cas 3 : 15 filles sur $n = 20$ **:** $75\% \rightarrow u_{obs} = 2.24$

Table loi normale : valeurs de u correspondant à $\alpha = 0.05$ $(= 1.96)$

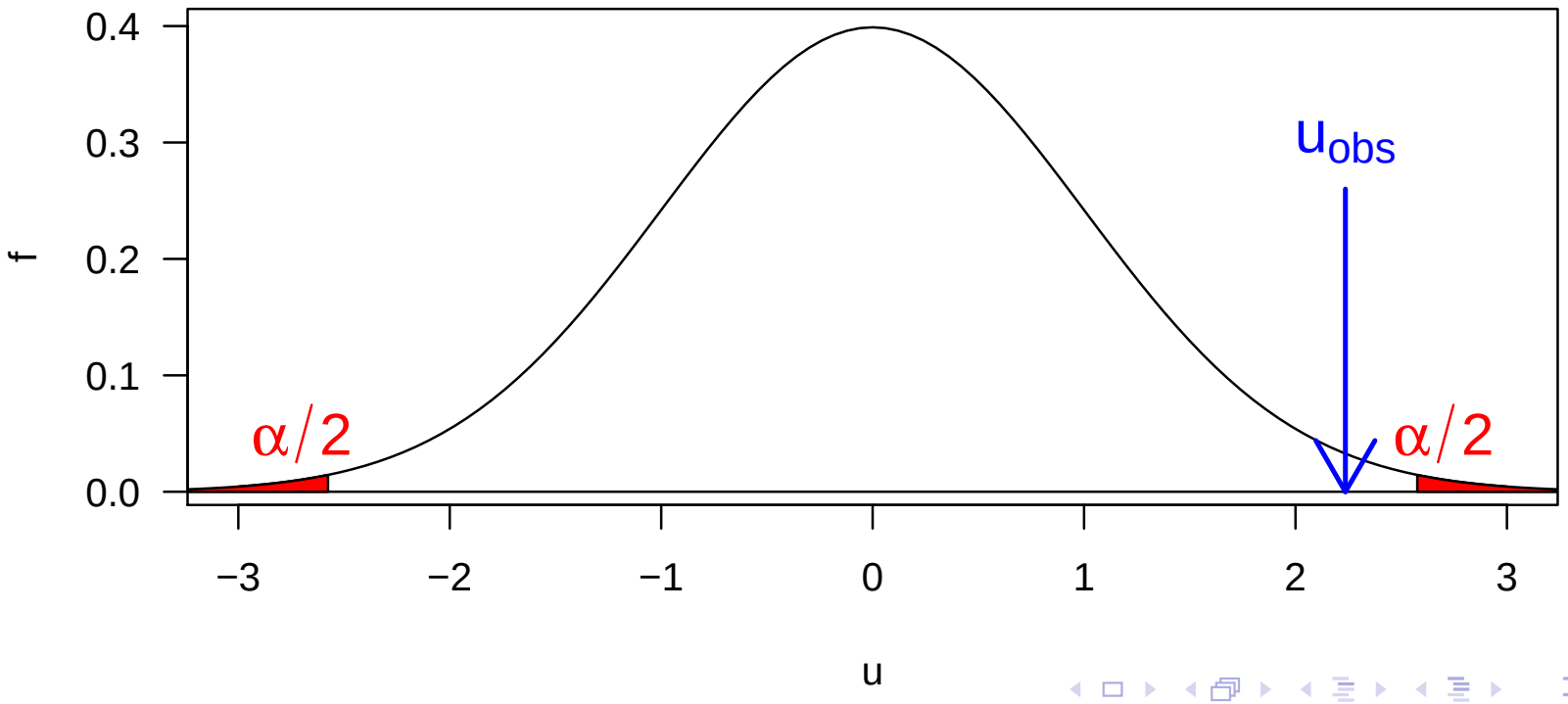
$\rightarrow p < 0.05 \rightarrow$ différence significative



Application sur notre ex. (cas 3 - encadrement de p)

Cas 3 : 15 filles sur $n = 20$ **:** $75\% \rightarrow u_{obs} = 2.24$

Table loi normale : valeurs de u correspondant à $\alpha = 0.01$ $(= 2.576) \rightarrow p > 0.01$ (information que l'on donnera en complément)



Application de la même technique aux autres cas

- **Cas 1 : 2 filles sur $n = 2$: 100%**
  $\rightarrow p > 0.05$ (test adapté aux petits effectifs) $\rightarrow$ non rejet de H_0
- **Cas 2 : 6 filles sur $n = 10$: 60%**
  $\rightarrow p > 0.05$ (test adapté aux petits effectifs) $\rightarrow$ non rejet de H_0
- Cas 3 : 15 filles sur $n = 20$: 75%
  $\rightarrow u_{obs} = 2.24 \rightarrow 0.01 < p < 0.05 \rightarrow$ rejet de H_0
- **Cas 4 : 37 filles sur $n = 50$: 74%**
  $\rightarrow u_{obs} = 3.39 \rightarrow p < 0.001 \rightarrow$ rejet de H_0
- **Cas 5 : 68 filles sur $n = 100$: 68%**
  $\rightarrow u_{obs} = 3.60 \rightarrow p < 0.001 \rightarrow$ rejet de H_0

Peut-on accepter H_0 lorsque p est élevé ?

Citation de R.A. Fisher en 1966
"The null hypothesis is never proved or established, but it is possibly disproved, in the course of experimentation"

Autrement dit **un test de signification peut conduire à rejeter H_0 dans certains cas, mais en aucun cas à l'accepter.**

Seriez-vous raisonnablement tenté d'accepter H_0 dans le cas 1 ?
Non bien sûr !

A RETENIR !

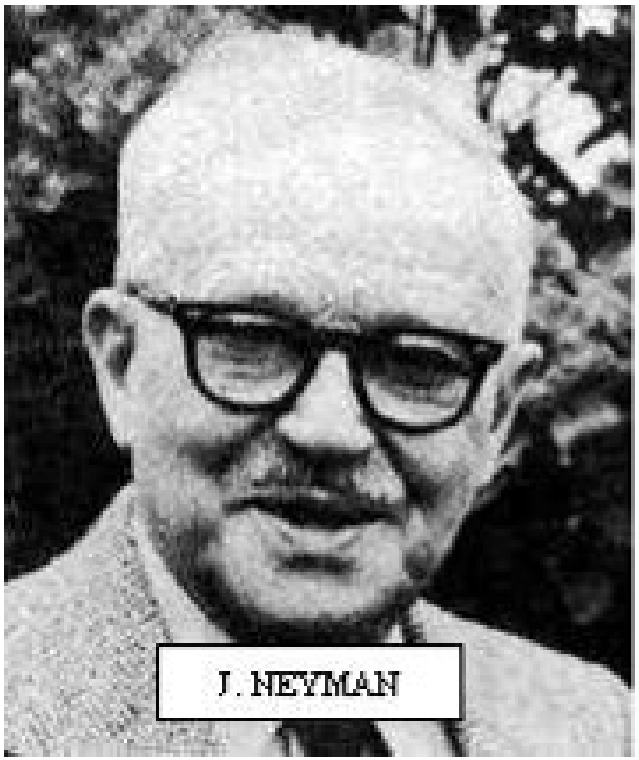
Objectif du test de signification : déterminer si une différence observée est significative (preuve d'une vraie différence et non simple reflet des fluctuations d'échantillonnage)

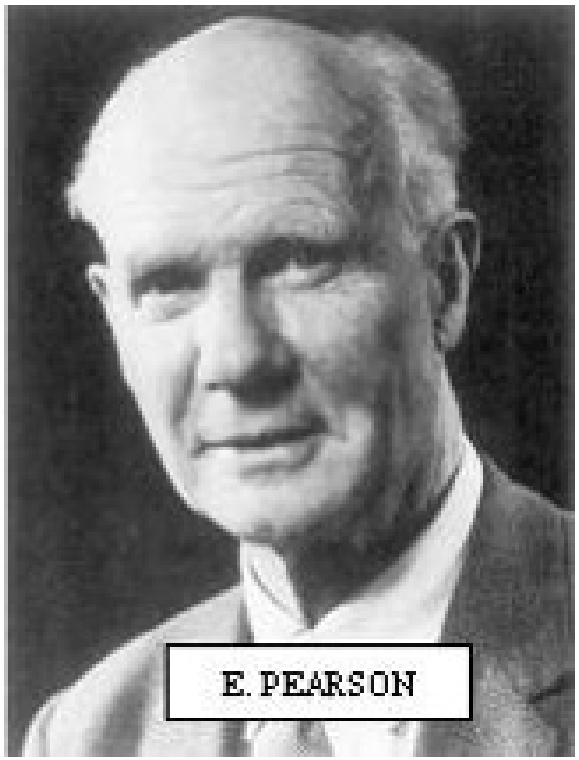
Principe :

- on fait l'hypothèse d'une différence nulle (H_0),
- à l'aide d'une variable de décision on calcule p la probabilité d'observer, sous H_0, une différence au moins aussi grande que celle observée,
- si $p < 0.05$ on rejette H_0 et on dit que la différence est significative,
- plus p est petit, plus on est convaincu qu'on a le droit de rejeter H_0. On ne peut jamais accepter H_0

Le test d'hypothèse

Deuxième vision proposée par Jerzy Neyman et Egon Pearson en 1928 et présentée comme une amélioration du test de signigication.



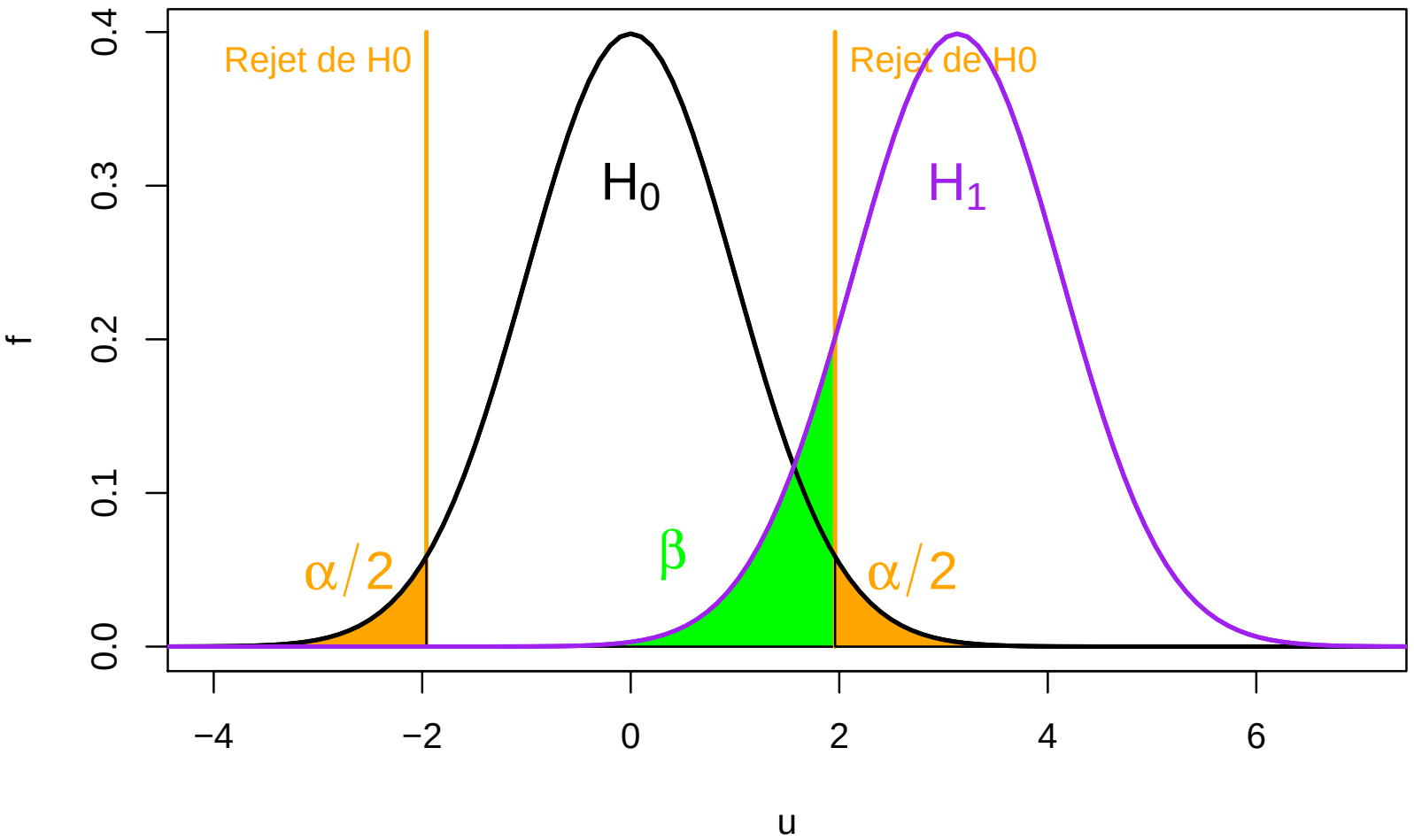


Le test d'hypothèse comme outil décisionnel

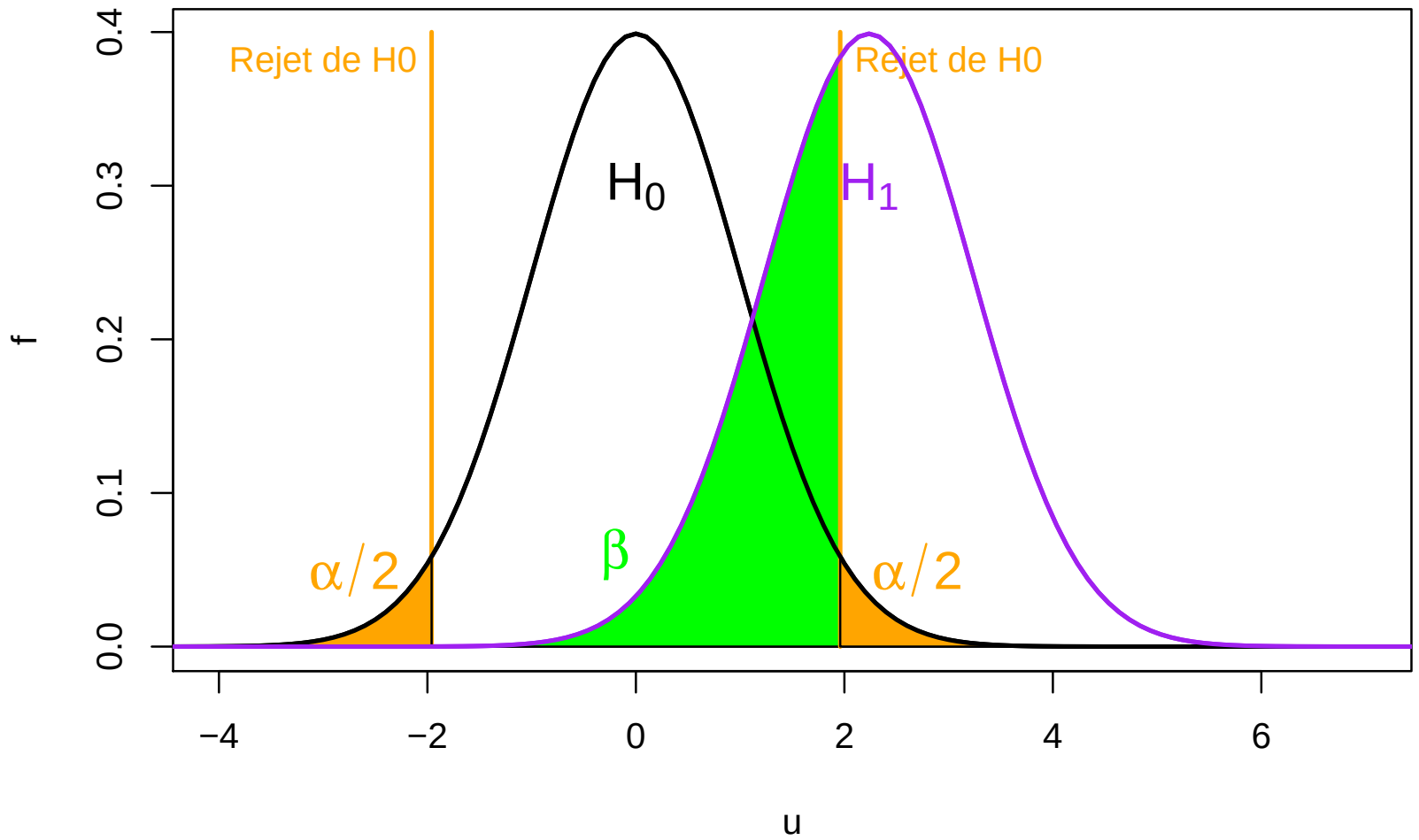
Introduction de la notion d'hypothèse alternative H_1 de différence non nulle.
Utilisation de p pour décider entre H_0 (si $p > 0.05$) et H_1 (si $p < 0.05$).
On a alors deux risques d'erreur :

- **risque de 1ère espèce α maîtrisé** ($\alpha = 0.05$) : risque de se tromper en rejetant H_0
- **risque de 2ème espèce β non maîtrisé** : risque de se tromper en acceptant H_0

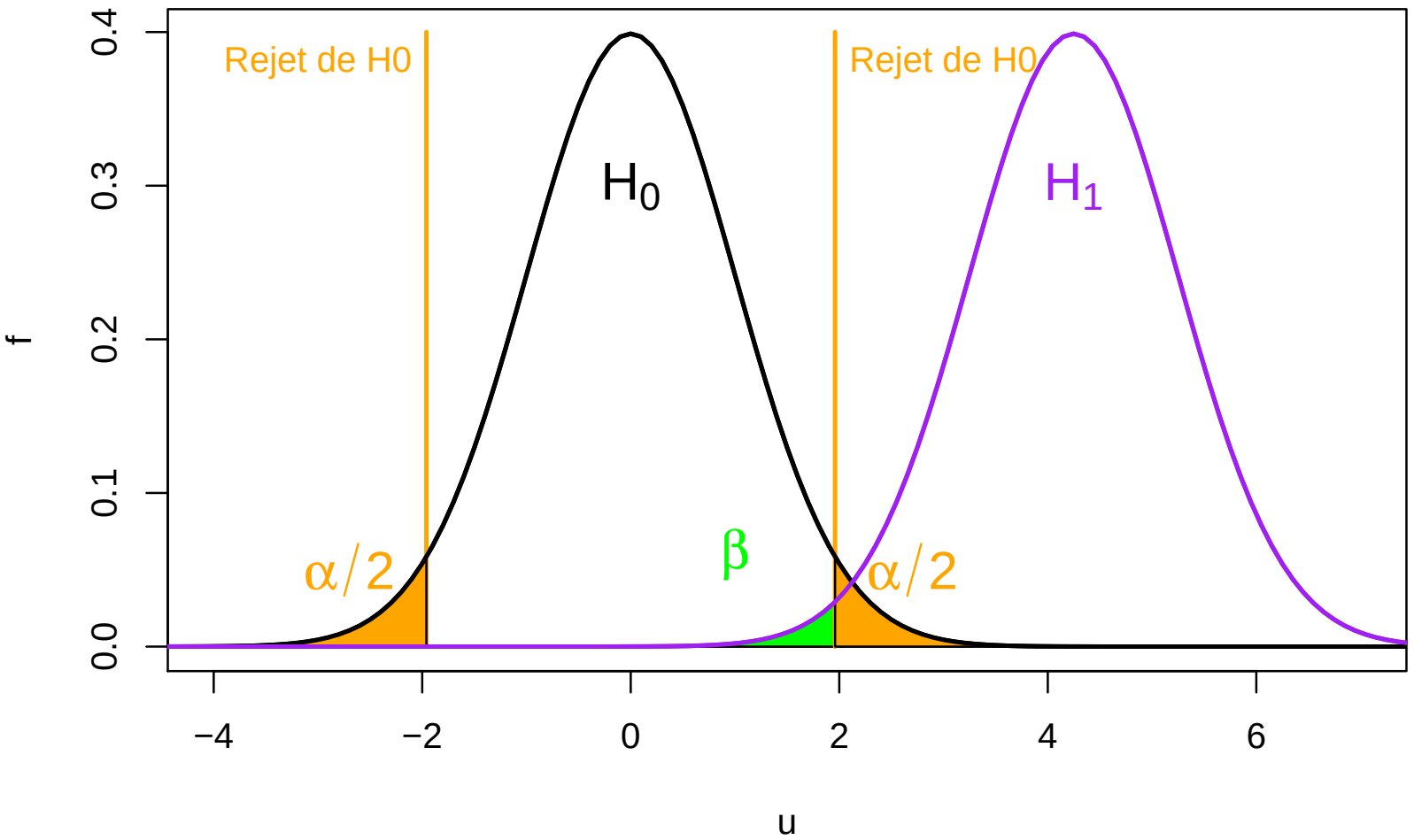
Visualisation du risque β dépendant de H_1 (1)



Visualisation du risque β dépendant de H_1 (2)



Visualisation du risque β dépendant de H_1 (3)



Risque β et puissance $1 - \beta$

On souhaite avoir un **risque** β faible donc une **puissance** $1 - \beta$ forte, mais β peut être élevé (donc puissance faible) du fait :

- d'une faible différence théorique (H_1 proche de H_0),
- d'une grande incertitude sur le paramètre estimé (faible effectif, forte variabilité)

Exemple de **simulations du nombre de rejets de H_0 sur 1000 échantillons** d'étudiants vétérinaires en supposant que la proportion de filles dans cette population est de 70%.

- sur 1000 échantillons de taille 10 : 161 rejets de H_0
- sur 1000 échantillons de taille 20 : 415 rejets de H_0
- sur 1000 échantillons de taille 50 : 784 rejets de H_0
- sur 1000 échantillons de taille 100 : 977 rejets de H_0

Test d'hypothèse à n'utiliser que très prudemment !

On ne peut raisonnablement utiliser un test d'hypothèse **que si la puissance est maîtrisée** donc si un calcul de puissance *a priori* a été réalisé :
calcul d'effectifs nécessaires pour atteindre une puissance donnée, c'est-à-dire une probabilité donnée de détecter une différence dépassant un seuil d'intérêt prédéfini.

Ce qu'en pensait R.A. Fisher :
"Errors of the second kind are committed only by those who misunderstand the nature and the application of tests of significance"

L'utilisation des tests statistiques : un sujet encore brûlant !

D'une ancienne discorde entre Fisher et Neyman et Pearson à un sujet qui fait encore couler beaucoup d'encre.

Quelques références :

- Hubbard R. 2011. The widespread misinterpretation of p-values as error probabilities.
- Goodman S. 2008. A Dirty Dozen : Twelve P-Value Misconceptions.
- Berger J.O. 2003. Could Fisher, Jeffreys and Neyman have agreed on testing ?
- Blume J. et al. 2003. What Your Statistician Never Told You about P-Values ?
- Haller H. et al. 2002. Misinterpretations of significance : A problem students share with their teachers.
- Gardner M.J. et al. 1986. **Confidence intervals rather than P values : Estimation rather than hypothesis testing.**

Un constat actuel qui semble donner raison à R.A. Fisher

L'**amalgame** courant entre les notions de **test de signification** et de **test d'hypothèse** semble à la source d'une mauvaise interprétation fréquente des résultats des tests et d'un malaise persistant à leur sujet.

Une référence consensuelle sur l'usage de la p-value

Taylor & Francis
Taylor & Francis Group

The American Statistician

ISSN: 0003-1305 (Print) 1537-2731 (Online) Journal homepage: http://amstat.tandfonline.com/loi/utas20

The ASA's Statement on p-Values: Context, Process, and Purpose

Ronald L. Wasserstein & Nicole A. Lazar

To cite this article: Ronald L. Wasserstein & Nicole A. Lazar (2016) The ASA's Statement on p-Values: Context, Process, and Purpose, The American Statistician, 70:2, 129-133, DOI: 10.1080/00031305.2016.1154108

"1. P-values can indicate how compatible the data are with a specified statistical model."

- Plus la valeur de p est petite et plus l'incompatibilité statistique entre les données et l'hypothèse nulle est grande.
- On peut voir la **valeur de p comme un indicateur de discordance entre les données et l'hypothèse nulle**.

"2. P-values do not measure the probability that the studied hypothesis is true."

La valeur de p ne doit surtout pas être interprétée comme la probabilité de l'hypothèse nulle connaissant les données, même si cela est très tentant.
On ne peut pas inverser les probabilités aussi facilement !

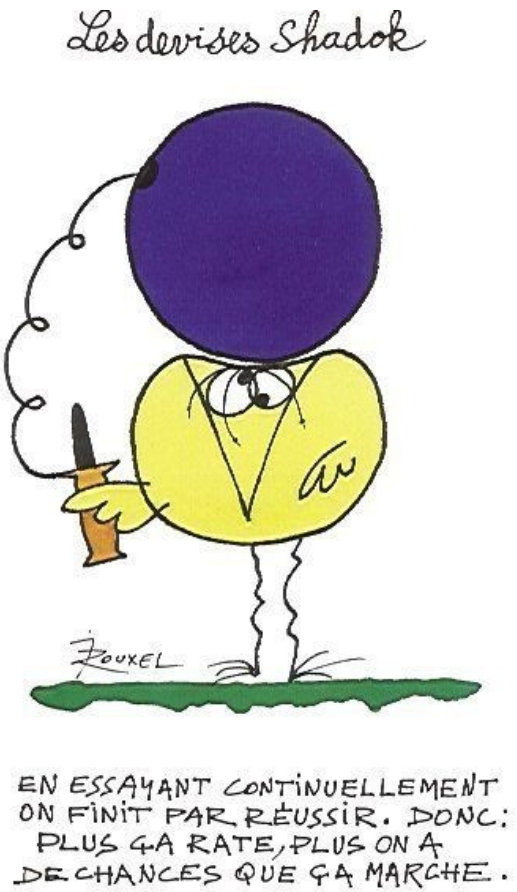
Si un jour vous en êtes tenté pensez au cas 1 de notre exemple (avec deux filles sur un échantillon aléatoire de deux étudiants vétérinaires concluerait-on qu'il y a autant de filles que de garçons parmi les étudiants vétérinaires ?)

"3. Scientific conclusions and decisions should not be based only on whether a p-value passes a specific threshold."

- Actuellement les scientifiques donnent souvent trop de poids à la valeur de p et au résultat du test en terme de différence significative ou non, parfois sans même regarder la différence estimée.
- Il convient plutôt de considérer le **test comme un garde fou, nous empêchant d'interpréter hâtivement une différence qui ne serait pas significative**.

"4. Proper inference requires full reporting and transparency."

- Les résultats de **tous les tests réalisés doivent être reportés**, et non seuls les résultats significatifs.
- En moyenne dans tous les cas où H_0 est vraie, une fois sur 20 on a $p < 0.05$. *A force de chercher on finit par trouver !*



"5. A p-value does not measure the size of an effect or the importance of a result."

- Une valeur de p petite n'implique pas forcément la mise en évidence d'une différence d'intérêt biologique.
- Une différence importante peut ne pas apparaître significative du fait du manque de puissance de l'analyse (par ex. en cas d'effectifs faibles).

Il est capital, lorsque cela est possible, **d'interpréter in fine l'estimation de la différence (estimation ponctuelle et intervalle de confiance)**.

"6. By itself, a p-value dose not provide a good measure of evidence regarding a hypothesis."

Ne jamais utiliser un test d'hypothèse pour montrer pour montrer une hypothèse et en particulier pour montrer une équivalence mais privilégier les tests d'équivalence basés sur les intervalles de confiance dans ce cas.

Principe des tests d'équivalence :

- On définit une zone d'équivalence sur des critères biologiques ("quelle différence maximum sera considérée comme négligeable ?").
- On conclut à l'équivalence si l'intervalle de confiance sur la différence observée est entièrement contenu dans cette zone.