UNIVERSITÉ DE NANTES

Introduction aux tests statistiques

Module HPS3-32
Année 2021-22

GALHARRET J-M,
Laboratoire de Mathématiques Jean Leray
Faculté de Psychologie.

Rappels de L1

Type de variables

Qualitatives ou **Nominales** ou **Catégorielles**

Exemples :
Sexe (H/F)
Groupe (Traité/Non traité)
CSP (…)

Les valeurs sont appelées **Modalités**

Quantitatives ou **Numériques** ou **Mesurées**

Exemples :
Taille (entre 100cm et 250cm)
Score d'Anxiété (Entre 0 et 20)
Âge (Entre 18 et 100)

Exemple

Référence : JASP (Data Library)

Description:

This data set, "Directed Reading Activities", provides reading performance of two groups of pupils - one control group and one group that was given Directed Reading Activities (Moore et al, 2012, p. 432).

Variables:

- **id** - Identification number of a pupil.
- **group** - Experimental group indicator ('Treatment' = participation in the Directed Reading Activities, 'Control' = Control group).
- **g** - Experimental group indicator expressed as a binary variable (0= Directed Reading Activities, 1= Control group).
- **drp** - The performance on Degree of Reading Power test.

In this example JASP file, we will compare two classrooms of students with the aim of testing the null hypothesis that Directed Reading Activities (only in one of the classes) do not enhance the performance of pupils on Degree of Reading Power test (DRP).

id	group	g	drp
1	Treat	0	24
2	Treat	0	56
3	Treat	0	43
4	Treat	0	59
5	Treat	0	58
6	Treat	0	52
7	Treat	0	71
8	Treat	0	62
9	Treat	0	43
10	Treat	0	54
11	Treat	0	49
12	Treat	0	57
13	Treat	0	61
14	Treat	0	33
15	Treat	0	44
16	Treat	0	46
17	Treat	0	67
18	Treat	0	43
19	Treat	0	49
20	Treat	0	57
21	Treat	0	53
22	Control	1	42
23	Control	1	46
24	Control	1	43
25	Control	1	10
26	Control	1	55

Résumé d'une variable

Variable nominale :

Frequencies for group		
group	**Frequency**	**Percent**
Control	23	52.27
Treat	21	47.73
Missing	0	0.00
Total	44	100.00

- Effectif (**Frequency**) : nombre d'élèves dans chaque groupe.
- Pourcentage (**Percent**) : % d'élèves dans chaque groupe.

Variable numérique :

Descriptive Statistics	
	drp
Valid	44
Missing	0
Mean (Moyenne)	46.273
Median (Médiane)	47.000
Standard Deviation (Ecart type)	15.235
IQR (Intervalle Inter-Quartiles)	16.250
Minimum	10.000
Maximum	85.000

- La moyenne et la médiane sont des indicateurs de **position.**
- L'écart type et l'IQR sont des indicateurs de **dispersion.**

Le problème des expérimentateurs

(Significativité statistique)

	group	drp
N	Treat	21
	Control	23
Missing	Treat	0
	Control	0
Mean	Treat	51.5
	Control	41.5
Median	Treat	53.0
	Control	42.0
Standard deviation	Treat	11.0
	Control	17.1
Minimum	Treat	24.0
	Control	10.0
Maximum	Treat	71.0
	Control	85.0

Le traitement peut-il être considéré comme efficace ?

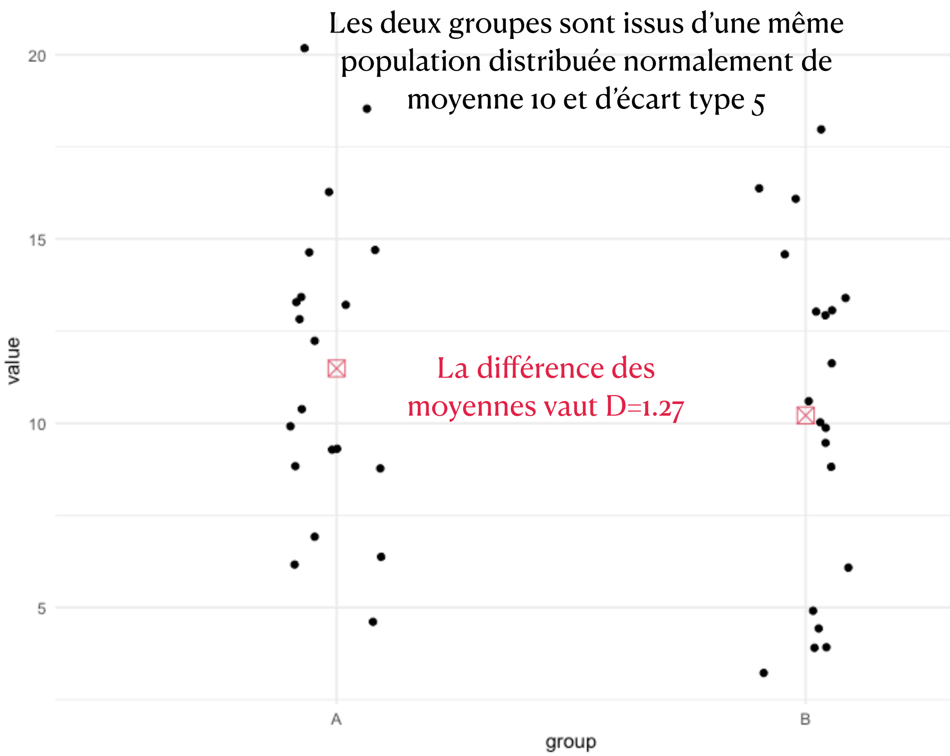
- Sur l'échantillon considéré on a :
  - Le groupe contrôle (n=23) obtient un drp moyen de 41.5.
  - Le groupe traité (n=21) obtient un drp moyen de 51.5.
- Ainsi, le groupe traité obtient des résultats en moyenne 10 points supérieurs à ceux du groupe contrôle.

Question :

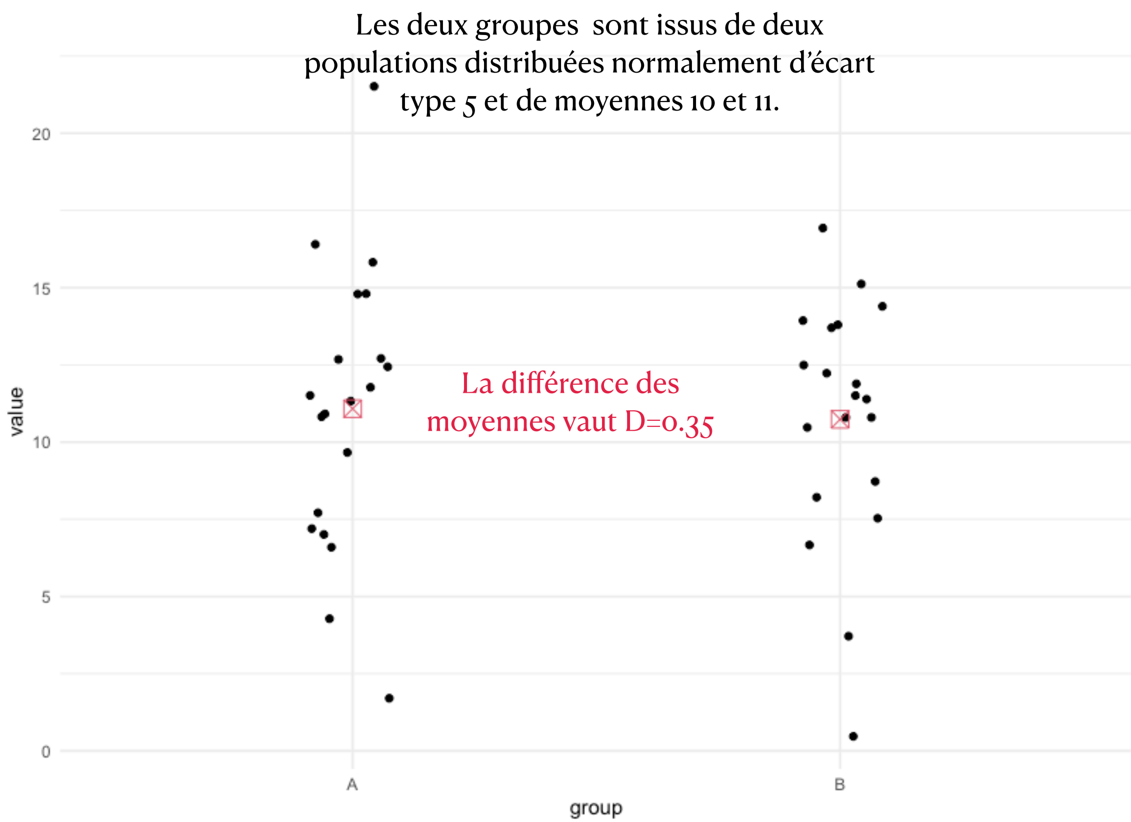
- Cette différence de 10 points est-elle due aux fluctuations d'échantillonnage (c'est à dire au hasard) ?
- Ou bien est-elle due au fait que le traitement a vraiment eu un impact sur les capacités de lecture des enfants ? (**significativité statistique**) et ainsi cette différence ne peut donc pas être due au simple hasard.

Echantillonage

Comparaison de deux groupes sur une réalisation (n petit)



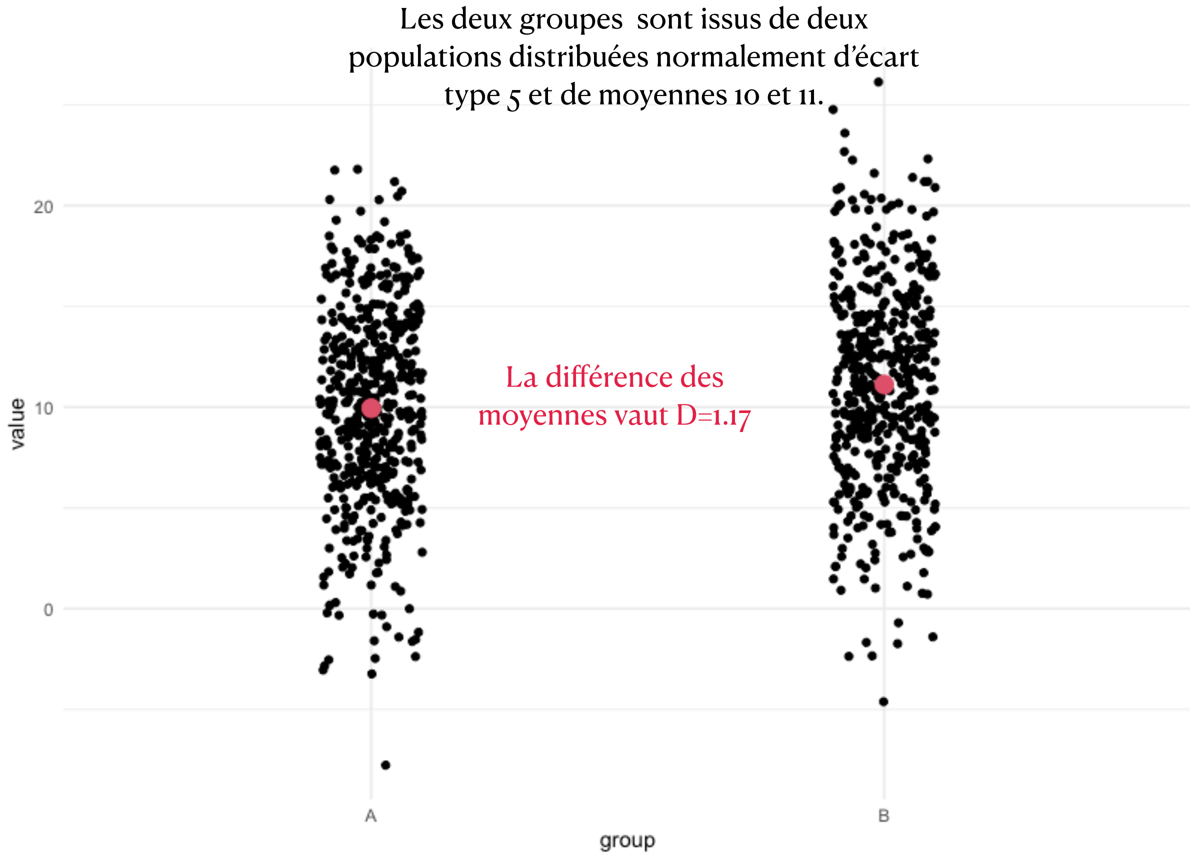
n=20



n=20

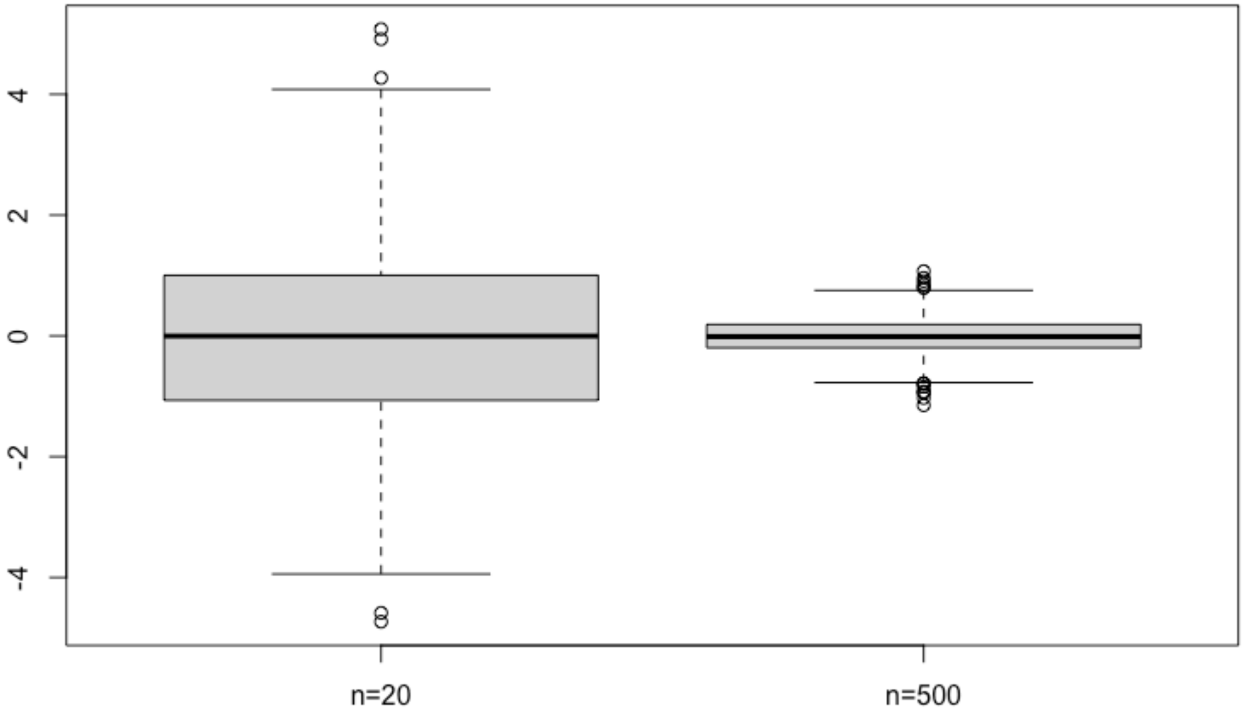
Comparaison de deux groupes sur une réalisation (n grand)

n=500



n=500

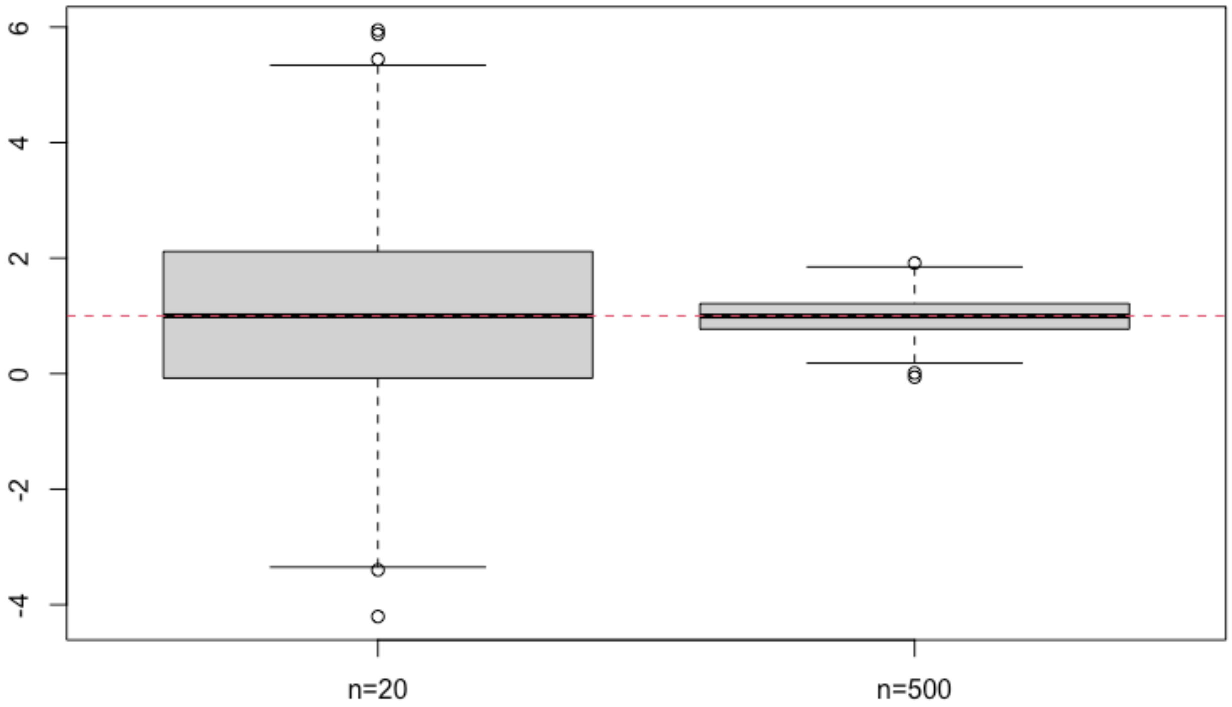
Réplication d'une expérience



Dans ce premier exemple on tire aléatoirement B=1000 fois deux groupes de même taille n dans une même population distribuée normalement de moyenne 10 et d'écart type 5.
On calcule à chaque fois la différence des deux moyennes et on dessine les boxplot correspondantes.
Lorsque n=20, la moyenne des 1000 différences vaut 0.07 et son écart type vaut 1.53.
Lorsque n=500, la moyenne des 1000 différences vaut 0.01 et son écart type vaut 0.3

Réplication d'une expérience

Cas où les groupes ne sont pas issus de la même population



Dans ce premier exemple on tire aléatoirement B=1000 fois deux groupes de même taille n dans deux populations différentes distribuées normalement d'écart type 5 et de moyenne 10 et 11.
Lorsque n=20, la moyenne des 1000 différences vaut 1.02 et son écart type vaut 1.57.
Lorsque n=500, la moyenne des 1000 différences vaut 1.00 et son écart type vaut 0.3.

Procédure de test

Les hypothèses de test

Traduction de la significativité statistique

- Le test d'hypothèse va confronter des hypothèses contraires l'une de l'autre.
  - L'hypothèse nulle (Ho) qui traduira le fait que le phénomène observé n'est dû qu'au hasard.
  - L'hypothèse alternative (H1) qui traduira **la significativité statistique,** c'est à dire que le phénomène observé ne peut pas être dû qu'au seul hasard.

Remarques :

- vues les définitions de Ho et H1 la procédure de test ne permettra que de rejeter ou de pas rejeter Ho. On ne peut pas prouver Ho !
- L'expérimentateur est intéressé par le fait de prouver H1.

La p-value

Le critère de rejet de H0

On teste : $H_0 : \mu_{control} = \mu_{treat}$ versus $H_1 : \mu_{control} \neq \mu_{treat}$

- On définit une version « centrée réduite » de la différence des moyennes, que l'on note *T*.
- On connait la loi de probabilité de *T* sous Ho.
- On calcule sa valeur *t*=2.267 sur les observations considérées (n=44).
- On calcule **la probabilité que, lorsque Ho est vraie, la différence attendue *T* soit aussi extrême que celle observée *t*.** Cette probabilité est la **p-value.**
- Dans l'exemple précédent, p-value=0.029. On interprète le résultat : si la différence n'est due qu'au hasard (Ho), il y a 2.9% de chance d'observer une différence aussi extrême que D=10 (différence observée entre les moyennes des 2 groupes).

Les deux types d'erreurs

Erreurs de première et deuxième espèce

Réalité / Décision du test	**H0 vraie (H1 fausse)**	**H0 fausse (H1 vraie)**
Rejet de H0 (H1 validée)	Erreur de 1ère espèce α	Puissance du test $1-\beta$
Non Rejet de H0 (H1 non validée)	Niveau de confiance $1-\alpha$	Erreur de 2ème espèce β

- L'erreur de 1ère espèce (respectivement de 2ème) peut aussi être vu comme la proportion de faux positifs (respectivement de faux négatifs)
- On ne peut fixer que l'une des deux erreurs et on choisit de fixer l'erreur de 1ère espèce.
- En général on fixe :

$$\alpha = 0.05$$

Il s'agit du niveau de signiﬁcativité d'un test.

Règle de décision

On considère une hypothèse nulle H0 et son alternative H1. On se fixe un niveau de significativité $\alpha \in]0,1[$. On suppose que :

- Une procédure de test est disponible c'est à dire qu'on a défini une variable (notée S) dont on connait la loi de probabilité sous H0.
- On a prélevé un échantillon d'observations et calculé la valeur s de S sur cet échantillon.
- On a calculé $p = \mathbb{P}(S \geq s \mid H_0$ est vraie) (p est la p-value du test).

Alors au risque $\alpha \in]0,1[$,

➡On peut rejeter H0 lorsque $p < \alpha$, c'est à dire que le phénomène est statistiquement significatif au risque α.

➡On ne peut pas rejeter H0 lorsque $p \geq \alpha$ c'est à dire que l'on a pas trouvé de preuves qui permettent d'affirmer que le phénomène est attribuable à d'autres choses que le hasard.

Retour sur l'exemple

Interprétation des résultats

- Rappel : t=2.267 et p=.029 sur les observations considérées (n=44). On a p=.029, donc au risque de 5% on peut affirmer que :
  - la différence de D=10 entre le groupe traité et le groupe contrôle n'est pas uniquement due au hasard.
  - Les résultats en lecture ont un lien significatif avec le fait que les enfants se sont exercés.

Interprétation de la p-value

On rappelle qu'on a trouvé $p=.029$ dans l'exemple précédent.

- Attention aux mauvaises interprétations de la p-value :
  - Il est faux de dire que : « il y a $p=2.9\%$ de chance que Ho soit vraie. »
  - Il est faux de dire que : « il y a $1-p=97.1\%$ de chance que H1 soit vraie. »
- On peut dire : si Ho est vraie alors il y a 2.9% d'observer sur une nouvelle expérience une différence entre les deux moyennes qui sera au moins aussi grande que $D=10$.

Puissance du test

Généralités

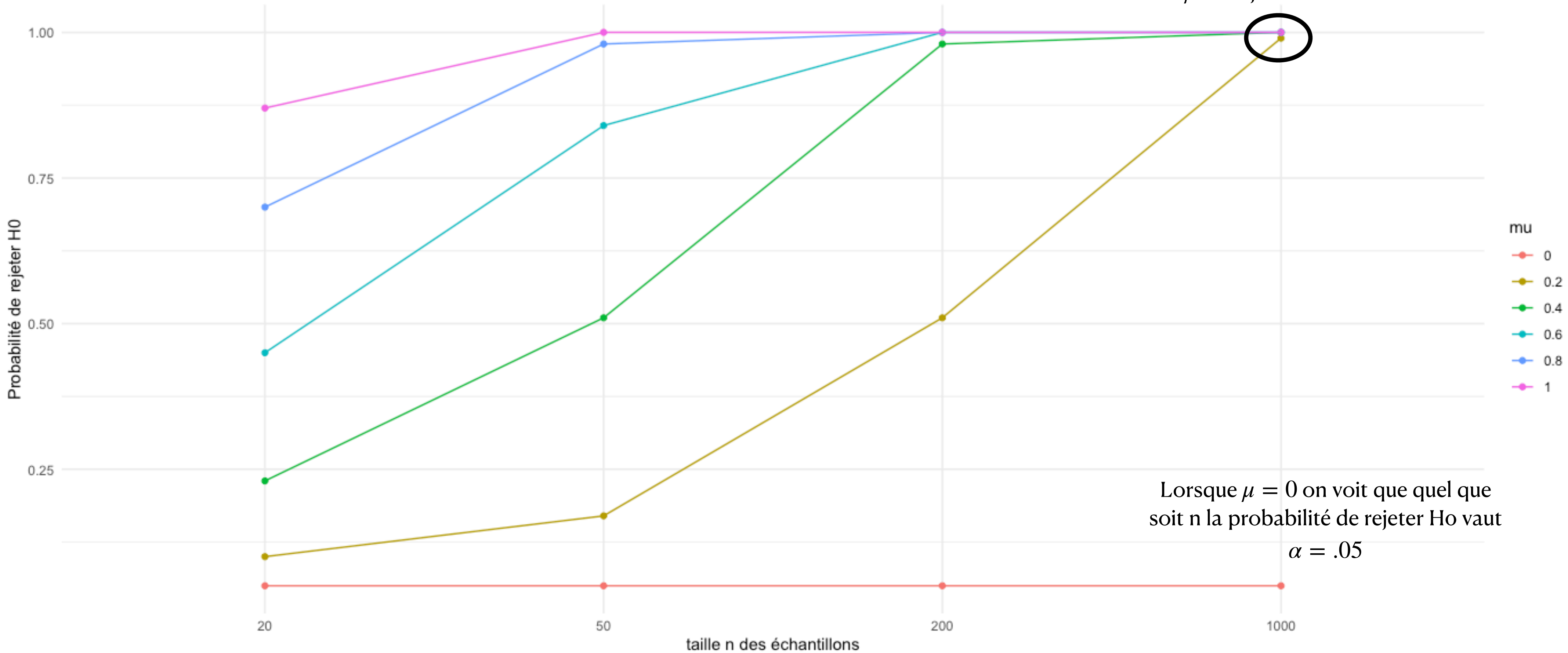
- La puissance du test est sa capacité à rejeter Ho lorsque celle-ci est fausse, autrement dit, c'est la capacité du test à identifier parmi des phénomènes ceux qui ne sont pas dus au hasard.
- On sait simuler des échantillons (méthode de Monte-Carlo) ce qui nous permet de calculer la puissance empirique des tests.
- Exemple : on considère une population normale $X \sim \mathcal{N}(\mu, 1)$ et on veut tester

$$H_0 : \mu = 0 \text{ versus } H_1 : \mu \neq 0.$$

  - On fixe une valeur de μ (entre 0 et 1 par exemple)
  - On prélève aléatoirement 100 000 échantillons de taille n dans la population et pour chaque échantillon on calcule la p-value (avec une procédure de test étudiée par la suite).
  - On calcule la proportion P d'échantillon dont la p-value est inférieure à $\alpha = 5\,\%$. P est une estimation de la puissance du test

Graphe de puissance

Lorsque $n = 1000$ on voit que quel que soit μ on rejette Ho dans 100% des cas.



Lorsque $\mu = 0$ on voit que quel que soit n la probabilité de rejeter Ho vaut $\alpha = .05$

Résumé du graphique

Propriété de la puissance statistique

- Plus μ augmente, plus la puissance du test augmente pour n fixé. Plus la différence réelle est grande plus on va rejeter Ho (rien d'étonnant !)
- De même pour une différence $\mu \neq 0$ fixée la puissance du test augmente lorsque n augmente.
- Lorsque n=1000, quelle que soit la différence réelle observée μ on va toujours rejeter Ho (P=1).

Cette dernière propriété caractérise les procédures de tests que nous verrons par la suite :

Lorsque n est suffisamment grand, toute différence $\mu \neq 0$ (aussi petite soit-elle) est statistiquement significative.