TESTS D'HYPOTHÈSE

Introduction – Raisonnement général – Données binaires

PASS Lyon Est

Dr. Nicolas ROMAIN-SCELLE

EXEMPLE INTRODUCTIF

Le dilemme

- Un proche vous propose de jouer à Pile ou Face, avec une mise d’argent
- Pile : vous remportez la mise ; Face : vous perdez
- La pièce de jeu est apportée par votre adversaire : vous ne savez pas si la pièce n’est pas truquée

- L’enjeu monétaire est limité, mais pas négligeable non plus
- Vous connaissez le proche, et savez qu’il n’est pas toujours très digne de confiance

- **Est-ce que vous acceptez de jouer ?**

Les éléments du problème

- Un phénomène **aléatoire** : lancer de pièce
- Un comportement **théorique désiré** de la pièce : 50/50
- Un comportement **effectif** de la pièce, observable
- Un risque à déclarer **la pièce équilibrée à tort** : vous perdrez de l’argent
- Un risque à déclarer **la pièce déséquilibrée à tort** : vous perdrez la confiance de votre adversaire

FORMALISATION DU PROBLÈME

Modéliser un lancer de pile ou face

- Deux résultats possibles : Pile et Face
- On modélise un lancer : X la variable aléatoire donnant
  - $X(Pile) = 1$
  - $X(Face) = 0$

- X suit une loi de Bernouilli de paramètre π :
  - $X \sim Bern(\pi)$ avec
  - $E(X) = \pi$
  - $Var(X) = \pi(1 - \pi)$

Rappel : théorème central limite

Soit $(X_n)_{n\geq} = (X_1, \ldots, X_n)$ la suite de n VA X *i.i.d* (identiquement et indépendamment distribuées), de même loi d'espérance $E(X) = \mu$ et variance $Var(X) = \sigma^2$.

Soit $\overline{X}$ la VA moyenne arithmétique de cette suite :

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{i=n} X_i$$

Alors pour $\mathbf{n}$ **tendant vers l'infini**, la variable $\overline{X}$ **converge** vers la **loi Normale** $\mathcal{N}$:

$$\overline{X} \overset{n\to+\infty}{\sim} \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

En posant $\overline{Z} = \frac{\overline{X}-\mu}{\sqrt{\frac{\sigma^2}{n}}}$ la variable standardisée, par analogie :

$$\overline{Z} \underset{n\to+\infty}{\sim} \mathcal{N}(0,1)$$

Résultat théorique d’une pièce équilibrée

- On note π_0 la probabilité de Pile d’une pièce équilibrée théorique
- On attend autant de Pile que de Face :
  - $E(X) = \pi_0 = 1/2$
  - $Var(X) = \pi_0(1 - \pi_0) = 1/2^2 = 1/4$

- Avec n lancers suffisamment grand, on peut appliquer le théorème central limite :

$$F \underset{n \to +\infty}{\sim} \mathcal{N}\left(\pi, \frac{\pi(1-\pi)}{n}\right)$$

Résultat théorique d’une pièce équilibrée

- On a précédemment posé une hypothèse sur le comportement de notre pièce lors d’une expérience donnée : $\pi = \pi_0 = 1/2$.

- On appellera cette hypothèse **hypothèse nulle** ou H_0.

- Si notre hypothèse nulle s’avère fausse : $\pi \neq 0.5$, on sera en présence de l’**hypothèse alternative**, notée H_1.

Résultat théorique d'une pièce équilibrée

- Notez que nos hypothèses sont posées sur la pièce en elle-même, et non sur le résultat d'une série de lancers issus de cette pièce.
- Sous l'**hypothèse nulle**, on connait ainsi en détail la loi de l'estimateur de la probabilité π, noté F:

$$F \underset{n\to+\infty}{\overset{H_0}{\sim}} \mathcal{N}\left(\pi_0 = \frac{1}{2}, \frac{\pi_0(1-\pi_0)}{n} = \frac{1}{4n}\right)$$

Aparté : écart-type vs. erreur-type

- On parle d'**écart-type** pour mesurer la variabilité d'une loi de probabilité : à quel point les tirages dans cette loi sont dispersés autour de son espérance. Généralement σ

- On parle d'**erreur-type** pour désigner l'écart-type d'une statistique. Ici : $ET = \frac{\sigma}{\sqrt{n}}$

- La nuance n'est pas fondamentale pour la compréhension du cours

Notation

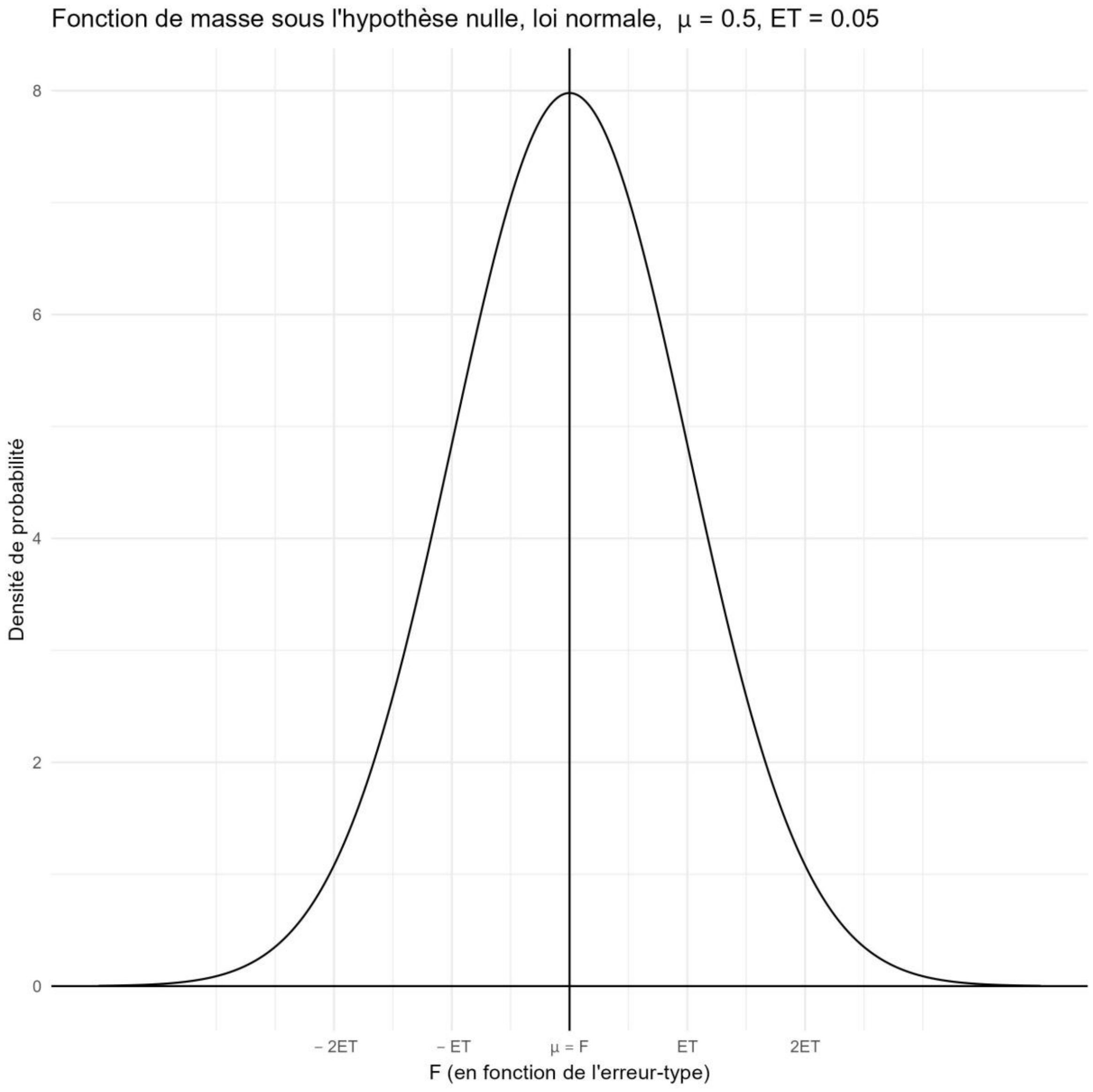
- π : **valeur théorique** : la vraie valeur, inobservable, que l'on cherche à estimer

- F **: l'estimateur** : variable aléatoire permettant, avec les données d'une expérience, d'estimer π

- f **: l'estimation** : une valeur numérique estimant π, potentiellement biaisée, et variant à chaque répétition de l'expérience

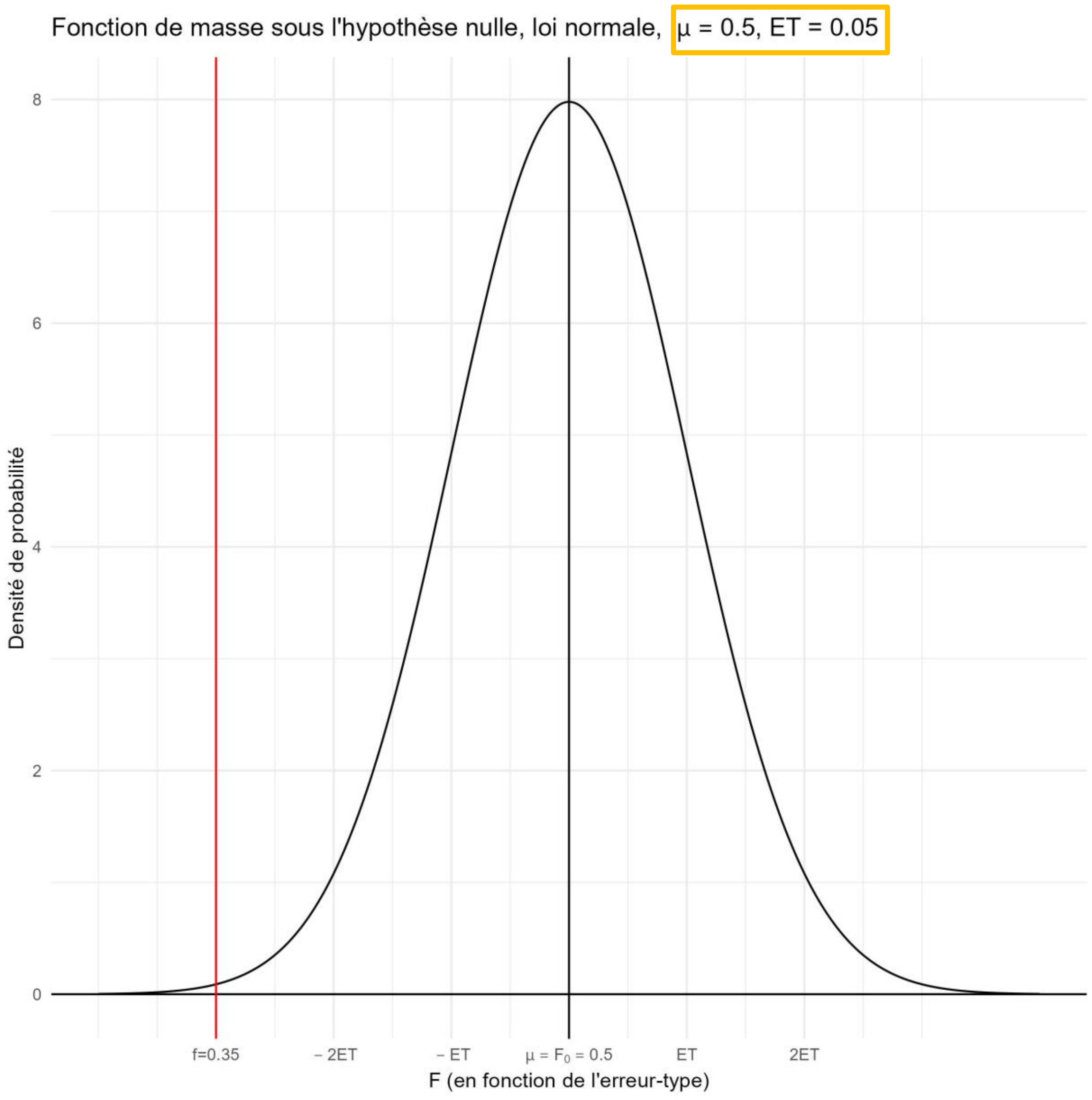
Comportement de la pièce sous H_0

Vu la définition de F donné supra, on peut représenter sa densité de probabilité ci-contre.



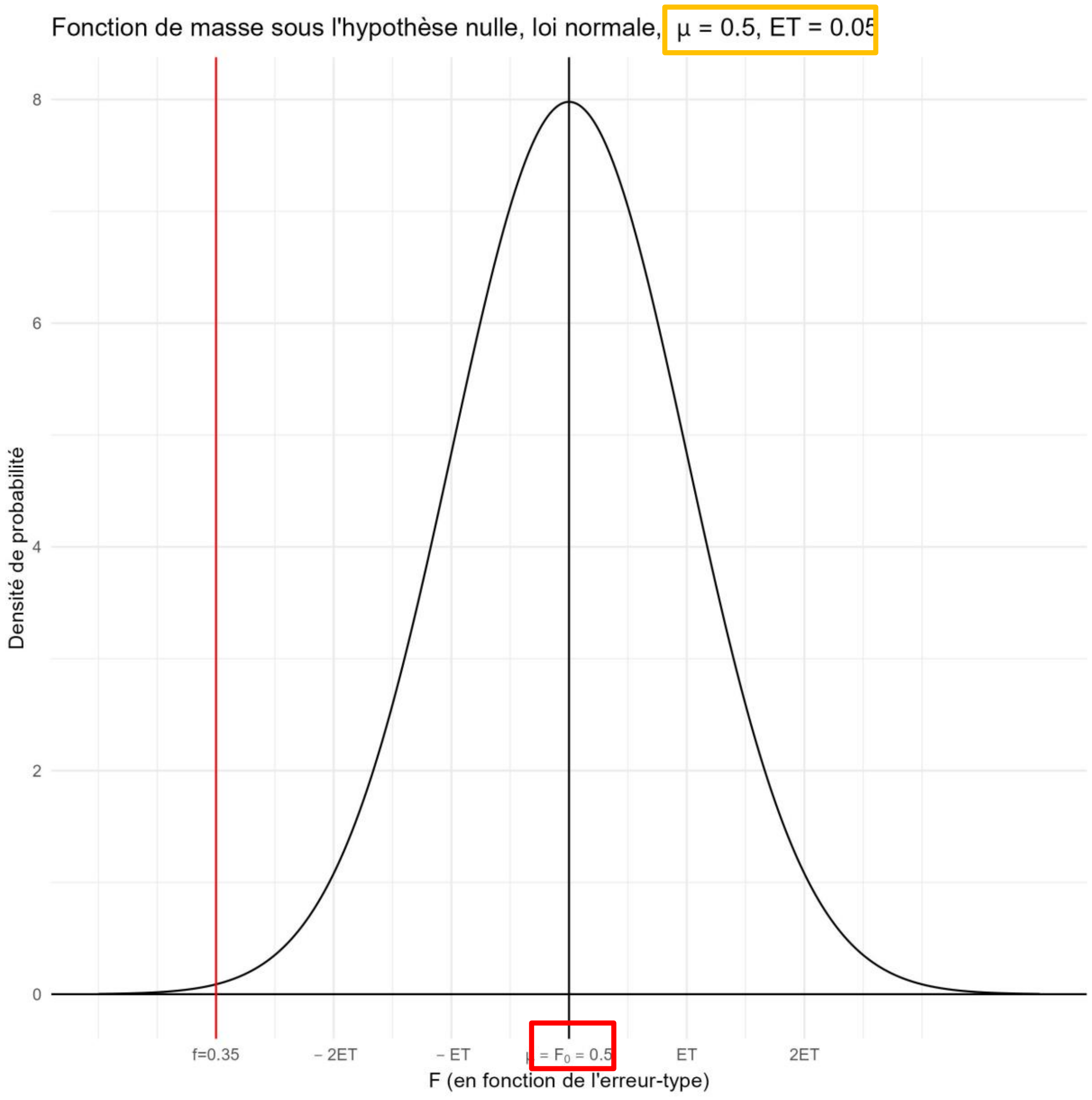
Comportement de la pièce sous H_0

Vu la définition de F donné supra, on peut représenter sa densité de probabilité ci-contre.



Comportement de la pièce sous H_0

Vu la définition de F donné supra, on peut représenter sa densité de probabilité ci-contre.



Résultat effectif de la pièce observée

- Par essence, on ne connaît pas $E(X) = \pi$ pour la pièce observée
- On doit l'**estimer**

- On réalise donc 100 épreuves aléatoires
- Les résultats :

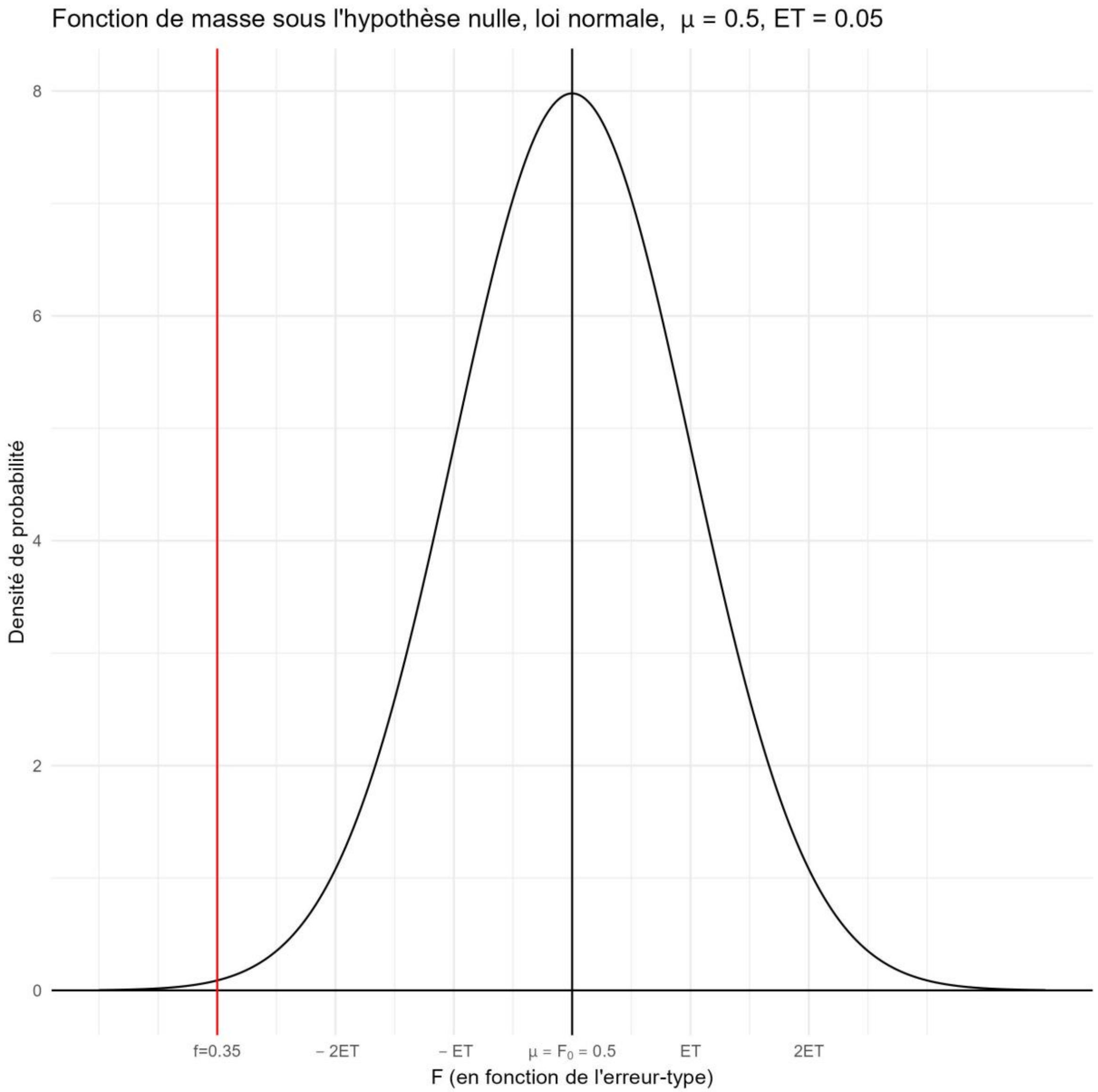
Pile	Face
35	65

Résultat effectif de la pièce observée

- On applique notre estimateur F au résultat de l'expérience
- $f = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{35}{35+65} = 0.35$
- **f est une estimation ponctuelle, sur un échantillon unique de 100 lancers**

Résultat effectif rapporté au théorique

On peut positionner f en relation avec la loi de F sur la figure précédente



Comparaison observé vs. théorique sous H_0

- On sent intuitivement, vu la figure précédente, que la pièce proposée par l’adversaire n’est pas équilibrée : le résultat de l’expérience (100 lancers) est loin de la moyenne attendue sous l’hypothèse nulle.

- Mais nous n’avons pas de quoi en faire la preuve en l’état.

Comparaison observé vs. théorique sous H_0

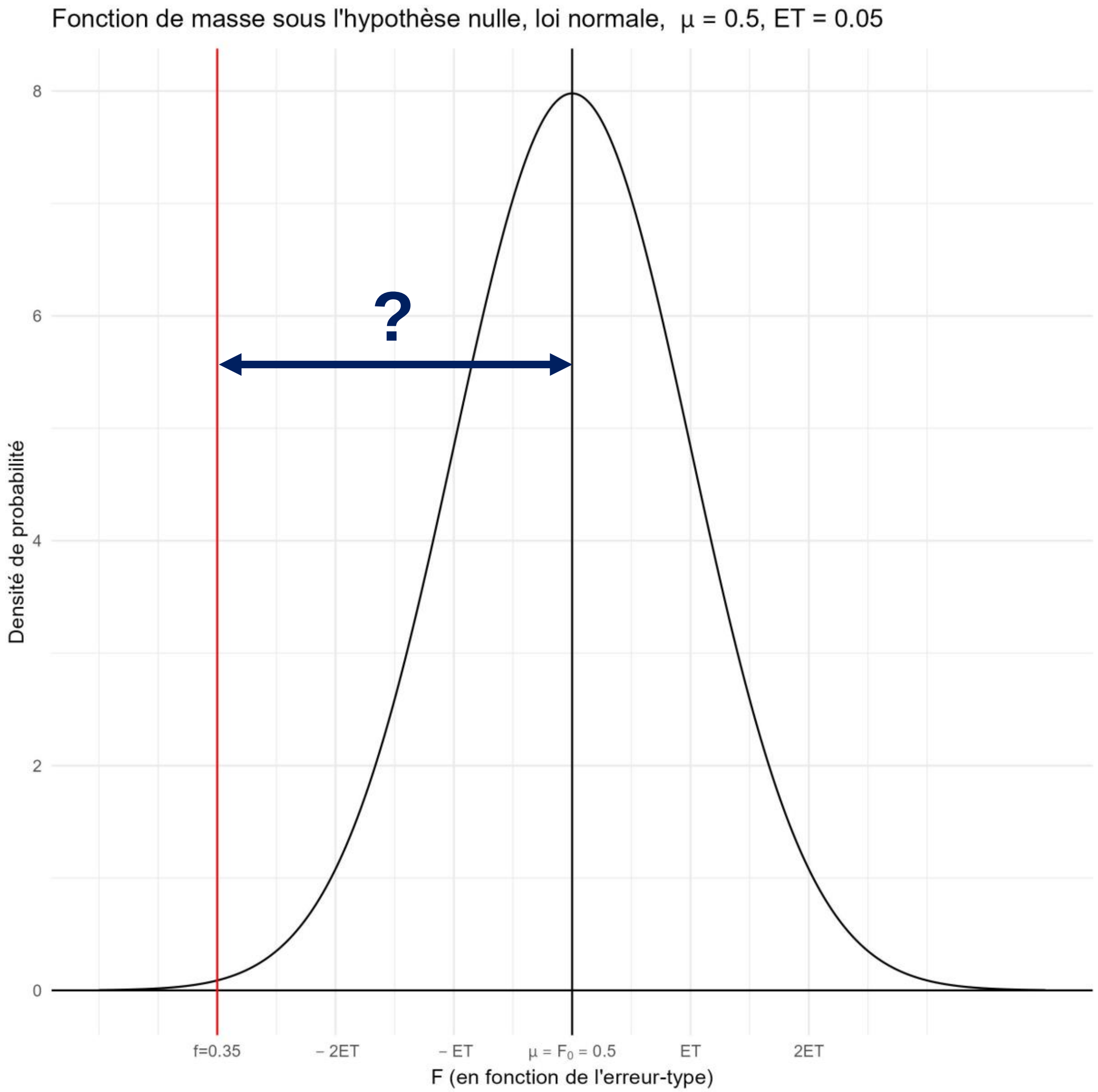
- Gardons le fil du raisonnement empirique : peut-on mettre une valeur chiffrée sur cette notion de « loin »?

- On peut essayer de mesurer la « distance » entre notre observé et notre attendu

- Autrement dit, on a besoin d’une mesure, ou d’une **statistique** pour mettre un nombre sur cette distance

Comparaison observé vs. théorique sous H_0

Comment mesurer correctement l'écart entre f estimé et le théorique sous H_0 ?



Définition de la statistique S

- On définit la statistique S ainsi :

$$S = \frac{|F - \pi|}{\sqrt{\frac{\pi(1-\pi)}{n}}}$$

Définition de la statistique S

- Propriétés de S

$$S = \frac{|\boldsymbol{F} - \boldsymbol{\pi}|}{\sqrt{\frac{\pi(1-\pi)}{n}}}$$

- Différence en valeur absolue entre l'estimateur F et la valeur théorique π
- *Plus l'estimation f est « loin » de π, plus S est élevé*

Définition de la statistique S

- On définit la statistique S ainsi :

$$S = \frac{|F - \pi|}{\sqrt{\boldsymbol{\frac{\pi(1-\pi)}{n}}}}$$

- Racine carrée de la variance $\pi(1-\pi)$ de la loi sur la pièce théorique divisée par l'effectif n : **erreur-type** de S
- *Plus la variance théorique est élevée, plus S est faible*
- *Plus l'effectif de l'expérience est élevé (nombre de tirages), plus S est élevé*

Définition de la statistique S

- Par ailleurs, formulée ainsi, notre statistique S est une variable aléatoire centrée réduite (F notre estimateur, moins son espérance divisé par son erreur-type).

$$S = \frac{|F - \pi|}{\sqrt{\frac{\pi(1-\pi)}{n}}} = \frac{|F - E(F)|}{ET(F)}$$

- On sait alors que $S \sim \mathcal{N}(0,1)$ (Loi Normale Centrée Réduite)

Calcul de la statistique $s_{H_0}^{obs}$

- On peut calculer notre statistique de test $s_{H_0}^{obs}$ sur la série de lancers expérimentaux observée, en tenant l'hypothèse nulle pour vraie ($H_0 : \pi = \pi_0 = 0{,}5$) :

$$s_{H_0}^{obs} = \frac{|f - \pi_0|}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} = \frac{|f - 1/2|}{\sqrt{\frac{1}{100 * 4}}} = \frac{|0.35 - 0.5|}{0.05} = 3$$

Interprétation de $s_{H_0}^{obs}$

- On obtient donc une distance entre les résultats expérimentaux et notre pièce théorique sous l'hypothèse nulle de 3.
- Cette distance n'est pas vraiment interprétable prise isolément…
- On sait en revanche comment elle se comporterai si la pièce est équilibrée :

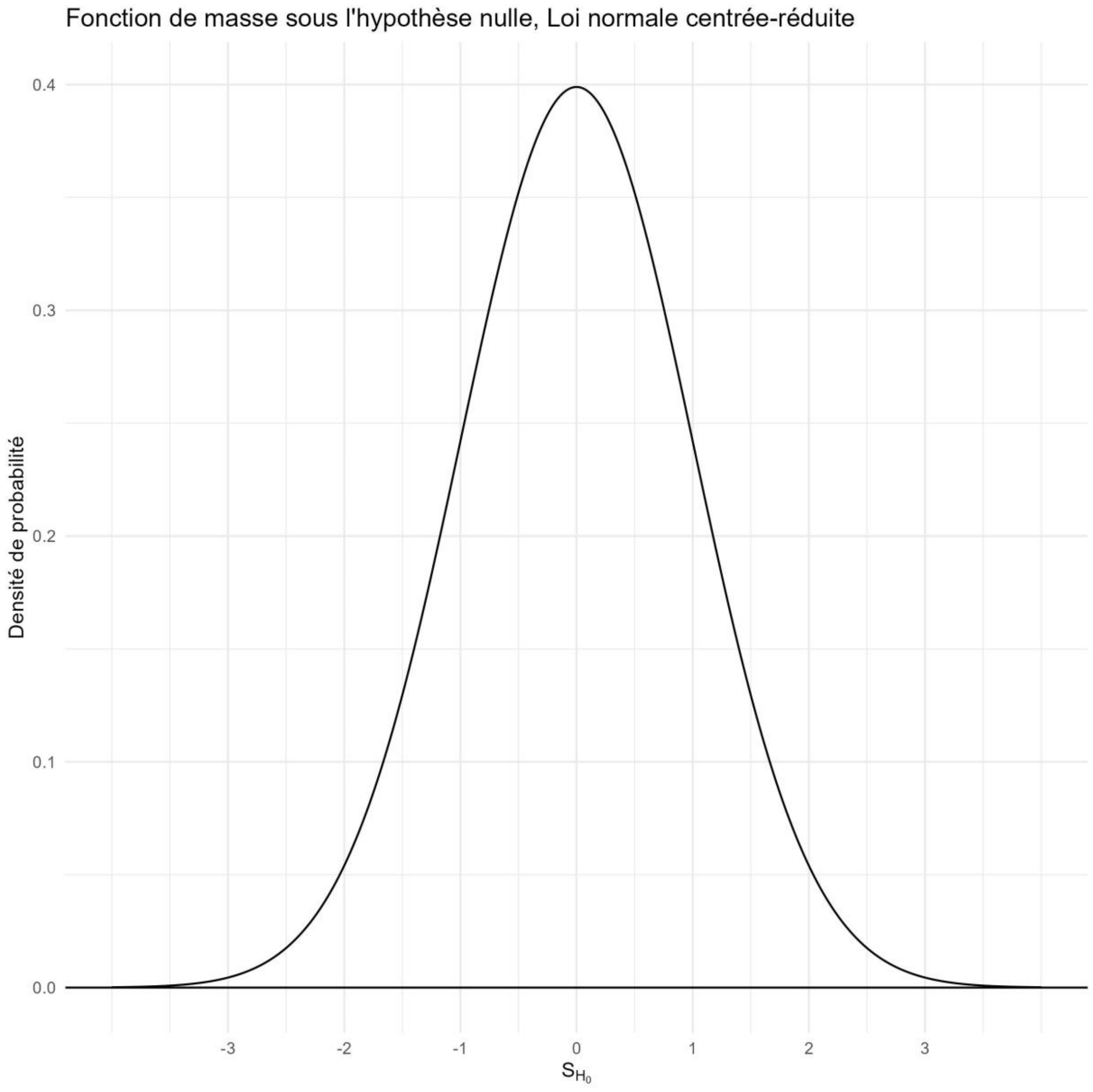
$$S_{H_0} = \frac{|F - 1/2|}{\sqrt{\frac{1}{100 * 4}}} avec\ S_{H_0} \sim \mathcal{N}(0,1)$$

Risques acceptés

- Nous avons deux risques à considérer, avec des conséquences différentes (perte d'argent vs. perte de confiance).
- On préférera contrôler pour la perte de confiance que la perte d'argent dans ce cas.
- On note la probabilité α le risque admissible de perdre la confiance de notre adversaire.
- α est également appelé **risque de première espèce**, **seuil de significativité**
- On aimerait comparer $s_{H_0}^{obs}$ avec α.

Distribution de S_{H_0}

On a $S_{H_0} \sim \mathcal{N}(0,1)$.

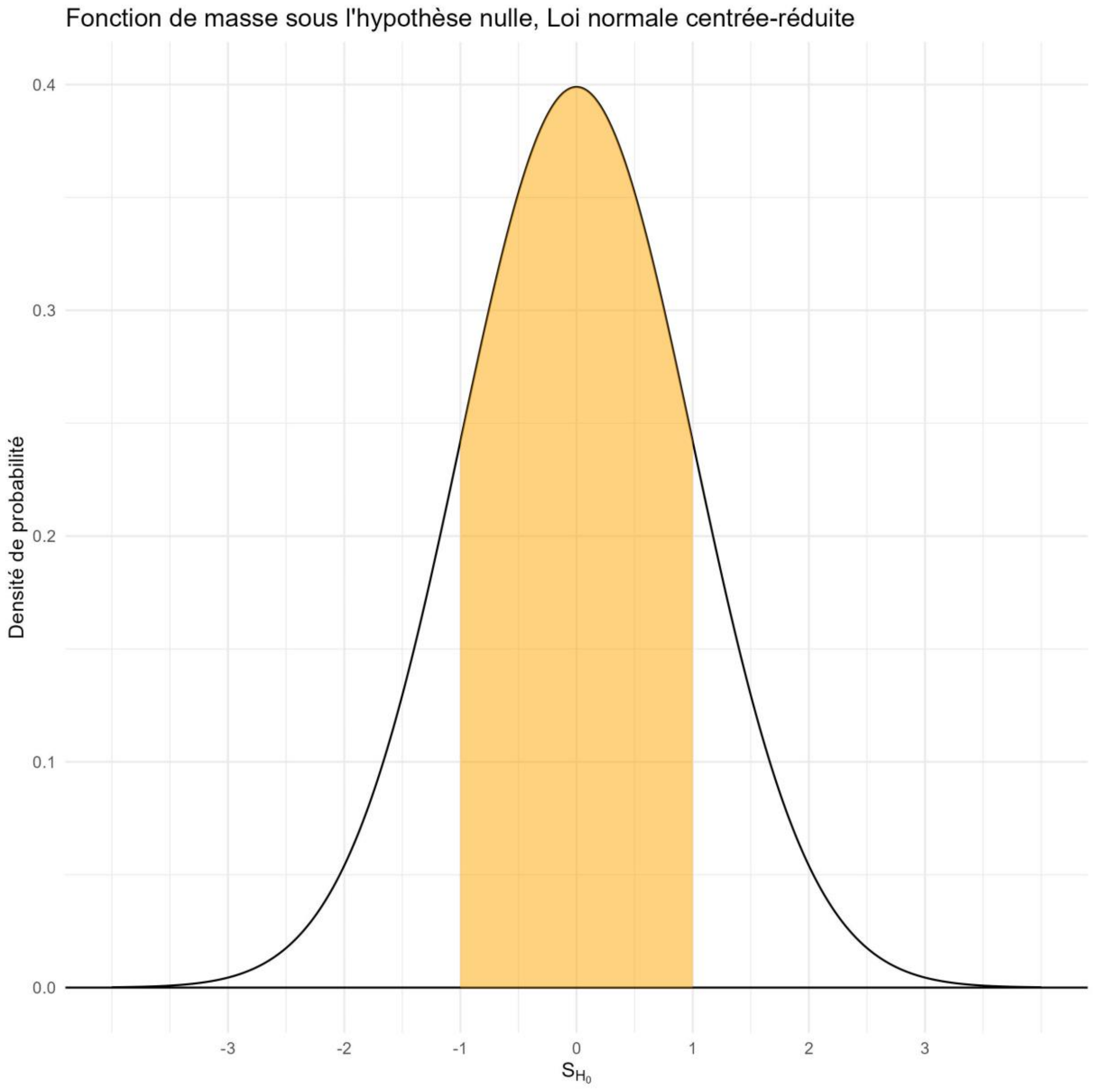


Distribution de S_{H_0}

On a $S_{H_0} \sim \mathcal{N}(0,1)$.

Pour rappel :

- Aire d'une tranche rectangulaire sous la courbe = probabilité
- Aire totale sous la courbe = probabilité totale (1)



Risque α

Aire rouge = α

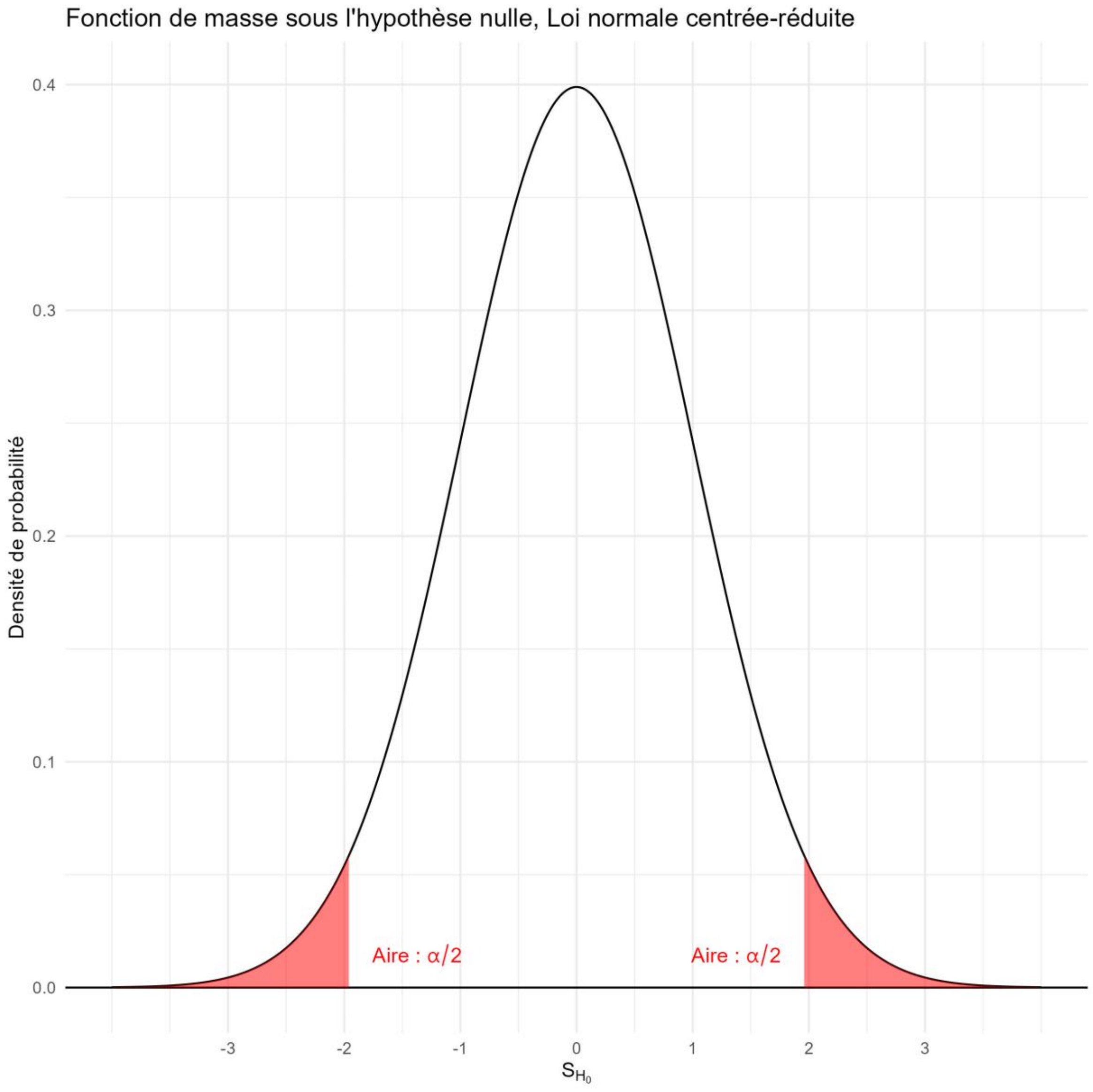
Rappel : $H_0 : \pi = \pi_0$
Donc :

- $H_1 : \pi < \pi_0$ ou
- $H_1 : \pi > \pi_0$

Nécessité de distribuer α pour les deux possibilités

$\alpha/2$ de chaque côté

Test bilatéral : π peut être plus grand ou plus petit que π_0

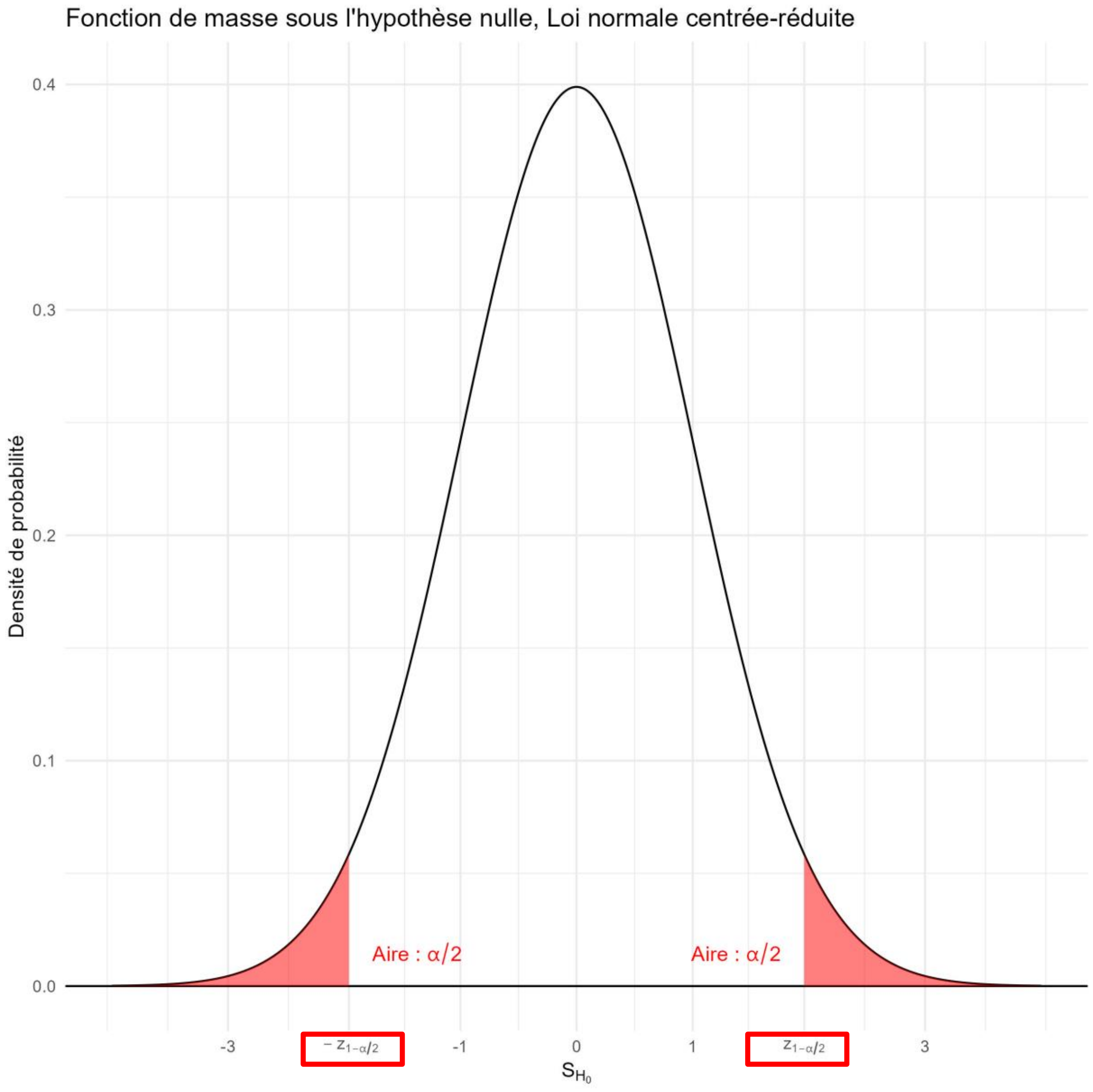


Risque α

On cherche le quantile $z_{1-\alpha/2}$ de la loi normale centrée réduite

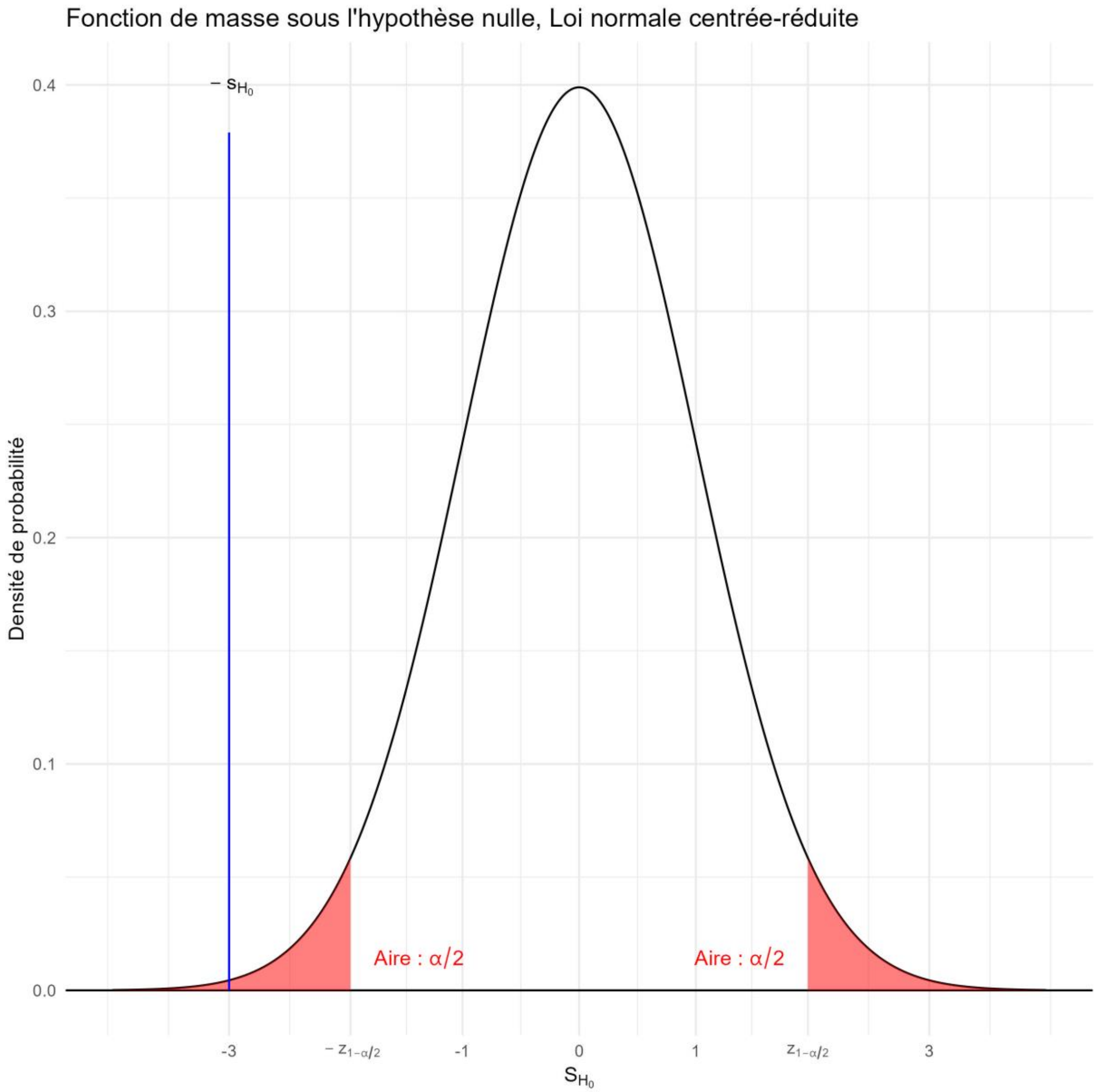
Tel que $P\left(\left|S_{H_0}\right| > z_{1-(\alpha/2)}\right) = \alpha$

Ce quantile est appelé **seuil de rejet**



Probabilité associée à $s_{H_0}^{obs}$

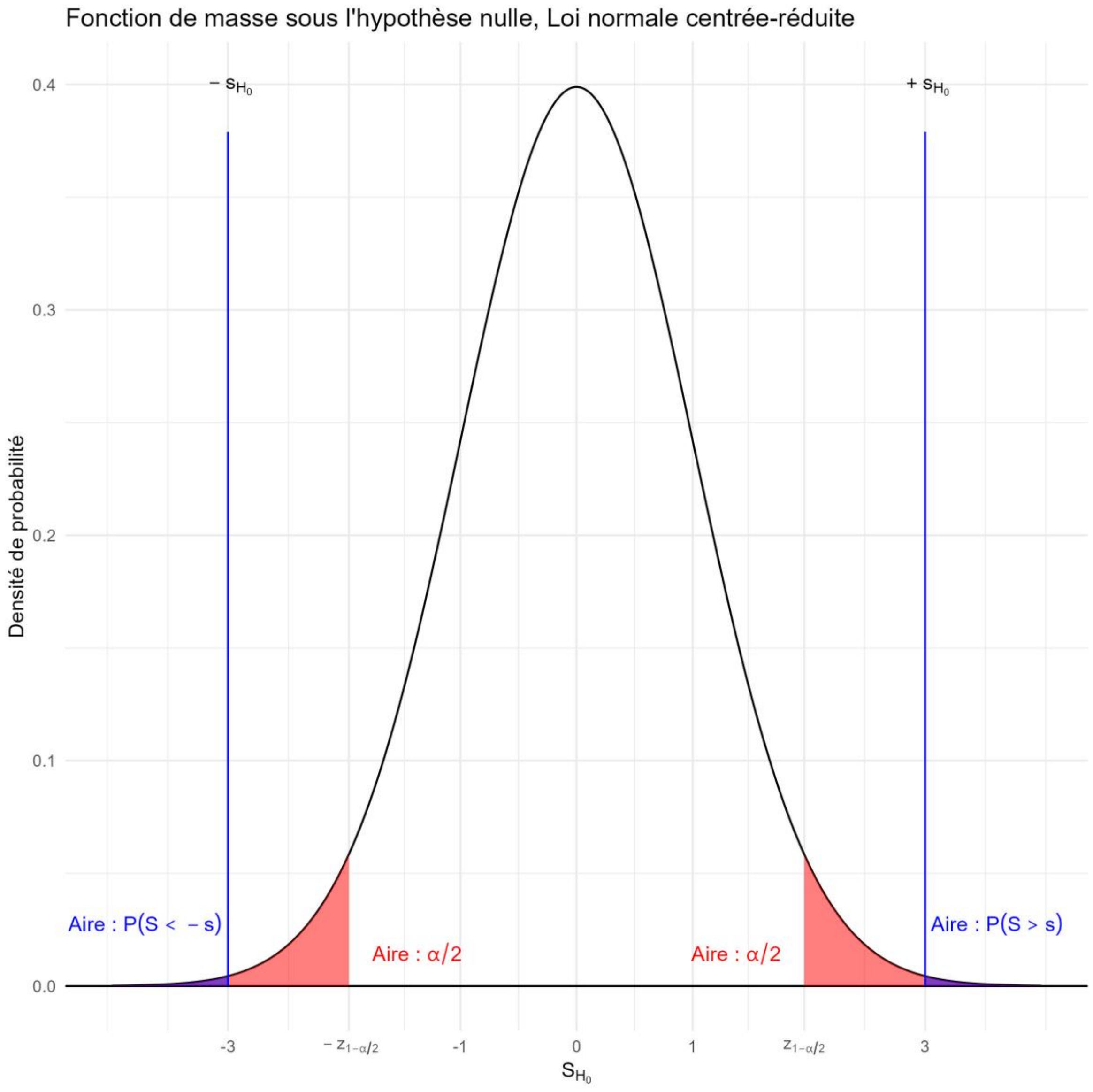
La statistique calculée sur notre échantillon peut être placée en rapport à la distribution de S_{H_0}



Probabilité associée à $s_{H_0}^{obs}$

La loi normale est symétriquement distribuée autour de son espérance : la démarche est strictement identique que S_{H_0} soit positif ou négatif.

On doit en revanche bien considérer les deux cas dans la démarche



Probabilité associée à $s_{H_0}^{obs}$

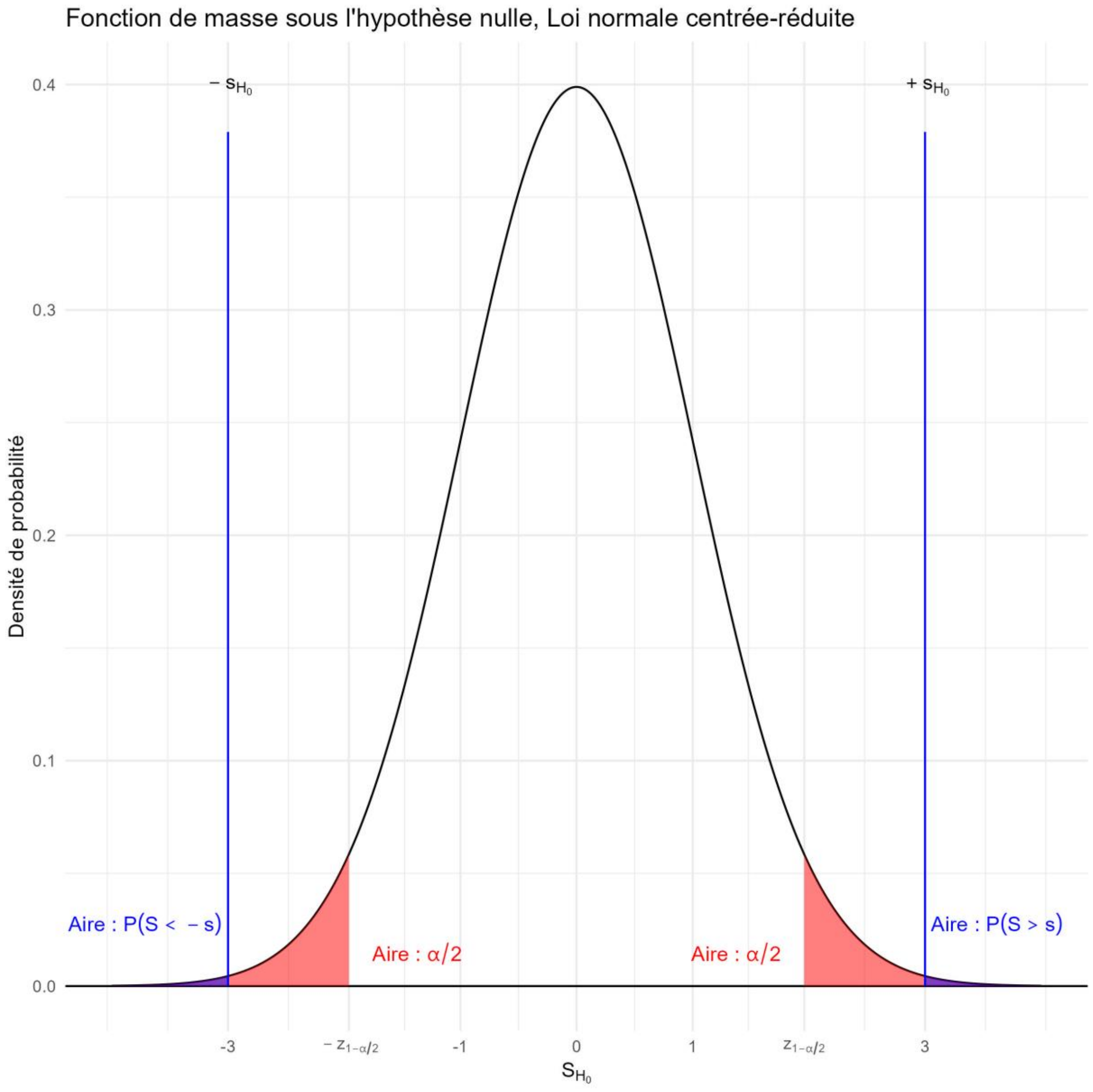
Aire bleue :

$$P\left(\left|S_{H_0}\right| > \left|s_{H_0}^{obs}\right|\right)$$

Si $\left|s_{H_0}^{obs}\right| > z_{1-(\alpha/2)}$

Alors

$$P\left(\left|S_{H_0}\right| > \left|s_{H_0}^{obs}\right|\right) < \alpha$$



Décision sur le test d'hypothèse : valeur du risque α

- On fixe $\alpha = 0{,}05$ ou 5%.
- Cette valeur de α est le défaut en recherche biomédicale.
- Le quantile $z_{1-(\alpha/2)}$ de la loi normale centrée réduite vaut environ 1,96. On appelle ce quantile le **seuil de décision**
- On a donc $\left|s_{H_0}^{obs}\right| = 3 > z_{1-(\alpha/2)} \cong 1{,}96$.

Décision sur le test d’hypothèse : interprétation

« Si l’hypothèse nulle était vraie, la probabilité d’une distance entre une expérience aléatoire de la pièce théorique et l’espérance du résultat de la pièce théorique égale ou supérieure à celle observée est inférieure à 5%. »

Décision sur le test d’hypothèse : interprétation

Si l’hypothèse nulle était vraie, la probabilité d’une distance entre le résultat d’**une expérience aléatoire de la pièce théorique** et **l’espérance du résultat de la pièce théorique** égale ou supérieure à celle observée est inférieure à 5%

- Une série de 100 lancers d’une pièce théorique
- Le résultat moyen à l’issue d’une infinité de séries de 100 lancers d’une pièce théorique

Décision sur le test d’hypothèse : interprétation

Si l’hypothèse nulle était vraie, la probabilité d’une **distance** entre le résultat d’une expérience aléatoire de la pièce théorique et l’espérance du résultat de la pièce théorique égale ou supérieure à celle observée est inférieure à 5%

- La statistique de test calculée à partir de l’espérance d’une pièce théorique et une série unique issue de cette même pièce

Décision sur le test d'hypothèse : interprétation

Si l'**hypothèse nulle était vraie**, la probabilité d'une distance entre une expérience aléatoire de la pièce théorique et l'espérance du résultat de la pièce théorique égale ou supérieure à celle observée est inférieure à 5%

- L'univers hypothétique dans lequel ces expériences sont menées : la pièce est équilibrée

Décision sur le test d'hypothèse : interprétation

Si l'hypothèse nulle était vraie, la probabilité d'une distance entre une expérience aléatoire de la pièce théorique et l'espérance du résultat de la pièce théorique égale ou supérieure à celle observée est **inférieure à 5%**

- Dans ce monde hypothétique, la probabilité de la distance observée si H_0 était vraie est inférieure à notre risque admissible de dire à tort que notre ami est un tricheur.

Décision sur le test d’hypothèse : interprétation

- Une pièce équilibré a moins de 5% de chance de produire un résultat au moins aussi extrême que celui observé.

- On peut donc dire, avec un **risque de faire erreur de 5% au maximum**, que **l’hypothèse nulle est fausse** (la pièce n’est pas équilibrée).

- Si l’hypothèse nulle est fausse, **l’hypothèse alternative est vraie** (H_0 et H_1 sont incompatibles)

- On affirme que la pièce n’est pas équilibrée ($\pi \neq 0{,}5$) avec un risque de 5% de faire erreur

Conclusion

Vous pouvez affirmer, avec 5% de chance de vous tromper, que votre adversaire est un tricheur, et refuser le jeu en bonne conscience.

DEGRÉ DE SIGNIFICATIVITÉ

Degré de significativité

- Prendre la décision via la statistique de test et le quantile $(1 - \alpha/2)$ de la loi normale n'est pas la seule approche.

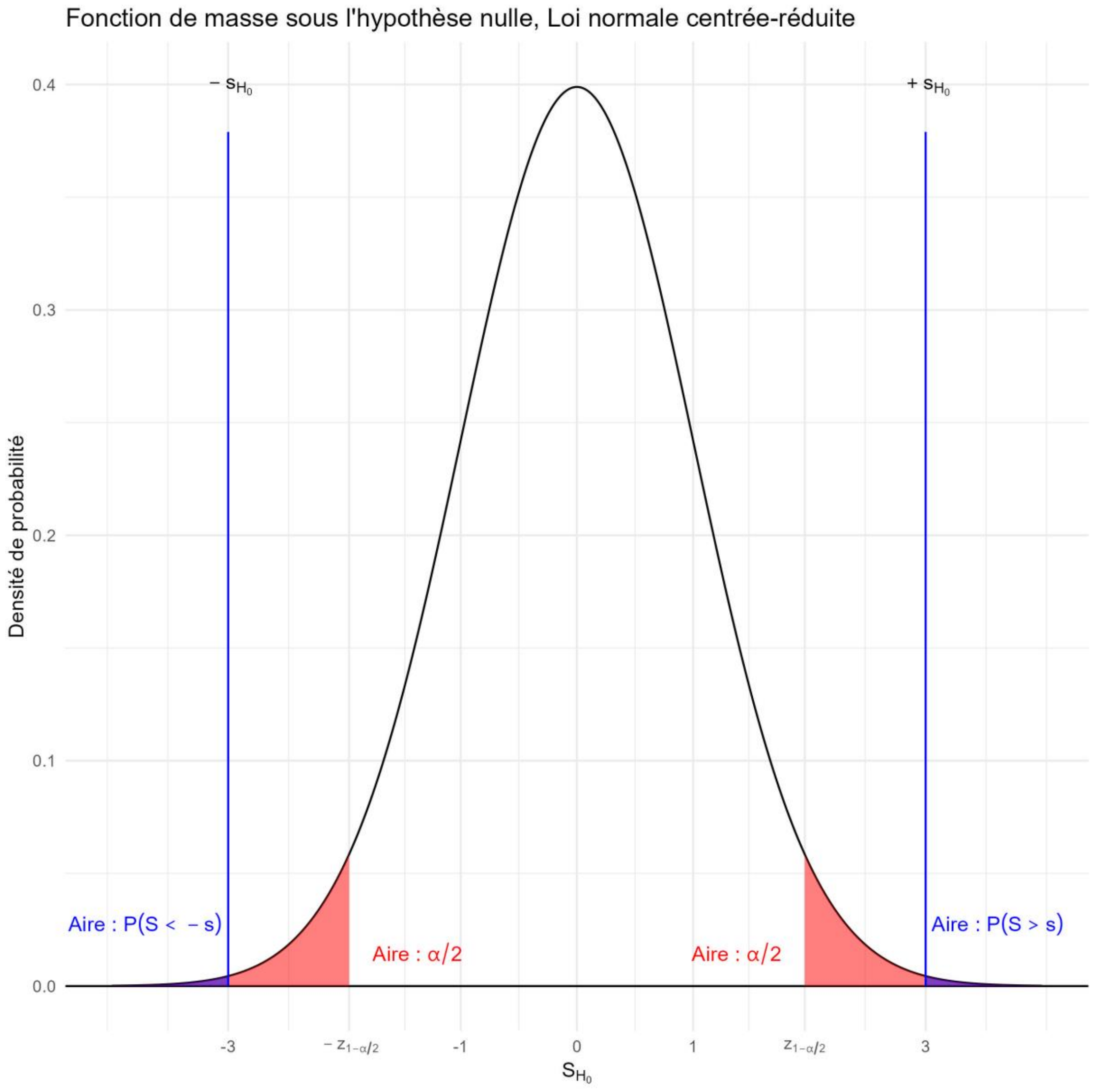
- On peut calculer la probabilité d'obtenir, sous H_0, une statistique de test au moins aussi grande que celle observée.

- On appelle cette probabilité la **p-value** (aussi *petit p*, *p* ou *degré de significativité du test*).

Calcul de la p-value

L'approche de la statistique de test effectue une comparaison sur l'axe des abscisses : « Est-ce que $s_{H_0}^{obs}$ est plus loin du centre que le quantile ? »

L'approche de la p-value effectue une comparaison sur les aires sous la courbe : « Est-ce que l'aire sous la courbe définie par $s_{H_0}^{obs}$ (p-value) est plus petite que α ? »



Calcul de la p-value

- On cherche : $P_{H_0}\left(\left|S_{H_0}\right| > \left|s_{H_0}^{obs}\right|\right)$.
- Deux cas sont à considérer : si $s_{H_0}^{obs}$ est négatif ou positif.
- Le cas $\left|s_{H_0}^{obs}\right| = 0$ est traité avec le négatif.
- *La mise en œuvre de ce calcul n'est pas à connaître*

Calcul de la p-value

s est négatif ou nul

$$P\left(S_{H_0} \le -s_{H_0}^{obs}\right) = \phi\left(-s_{H_0}^{obs}\right) = 1 - \phi\left(\left|s_{H_0}^{obs}\right|\right)$$

s est positif

$$P\left(S_{H_0} > s_{H_0}^{obs}\right) = 1 - P\left(S_{H_0} \le s_{H_0}^{obs}\right) = 1 - \phi\left(\left|s_{H_0}^{obs}\right|\right)$$

Avec $\phi(x)$ la fonction de répartition de la loi normale telle que $P(X \le x) = 1 - P(X > x) = \phi(x) = 1 - \phi(-x)$

$$P\left(\left|S_{H_0}\right| > \left|s_{H_0}^{obs}\right|\right) = P\left(S_{H_0} \le -s_{H_0}^{obs}\right) + P\left(S_{H_0} > s_{H_0}^{obs}\right) = 1 - \phi\left(\left|s_{H_0}^{obs}\right|\right) + 1 - \phi\left(\left|s_{H_0}^{obs}\right|\right) = 2 - 2\phi\left(\left|s_{H_0}^{obs}\right|\right)$$

Décision sur la p-value

- $\phi(x)$ est obtenu dans les tables de répartition de la loi normale centrée réduite (1ère table : $F(X \leq x) = p$).

- On obtient $P\left(\left|S_{H_0}\right| \leq \left|s_{H_0}^{obs}\right|\right) = 0{,}0026 < \alpha = 0{,}05$.

- De façon analogue, on rejette H_0 et on déclare H_1 vraie (avec un risque de 5% de faire erreur).

Interprétation de la p-value

- La **p-value** est définie comme la probabilité d'obtenir un résultat expérimental si l'hypothèse nulle était vraie au moins aussi extrême que le résultat observé dans la réalité.
- **=> Probabilité, post-test, de rejeter H_0 à tort**

- **Cette quantité n'a pas d'autres interprétation formellement correcte.**

- *Une interprétation formellement fausse, mais plus intuitive, est la suivante : la p-value représente la plausibilité du résultat expérimental si l'hypothèse nulle était vraie.*

RAISONNEMENT GÉNÉRAL DES TESTS D’HYPOTHÈSE

Tests d'hypothèse : raisonnement

Monde réel (nous)
π est inconnu
f est estimé via l'expérience
f estime π

Dans notre réalité, l'espérance d'obtenir un pile π n'est pas connue directement : on doit l'estimer à partir des résultats de l'expérience (100 lancers de la pièce)

Tests d'hypothèse : raisonnement

Monde réel (nous)
π est inconnu
f est calculé via l'expérience
f estime π

Monde sous $\boldsymbol{H_0}$
$\pi_0 = 0{,}5$

Dans un monde fictif, simulé pour les besoins du test, nous définissions $\pi = 0{,}5$: la pièce est équilibrée

Tests d'hypothèse : raisonnement

Monde réel (nous)
π est inconnu
f est calculé via l'expérience
f estime π

Monde sous $\boldsymbol{H_0}$
$\pi_0 = 0{,}5$
On définit S la statistique de test
On connait la probabilité p
d'observer une valeur s donnée

Dans le monde fictif, nous sommes **omniscients** par définition (nous décidons que $\pi = 0{,}5$). Nous connaissons la **distribution de la statistique de test**, donc la probabilité d'observer un résultat expérimental donné

Tests d’hypothèse : raisonnement

Monde réel (nous)
π est inconnu
f est calculé via l’expérience
f estime π

Monde sous H_0
$\pi_0 = 0{,}5$
On définit S la statistique de test
On connait la probabilité p
d’observer une valeur s donnée

Calcul de $s_{H_0}^{obs}$

On calcule la **statistique de test s^{obs} dans le monde sous H_0**

Tests d’hypothèse : raisonnement

Monde réel (nous)
π est inconnu
f est calculé via l’expérience
f estime π

$\boldsymbol{s_{H_0}^{obs}}$
« Distance » entre le monde réel et le monde sous H_0

Monde sous $\boldsymbol{H_0}$
$\pi_0 = 0{,}5$
On définit S la statistique de test
On connait la probabilité p d’observer une valeur s donnée

Calcul de $\boldsymbol{s_{H_0}^{obs}}$

$\boldsymbol{s^{obs}}$ peut s’interpréter comme la distance entre le monde de l’expérience réel et le monde si H_0 était vraie

Tests d’hypothèse : raisonnement

Monde réel (nous)
π est inconnu
f est calculé via l’expérience
f estime π

$\boldsymbol{s_{H_0}^{obs}}$
« Distance » entre le monde réel et le monde sous H_0

Monde sous $\boldsymbol{H_0}$
$\pi_0 = 0{,}5$
On définit S la statistique de test
On connait la probabilité p d’observer une valeur s donnée

P-value = $\boldsymbol{P_{H_0}\left(\left|S_{H_0}\right| > \left|s_{H_0}^{obs}\right|\right)}$

La p-value mesure la **plausibilité d’un résultat expérimental dans le monde sous $\boldsymbol{H_0}$ au moins aussi extrême** que celui observé dans le monde réel

Tests d’hypothèse : raisonnement

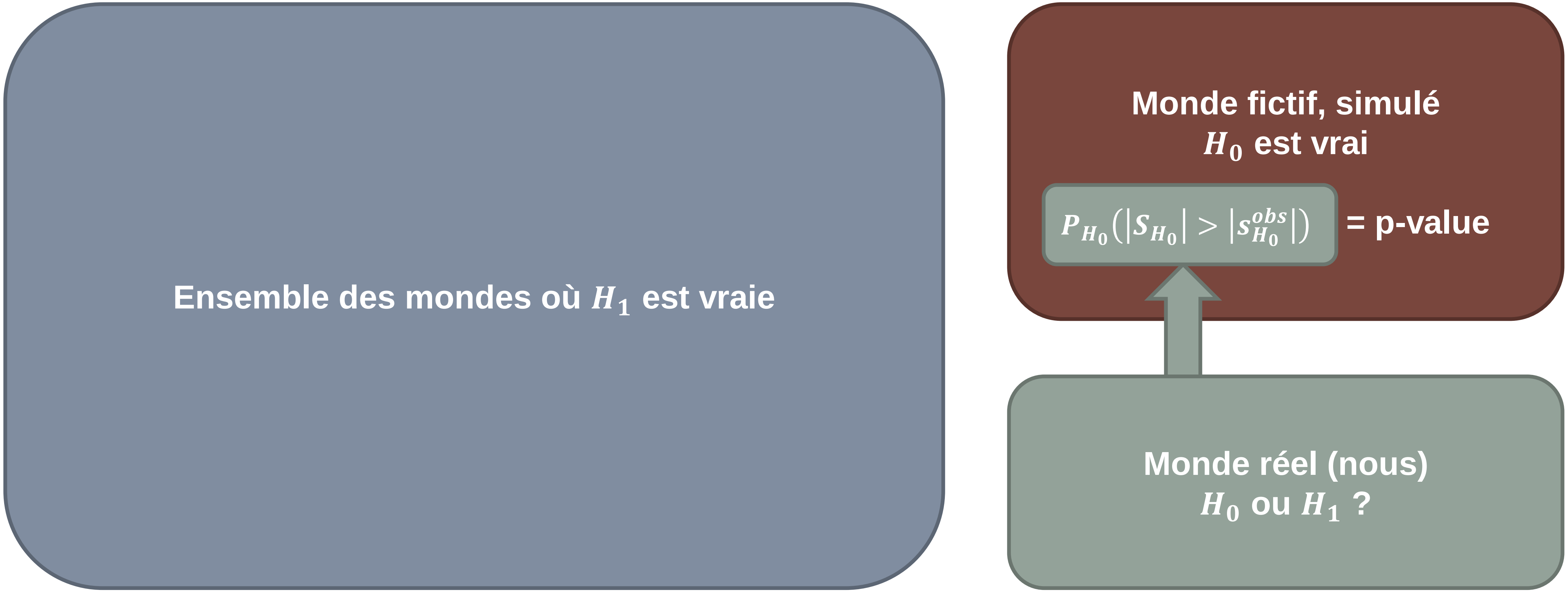
Ensemble des mondes où H_1 est vraie

Monde fictif, simulé
H_0 est vrai

Monde réel (nous)
H_0 ou H_1 ?

Avant le test : on ne sait pas quelle hypothèse est vraie dans notre monde réel

Tests d’hypothèse : raisonnement



P-value > α : on ne sait toujours pas (l’hypothèse nulle est plausible dans notre monde)

Tests d’hypothèse : raisonnement

Ensemble des mondes où H_1 est vraie

Monde réel (nous)
H_0 est fausse donc H_1 est vraie

Monde fictif, simulé
H_0 est vrai

$P_{H_0}\left(\left|S_{H_0}\right| > \left|s_{H_0}^{obs}\right|\right)$ **= p-value**

P-value < α : on rejette H_0 dans notre monde, et on admet H_1 par défaut avec une probabilité α de faire erreur

Tests d'hypothèse en pratique

1. Identifier les hypothèses H_0 et H_1
2. Sélectionner un risque alpha
3. Identifier le type de variables comparées : continues, binaires, de survie
4. Sélectionner le test adéquat
5. Effectuer le calcul de la statistique de test
6. Confronter la p-value au risque alpha
   1. Si p-value < alpha : on rejette H_0 et on accepte H_1 pour vraie => différence statistiquement significative
   2. Si p-value > alpha : on ne peut pas conclure

Différence statistiquement significative et « preuve »

- **On ne parle pas de preuve par le simple résultat du test**

- Le concept de preuve scientifique dépend :
  - Du design de l’expérience (essai clinique)
  - Du bon choix du test
  - De la représentativité de l’échantillon
  - …

- Le test statistique est un des outils au service d’une démarche scientifique

COMPARAISON DE PROPORTIONS

Deux échantillons indépendants : test du Chi-deux

Données de l’exemple

- On interroge au hasard 1000 personnes dans la population résidant en France par téléphone.
- On récupère deux informations sur chaque personne : le sexe, et leur consommation de tabac (fumeur vs. non-fumeur)

Observation	Femme	Homme	
Non-fumeur	408	367	775
Fumeur	102	123	225
	510	490	1000

Question posée

- Existe-t-il une différence **statistiquement significative** dans l'échantillon, au risque $\alpha = 0{,}05$ entre la consommation de tabac chez les hommes et les femmes dans **l'échantillon** ?

- Existe-t-il une différence, au risque $\alpha = 0{,}05$ entre la consommation de tabac chez les hommes et les femmes dans la population française ?

Hypothèses nulle et alternative

- Hypothèse nulle : il n'existe pas de différence.
- $H_0 : \pi_h = \pi_f = \pi_0$ ou alternativement,

- Hypothèse alternative : il existe une différence.
- $H_1 : \pi_h \neq \pi_f \neq \pi_0$

- La statistique de test nécessite de poser l'hypothèse nulle en rapport avec la table de contingence.

Distribution des effectifs attendus sous H_0

- Sous H_0, on a défini $\pi_f = \pi_h = \pi_0$

Sous H_0	Femme	Homme	
Non-fumeur			
Fumeur			
			N

Distribution des proportions sous H_0

- Sous H_0, on a défini $\pi_f = \pi_h = \pi_0$

Sous H_0	Femme	Homme	
Non-fumeur			
Fumeur	π_0	π_0	
			N

Distribution des proportions sous H_0

- Sous H_0, on a défini $\pi_f = \pi_h = \pi_0$

Sous H_0	**Femme**	**Homme**	
Non-fumeur	$1 - \pi_0$	$1 - \pi_0$	
Fumeur	π_0	π_0	
			$\boldsymbol{N}$

Distribution des proportions sous H_0

- Sous H_0, on a défini $\pi_f = \pi_h = \pi_0$

Sous H_0	Femme	Homme	
Non-fumeur	$1 - \pi_0$	$1 - \pi_0$	$\mathbf{1 - \pi_0}$
Fumeur	π_0	π_0	$\boldsymbol{\pi_0}$
			$\boldsymbol{N}$

Distribution des proportions sous H_0

- π_0 n'est pas connu. On utilise son estimateur F_0
- On calcule son estimation : $f_0 = 0{,}2245$

Sous H_0	Femme	Homme	
Non-fumeur	$1 - f_0$	$1 - f_0$	$\mathbf{1 - f_0}$
Fumeur	f_0	f_0	$\boldsymbol{f_0}$
			$\boldsymbol{N}$

Distribution des proportions sous H_0

- Sous H_0, on connait les probabilités dans chaque catégorie. On peut calculer des effectifs attendus :

Sous H_0	Femme	Homme	
Non-fumeur	$\mathrm{E}_{f0} = n_f(1 - f_0)$	$\mathrm{E}_{h0} = n_h(1 - f_0)$	$1 - f_0$
Fumeur	$\mathrm{E}_{f1} = n_f f_0$	$\mathrm{E}_{h1} = n_h f_0$	f_0
	$n_f = 510$	$n_h = 490$	$n = 1000$

Distribution des proportions sous H_0

- Sous H_0, on connait les probabilités dans chaque catégorie. On peut calculer des effectifs attendus :

Sous H_0	**Femme**	**Homme**	
Non-fumeur	$\mathrm{E}_{f0} = 395{,}5$	$\mathrm{E}_{h0} = 380$	$\mathbf{775}$
Fumeur	$\mathrm{E}_{f1} = 114{,}5$	$\mathrm{E}_{h1} = 110$	$\mathbf{225}$
	$\boldsymbol{n_f = 510}$	$\boldsymbol{n_h = 490}$	$\boldsymbol{n = 1000}$

Ecart entre l'observé et l'attendu sous H_0

- Comment mesurer une distance entre la distribution empirique et la distribution théorique sous H_0 ?

Observation	Femme	Homme	
Non-fumeur	408	367	775
Fumeur	102	123	225
	510	490	1000

Sous H_0	Femme	Homme	
Non-fumeur	395,5	380	775
Fumeur	114,5	110	225
	510	490	1000

Statistique du χ^2

- La statistique du χ^2 (lire « Chi-deux » ou « Chi-carré ») est définie ainsi :

$$S = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}} \sim \chi^2_{(I-1)\times(J-1)}$$

Statistique du χ^2

- La statistique du χ^2 (lire « Chi-deux » ou « Chi-carré ») est définie ainsi :

$$S = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}} \sim \chi^2_{(I-1)\times(J-1)}$$

- $(O - E)^2$: distance positive, cellule par cellule, entre l'observé et l'attendu
- E : effectif attendu

Statistique du χ^2

- La statistique du χ^2 (lire « Chi-deux » ou « Chi-carré ») est définie ainsi :

$$S = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}} \sim \chi^2_{(I-1)\times(J-1)}$$

La démonstration n'est pas au programme et n'est pas traitée

Statistique du $\mathcal{X}^2$

$$S = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{\left(O_{ij}-E_{ij}\right)^2}{E_{ij}} \sim \mathcal{X}^2_{(I-1)\times(J-1)}$$

- Condition d'application : $E_{ij} \geq 5$ nécessaire pour l'usage de la loi du Chi-deux

- *Note : une VA suivant une loi du χ^2 à k degrés de liberté est une somme de k VAs suivant une LNCR, $E_{ij} \geq 5$ est dérivé des conditions d'application du TCL*

Statistique du χ^2

$$S = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}} \sim \chi^2_{(I-1)\times(J-1)}$$

- Notations :
  - I nombre de modalités de la première variable
  - J nombre de modalités de la deuxième variable
  - O_{ij} effectif observé dans la cellule ij de la table de contingence
  - E_{ij} effectif attendu sous H_0 dans la cellule ij de la table de contingence
  - $\chi^2_{(I-1)\times(J-1)}$ loi du Chi-deux à $(I-1)\times(J-1)$ degrés de liberté

Statistique du χ^2

- Degrés de liberté (ddl) : combien de modalités d'une variable doivent être connues pour déterminer complètement les autres ?

- Sexe : 2 modalités : $I = 2$
- Consommation de tabac : 2 modalités : $J = 2$

Observation	Femme	Homme	
Non-fumeur			$a + b = n_0$
Fumeur			$c + d = n_1$
	$a + c = n_f$	$b + d = n_h$	$n_f + n_h = n_0 + n_1 = N$

Statistique du $\mathcal{X}^2$

- Exemple : si n_f est connu, on détermine le reste des effectifs pour la variable Sexe : $n_h = N - n_f$
- La relation $n_f + n_h = N$ est une **contrainte**
- En règle générale, le degré de liberté peut se définir comme le nombre de modalités moins le nombre de contraintes

Observation	Femme	Homme	
Non-fumeur			$a + b = n_0$
Fumeur			$c + d = n_1$
	$a + c = n_f$	$b + d = n_h$	$n_f + n_h = n_0 + n_1 = N$

Statistique du χ^2

- Dans le cadre du test du χ^2, chaque variable dispose **d'une seule contrainte** : la distribution marginale des effectifs de cette variable ($n_0 + n_1 = n_f + n_h = N$)
- Donc :
  - Pour le Sexe : 2 modalités et 1 contrainte : $I - 1$
  - Pour la Consommation de tabac : 2 modalités et 1 contrainte : $J - 1$
  - Au total pour la table : $(I - 1) \times (J - 1) = 1 \times 1 = 1\ ddl$

Observation	Femme	Homme	
Non-fumeur			$a + b = n_0$
Fumeur			$c + d = n_1$
	$a + c = n_f$	$b + d = n_h$	$n_f + n_h = n_0 + n_1 = N$

Calcul de la statistique de test

- On dispose donc du degré de liberté du test $ddl = 1$
- On dispose de l'ensemble des comptes attendus : table de contingence sous H_0
- On peut fixer la statistique de test :

$$S = \sum_{i=1}^{I=2} \sum_{j=1}^{J=2} \frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}} \sim \mathcal{X}^2_{1\,ddl}$$

Calcul de la statistique de test

$$S = \frac{(O_{11} - E_{11})^2}{E_{11}} + \frac{(O_{12} - E_{12})^2}{E_{12}} + \frac{(O_{21} - E_{21})^2}{E_{21}} + \frac{(O_{22} - E_{22})^2}{E_{22}}$$

$$s^{obs} = \frac{(408 - 395{,}5)^2}{395{,}5} + \frac{(102 - 114{,}5)^2}{114{,}5} + \frac{(367 - 380)^2}{380} + \frac{(123 - 110)^2}{110}$$

$$s^{obs} = 3{,}74$$

Calcul de la statistique de test

- On dispose de la valeur de la statistique de test pour notre expérience : $s^{obs} = 3,74$
- On a besoin du quantile d'ordre $q_{1-\alpha}$ de la loi $\mathcal{X}_1^2$, ou loi du Chi-deux à 1 degré de liberté
- Ce quantile $q_{1-\alpha} \cong 3,84$ (obtenu dans la table de la fonction de répartition de la loi du Chi-deux)

- Donc $\boldsymbol{s^{obs} < q_{1-\alpha}}$, on ne rejette pas l'hypothèse nulle.

Déterminer la valeur seuil

Table 4 du formulaire

Fractiles de la loi du χ^2

Soit X une variable aléatoire suivant une loi du χ^2 à n degrés de liberté. Pour une probabilité p donnée, la table donne la valeur x telle que $P(X \leq x) = p$

ddl	p	0,005	0,010	0,025	0,050	0,100	0,250	0,500	0,750	0,900	0,950	0,975	0,990	0,999
1		0,0000	0,0002	0,0010	0,0039	0,0158	0,1015	0,4549	1,3233	2,7055	3,8415	5,0239	6,6349	10,8276
2		0,0100	0,0201	0,0506	0,1026	0,2107	0,5754	1,3863	2,7726	4,6052	5,9915	7,3778	9,2103	13,8155
3		0,0717	0,1148	0,2158	0,3518	0,5844	1,2125	2,3660	4,1083	6,2514	7,8147	9,3484	11,3449	16,2662
4		0,2070	0,2971	0,4844	0,7107	1,0636	1,9226	3,3567	5,3853	7,7794	9,4877	11,1433	13,2767	18,4668
5		0,4117	0,5543	0,8312	1,1455	1,6103	2,6746	4,3515	6,6257	9,2364	11,0705	12,8325	15,0863	20,5150
6		0,6757	0,8721	1,2373	1,6354	2,2041	3,4546	5,3481	7,8408	10,6446	12,5916	14,4494	16,8119	22,4577
7		0,9893	1,2390	1,6899	2,1673	2,8331	4,2549	6,3458	9,0371	12,0170	14,0671	16,0128	18,4753	24,3219
8		1,3444	1,6465	2,1797	2,7326	3,4895	5,0706	7,3441	10,2189	13,3616	15,5073	17,5345	20,0902	26,1245
9		1,7349	2,0879	2,7004	3,3251	4,1682	5,8988	8,3428	11,3888	14,6837	16,9190	19,0228	21,6660	27,8772
10		2,1559	2,5582	3,2470	3,9403	4,8652	6,7372	9,3418	12,5489	15,9872	18,3070	20,4832	23,2093	29,5883
11		2,6032	3,0535	3,8157	4,5748	5,5778	7,5841	10,3410	13,7007	17,2750	19,6751	21,9200	24,7250	31,2641
12		3,0738	3,5706	4,4038	5,2260	6,3038	8,4384	11,3403	14,8454	18,5493	21,0261	23,3367	26,2170	32,9095
13		3,5650	4,1069	5,0088	5,8919	7,0415	9,2991	12,3398	15,9839	19,8119	22,3620	24,7356	27,6882	34,5282

Déterminer la valeur seuil

Table 4 du formulaire

Les lignes correspondent au degré de liberté du test

Fractiles de la loi du χ^2

Soit X une variable aléatoire suivant une loi du χ^2 à n degrés de liberté. Pour une probabilité p donnée, la table donne la valeur x telle que $P(X \leq x) = p$

ddl	p	0,005	0,010	0,025	0,050	0,100	0,250	0,500	0,750	0,900	0,950	0,975	0,990	0,999
1		0,0000	0,0002	0,0010	0,0039	0,0158	0,1015	0,4549	1,3233	2,7055	3,8415	5,0239	6,6349	10,8276
2		0,0100	0,0201	0,0506	0,1026	0,2107	0,5754	1,3863	2,7726	4,6052	5,9915	7,3778	9,2103	13,8155
3		0,0717	0,1148	0,2158	0,3518	0,5844	1,2125	2,3660	4,1083	6,2514	7,8147	9,3484	11,3449	16,2662
4		0,2070	0,2971	0,4844	0,7107	1,0636	1,9226	3,3567	5,3853	7,7794	9,4877	11,1433	13,2767	18,4668
5		0,4117	0,5543	0,8312	1,1455	1,6103	2,6746	4,3515	6,6257	9,2364	11,0705	12,8325	15,0863	20,5150
6		0,6757	0,8721	1,2373	1,6354	2,2041	3,4546	5,3481	7,8408	10,6446	12,5916	14,4494	16,8119	22,4577
7		0,9893	1,2390	1,6899	2,1673	2,8331	4,2549	6,3458	9,0371	12,0170	14,0671	16,0128	18,4753	24,3219
8		1,3444	1,6465	2,1797	2,7326	3,4895	5,0706	7,3441	10,2189	13,3616	15,5073	17,5345	20,0902	26,1245
9		1,7349	2,0879	2,7004	3,3251	4,1682	5,8988	8,3428	11,3888	14,6837	16,9190	19,0228	21,6660	27,8772
10		2,1559	2,5582	3,2470	3,9403	4,8652	6,7372	9,3418	12,5489	15,9872	18,3070	20,4832	23,2093	29,5883
11		2,6032	3,0535	3,8157	4,5748	5,5778	7,5841	10,3410	13,7007	17,2750	19,6751	21,9200	24,7250	31,2641
12		3,0738	3,5706	4,4038	5,2260	6,3038	8,4384	11,3403	14,8454	18,5493	21,0261	23,3367	26,2170	32,9095
13		3,5650	4,1069	5,0088	5,8919	7,0415	9,2991	12,3398	15,9839	19,8119	22,3620	24,7356	27,6882	34,5282

Déterminer la valeur seuil

Table 4 du formulaire

Les lignes correspondent au degré de liberté du test

Les colonnes correspondent au risque α retenu

Le quantile $q_{1-\alpha} \cong 3{,}8415$ est notre valeur seuil pour un test du χ^2 à 1 degré de liberté

Note : $z^2_{1-\frac{\alpha}{2}} = q^{\chi^2_{1\,ddl}}_{1-\alpha}$

Fractiles de la loi du χ^2

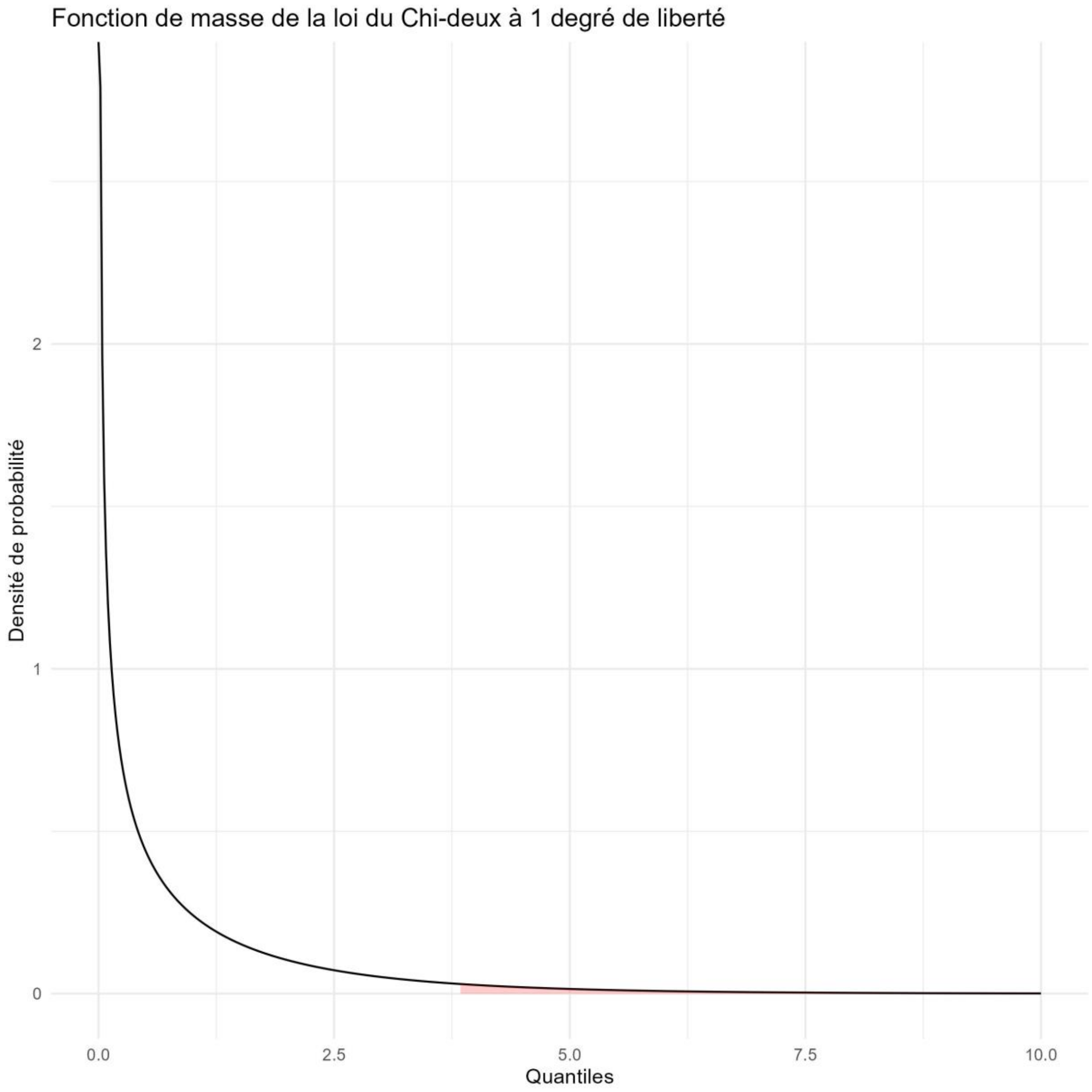
Soit X une variable aléatoire suivant une loi du χ^2 à n degrés de liberté. Pour une probabilité p donnée, la table donne la valeur x telle que $P(X \leq x) = p$

ddl	p	0,005	0,010	0,025	0,050	0,100	0,250	0,500	0,750	0,900	0,950	0,975	0,990	0,999
1		0,0000	0,0002	0,0010	0,0039	0,0158	0,1015	0,4549	1,3233	2,7055	3,8415	5,0239	6,6349	10,8276
2		0,0100	0,0201	0,0506	0,1026	0,2107	0,5754	1,3863	2,7726	4,6052	5,9915	7,3778	9,2103	13,8155
3		0,0717	0,1148	0,2158	0,3518	0,5844	1,2125	2,3660	4,1083	6,2514	7,8147	9,3484	11,3449	16,2662
4		0,2070	0,2971	0,4844	0,7107	1,0636	1,9226	3,3567	5,3853	7,7794	9,4877	11,1433	13,2767	18,4668
5		0,4117	0,5543	0,8312	1,1455	1,6103	2,6746	4,3515	6,6257	9,2364	11,0705	12,8325	15,0863	20,5150
6		0,6757	0,8721	1,2373	1,6354	2,2041	3,4546	5,3481	7,8408	10,6446	12,5916	14,4494	16,8119	22,4577
7		0,9893	1,2390	1,6899	2,1673	2,8331	4,2549	6,3458	9,0371	12,0170	14,0671	16,0128	18,4753	24,3219
8		1,3444	1,6465	2,1797	2,7326	3,4895	5,0706	7,3441	10,2189	13,3616	15,5073	17,5345	20,0902	26,1245
9		1,7349	2,0879	2,7004	3,3251	4,1682	5,8988	8,3428	11,3888	14,6837	16,9190	19,0228	21,6660	27,8772
10		2,1559	2,5582	3,2470	3,9403	4,8652	6,7372	9,3418	12,5489	15,9872	18,3070	20,4832	23,2093	29,5883
11		2,6032	3,0535	3,8157	4,5748	5,5778	7,5841	10,3410	13,7007	17,2750	19,6751	21,9200	24,7250	31,2641
12		3,0738	3,5706	4,4038	5,2260	6,3038	8,4384	11,3403	14,8454	18,5493	21,0261	23,3367	26,2170	32,9095
13		3,5650	4,1069	5,0088	5,8919	7,0415	9,2991	12,3398	15,9839	19,8119	22,3620	24,7356	27,6882	34,5282

Risque α sous la loi du $\mathcal{X}_1^2$

Aire en rouge : $\alpha = 5\%$

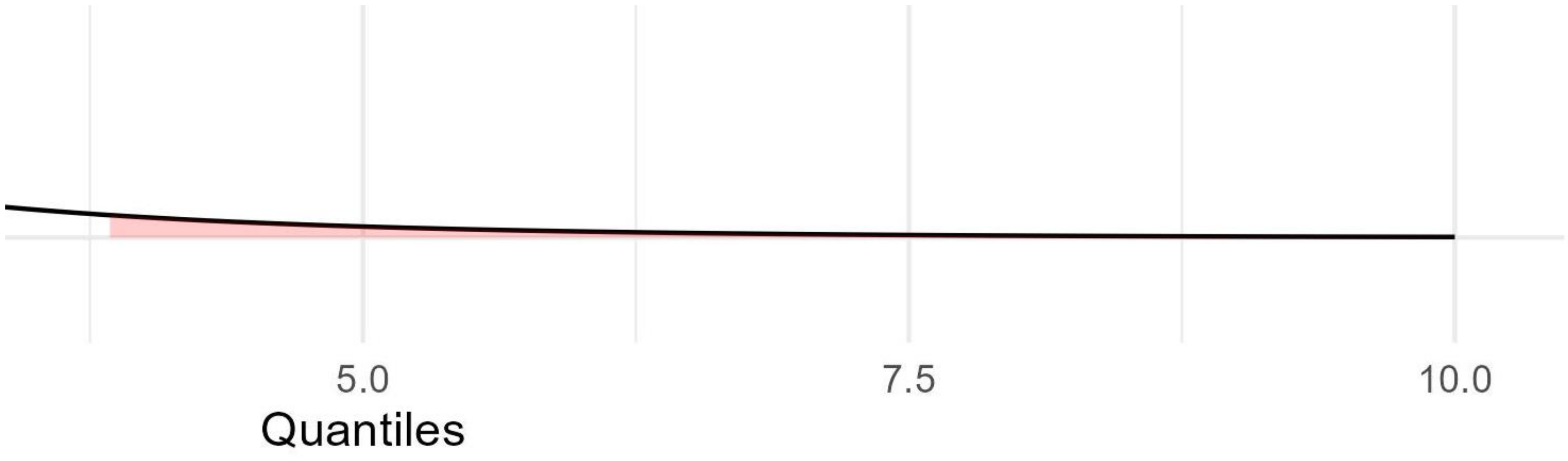
Loi asymétrique => risque α à droite



Risque α sous la loi du $\mathcal{X}_1^2$

Aire en rouge : $\alpha = 5\%$

Loi asymétrique => risque α à droite



Test du Chi-deux pour la comparaison de deux proportions observées

Statistique de test :

$$S_{H_0} = \sum_{i=1}^{I=2} \sum_{j=1}^{J=2} \frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}} \sim \chi^2_{1\,ddl}$$

$$q_{1-\alpha} \cong 3{,}84$$

Note : on cherche le quantile au niveau $1-\alpha$ ici, et non $1-\frac{\alpha}{2}$ du fait des propriétés de la loi du χ^2

- O_{ij} : effectif observé dans la cellule ij du tableau de contingence
- E_{ij} : effectif attendu sous H_0 dans la cellule ij du tableau de contingence

- **Conditions** : $E_{ij} \geq 5$ dans toutes les cellules

COMPARAISON DE PROPORTIONS

Un échantillon à une valeur de référence : test du Chi-deux

Données de l’exemple : reformulation

- On retourne à l’exemple introductif : 100 lancers successifs d’une pièce
- Les observations sont les suivantes

Observation	Pile	Face	
n	35	65	100

Question posée

- Peut-on affirmer que la pièce est déséquilibrée au risque $\alpha = 0{,}05$?

Hypothèses nulle et alternative

- Hypothèse nulle : il n'existe pas de différence.
- $H_0 : \pi = \pi_0 = 0{,}5$ ou alternativement,

- Hypothèse alternative : il existe une différence.
- $H_1 : \pi \neq \pi_0 \neq 0{,}5$

Distribution des proportions sous H_0

- Sous H_0, on a défini $\pi = \pi_0 = 0{,}5$

Observation	Pile	Face	
n	35	65	100

Sous H_0	Pile	Face	
n	$E_1 = n\pi_0 = 50$	$E_2 = n(1 - \pi_0) = 50$	100

- Le nombre de degrés de liberté est de 1 : 2 modalités (Pile ou Face) pour 1 contrainte (N)

Calcul de la statistique de test

- On dispose donc du degré de liberté du test $ddl = 1$
- On dispose de l'ensemble des comptes attendus : table de contingence sous H_0
- On peut fixer la statistique de test :

$$S = \frac{(O_2 - E_2)^2}{E_2} + \frac{(O_1 - E_1)^2}{E_1} \sim \mathcal{X}^2_{1\,ddl}$$

Calcul de la statistique de test

$$S = \frac{(O_2 - E_2)^2}{E_2} + \frac{(O_1 - E_1)^2}{E_1}$$

$$s^{obs} = \frac{(65 - 50)^2}{50} + \frac{(35 - 50)^2}{50}$$

$$s^{obs} = 9$$

Calcul de la statistique de test

- On dispose de la valeur de la statistique de test pour notre expérience : $s^{obs} = 9$
- On a besoin du quantile d'ordre $q_{1-\alpha}$ de la loi $\mathcal{X}_1^2$, ou loi du Chi-deux à 1 degré de liberté
- Ce quantile $q_{1-\alpha} \cong 3{,}84$ (obtenu dans la table de la fonction de répartition de la loi du Chi-deux)

- Donc $\boldsymbol{s^{obs} > q_{1-\alpha}}$, on rejette l'hypothèse nulle au risque α de faire erreur.

Test du Chi-deux pour la comparaison de deux proportions observées

Statistique de test :

$$S_{H_0} = \frac{(O_2 - E_2)^2}{E_2} + \frac{(O_1 - E_1)^2}{E_1} \sim \mathcal{X}^2_{1\,ddl}$$

$$q_{1-\alpha} \cong 3{,}84$$

Note : Cette formule est disponible dans votre formulaire

- O_i : effectif observé dans la cellule i du tableau de contingence
- E_i : effectif attendu sous H_0 dans la cellule i du tableau de contingence
- **Conditions** : $E_i \geq 5$ dans toutes les cellules

MERCI POUR VOTRE ATTENTION