Statistique

Benjamin Bobbia

ISAE

-1.1 Tests statistiques

raisonnement par l'absurde mais probabiliste

Motivation à partir d'un intervalle de confiance

Reprenons l'étude de la proportion $p \in]0,1[$ d'une population qui présente une mutation génétique particulière. La généticienne a de bonnes raisons de penser que la moitié de la population a cette mutation dans son génome. Elle souhaite donc **valider** cette hypothèse « $p = 1/2$ ». Si elle est amenée à **rejeter** son hypothèse de travail, elle en conclura que la mutation est sur-représentée ($p > 1/2$) ou sous-représentée ($p < 1/2$), *i.e.* « $p \neq 1/2$ ».

En termes statistiques, nous disons qu'elle veut **tester** l'**hypothèse nulle**

$$H_0 \ : \ p = \frac{1}{2}$$

contre l'**hypothèse alternative**

$$H_1 \ : \ p \neq \frac{1}{2}.$$

Motivation à partir d'un intervalle de confiance

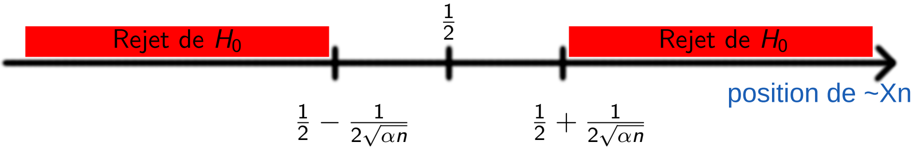
Grâce à son étude statistique précédente effectuée sur n individus tirés uniformément avec remise, elle dispose de l'estimateur $\overline{X}_n$ de p et de l'intervalle de confiance de niveau $1-\alpha \in]0,1[$ donné par Bienaymé-Tchebychev,

$$\left]\overline{X}_n - \frac{1}{2\sqrt{\alpha n}} ; \overline{X}_n + \frac{1}{2\sqrt{\alpha n}}\right[.$$

Elle décide donc d'**accepter l'hypothèse nulle** si

$$\left|\overline{X}_n - \frac{1}{2}\right| < \frac{1}{2\sqrt{\alpha n}}$$

et de **rejeter l'hypothèse nulle** si ce n'est pas le cas.



Motivation à partir d'un intervalle de confiance

Si l'hypothèse nulle H_0 est vraie, alors $p = 1/2$ et la probabilité de prendre la bonne décision est donnée par

$$\mathbb{P}_{H_0}\left(\frac{1}{2} \in \left]\overline{X}_n - \frac{1}{2\sqrt{\alpha n}}; \overline{X}_n + \frac{1}{2\sqrt{\alpha n}}\right[\right) \geqslant 1 - \alpha.$$

Autrement dit, la généticienne fera une **erreur** en rejetant l'hypothèse H_0 si elle est vraie avec une probabilité inférieure à α.

Si l'hypothèse alternative H_1 est vraie, alors $p \neq 1/2$ et il faudrait idéalement rejeter l'hypothèse nulle. Cela a lieu avec une probabilité donnée par

$$\mathbb{P}_{H_1}\left(\frac{1}{2} \notin \left]\overline{X}_n - \frac{1}{2\sqrt{\alpha n}}; \overline{X}_n + \frac{1}{2\sqrt{\alpha n}}\right[\right)$$

noté comme une proba conditionnelle
mais ne l'est pas car H1 n'est pas aléatoire,
c'est une hypothèse

$$= 1 - \mathbb{P}_{H_1}\left(\frac{1}{2} - \frac{1}{2\sqrt{\alpha n}} \leqslant \overline{X}_n \leqslant \frac{1}{2} + \frac{1}{2\sqrt{\alpha n}}\right).$$

Motivation à partir d'un intervalle de confiance

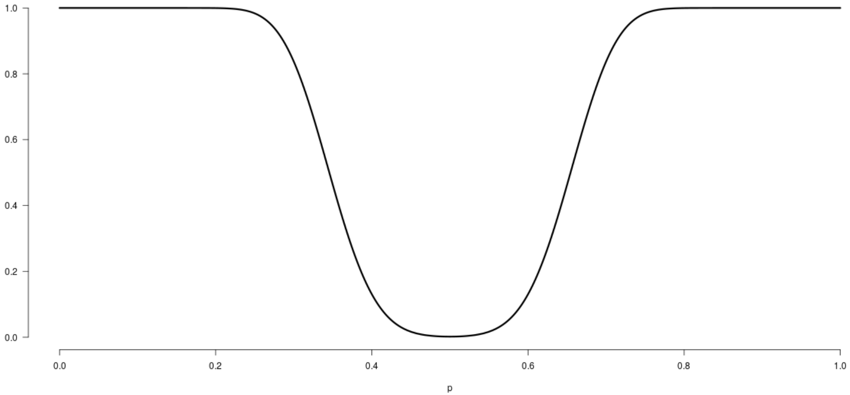
Par définition, $n\overline{X}_n$ est le nombre d'individus présentant la mutation génétique parmi les n individus tirés uniformément avec remise. La loi de cette variable aléatoire à valeurs dans $\{0, \ldots, n\}$ est une loi binomiale $\mathcal{B}(n, p)$.

En notant $F_{n,p}$ la fonction de répartition de la loi $\mathcal{B}(n, p)$, nous obtenons que la probabilité de **rejeter l'hypothèse nulle H_0 si l'hypothèse alternative H_1 est vraie** vaut

$$\begin{aligned}
&\mathbb{P}_{H_1}\left(\frac{1}{2} \notin \left]\overline{X}_n - \frac{1}{2\sqrt{\alpha n}} ; \overline{X}_n + \frac{1}{2\sqrt{\alpha n}}\right[\right) \\
&= 1 - F_{n,p}\left(\frac{n}{2} + \frac{\sqrt{n}}{2\sqrt{\alpha}}\right) + F_{n,p}\left(\frac{n}{2} - \frac{\sqrt{n}}{2\sqrt{\alpha}}\right).
\end{aligned}$$

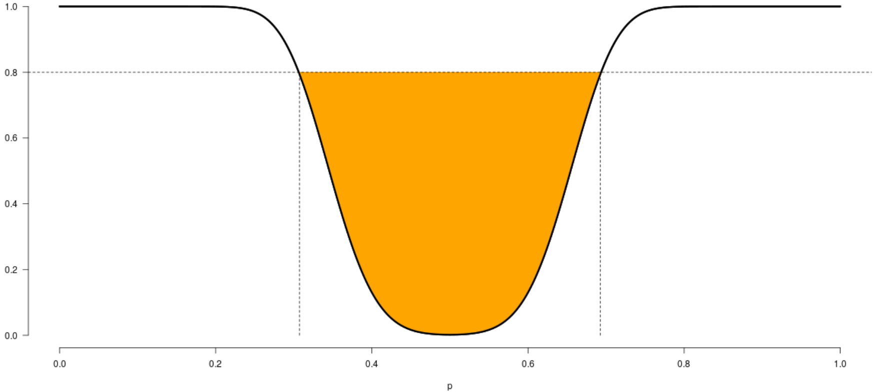
Il s'agit d'une fonction de p.

Motivation à partir d'un intervalle de confiance



$$p \in]0,1[\longmapsto \; 1 - F_{100,p}\left(\frac{100}{2} + \frac{\sqrt{100}}{2\sqrt{0.1}}\right) + F_{100,p}\left(\frac{100}{2} - \frac{\sqrt{100}}{2\sqrt{0.1}}\right)$$

Motivation à partir d'un intervalle de confiance



La probabilité de rejeter l'hypothèse nulle H_0 lorsque l'hypothèse alternative H_1 est vraie dépend de p et est d'autant plus grande que p est « loin » de $1/2$. Cette fonction est appelée la **puissance** du test.

Principe d'un test statistique

choix de H0 : hyp que l'on veut tester et régime sous lequel on sait faire des calculs

1. Définir une **hypothèse nulle** H_0 et une **hypothèse alternative** H_1.
2. Choisir un **niveau de confiance** $1-\alpha \in]0,1[$.
3. Proposer une **règle de décision** telle que

$$\mathbb{P}_{H_0}\left(\text{« Rejeter } H_0 \text{ »}\right) \leqslant \alpha.$$ proba de rejeter H0 sachant que H0 est vrai

4. Appliquer la règle de décision avec les **valeurs observées** dans l'échantillon.
5. Conclure si nous acceptons ou rejetons l'hypothèse H_0.

Remarque : pour deux tests de H_0 contre H_1 même niveau de confiance, il est préférable de choisir celui qui a tendance à avoir la plus grande fonction de puissance. Ce n'est pas un ordre total ...

Différentes erreurs

	Accepter H_0	Rejeter H_0
H_0 est vraie	Bonne décision	**Erreur de 1ère espèce**
H_1 est vraie	**Erreur de 2ème espèce**	Bonne décision

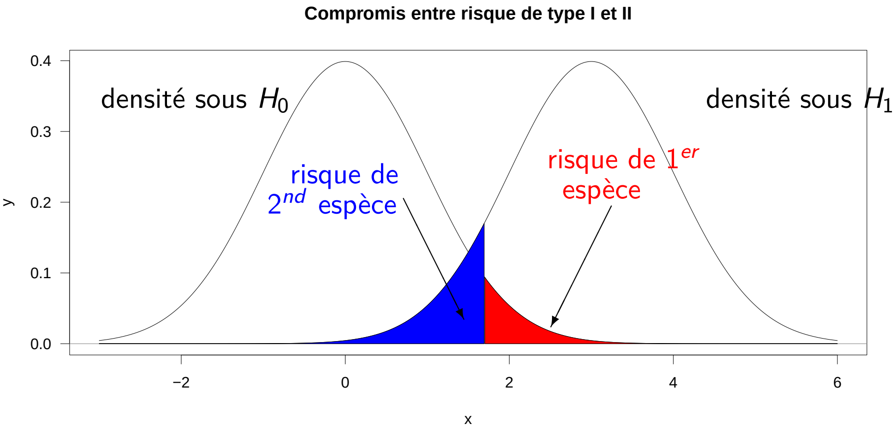
erreur de 1ere espèce = niveau du test (alpha)
erreur de 2ème espèce = puissance du test (

Exemple (extrême) : une étude médicale doit être menée pour valider la non dangerosité d'un nouveau médicament proposé par un laboratoire médical. L'utilisation d'un test de l'hypothèse nulle H_0 : « Le médicament est mortel » contre l'hypothèse alternative H_1 : « Le médicament n'est pas mortel » peut conduire aux deux erreurs suivantes :

- **1ère espèce :** médicament déclaré sain alors qu'il est mortel.
  $\Rightarrow$ C'est grave, des gens vont mourir !
- **2ème espèce :** médicament déclaré mortel alors qu'il est sans danger.
  $\Rightarrow$ C'est dommage pour le laboratoire mais personne ne va mourir.

Différentes erreurs, un compromis (encore...)

Il est **impossible** d'avoir à la fois un risque de 1ère espère **et** de 2ème espèce faible.



Exemple du pain d'épice

Une usine agroalimentaire produit du pain d'épice en tranches. Un des critères de qualité est que l'angle de rupture d'une tranche doit être supérieur à 42°. De nombreux facteurs (humidité ambiante, dosage des ingrédients, ...) rendent cette mesure d'angle aléatoire et cette variabilité semble bien modélisée par une loi normale (ce genre d'hypothèse aussi peut faire l'objet d'un test comme nous le verrons plus tard).

En piochant aléatoirement n tranches produites, nous disposons des réalisations de *v.a.i.i.d.* $X_1, \ldots, X_n$ de loi normale $\mathcal{N}(m, \sigma^2)$ de moyenne $m \in \mathbb{R}$ **inconnue** et de variance $\sigma^2 > 0$ **connue** (idem, il y a des tests pour valider cela). Nous voulons ainsi tester

$$H_0 \;:\; m > 42 \qquad \text{contre} \qquad H_1 \;:\; m \leqslant 42$$

pour un niveau de confiance $1 - \alpha \in]0, 1[$.

Exemple du pain d'épice

$$H_0 \;:\; m > 42 \qquad \text{contre} \qquad H_1 \;:\; m \leqslant 42$$

Pour construire une règle de décision, nous considérons la moyenne empirique

$$\overline{X}_n = \frac{1}{n}\sum_{k=1}^{n} X_k$$

qui est un estimateur sans biais et consistant de m.

Loi de $\overline{X}_n$

Une combinaison linéaire de variables normales **indépendantes** suit une loi normale. Pour la moyenne empirique, nous obtenons

$$\overline{X}_n \text{ suit la loi } \mathcal{N}\left(m, \frac{\sigma^2}{n}\right).$$

Exemple du pain d'épice

$$H_0 \ : \ m > 42 \qquad \text{contre} \qquad H_1 \ : \ m \leqslant 42$$

Guidés par l'idée d'un intervalle de confiance de niveau $1-\alpha$, nous proposons la règle de décision suivante :

$$\text{Rejeter } H_0 \iff \overline{X}_n \leqslant x_\alpha$$

où $x_\alpha \in \mathbb{R}$ est tel que

$$\mathbb{P}_{H_0}(\overline{X}_n \leqslant x_\alpha) \leqslant \alpha.$$

Autrement dit, nous souhaitons rejeter l'hypothèse d'une moyenne supérieure à 42 lorsque la moyenne empirique observée sur notre échantillon est « trop petite ».

Question : comment **calibrer** x_α pour assurer la probabilité d'erreur de 1ère espèce ?

Exemple du pain d'épice

$$H_0 \ : \ m > 42 \qquad \text{contre} \qquad H_1 \ : \ m \leqslant 42$$

Une première étape consiste à introduire la version Z **centrée et réduite** de la moyenne empirique $\overline{X}_n$,

$$Z = \sqrt{n}\frac{\overline{X}_n - m}{\sqrt{\sigma^2}}.$$

Ainsi, nous avons $\overline{X}_n = m + \sqrt{\frac{\sigma^2}{n}} Z$ avec Z qui suit une loi $\mathcal{N}(0,1)$.

De cette façon, nous avons exhibé une **loi bien connue qui ne dépend pas de** m et nous cherchons donc $x_\alpha \in \mathbb{R}$ tel que

$$\mathbb{P}_{H_0}\left(Z \leqslant \sqrt{n}\frac{x_\alpha - m}{\sqrt{\sigma^2}}\right) \leqslant \alpha.$$

Exemple du pain d'épice

$$H_0 \ : \ m > 42 \qquad \text{contre} \qquad H_1 \ : \ m \leqslant 42$$

La borne dans la probabilité dépend de m qui reste inconnue **même sous l'hypothèse** H_0. Cependant, si $m > 42$ alors nous savons que

$$\sqrt{n}\frac{x_\alpha - m}{\sqrt{\sigma^2}} < \sqrt{n}\frac{x_\alpha - 42}{\sqrt{\sigma^2}} = z_\alpha$$

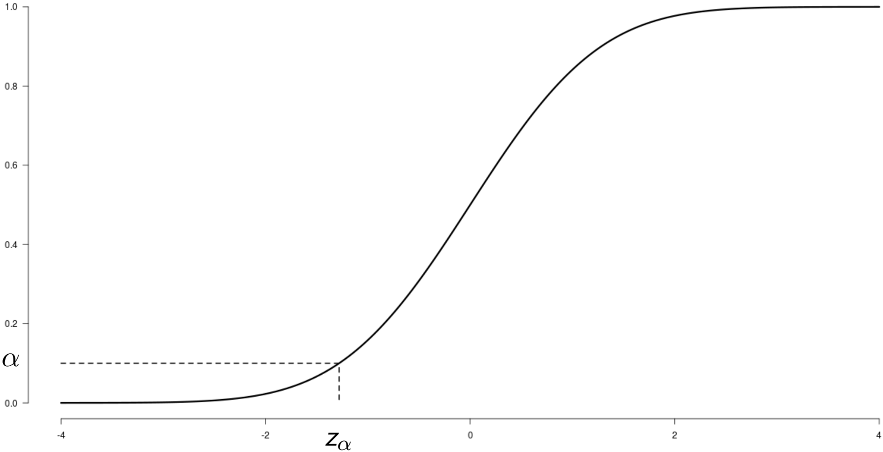
et donc

$$\mathbb{P}_{H_0}\left(Z \leqslant \sqrt{n}\frac{x_\alpha - m}{\sqrt{\sigma^2}}\right) \leqslant \mathbb{P}\left(Z \leqslant z_\alpha\right) \qquad \text{(probabilité libre de } H_0\text{)}$$

Il suffit de prendre $z_\alpha \in \mathbb{R}$ comme le **quantile** de niveau α de la loi $\mathcal{N}(0,1)$,

$$F_{\mathcal{N}(0,1)}(z_\alpha) = \int_{-\infty}^{z_\alpha} \frac{e^{-t^2/2}}{\sqrt{2\pi}} dt = \alpha \qquad \text{(Merci monsieur l'ordinateur !)}$$

Exemple du pain d'épice



Fonction de répartition de la loi $\mathcal{N}(0,1)$.

Exemple : pour $\alpha = 5\%$, nous obtenons $z_\alpha = -1.644854\dots$

Exemple du pain d'épice

$$H_0 \; : \; m > 42 \qquad \text{contre} \qquad H_1 \; : \; m \leqslant 42$$

À partir de cette valeur de z_α, nous déduisons le seuil de rejet x_α,

$$\sqrt{n}\frac{x_\alpha - 42}{\sqrt{\sigma^2}} = z_\alpha \iff x_\alpha = 42 + z_\alpha\sqrt{\frac{\sigma^2}{n}}.$$

La règle de décision est donc donnée par

$$\text{Rejeter } H_0 \iff \overline{X}_n \leqslant 42 + z_\alpha\sqrt{\frac{\sigma^2}{n}}.$$

Remarque : il s'agit bien d'une règle **statistique** car le seuil est **calculable** puisque tout est connu, en particulier la variance σ^2 dans cet exemple.

Exemple du pain d'épice

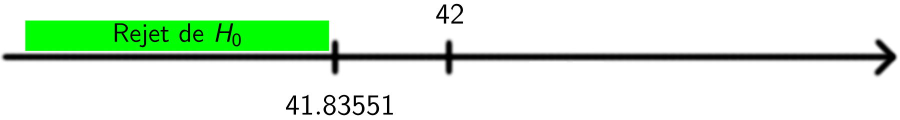
$$H_0 \ : \ m > 42 \qquad \text{contre} \qquad H_1 \ : \ m \leqslant 42$$

$$\text{Rejeter } H_0 \iff \overline{X}_n \leqslant 42 + z_\alpha \sqrt{\frac{\sigma^2}{n}}$$

Appliquons ce test avec $\alpha = 5\%$ (donc $z_\alpha = -1.644854$), $n = 100$ tranches de pain d'épice et une variance $\sigma^2 = 1$.
La règle de décision devient

$$\text{Rejeter } H_0 \iff \overline{X}_{100} \leqslant 41.83551.$$

Si la **réalisation** $\overline{x}_{100}$ de la **variable aléatoire** $\overline{X}_{100}$ sur notre échantillon donne $\overline{x}_{100} = 42.126$, alors **nous acceptons l'hypothèse** « $m > 42$ ».



Exemple du pain d'épice

$$H_0 \;:\; m > 42 \qquad \text{contre} \qquad H_1 \;:\; m \leqslant 42$$

$$\text{Rejeter } H_0 \iff \overline{X}_n \leqslant 42 + z_\alpha \sqrt{\frac{\sigma^2}{n}}$$

- **Erreur de 1ère espèce :** dans **moins de** 5% **des cas**, nous rejetons la production de pain d'épice alors qu'elle est de qualité, gaspillage !
  $\Rightarrow$ Le patron ne va pas être content (en fait, il ne le saura pas ...).
- **Erreur de 2ème espèce :** avec une **probabilité d'autant plus petite** que m est inférieure à 42 (puissance du test), nous vendons des tranches de mauvaise qualité.
  $\Rightarrow$ Le client ne va pas être content (lui, il le saura ...).

Exemple du pain d'épice

$$H_0 \ : \ m > 42 \qquad \text{contre} \qquad H_1 \ : \ m \leqslant 42$$

$$\text{Rejeter } H_0 \iff \overline{X}_n \leqslant 42 + z_\alpha \sqrt{\frac{\sigma^2}{n}}$$

- **Erreur de 1ère espèce :** dans **moins de** 5% **des cas**, nous rejetons la production de pain d'épice alors qu'elle est de qualité, gaspillage !
  $\Rightarrow$ Le patron ne va pas être content (en fait, il ne le saura pas ...).
- **Erreur de 2ème espèce :** avec une **probabilité d'autant plus petite** que m est inférieure à 42 (puissance du test), nous vendons des tranches de mauvaise qualité.
  $\Rightarrow$ Le client ne va pas être content (lui, il le saura ...).

Et si nous prenions le point de vue du client ?

Exemple du pain d'épice (point de vue du client)

Pour la personne qui achète notre pain d'épice, l'important est surtout de ne pas trouver des tranches de mauvaise qualité. En termes de test statistique, cela signifie qu'elle veut s'assurer que l'erreur contrôlée est celle commise sur l'hypothèse « $m \leqslant 42$ » qui servait d'alternative pour notre test initial.

Le point de vue du client consiste donc à échanger les hypothèses précédentes et à tester

$$H_0' \;:\; m \leqslant 42 \qquad \text{contre} \qquad H_1' \;:\; m > 42$$

pour un niveau de confiance $1 - \alpha \in]0, 1[$.
Les mêmes idées que précédemment nous conduisent à proposer la règle de décision

$$\text{Rejeter } H_0' \iff \overline{X}_n > x_\alpha'$$

où $x_\alpha' \in \mathbb{R}$ est tel que

$$\mathbb{P}_{H_0'}\left(\overline{X}_n > x_\alpha'\right) \leqslant \alpha.$$

Exemple du pain d'épice (point de vue du client)

$$H_0' \ : \ m \leqslant 42 \qquad \text{contre} \qquad H_1' \ : \ m > 42$$

Afin de calibrer le seuil x'_α, nous procédons de la même façon en introduisant le nombre $z'_\alpha \in \mathbb{R}$ tel que

$$\int_{z'_\alpha}^{+\infty} \frac{e^{-t^2/2}}{\sqrt{2\pi}} dt = \alpha \iff F_{\mathcal{N}(0,1)}(z'_\alpha) = \int_{-\infty}^{z'_\alpha} \frac{e^{-t^2/2}}{\sqrt{2\pi}} dt = 1 - \alpha.$$

Sous l'hypothèse que $m \leqslant 42$, nous savons

$$\sqrt{n}\frac{x'_\alpha - m}{\sqrt{\sigma^2}} \geqslant \sqrt{n}\frac{x'_\alpha - 42}{\sqrt{\sigma^2}}$$

et cela conduit à

$$x'_\alpha = 42 + z'_\alpha\sqrt{\frac{\sigma^2}{n}}.$$

Exemple du pain d'épice (point de vue du client)

$$H_0' \ : \ m \leqslant 42 \qquad \text{contre} \qquad H_1' \ : \ m > 42$$

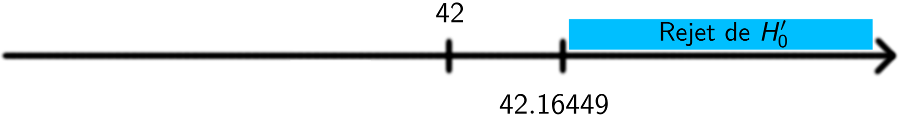
$$\text{Rejeter } H_0' \iff \overline{X}_n > 42 + z_\alpha' \sqrt{\frac{\sigma^2}{n}}$$

Appliquons ce test avec les mêmes valeurs que dans l'exemple précédent : $\alpha = 5\%$ (et $z_\alpha' = 1.644854$), $n = 100$ et $\sigma^2 = 1$.

La règle de décision devient

$$\text{Rejeter } H_0' \iff \overline{X}_n > 42.16449.$$

Si la **moyenne observée** vaut $\overline{x}_{100} = 42.126$, alors **nous acceptons l'hypothèse** « $m \leqslant 42$ ».



Exemple du pain d'épice (point de vue du client)

$$H_0 \ : \ m > 42 \qquad \text{contre} \qquad H_1 \ : \ m \leqslant 42$$

$$\text{Rejeter } H_0 \iff \overline{X}_n \leqslant 42 + z_\alpha \sqrt{\frac{\sigma^2}{n}} \ (= 41.83551)$$

$$H_0' \ : \ m \leqslant 42 \qquad \text{contre} \qquad H_1' \ : \ m > 42$$

$$\text{Rejeter } H_0' \iff \overline{X}_n > 42 + z_\alpha' \sqrt{\frac{\sigma^2}{n}} \ (= 42.16449)$$

Nous venons de construire deux tests de même niveau de confiance qui donnent des **réponses opposées** sur les **mêmes données** ...

Est-ce contradictoire ?

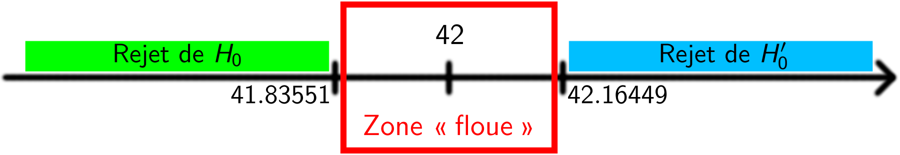
Exemple du pain d'épice (point de vue du client)

$$H_0 \;:\; m > 42 \qquad \text{contre} \qquad H_1 \;:\; m \leqslant 42$$

$$\text{Rejeter } H_0 \iff \overline{X}_n \leqslant 42 + z_\alpha \sqrt{\frac{\sigma^2}{n}} \; (= 41.83551)$$

$$H_0' \;:\; m \leqslant 42 \qquad \text{contre} \qquad H_1' \;:\; m > 42$$

$$\text{Rejeter } H_0' \iff \overline{X}_n > 42 + z_\alpha' \sqrt{\frac{\sigma^2}{n}} \; (= 42.16449)$$



Choix de l'hypothèse nulle H_0

Dans le « doute », un test statistique favorise **toujours** son hypothèse nulle. Les hypothèses ne sont pas **symétriques** et **la décision dépend du parti pris de départ**.

Idée générale : un test statistique ne rejette son hypothèse nulle que si celle-ci n'est vraiment **pas vraisemblable**. Nous parlons alors de **significativité** du test statistique.

Choix de l'hypothèse nulle H_0

Dans le « doute », un test statistique favorise **toujours** son hypothèse nulle. Les hypothèses ne sont pas **symétriques** et **la décision dépend du parti pris de départ**.

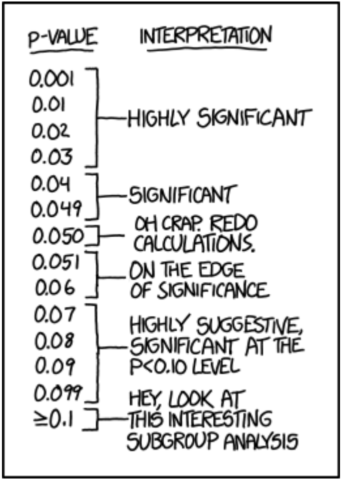
Idée générale : un test statistique ne rejette son hypothèse nulle que si celle-ci n'est vraiment **pas vraisemblable**. Nous parlons alors de **significativité** du test statistique.

Comment choisir H_0 ?

1. H_0 est l'hypothèse la plus grave (le pont s'écroule, le médicament est mortel, ...).
2. H_0 est l'hypothèse la plus communément admise.
3. Les calculs de calibration ne peuvent être faits que sous H_0.

Notion de *p*-valeur

L'utilisation de la *p*-valeur est très criticable en pratique.



« We teach it because it's what we do ; we do it because it's what we teach. »

George Cobb

P-Values, XKCD, xkcd.com/1478

Notion de p-valeur

Bien que la définition de la p-valeur fasse **encore débat**, il s'agit d'une quantité couramment utilisée dans de nombreux domaines de recherche.

L'objectif de la p-valeur est de quantifier le « **degré de significativité** » d'un test statistique.

Une façon commune de présenter la p-valeur est de la définir comme la **plus petite valeur** de l'erreur de 1ère espèce $\alpha \in]0, 1[$ pour laquelle les observations conduisent au **rejet de l'hypothèse nulle** H_0.

La p-valeur est donc la probabilité, sous l'hypothèse H_0, d'observer les données les « plus extrêmes ».

Lien entre le niveau et la p-valeur

$$\text{Rejet de } H_0 \text{ au niveau } 1-\alpha \iff p\text{-valeur} < \alpha.$$

Plus la p-valeur est faible, plus le risque de rejeter H_0 à tort est faible.

-1.2 Quelques tests statistiques classiques

Tests sur la moyenne (loi normale)

Cadre : $X_1, \ldots, X_n$ *v.a.i.i.d.* de loi $\mathcal{N}(m, \sigma^2)$

Exemples d'hypothèses :

$$
\begin{array}{lllll}
H_0 & : & m = m_0 & \text{contre} & H_1 : m = m_1 \text{ (avec } m_0 \neq m_1) \\
H_0 & : & m = m_0 & \text{contre} & H_1 : m > m_0 \text{ (ou } m < m_0) \\
H_0 & : & m = m_0 & \text{contre} & H_1 : m \neq m_0 \\
H_0 & : & m \leqslant m_0 & \text{contre} & H_1 : m > m_0 \\
H_0 & : & m \geqslant m_0 & \text{contre} & H_1 : m < m_0
\end{array}
$$

Et bien d'autres ...

Tests sur la moyenne (loi normale)

Cadre : $X_1, \ldots, X_n$ *v.a.i.i.d.* de loi $\mathcal{N}(m, \sigma^2)$

Si la variance σ^2 est **connue**, la règle décision se construit à partir de la moyenne empirique $\overline{X}_n$ et se calibre avec

$$\sqrt{n}\frac{\overline{X}_n - m}{\sqrt{\sigma^2}} \text{ suit la loi } \mathcal{N}(0, 1).$$

Voir l'exemple du pain d'épice . . .

Tests sur la moyenne (loi normale)

Cadre : $X_1, \ldots, X_n$ *v.a.i.i.d.* de loi $\mathcal{N}(m, \sigma^2)$

Si la variance σ^2 est **inconnue**, elle peut être estimée par

$$\hat{\sigma}_n^2 = \frac{1}{n} \sum_{k=1}^{n} \left(X_k - \overline{X}_n \right)^2 .$$

Cette définition fait intervenir la somme des carrés de variables normales indépendantes recentrées par la moyenne empirique. À la variance σ^2 près, une telle variable admet une loi dite du χ^2 à $n-1$ degrés de liberté (et non pas n à cause du **recentrage empirique**),

$$\frac{1}{\sigma^2} \sum_{k=1}^{n} \left(X_k - \overline{X}_n \right)^2 \text{ suit la loi } \chi^2(n-1).$$

Conséquence : $\mathbb{E}[\hat{\sigma}_n^2] = \dfrac{n-1}{n}\sigma^2$ donc $b(\hat{\sigma}_n^2) = -\sigma^2/n \neq 0$.

Tests sur la moyenne (loi normale)

Cadre : $X_1, \ldots, X_n$ *v.a.i.i.d.* de loi $\mathcal{N}(m, \sigma^2)$

Un estimateur **sans biais** de la variance σ^2 est donné par

$$\tilde{\sigma}_n^2 = \frac{n}{n-1}\hat{\sigma}^2{}_n = \frac{1}{n-1}\sum_{k=1}^{n}\left(X_k - \overline{X}_n\right)^2.$$

Le théorème de Cochran assure que la moyenne empirique $\overline{X}_n$ est **indépendante** de $\tilde{\sigma}_n^2$ (et de $\hat{\sigma}_n^2$). Cela permet de déduire que

$$\sqrt{n}\frac{\overline{X}_n - m}{\sqrt{\tilde{\sigma}_n^2}}$$

suit la loi de Student $\mathcal{T}(n-1)$ à $n-1$ degrés de liberté. Cette loi permet de calibrer la règle de décision de façon similaire au cas de variance connue.

Tests sur la moyenne (cas général)

Cadre : $X_1, \ldots, X_n$ *v.a.i.i.d.* avec $\mathbb{E}[X_1] = m \in \mathbb{R}$ et $\mathrm{Var}(X_1) = \sigma^2 > 0$.

Exemples d'hypothèses :

$$
\begin{array}{lllll}
H_0 : m = m_0 & \text{contre} & H_1 : m = m_1 \text{ (avec } m_0 \neq m_1) \\
H_0 : m = m_0 & \text{contre} & H_1 : m > m_0 \text{ (ou } m < m_0) \\
H_0 : m = m_0 & \text{contre} & H_1 : m \neq m_0 \\
H_0 : m \leqslant m_0 & \text{contre} & H_1 : m > m_0 \\
H_0 : m \geqslant m_0 & \text{contre} & H_1 : m < m_0
\end{array}
$$

Et bien d'autres ...

Remarque : bien qu'il soit possible dans certains cas de proposer des tests de niveau fixé, les résultats généraux sont tous **asymptotiques**.

Tests sur la moyenne (cas général)

Cadre : $X_1, \ldots, X_n$ *v.a.i.i.d.* avec $\mathbb{E}[X_1] = m \in \mathbb{R}$ et $\mathrm{Var}(X_1) = \sigma^2 > 0$.

Si la variance σ^2 est **connue**, la règle décision se construit à partir de la moyenne empirique $\overline{X}_n$ et se **calibre asymptotiquement** comme dans le cas normal grâce au théorème central limite,

$$\sqrt{n}\frac{\overline{X}_n - m}{\sqrt{\sigma^2}} \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0,1).$$

Si la variance σ^2 est **inconnue**, cette convergence en loi reste vrai en remplaçant la variance par $\hat{\sigma}_n^2$ (ou $\tilde{\sigma}_n^2$) car il s'agit d'un estimateur consistant et le lemme de Slutsky donne

$$\sqrt{n}\frac{\overline{X}_n - m}{\sqrt{\hat{\sigma}_n^2}} \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0,1).$$

Cela permet encore de **calibrer asymptotiquement** la règle de décision.

Tests sur la variance (loi normale)

Cadre : $X_1, \ldots, X_n$ *v.a.i.i.d.* de loi $\mathcal{N}(m, \sigma^2)$

Exemples d'hypothèses :

$$
\begin{array}{lllll}
H_0 : \sigma^2 = \sigma_0^2 & \text{contre} & H_1 : \sigma^2 = \sigma_1^2 \text{ (avec } \sigma_0^2 \neq \sigma_1^2) \\
H_0 : \sigma^2 = \sigma_0^2 & \text{contre} & H_1 : \sigma^2 > \sigma_0^2 \text{ (ou } \sigma^2 < \sigma_0^2) \\
H_0 : \sigma^2 = \sigma_0^2 & \text{contre} & H_1 : \sigma^2 \neq \sigma_0^2 \\
H_0 : \sigma^2 \leqslant \sigma_0^2 & \text{contre} & H_1 : \sigma^2 > \sigma_0^2 \\
H_0 : \sigma^2 \geqslant \sigma_0^2 & \text{contre} & H_1 : \sigma^2 < \sigma_0^2
\end{array}
$$

Et bien d'autres ...

Tests sur la variance (loi normale)

Cadre : $X_1, \ldots, X_n$ *v.a.i.i.d.* de loi $\mathcal{N}(m, \sigma^2)$

Si la moyenne m est **connue**, il est possible de recentrer les variables observées de façon **déterministe** et de calibrer la règle de décision avec

$$\frac{1}{\sigma^2} \sum_{k=1}^{n} (X_k - m)^2 \text{ suit la loi } \chi^2(n).$$

Si la moyenne m est **inconnue**, il faut **recentrer empiriquement** avec $\overline{X}_n$ et la calibration se fait grâce à

$$\frac{1}{\sigma^2} \sum_{k=1}^{n} \left(X_k - \overline{X}_n\right)^2 \text{ suit la loi } \chi^2(n-1).$$

Test d'égalité des variances (loi normale)

Cadre : deux groupes **indépendants** de variables $X_1, \dots, X_p$ *i.i.d.* de loi $\mathcal{N}(m_1, \sigma^2)$ et $Y_1, \dots, Y_q$ *i.i.d.* de loi $\mathcal{N}(m_2, \tau^2)$

$$H_0 \ : \ \sigma^2 = \tau^2 \qquad \text{contre} \qquad H_1 \ : \ \sigma^2 \neq \tau^2$$

Nous considérons les estimateurs sans biais des variances,

$$\tilde{\sigma}_p^2 = \frac{1}{p-1} \sum_{k=1}^{p} \left(X_k - \overline{X}_p \right)^2 \qquad \text{et} \qquad \tilde{\tau}_q^2 = \frac{1}{q-1} \sum_{k=1}^{q} \left(Y_k - \overline{Y}_q \right)^2 .$$

Sous l'hypothèse H_0 d'égalité des variances, nous avons

$$F = \frac{\tilde{\sigma}_p^2}{\tilde{\tau}_q^2} = \frac{\tilde{\sigma}_p^2/\sigma^2}{\tilde{\tau}_q^2/\tau^2} \quad \text{suit la loi} \quad \frac{\chi^2(p-1)/(p-1)}{\chi^2(q-1)/(q-1)}.$$

Il s'agit de la loi de Fisher $\mathcal{F}(p-1, q-1)$ à $p-1$ et $q-1$ degrés de liberté.

Test d'égalité des variances (loi normale)

Cadre : deux groupes **indépendants** de variables $X_1, \ldots, X_p$ *i.i.d.* de loi $\mathcal{N}(m_1, \sigma^2)$ et $Y_1, \ldots, Y_q$ *i.i.d.* de loi $\mathcal{N}(m_2, \tau^2)$

$$H_0 \;:\; \sigma^2 = \tau^2 \qquad \text{contre} \qquad H_1 \;:\; \sigma^2 \neq \tau^2$$

Le principe de la règle de décision est de rejeter H_0 si le rapport $F = \tilde{\sigma}_p^2 / \tilde{\tau}_q^2$ est « trop petit » ou « trop grand »,

$$\text{Rejeter } H_0 \iff F < u_\alpha \text{ ou } F > v_\alpha$$

avec u_α et v_α à calibrer pour un niveau $1 - \alpha \in]0, 1[$ donné.

Remarques : pour calibrer u_α et v_α, il faut utiliser le fait que, sous H_0,

$$F \text{ suit la loi } \mathcal{F}(p-1, q-1) \quad \text{et} \quad 1/F \text{ suit la loi } \mathcal{F}(q-1, p-1).$$

Test de comparaison des moyennes (loi normale)

Cadre : deux groupes **indépendants** de variables $X_1, \dots, X_p$ *i.i.d.* de loi $\mathcal{N}(m_1, \sigma^2)$ et $Y_1, \dots, Y_q$ *i.i.d.* de loi $\mathcal{N}(m_2, \sigma^2)$ de **même variance**
Exemples d'hypothèses :

$$\begin{array}{lllll}
H_0 : m_1 = m_2 & \text{contre} & H_1 : m_1 \neq m_2 \\
H_0 : m_1 = m_2 & \text{contre} & H_1 : m_1 > m_2 \\
H_0 : m_1 = m_2 & \text{contre} & H_1 : m_1 < m_2 \\
H_0 : m_1 \leqslant m_2 & \text{contre} & H_1 : m_1 > m_2 \\
H_0 : m_1 \geqslant m_2 & \text{contre} & H_1 : m_1 < m_2
\end{array}$$

Et bien d'autres ...

Pour estimer la variance σ^2 commune sans biais, nous disposons de

$$\tilde{\sigma}^2_{p,q} = \frac{(p-1)\tilde{\sigma}^2_{X,p} + (q-1)\tilde{\sigma}^2_{Y,q}}{p+q-2}.$$

Test de comparaison des moyennes (loi normale)

Cadre : deux groupes **indépendants** de variables $X_1, \dots, X_p$ *i.i.d.* de loi $\mathcal{N}(m_1, \sigma^2)$ et $Y_1, \dots, Y_q$ *i.i.d.* de loi $\mathcal{N}(m_2, \sigma^2)$ de **même variance**

Le théorème de Cochran donne encore que $\overline{X}_p$ et $\overline{Y}_q$ sont indépendantes de $\tilde{\sigma}^2_{p,q}$ et que

$$(p+q-2)\frac{\tilde{\sigma}^2_{p,q}}{\sigma^2} \text{ suit la loi } \chi^2(p+q-2).$$

La règle de décision se calibre alors grâce à

$$\sqrt{\frac{pq}{p+q}} \times \frac{(\overline{X}_p - m_1) - (\overline{Y}_q - m_2)}{\sqrt{\tilde{\sigma}^2_{p,q}}} \text{ suit la loi } \mathcal{T}(p+q-2).$$

Test de Shapiro-Wilk (non paramétrique)

Cadre : $X_1, \dots, X_n$ *v.a.i.i.d.* de loi $\mathcal{L}$ **inconnue**

Le test de normalité de Shapiro-Wilk considère

$$H_0 \ : \ \mathcal{L} \text{ est normale} \qquad \text{contre} \qquad H_1 \ : \ \mathcal{L} \text{ n'est pas normale.}$$

Pour cela, nous introduisons la **version ordonnée** des variables observées,

$$X_{(1)} \leqslant \cdots \leqslant X_{(n)}$$

et nous définissons la variable

$$W = \frac{\left(\sum_{k=1}^{n} a_k X_{(k)}\right)^2}{\sum_{k=1}^{n} (X_k - \overline{X}_n)^2}.$$

Test de Shapiro-Wilk (non paramétrique)

Les coefficients $a_1, \dots, a_n$ sont connus et disponibles dans tout bon logiciel de statistique. Ils sont donnés par

$$\begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix} = \frac{m^\top \Sigma^{-1}}{\sqrt{m^\top \Sigma^{-2} m}}$$

où $m \in \mathbb{R}^n$ est le vecteur des espérances de la version ordonnée de n *v.a.i.i.d.* normales centrées réduites et Σ est la matrice de covariance de ces mêmes variables normales ordonnées.

En pratique, plus la valeur de W est élevée, plus l'adéquation à la loi normale est acceptable,

$$\text{Rejeter } H_0 \iff W < w_\alpha.$$

Test de significativité en régression

On cherche à vérifier si une régression linéaire est pertinente, i.e. si le modèle

$$Y_i \sim \mathcal{N}(ax_i + b, \sigma^2) \quad i = 1, \ldots, n$$

est bien ajusté. Ce qui revient formellement à tester

$$H_0 : a = 0 \quad VS \quad H_1 : a \neq 0.$$

On considère la statistique

$$F = (n-2)\frac{R^2}{1-R^2} \sim \mathcal{F}(1, n-2).$$

Ce qui donne la règle de décision suivante :

$$\text{Rejeter } H_0 \iff F < f_\alpha.$$

Test de Kolmogorov-Smirnov (non paramétrique)

Cadre : $X_1, \ldots, X_n$ *v.a.i.i.d.* de loi $\mathcal{L}_X$ **inconnue**

Pour une loi $\mathcal{L}$ donnée, le test d'adéquation de Kolmogorov-Smirnov considère

$$H_0 \;:\; \mathcal{L}_X = \mathcal{L} \qquad \text{contre} \qquad H_1 \;:\; \mathcal{L}_X \neq \mathcal{L}.$$

Nous introduisons la **fonction de répartition empirique** des observations,

$$\forall x \in \mathbb{R},\; F_n(x) = \frac{1}{n} \sum_{k=1}^{n} \mathbf{1}_{X_k \leqslant t}.$$

Test de Kolmogorov-Smirnov (non paramétrique)

Cadre : $X_1, \dots, X_n$ *v.a.i.i.d.* de loi $\mathcal{L}_X$ **inconnue**

Pour une loi $\mathcal{L}$ donnée, le test d'adéquation de Kolmogorov-Smirnov considère

$$H_0 \ : \ \mathcal{L}_X = \mathcal{L} \qquad \text{contre} \qquad H_1 \ : \ \mathcal{L}_X \neq \mathcal{L}.$$

Si F est la fonction de répartition de $\mathcal{L}$, il est possible de montrer que, sous l'hypothèse H_0, la « **fonction aléatoire** »

$$x \in \mathbb{R} \mapsto \sqrt{n}(F_n(x) - F(x))$$

converge en loi vers un **pont brownien**. Cet objet aléatoire dépasse de loin le cadre de ce cours mais il est bien connu et permet de calibrer **asymptotiquement** la règle de décision suivante

$$\text{Rejeter } H_0 \iff \sup_{x \in \mathbb{R}} |F_n(x) - F(x)| > k_\alpha.$$

Test d'interdependence (cas Gaussien)

Cadre : $(Y_1, X_1), \dots, (X_n, Y_n)$ des vecteur gaussiens *i.i.d.*, i.e

$$(X_i, Y_i) \sim \mathcal{N}_2\left(\begin{pmatrix}\mu_X \\ \mu_Y\end{pmatrix}, \begin{pmatrix}\sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2\end{pmatrix}\right).$$

avec ρ le coefficient de correlation.

Dans le des vecteurs gaussiens, on sait que $X \perp Y \iff \rho = 0$. On considère alors le test

$$H_0 : \rho = 0 \quad \text{contre} \quad H_1 : \rho \neq 0.$$

Test d'interdependence (cas Gaussien)

Le coefficient de corrélation empirique est :

$$\hat{\rho} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \ \ \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}.$$

Il est possible de montrer que sous $H_0 : \rho = 0$, la statistique

$$\frac{\hat{\rho}\sqrt{n-2}}{\sqrt{1-\hat{\rho}^2}} \sim \mathcal{T}(n-2).$$

Ce qui donne la régle de décision suivante :

$$\text{Accepter } H_0 \iff t_{1-\alpha/2}(n-2) \leq \frac{\hat{\rho}\sqrt{n-2}}{\sqrt{1-\hat{\rho}^2}} \leq t_{\alpha/2}(n-2)$$

Analyse de la variance (ANOVA)

Cadre : $(X_{i,k})$ indépendants avec $k = 1, \dots, K$ et $i = 1, \dots n_k$ modélisé par

$$X_{i,k} = \mu + \alpha_k + \mathcal{N}(0, \sigma^2).$$

Avec $\mu \in \mathbb{R}$ la moyenne globale, les α_k un effet du groupe k.
on veut tester

$$H_0 : \alpha_1 = \dots = \alpha_K = 0 \quad \text{contre} \quad H_1 : \exists \ (k, l) \text{ tels que } k \neq l, \ \alpha_k \neq \alpha_l.$$

Remarques :

- Cela signifie qu'on cherche à tester si les K groupes ont la **même moyenne**.
- C'est beaucoup **plus puissant** que de tester $\alpha_l = \alpha_k$ pour toutes les paires (k, l).

Analyse de la variance (ANOVA)

On se sert de la décomposition de la variance (classique dans les modèles linéaires) pour construire la statistique de test : SCT = SCE + SCR, avec

- Somme des carrés totale (variance totale) :

$$\mathrm{SCT} = \sum_{k=1}^{K} \sum_{k=i}^{n_k} (X_{i,k} - \bar{X}_n)^2, \quad \text{avec } \bar{X}_n = \sum_{k=1}^{K} \sum_{k=i}^{n_k} X_{i,k}.$$

- Somme des carrées expliquée (variance interigroupes) :

$$\mathrm{SCE} = \sum_{k=1}^{K} n_k (\bar{X}_{(k)} - \bar{Y}_n)^2 \quad \text{avec } \bar{X}_{(k)} = \sum_{k=i}^{n_k} X_{i,k}, \text{ pour } k = 1, \dots, K.$$

- Somme des carrés résiduels (variance intra-groupes) :

$$\mathrm{SCR} = \sum_{k=1}^{K} \sum_{i=1}^{n_k} (X_{i,k} - \bar{X}_k)^2.$$

Analyse de la variance (ANOVA)

On a alors, sous H_0, en notant $N = \sum_{k=1}^{K} n_k$:

$$\frac{\text{SCE}}{\sigma^2} \sim \chi^2_{K-1} \quad \text{et} \quad \frac{\text{SCR}}{\sigma^2} \sim \chi^2_{N-K}, \text{ pour tout } k = 1, \dots, K.$$

On en déduit alors la statistique de test :

$$F = \frac{\text{SCE}}{K-1} \frac{N-K}{\text{SCR}} \sim \mathcal{F}_{K-1, N-K}.$$

Ce qui donne la régle de décision suivante :

$$\text{Accepter } H_0 \iff F < f_\alpha.$$