Power and more | *Puissance statistique et au-delà*

Macartan, Vin

2026-06-12

Section 1

Outline

Overview
Vue d'ensemble

1. Tests review
2. p values and significance
3. Power
4. Sources of power

1. Rappel sur les tests
2. p-values et significativité
3. Puissance
4. Sources de puissance

Section 2

Tests

Classical testing
Tests classiques

In the classical approach we ask:
How likely are we to see data like this if the hypothesis is true?

- If the answer is "not very likely", we treat the hypothesis as suspect.
- Otherwise we **maintain** it (not quite "accept" — we may maintain incompatible hypotheses).

How unlikely is "not very likely"?

Dans l'approche classique, on se demande :
Quelle est la probabilité d'observer de telles données si l'hypothèse est vraie ?

- Si la réponse est « peu probable », l'hypothèse devient suspecte.
- Sinon, on la **maintient** (pas vraiment « accepter » — on peut maintenir des hypothèses incompatibles).

Qu'est-ce que « peu probable » ?

Weighing evidence
Peser les preuves

Before testing, decide what evidence would convince you the hypothesis is unreliable.
Othello believes Desdemona is innocent. **Iago** offers evidence:

- Would she look at Cassio like that if innocent?
- Would she defend him like that?
- Would Cassio have her handkerchief?

Othello: the chance of all this if she were innocent is surely below 5%.

Avant de tester, fixez quelles preuves vous feraient douter de l'hypothèse.
Othello croit Desdemona innocente. **Iago** avance :

- La regarderait-elle ainsi si elle était innocente ?
- Le défendrait-elle ainsi ?
- Cassio aurait-il son mouchoir ?

Othello : la probabilité de tout cela si elle était innocente est sûrement inférieure à 5 %.

Section 3

Power

What is power?
Qu'est-ce que la puissance ?

Power is the probability of **rejecting** a hypothesis.
It presupposes:

- a well-defined hypothesis
- a stipulation of how the world actually works
- a procedure for producing results and deciding significance

La **puissance** est la probabilité de **rejeter** une hypothèse.
Elle présuppose :

- une hypothèse bien définie
- une description du monde tel qu'on le croit vrai
- une procédure pour produire des résultats et décider de la significativité

Dice: one roll

Dé : un lancer

Hypothesis: a six **never** comes up on this die.
Test:

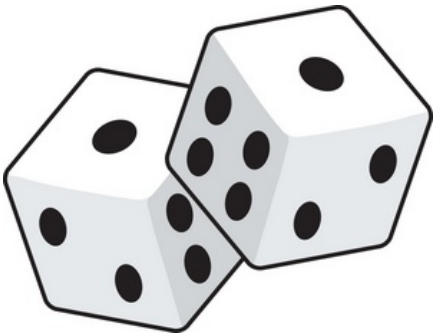
- Roll **once**.
- Reject if a six appears.

What is the power of this test?

Hypothèse : un six **ne sort jamais** sur ce dé.
Test :

- Lancer **une fois**.
- Rejeter si un six apparaît.

Quelle est la puissance de ce test ?

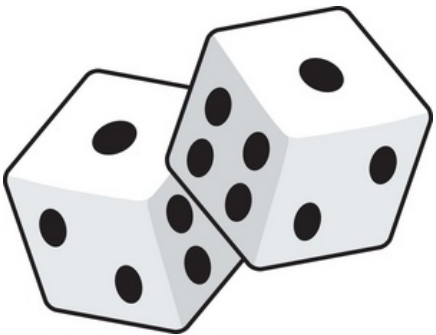


Dice: two rolls

Dé : deux lancers

Same hypothesis — but roll **twice**.
Reject if a six appears **either time**.
What is the power of **this** test?

Même hypothèse — mais lancer **deux fois**.
Rejeter si un six apparaît **l'une ou l'autre fois**.
Quelle est la puissance de **ce** test ?



Two probabilities
Deux probabilités

Power can seem harder because rejection involves a **calculated** probability — a probability of a probability.
Hypothesis: the die is **fair**.

- Roll **1000** times.
- Reject if fewer than x sixes or more than y sixes.

What should x and y be?

La puissance peut sembler plus difficile car le rejet implique une probabilité **calculée** — une probabilité de probabilité.
Hypothèse : le dé est **équitable**.

- Lancer **1000** fois.
- Rejeter si moins de x six ou plus de y six.

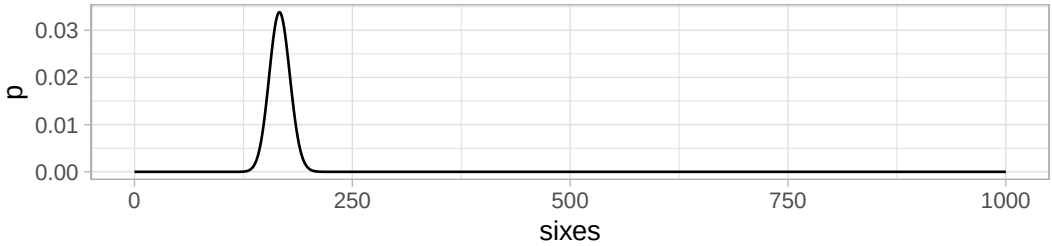
Que valent x et y ?

Step 1: rejection rule

Étape 1 : règle de rejet

- Set a rejection rule based on events unlikely under the hypothesis.
- Expected number of sixes if the die is fair is 166. What would be 'unusually many' 6s? What would be 'unusually few' 6s?

- Fixer une règle de rejet à partir d'événements improbables sous l'hypothèse.
- Nombre attendu de six si le dé est équitable :



Rejection thresholds
Seuils de rejet

143 or fewer is very few; 190 or more is very many:

143 ou moins, c'est très peu ; 190 ou plus, c'est très nombreux :

```
c(lower = pbinom(143, 1000, 1 / 6), upper = 1 - pbinom(189, 1000, 1 / 6))
```

```
     lower      upper
0.02302647 0.02785689
```

Step 2: power
Étape 2 : puissance

What is the power?
Now stipulate how the world **really** works — not the null you plan to reject.
Suppose sixes actually appear **20%** of the time.
What is the probability of seeing at least 190 sixes?

Quelle est la puissance ?
Maintenant, décrire comment le monde **fonctionne vraiment** — pas l'hypothèse nulle que l'on compte rejeter.
Supposons que les six sortent **20 %** du temps.
Quelle est la probabilité d'observer au moins 190 six ?

```
1 - pbinom(189, 1000, .2)
```

```
[1] 0.796066
```

If sixes appear 20% of the time, you will likely see 190 sixes and reject "fair die."

Si les six sortent 20 % du temps, vous verrez probablement 190 six et rejeterez « dé équitable ».

Section 4

Interpretations

Rule of thumb

Règle empirique

- **80%** or **90%** is a common rule of thumb for "sufficient" power.
- How much you need depends on the purpose.

- **80 %** ou **90 %** est une règle empirique courante pour une puissance « suffisante ».
- Le besoin réel dépend de l'objectif.

Think about
À réfléchir

- Are there other tests you could implement?
- Are there other ways to improve this test?

- D'autres tests seraient-ils possibles ?
- D'autres façons d'améliorer ce test ?

Section 5

Power analytics

Back to experiments

Retour aux expériences

- If we run an experiment we get an estimate b
- We form beliefs about a distribution of estimates b (sampling distribution)
- We also can think about the distribution of estimates we might get if there were no effect (the null distribution)
- Just like with the dice!

- Si nous menons une expérience, nous obtenons une estimation b
- Nous formons des croyances sur une distribution des estimations b (distribution d'échantillonnage)
- Nous pouvons aussi penser à la distribution des estimations que nous pourrions obtenir s'il n'y avait pas d'effet (la distribution nulle)
- Comme avec le dé !

Simple intuition
Intuition simple

If the sampling distribution is centered on b with standard error 1, what is the probability of a significant estimate?

- Below -1.96: $F(-1.96 \mid \tau)$
- Above $+1.96$: $1 - F(1.96 \mid \tau)$

Add them: probability above 1.96 or below -1.96.

Si la distribution d'échantillonnage est centrée sur b avec erreur-type 1, quelle est la probabilité d'une estimation significative ?

- En dessous de -1.96 : $F(-1.96 \mid \tau)$
- Au-dessus de $+1.96$: $1 - F(1.96 \mid \tau)$

Sommez : probabilité au-dessus de 1,96 ou en dessous de -1.96.

Power function

Fonction de puissance

```r
power <- function(b, alpha = 0.05, critical = qnorm(1 - alpha / 2)) {
  1 - pnorm(critical, mean = abs(b)) + pnorm(-critical, mean = abs(b))
}
```

```r
power(0)
```

```
[1] 0.05
```

```r
power(1.96)
```

```
[1] 0.5000586
```

```r
power(-1.96)
```

```
[1] 0.5000586
```

```r
power(3)
```

```
[1] 0.8508388
```

Graph: effect 1.96

Graphique : effet 1,96

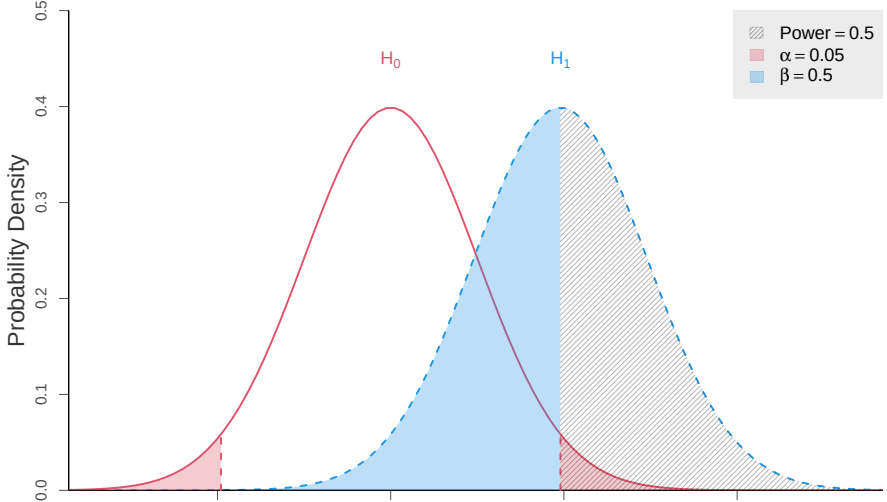
This is what `pwrss::power.z.test` does — with nice graphs.
If the true effect were 1.96, what would power be?

C'est ce que fait `pwrss::power.z.test` — avec de jolis graphiques.
Si l'effet vrai valait 1,96, quelle serait la puissance ?

```
pwrss::power.z.test(
  mean = 1.96, alpha = 0.05,
  alternative = "two.sided", plot = TRUE, verbose = FALSE
)
```

Graph: effect 1.96 (plot)

Graphique : effet 1,96 (graphique)



Graph: interpretation

Graphique : interprétation

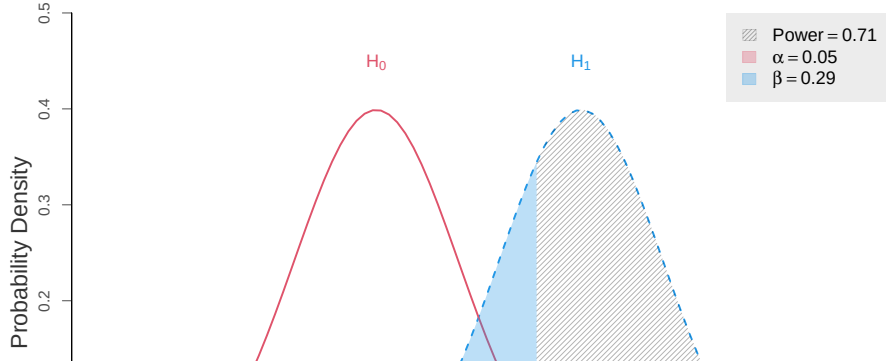
Substantively: if an estimate is **just** significant on average, power is **50%**.

Substantiellement : si une estimation est **à peine** significative en moyenne, la puissance est de **50 %**.

Graph: effect 2.5

Graphique : effet 2,5

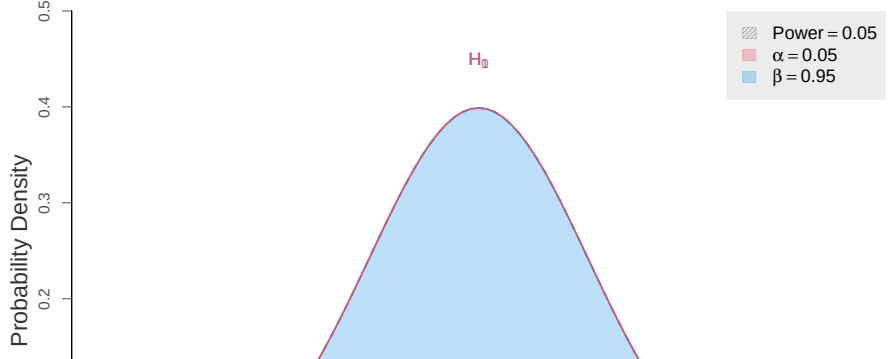
```
pwrss::power.z.test(
  mean = 2.5, alpha = 0.05,
  alternative = "two.sided", plot = TRUE, verbose = FALSE
)
```



Graph: effect 0

Graphique : effet 0

```
pwrss::power.z.test(
  mean = 0, alpha = 0.05,
  alternative = "two.sided", plot = TRUE, verbose = FALSE
)
```



Section 6

Math and functions

Trial: standard error

Essai : erreur-type

Standard error depends on N and outcome variance σ.
With N subjects in two equal groups:

$$Var(\tau) = \frac{\sigma^2}{N/2} + \frac{\sigma^2}{N/2} = 4\frac{\sigma^2}{N}$$

$$\sigma_\tau = \frac{2\sigma}{\sqrt{N}}$$

(Tests use an **estimated** standard error.)

L'erreur-type dépend de N et de la variance σ.
Avec N sujets en deux groupes égaux :

$$Var(\tau) = \frac{\sigma^2}{N/2} + \frac{\sigma^2}{N/2} = 4\frac{\sigma^2}{N}$$

$$\sigma_\tau = \frac{2\sigma}{\sqrt{N}}$$

(Les tests utilisent une erreur-type **estimée**.)

Trial: with pwrss

Essai : avec pwrss

Same idea with `pwrss`:

Même idée avec `pwrss` :

```
pwrss::pwrss.t.2means(
  mu1 = .2, mu2 = .1, sd1 = 1, sd2 = 1, n2 = 50, alpha = 0.05,
  alternative = "not equal", verbose = FALSE
)
```

Mostly: figure out the standard error.
Cluster shock ϵ_k, individual shock ν_i with variances σ_k^2, σ_i^2:

En pratique : calculer l'erreur-type.
Choc de grappe ϵ_k, choc individuel ν_i avec variances σ_k^2, σ_i^2 :

$$\sqrt{\frac{4\sigma_k^2}{K} + \frac{4\sigma_i^2}{nK}}$$

Design effect
Effet de conception

With $\rho = \frac{\sigma_k^2}{\sigma_k^2+\sigma_i^2}$:

$$\sqrt{((n-1)\rho+1)\frac{4\sigma^2}{nK}}$$

- $((n-1)\rho+1)$ = **design effect**
- $\frac{nK}{((n-1)\rho+1)}$ = **effective sample size**

Plug in and proceed as before.

Avec $\rho = \frac{\sigma_k^2}{\sigma_k^2+\sigma_i^2}$:

$$\sqrt{((n-1)\rho+1)\frac{4\sigma^2}{nK}}$$

- $((n-1)\rho+1)$ = **effet de conception**
- $\frac{nK}{((n-1)\rho+1)}$ = **taille d'échantillon effective**

Insérez dans le calcul et procédez comme avant.

Section 7

Power via diagnosis

Design 1: complete randomization

Conception 1 : randomisation complète

```
# Simple
N <- 100

complete_design <-
  declare_model(
    N = N,
    State = sample(1:4, N, replace = TRUE),
    U = rnorm(N)/4,
    Y0 = State + U,
    Y1 = State + U + .2
  ) +
  declare_inquiry(ate = mean(Y1 - Y0)) +
  declare_assignment(
    Z = complete_ra(N),
    Y = Z*Y1 + (1-Z)*Y0
  ) +
  declare_estimator(Y ~ Z, label = "complete")
```

Design 2: cluster randomization

Conception 2 : randomisation par grappes

```
# Simple
N <- 100

clustered_design <-
  declare_model(
    N = N,
    State = sample(1:4, N, replace = TRUE),
    U = rnorm(N)/4,
    Y0 = State + U,
    Y1 = State + U + .2
  ) +
  declare_inquiry(ate = mean(Y1 - Y0)) +
  declare_assignment(
    Z = cluster_ra(clusters = State),
    Y = Z*Y1 + (1-Z)*Y0
  ) +
  declare_estimator(Y ~ Z,  clusters = State, label = "clustered")
```

Design 3: blocked randomization

Conception 3 : randomisation par blocs

```
# Simple
N <- 100

blocked_design <-
  declare_model(
    N = N,
    State = sample(1:4, N, replace = TRUE),
    U = rnorm(N)/4,
    Y0 = State + U,
    Y1 = State + U + .2
  ) +
  declare_inquiry(ate = mean(Y1 - Y0)) +
  declare_assignment(
    Z = block_ra(blocks = State),
    Y = Z*Y1 + (1-Z)*Y0
  ) +
  declare_estimator(Y ~ Z + State, label = "blocked")
```

Design diagnosis

Diagnostic de conception

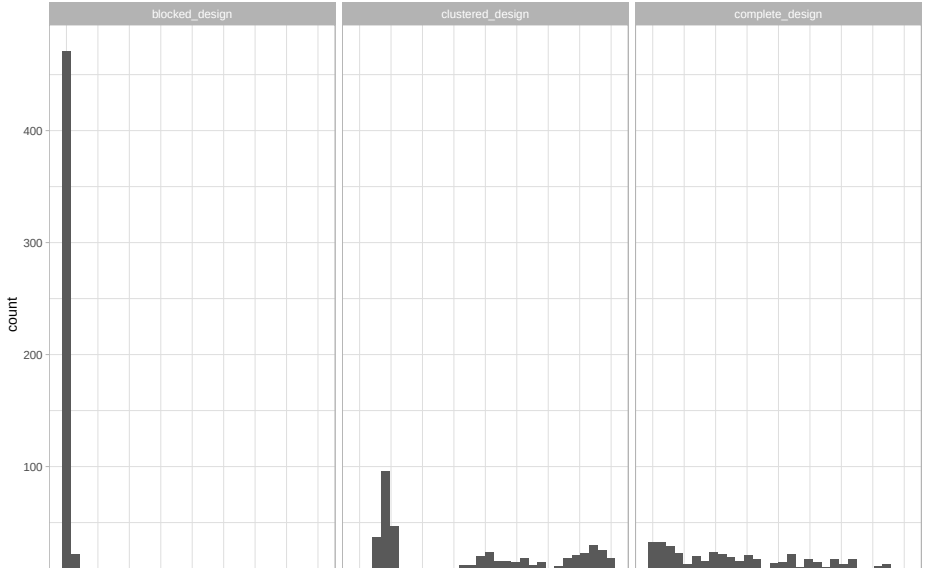
Design diagnosis is **arbitrarily flexible**.

Le diagnostic de conception est **arbitrairement flexible**.

```
diagnoses <- diagnose_designs(
  complete_design, clustered_design, blocked_design,
  sims = 500)
```

Distribution of p-values

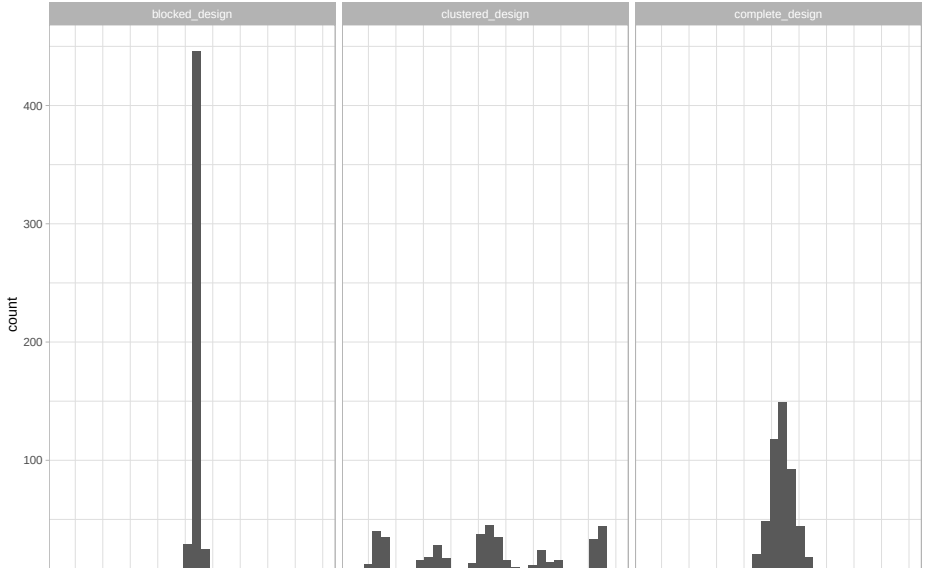
Distribution des p-values



Design	Mean Estimate	Bias	SD Estimate	RMSE	Power	Coverage
complete_design	0.19	-0.01	0.23	0.23	0.13	0.94
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
clustered_design	0.19	-0.01	1.35	1.34	0.00	1.00
	(0.06)	(0.06)	(0.02)	(0.02)	(0.00)	(0.00)
blocked_design	0.20	0.00	0.05	0.05	0.99	0.95
	(0.00)	(0.00)	(0.00)	(0.00)	(0.01)	(0.01)

Distribution of estimates

Distribution des estimations



Varying assignment and N

Varier l'assignation et N

```
diagnosis <-
  list(redesign(complete_design, N = 100),
       redesign(complete_design, N = 200),
       redesign(complete_design, N = 600),
       redesign(clustered_design, N = 100),
       redesign(clustered_design, N = 200),
       redesign(clustered_design, N = 600),
       redesign(blocked_design, N = 100),
       redesign(blocked_design, N = 200),
       redesign(blocked_design, N = 600)) |>
  diagnose_design(sims = 500)
```

Power over N and strategy

Puissance selon N et la stratégie

Big takeaways

Messages clés

- Power depends on sample size, variability, effect size — **and** data and analysis strategies.
- Estimate power under **multiple scenarios**.
- Use the **same code** for power as for your final analysis.
- If you can declare a design and have a test, you can calculate power.

- La puissance dépend de N, de la variabilité, de la taille d'effet — **et** des stratégies de données et d'analyse.
- Estimez la puissance sous **plusieurs scénarios**.
- Utilisez le **même code** pour la puissance et l'analyse finale.
- Si vous pouvez déclarer une conception et tester, vous pouvez calculer la puissance.

Big takeaways (2)
Messages clés (2)

- Power can be right but **misleading**. For confidence:
  - check **bias** and **coverage**, not just power
  - check power especially **under the null**
- Don't let power distract from more **substantive** diagnosands.

- La puissance peut être correcte mais **trompeuse**. Pour être sûr :
  - vérifiez **biais** et **couverture**, pas seulement la puissance
  - vérifiez la puissance surtout **sous l'hypothèse nulle**
- Ne laissez pas la puissance masquer des diagnosandes plus **substantiels**.