# Tests statistiques

Références :

- **RS** *Statistique mathématique en action*, V. Rivoirard et G. Stoltz
- **VN** Chaîne Youtube Very Normal (nos doctorants ont du talent)

## Objectifs

Le but de ce court document est de donner une intuition des tests statistiques. Il ne comprendra que très peu de mathématiques, et se concentrera sur l'aspect pratique et concret des objets. D'ailleurs, à peu près la moitié du document est occupée par les exemples et graphiques.
En particulier, il n'est pas nécessaire d'avoir étudié des statistiques pour le comprendre. Il suffit d'un peu de probas (et encore, réduit au maximum), et de bon sens (mais ça je ne m'inquiète pas).
J'ai fait ce document pour rassembler des idées issues de mon stage de L3, des cours qu'on a reçus et des livres/vidéos que j'ai consultés. Je voulais me mettre les idées au clair pour l'épreuve de modélisation, pour avoir des réponses "instinctives" à donner et des interprétations concrètes, ce qui me semble être un point valorisé.

*Il vous suffit d'un guide (c'est moi) et d'un cerveau en état de marche (ça c'est vous).*
*On ne va pas vraiment faire de maths, juste jeter un oeil au paysage.*
Voyage au pays des maths, Arte

En caractères grisés, il y a quelques exemples qui servent de fil conducteur au document. Ils présentent l'inconvénient de rendre ce dernier franchement plus dense. Ils peuvent peut-être être passés pour une première lecture, et éclaireront une éventuelle deuxième lecture. Bien sûr, ils font partie intégrante du processus de compréhension, mais si comme moi vous avez une attention et une mémoire de travail limitées, n'hésitez pas à passer les exemples si vous en ressentez le besoin.

La section suivante sert à la fois à donner une première intuition en quelques lignes, et à établir un sommaire à ce document.
Afin de teaser un peu le plan, sont indiquées en violet les notions abordées dans chaque partie.

## Les tests statistiques en 30 secondes

1. On choisit une hypothèse $H_0$ qu'on cherche à réfuter, et une hypothèse $H_1$ qu'on voudrait lui substituer. La quasi-totalité du processus se déroulera alors en supposant que $H_0$ est vraie. (Hypothèse nulle, hypothèse alternative, asymétrie)
2. On trouve une "quantité" (calculée à partir d'une observation quelconque) dont on connaît le comportement sous $H_0$. (statistique de test)
3. On calcule la valeur de cette quantité pour notre observation concrète.
4. On regarde si, sous l'hypothèse $H_0$, il était probable ou non que l'observation prenne une telle valeur : si c'était vraiment improbable, c'est qu'on n'était sûrement pas sous $H_0$. ($p$-value, risque de première espèce)

# 1 Mise en place des hypothèses et asymétrie du problème

La raison d'exister des tests statistiques est de *rejeter une hypothèse*. Cela peut prendre différentes formes, mais c'est vraiment l'idée à avoir en tête. On va donc noter $H_0$ une hypothèse qu'on souhaite rejeter, appelée hypothèse nulle.

On va également choisir une hypothèse alternative $H_1$ qui va être notre candidate pour remplacer $H_0$ si on arrive à rejeter cette dernière. Si on n'a pas trop d'idée, $H_1$ peut-être le complémentaire de $H_0$, mais des fois on pourra être plus précis.

Il est important de remarquer que la structure-même du test n'est pas symétrique en $H_0$ et $H_1$. Comme on le verra, un test est construit pour nier une hypothèse, mais jamais il ne la valide. Les deux seules réponses qu'un test peut donner sont "On rejette $H_0$" et "On ne sait pas" (ou "on n'a pas assez d'information pour rejeter $H_0$").
Cela donne une importance particulière à $H_0$, puisque tant qu'on n'aura pas "prouvé" (c'est un bien grand mot, mais c'est l'idée) qu'elle est fausse, on gardera cette hypothèse. C'est donc souvent l'hypothèse la plus probable qui est prise comme $H_0$, ou alors celle qui a le moins de conséquence négative si on la suppose vraie à tort.
En particulier, le même test mené par deux personnes différentes peut avoir des hypothèses inversées : l'exemple classique est celui du laboratoire pharmaceutique qui préfère supposer que son médicament est sans danger jusqu'à preuve du contraire, alors que le patient préférera le supposer nocif jusqu'à être sûr qu'il est safe.

En grisé, on verra deux exemples sur lesquels on appliquera les notions théoriques.

Le premier exemple **(E1)** (inspiré de **VN**) est celui d'un Youtuber qui a l'impression que sa nouvelle vidéo a bénéficié d'un temps de visionnage moyen plus élevé que d'habitude, et qui veut vérifier cela au moyen d'un test statistique. Pour cela, admettons qu'il connaisse le temps de visionnage moyen sur ses dernières vidéos (il sait d'ailleurs que ces temps de visionnages suivent des lois gaussiennes $\mathcal{N}(50, 10^2)$) et qu'il dispose des temps de visionnage de sa dernière vidéo par $n$ personnes (on notera **(E1a)** quand $n = 10$ et **(E1b)** quand $n = 100$). L'hypothèse $H_0$ est donc que la vidéo a marché comme d'habitude, et l'hypothèse $H_1$ est qu'elle a mieux marché. En particulier, il considère comme acquis le fait qu'elle n'a pas fait significativement moins de temps de visionnages, ce qui aura des conséquences sur le résultat du test.

Le second exemple sera en 2 parties : **(E2a)** (un exemple classique) et **(E2b)** (tiré de **RS**). **(E2a)** est un joueur qui joue à pile ou face, et qui au bout de 200 parties se demande si la pièce est équilibrée. Son hypothèse nulle est que la pièce est équilibrée, car il n'a pas de raison de remettre ça en question sans élément contradictoire. **(E2b)** est un professeur qui demande à ses élèves de tirer 200 fois à pile ou face en devoir maison, et qui veut déterminer, étant donné le tirage d'un élève, s'il l'a fait sérieusement ou s'il a inventé les chiffres. L'hypothèse $H_0$ est que l'élève n'a pas triché (car il croit en la présomption d'innocence). Dans les deux cas, $H_0$ est "L'observation est issue de 200 lois de Bernoulli indépendantes de paramètre $1/2$", et $H_1$ est le complémentaire de $H_0$, mais on verra que ces deux situations vont ammener deux tests différents.

## 2 Statistique de test

La première étape est donc de choisir la manière de s'y prendre pour faire notre choix sur $H_0$, c'est-à-dire de choisir une quantité à regarder. C'est à cette étape qu'on fait le gros du travail finalement, mais comme c'est un travail théorique on ne va pas rentrer dans les détails.
Ce qu'il faut avoir en tête, c'est qu'on va chercher une fonction $\mathcal{T}$ (appelée parfois statistique de test, mais attention des fois c'est autre chose qui est appelé comme ça je crois) qui prend en entrée une observation d'une variable aléatoire, et qui renvoie une valeur réelle. Souvent, l'observation sera un $n-$échantillon, notée $\boldsymbol{x} = (x_1, \dots, x_n)$, c'est-à-dire $n$ réalisations indépendantes d'une variable aléatoire.

Cette fonction (qu'on peut voir comme $\mathcal{T} \circ X$ en se rappelant qu'une v.a. est une fonction mesurable) est alors elle-même une variable aléatoire réelle. Elle possède donc tous les attributs d'une variable aléatoire réelle, notamment une loi de probabilité (qu'on représentera souvent en traçant sa densité, quand elle existe ; pour simplifier, on parlera de "densité" même pour parler de la fonction de masse d'une variable discrète). On essaiera d'ailleurs de bien distinguer la variable aléatoire $\mathcal{T}(X)$, de la valeur $\mathcal{T}(\boldsymbol{x})$ de $\mathcal{T}$ prise en une observation $\boldsymbol{x}$.
Intervient alors un point important de cette fonction/v.a. : on connaît son comportement sous $H_0$. On peut donc par exemple tracer sa densité sachant $H_0$.

> **(E1)** On peut ici prendre l'exemple classique de la fonction $\mathcal{T} : (X_1, \dots, X_n) \mapsto \dfrac{1}{n}\displaystyle\sum_{k=1}^{n} X_k$, qui a un échantillon associe sa moyenne (empirique). Sous $H_0$, comme chaque $X_i$ suit une loi $\mathcal{N}(50, 10^2)$, $\mathcal{T}(X)$ suit la loi $\mathcal{N}(50, 10^2/n)$ (c'est une somme de loi normales indépendantes).
> **(E2a)** Ici, c'est un peu la même chose : on regarde la moyenne empirique (en comptant par exemple 0 pour une pile et 1 pour une face), qui sous $H_0$ suit une loi binomiale $\mathcal{B}(n, 1/2)$.
> **(E2b)** Là, prendre la moyenne n'aurait que peu d'intérêt, car un éventuel tricheur ferait attention à ce que la moyenne des lancers soit proche de $0,5$. On peut prendre par exemple $\mathcal{T} = \mathbb{1}_{\{\text{Il y a au moins une séquence de 6 lancers égaux à la suite}\}}$. Sous $H_0$, donc si les lancers ont vraiment été effectués, on peut montrer que la variable aléatoire $\mathcal{T}(X)$ prend la valeur 1 avec probabilité $0,97$ et la valeur 0 avec probabilité $0,03$ environ.

Une fois cette densité sous $H_0$ connue, il ne nous reste plus qu'à voir où tombe notre échantillon et on aura déjà des choses à dire sur notre problème.

## 3 Appliquer notre fonction $\mathcal{T}$ à notre échantillon

Pas grand chose à faire à cette étape, donc plutôt qu'un long discours on va se donner un échantillon $\boldsymbol{x}$ pour chaque exemple, tracer la densité de $\mathcal{T}(X)$ sous $H_0$, et regarder où tombe $\mathcal{T}(\boldsymbol{x})$ dans chaque cas.

**(E1a)** Sur 10 viewers au hasard, le Youtuber relève les durées de visionnage suivantes : $\boldsymbol{x} = \{48, 18, 67, 75, 37, 59, 65, 57, 29, 65\}$. On a donc $\mathcal{T}(\boldsymbol{x}) = 52$.
**(E1b)** Après avoir sondé plus de viewiers, le Youtuber a relevé 90 valeurs en plus des 10 premières. Il remarque que la moyenne des 100 valeurs est également 52.
**(E2a)** Bon on ne va pas se donner un vrai échantillon avec 200 tirages, surtout que ce qui nous importe ici est simplement la moyenne, donc disons que notre joueur a observé une fréquence de 0.52 pour les faces.
**(E2b)** Idem ici, on ne va pas sortir 200 nombres, supposons donc qu'un élève présente un échantillon qui n'a jamais 6 répétitions (c'est-à-dire que ni la séquence "PPPPPP", ni la séquence "FFFFFF" n'apparaisse).

Pour anticiper un peu sur la prochaine partie, on peut déjà essayer de tirer des conclusions de ces graphiques...

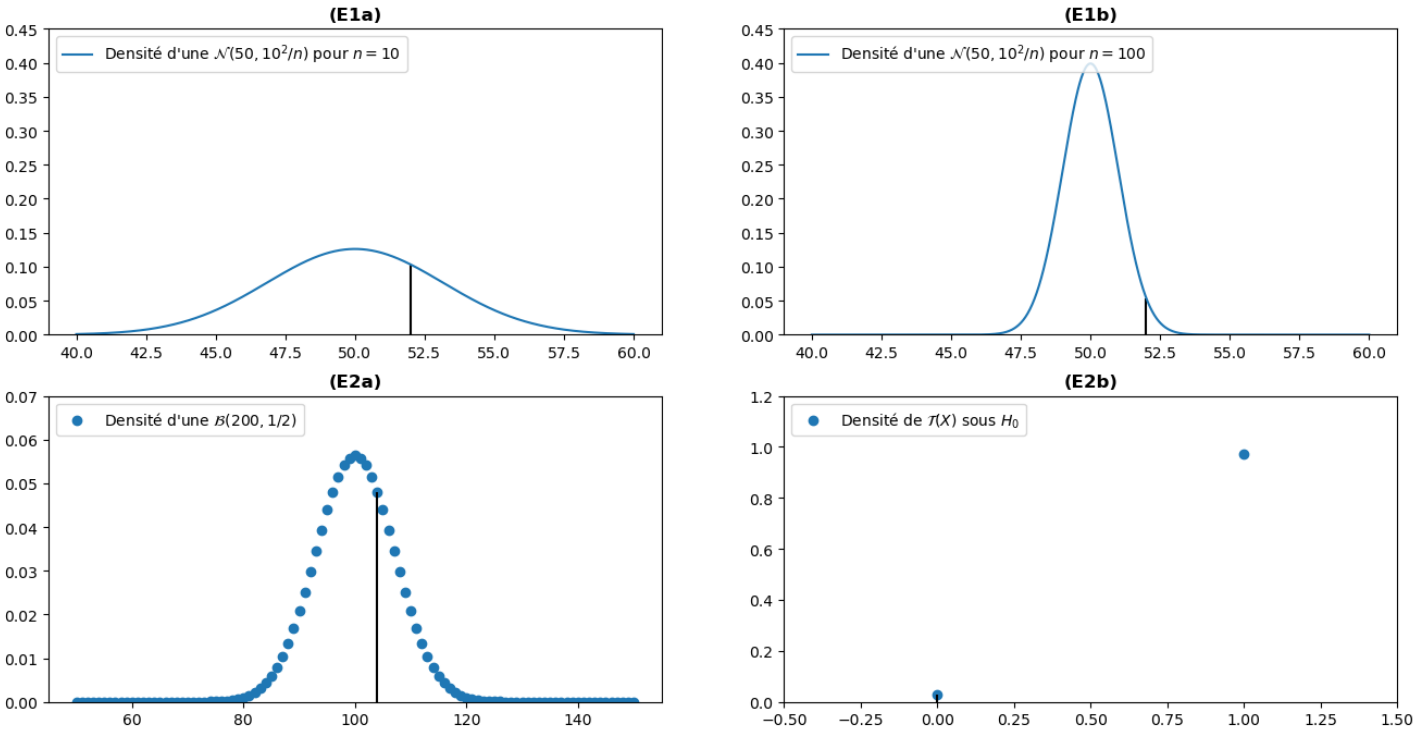


FIGURE 1 – Les densités de $\mathcal{T}(X)$ sous $H_0$ et les valeurs obtenues pour $\mathcal{T}(\boldsymbol{x})$

# 4 Interprétation, conclusion et risques

Ici, on va faire intervenir un outil que je trouve hyper important, et qui pourtant ne semble pas privilégié à notre niveau (mais qui vient de faire son entrée dans le programme, alors peut-être que c'est en train de changer ?) : la $p$-value, ou valeur $p$. En tout cas, dans les publications scientifiques, avoir une petite $p$-value est le nerf de la guerre.
Rappelons-nous le contexte : on a une hypothèse $H_0$ qu'on a supposée vraie, et on a observé des données dont on connaît le comportement sous cette hypothèse. Si maintenant, on observe quelque chose de hautement improbable, c'est sans doute que cette hypothèse $H_0$ ne tient pas la route, puisque nos observations ne semblent pas, probabilistiquement parlant, être issues d'une telle loi.

C'est comme ça qu'on définit la $p$-value : c'est la probabilité, sous l'hypothèse $H_0$, qu'une observation aléatoire prenne des valeurs aussi extrêmes que notre observation à nous.

De manière rigoureuse, la notion de "valeur extrême" dépend du contexte (et notamment de $H_1$), mais elle est assez instinctive. Pour comprendre un peu mieux, on va manipuler l'exemple très simple qui suit :

> **(E3)** Si notre observation est un unique nombre et que l'hypothèse $H_0$ est que ce nombre suit une loi $\mathcal{N}(0,1)$, alors on peut alors penser à la définition suivante de "valeur extrême" : une valeur est plus extrême qu'une autre quand sa valeur absolue est plus grande.
>
> L'observation $\boldsymbol{x} = \{0.3\}$ aura alors une grande $p$-value (environ 0.76) car c'est très probable d'avoir un résultat au moins aussi extrême en tirant selon un $\mathcal{N}(0,1)$. L'observation $\boldsymbol{x} = \{-1.6\}$ aura une $p$-value petite mais pas négligeable (environ 0.1), tandis que $\boldsymbol{x} = \{1000\}$ va prendre une $p$-value ridiculement petite ($10^{-155}$ apparemment), traduction directe du fait que c'est ridiculement peu probable de tirer 1000 à partir d'une normale centrée réduite.

Ainsi, si on est vraiment sous $H_0$, la $p$-value nous dit si notre observation était un tirage plutôt probable (grande $p$-value) ou plutôt improbable (petite $p$-value).
Sauf que se placer sous $H_0$ était juste une hypothèse, et on va se servir de la $p$-value pour estimer la vraisemblance de cette hypothèse.

> **(E3)** Pour continuer sur l'exemple précédent, si $H_0$ est que vous tirez selon une $\mathcal{N}(0,1)$ et que vous observez 1000, vous n'allez pas vous dire que c'est le moment de jouer au loto parce que vous êtes hyper chanceux : vous allez plutôt conclure que c'est l'hypothèse $H_0$ qui était fausse.
> Mais tout ça est une histoire de probabilités : c'est possible qu'aujourd'hui était vraiment votre jour de chance et que vous passiez à côté du jackpot. Extrêmement peu probable, et d'autant moins probable que $p$ est petite, mais possible. Vive les probas :)

Toute la subtilité des tests statistiques est là-dedans : j'ai observé quelque chose, c'est peu probable au vu de mon hypothèse, mais est-ce que c'est suffisamment peu probable pour que je puisse en conclure que l'hypothèse était fausse ?

Dans tout ça, la $p$-value me paraît fondamentale, puisqu'elle quantifie cette (im)probabilité de notre observation. L'information apportée par la $p$-value est bien plus subtile que l'information binaire de rejet ou non de $H_0$, et elle est aussi plus générale car elle ne dépend pas des mécanismes de décision.
En particulier, elle permet de confier la tâche de tester et celle de décider à deux personnes différentes : le statisticien estime la $p$-value, et le décideur s'en sert pour prendre une décision.

À partir de là, on peut retourner voir nos graphiques de la partie précédente et se demander où apparaissent les $p$-value. En fonction des cas, elles sont indiquées grâce aux parties grisées ci-dessous :

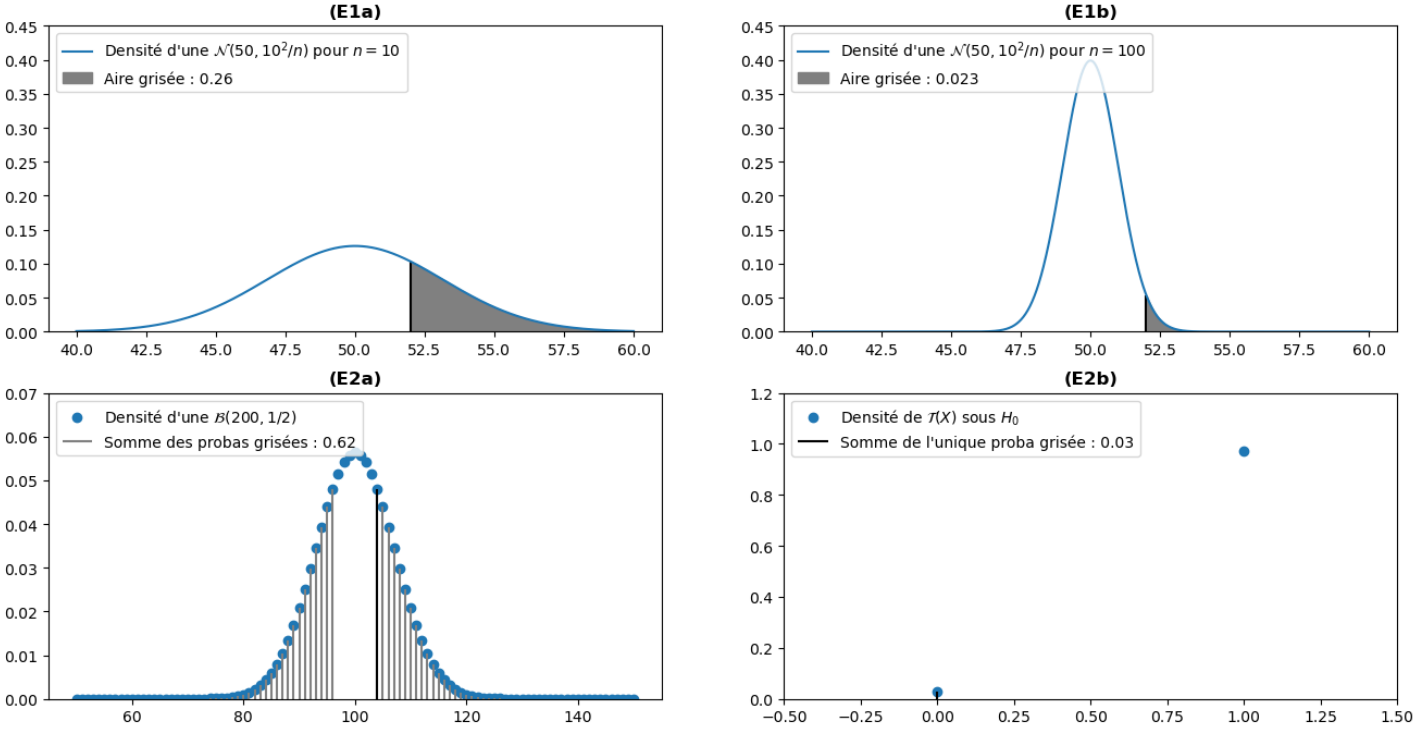


FIGURE 2 – La $p$-value est l'aire/la somme de la partie grisée (des valeurs au moins aussi extrêmes que l'observation)

L'interprétation est donc la suivante (et ce, dans tous les cas, pour n'importe quel test !) : si jamais $H_0$ est vraie, alors l'observation (ou une observation au moins aussi extrême) aurait une probabilité $p$ de se produire (où $p$ est la $p$-value obtenue).
Pour prendre la décision finale et savoir si ce $p$ est suffisamment petit, on va avoir d'abord besoin du paragraphe suivant.

Vient maintenant le moment de parler du risque. J'ai repoussé ce moment parce que, bien que fixer le risque soit une des premières choses qu'on fait quand on fait un test, il me paraît fondamental de comprendre comment ce risque va intervenir pour bien le choisir.
Ça reste mon avis et mon interprétation personnelle, et je vous encourage évidemment à ne pas me croire sur parole, mais ça explique la maigre part allouée au risque dans ce document (petite pensée émue au risque de seconde espèce d'ailleurs).
On appelle risque de première espèce, souvent noté $\alpha$, la probabilité maximale, sous l'hypothèse que $H_0$ est vraie, de rejeter $H_0$ (à tort, donc). Il y a une vraie définition rigoureuse, et même peut-être légèrement différente (voir **RS** p30) mais bien plus sophistiquée et celle-ci suffit pour ce qu'on veut en faire ici ; si besoin, faites vos propres recherches.

L'idée du test est donc de se fixer $\alpha$ *avant* de faire le test, en fonction des risques qu'on est prêt à prendre pour rejeter $H_0$ (puisque, vous l'aurez maintenant compris, rejeter $H_0$ n'est (presque[1]) jamais une action 100% sûre, notre observation peut juste être hautement improbable). Et une fois ce $\alpha$ fixé, il y a une règle de décision assez simple qui est utilisée en pratique : on rejette $H_0$ quand $p \leqslant \alpha$.
Attention ! La $p$-value ne représente pas la probabilité de se tromper sous l'hypothèse $H_0$ ! Je ne vais pas rentrer dans les détails ici, parce que je pense que c'est des discussions subtiles et que ce n'est pas très clair dans ma tête, mais il faut vraiment éviter les raccourcis.
Terminons en concluant nos exemples, bien que ce ne soit pas le plus important :

> On se fixe partout un risque $\alpha = 0.05 = 5\%$ (officiellement ce n'est pas bien, il faudrait étudier au cas par cas, mais c'est plus simple).
> **(E1)** On voit qu'avec seulement $n = 10$ valeurs, notre Youtuber n'a pas assez de données pour rejeter $H_0$. C'est logique : sur si peu de valeurs, il n'est pas improbable que la moyenne de l'échantillon s'écarte un peu de la moyenne de la population, et c'est confirmé par cette $p$-value : ça avait environ une chance sur 4. On a $p = 0.26 > 0.05$, donc on ne rejette pas $H_0$. Mais comme il avait quand même l'intuition que cette vidéo avait mieux marché, il a récolté plus de données, et bingo : avec $n = 100$ valeurs, il obtient $p = 0.23 < 0.05$. Avec plus de valeurs, ça devient de moins en moins probable que la moyenne empirique s'écarte autant de l'espérance, et cette baisse de probabilité est traduite par la $p$-value. Il peut donc conclure que cette dernière vidéo a bien marché.
> **(E2a)** Notre joueur qui a obtenu 104 faces sur 200 lancers s'est sans doute inquiété pour rien : avec son $p = 0,62$, il est bien loin de pouvoir rejeter l'hypothèse d'une pièce équilibrée puisque plus d'une fois sur 2, il aurait obtenu un résultat au moins aussi déséquilibré. Peut-être qu'avec (beaucoup) plus de tirages, il aura une différence significative, souhaitons-lui bonne chance dans cette quête...
> **(E2b)** Bon là l'interprétation est facile puisque le test est construit de manière adaptée au problème : le professeur va punir tous les élèves qui n'ont pas 6 tirages identiques à la suite. Ce faisant, il punira environ 3% des honnêtes élèves (mais il est impossible de prévoir combien de tricheurs passeront sous les radars, car cela dépend de la méthode qu'ils ont employé pour tricher).

S'il vous reste un peu de motivation, vous pouvez maintenant relire la section **Les tests statistiques en 30 secondes** pour avoir l'éclairage nouveau apporté par la lecture du document entier.
Pour celles et ceux qui veulent aller un peu plus loin, mais toujours dans cet esprit de ne pas rentrer dans les détails mathématiques, la section suivante est là pour ça.

Bonne journée aux autres !

1. Si mon hypothèse est que mon échantillon suit une loi exponentielle et que mes données font apparaître un nombre strictement négatif, là je suis sûr de pouvoir rejeter $H_0$, mais la plupart du temps on rejette avec une certaine probabilité.

# Quelques remarques moins importantes en vrac

— Faire plein de tests statistiques Admettons que je veuille prouver qu'une affirmation couramment admise est fausse. Par exemple, je veux montrer envers et contre tout que la taille moyenne d'un être humain est supérieure à 1m80. Je vais donc rassembler des données (c'est-à-dire mesurer des gens, disons une dizaine), et regarder si un test statistique me permet de conclure. Ah pas de bol, je trouve une taille moyenne sur mon échantillon de 1,70m. Pas grave, je suis motivé, je rassemble un nouvel échantillon de 10 personnes, et zut, 1,64m de moyenne. Bon le monde ne veut pas laisser la vérité éclater au grand jour, alors je vais continuer de tester jusqu'à ce qu'au hasard des constitutions de mes échantillons, je trouve enfin une taille moyenne qui contredit significativement l'hypothèse de taille moyenne inférieure à 1,80m ! Je m'empresse alors de contacter ma revue scientifique préférée pour publier cet article révolutionnaire, cautionné par cette magnifique $p$-value de 0,047.

On voit bien ici que, à cause de comment sont construits les tests, si on en répète plusieurs à la suite on va finir par tomber sur une observation peu probable et conclure au rejet de $H_0$ alors que les précédents ne nous permettaient pas de conclure. Et si on est de bonne foi mais qu'on n'a pas de chance, le premier échantillon peut être significativement différent mais quand même être issu de $H_0$. C'est pourquoi l'élément de preuve scientifique le plus solide est la meta-analyse : elle permet d'éviter ce biais de sélection des échantillons improbables en prenant en compte plusieurs études.

Ce paragraphe n'est pas là pour critiquer le système des tests statistiques, mais pour ne pas lui attribuer des avantages qu'il n'a pas. Ça reste une méthode *probabiliste*, et en tant que telle elle produit forcément des résultats faux (à une fréquence qu'on peut heureusement contrôler). Je pense que c'est important de garder en tête les limites des outils qu'on utilise.

— Une grande $p$-value On a vu qu'une petite $p$-value est la clé pour pouvoir conclure au rejet de $H_0$, mais alors ne peut-on pas conclure à l'acceptation de $H_0$ si $p$ est proche de 1 ? Prenez le temps d'y réflechir si besoin, et j'écris cette phrase un peu longue pour temporiser et vous laisser le temps d'arrêter votre lecture si vous voulez chercher, mais la réponse est non. Comme on regarde l'image par $\mathcal{T}$ de l'échantillon, on se prive d'une partie de l'information. L'information contenue dans $\mathcal{T}(\boldsymbol{x})$ peut être compatible avec $H_0$ et que pourtant $\boldsymbol{x}$ soit issu d'une loi complétement différente. Il suffit par exemple que l'échantillon soit tiré selon une loi de même espérance que la loi sous $H_0$ pour que le test utilisant la moyenne empirique ne détecte pas d'écart (alors qu'une $\mathcal{N}(0,1)$ et une $\mathcal{U}([-1,1])$ sont deux lois bien différentes !). Ce serait un raisonnement du type "Je fais une hypothèse, je trouve un résultat cohérent sous l'hypothèse, j'en conclus que l'hypothèse est vraie" que celles et ceux qui ont fait prépa ont écrit contraint(e)s et forcé(e)s dans leurs copies de chimie.

— Bien choisir $H_1$ Dans l'exemple **(E1)**, le fait que le Youtuber présuppose que la vidéo a au moins aussi bien marché lui permet de réduire sa $p$-value en "éliminant" des valeurs plus extrêmes. Sur le dernier graphique, on voit que s'il n'avait pas fait cette hypothèse, la partie grisée aurait été symétrique par rapport à 50, et donc sa $p$-value aurait été doublée, ce qui l'aurait presque empêché de rejeter $H_0$ au risque $\alpha = 5\%$.

Le choix de $H_1$ a donc quand même des conséquences sur le résultat du test

— Auto-promo Si maintenant vous voulez savoir comment adapter tout ça à de la comparaison de *courbes*, vous pouvez aller jeter un oeil à mon rapport de stage de L3 :) (disclaimer : j'étais jeune, je ne garantis ni l'intérêt ni la justesse de ce rapport)