Probabilité et Statistiques

II – Statistiques

Richard Wilson

(richard.wilson@upmc.fr)

LATMOS/IPSL
Université Pierre & Marie Curie

27 octobre 2020

Sommaire

Généralités

Généralités

- Nous cherchons à décrire certaines propriétés d'une **population** (personnes, poissons rouges ou multiples résultats d'une mesure physique).
- Quelle que soit la nature de cette population, les éléments qui la constituent sont appelés **individus**.
- Certains individus de la population possèdent un caractère particulier qu'on appellera C. On cherche à connaître la proportion des individus possédant le caractère C.
- Un recensement de toute la population étant le plus souvent inenvisageable, on prélève un **échantillon**, c'est à dire un sous-ensemble de la population. La proportion d'individus de l'échantillon possédant la caractéristique C est P.
- Peut-on en conclure que la proportion des individus dans la population possédant le caractère X est également P ?
- Ou mieux : peut on donner un intervalle I dans lequel la proportion P se trouve presque sûrement ?

- La question posée est de nature radicalement différente de celles abordées en probabilité.
- Jusqu'ici, nous supposions connue la loi de probabilité du caractère C. À partir de cette loi nous calculions les probabilités p d'occurrence de C dans un échantillon, i.e. nous calculions la probabilité d'avoir k succès sur n épreuves, autrement dit d'observer k individus ayant le caractère C dans un échantillon de n individus.
- La question que pose le statisticien est inversée : il connaît un échantillon de taille n dans lequel il observe k succès. Il cherche à en déduire la proportion p au sein la population.
- Cette dernière question se pose dans de nombreux domaines d'activité.
- Il s'agit non pas de **probabilité déductive** (de la loi de probabilité vers l'échantillon) mais d'**inférence statistique** (de l'échantillon vers la population).

Les méthodes présentées dans les chapitres qui suivent visent :

1. à décrire les propriétés de quelques statistiques d'échantillon ;
2. à **estimer** les propriétés d'une **population** à partir d'**échantillons** de celle-ci ;
3. à **tester une hypothèse** concernant une population à partir de la connaissance d'un échantillon.
4. à **modéliser** une relation linéaire entre deux variables.

Statistiques d'échantillon

- Considérons tous les échantillons pouvant être constitués à partir d'une population.
- Pour chaque échantillon, on peut calculer une statistique, telle la moyenne ou l'écart type, dont la valeur variera d'un échantillon à l'autre.
- De cette manière on obtient une distribution d'échantillonnage de la statistique considérée.
- Les statistiques (moyenne, écart type...) sont donc des variables aléatoires dont les propriétés sont décrites par les distributions d'échantillonnage.

La statistique de la moyenne $\overline{X}$

- L'échantillon $(x_1, \ldots, x_n)$ constitue une réalisation d'un n-uplet de variables aléatoires $(X_1, \ldots, X_n)$ dont les éléments sont indépendants et identiquement distribués (i.i.d.) selon la loi parente.
- La loi décrivant la population n'est pas connue, en particulier son espérance μ et l'écart type σ.
- On cherche dans un premier temps à estimer l'espérance μ des X_i à partir de la seule connaissance des x_i de l'échantillon.
- Si la probabilité de chaque observation $X = x_i$ est uniforme (échantillonnage aléatoire simple), une façon naturelle d'évaluer l'espérance des X_i est de prendre la moyenne de l'échantillon.

  Définition :La statistique $\overline{X}$ est la moyenne empirique de l'échantillon :

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

La statistique de la moyenne $\overline{X}$

- Espérance de $\overline{X}$

$$\mathrm{E}\left[\overline{X}\right] = \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\,\mathrm{E}\left[\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\left[X_i\right] = \frac{1}{n}\sum_{i=1}^{n}\mu = \mu$$

- On déduit le résultat fondamental :

  L'espérance de la moyenne d'échantillon est égale à l'espérance de la population.

$$\mathrm{E}\left[\overline{X}\right] = \mu$$

- On dira que la moyenne de l'échantillon est un estimateur non biaisé de la moyenne de la population entière.

La statistique de la moyenne $\overline{X}$

- Variance $\mathrm{Var}\left[\overline{X}\right]$

$$\mathrm{Var}\left[\overline{X}\right] = \mathrm{Var}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}\,\mathrm{Var}\left[\sum_{i=1}^{n} X_i\right]$$

- Or, si les X_i sont indépendants : $\mathrm{Var}\left[\sum_{i=1}^{n} X_i\right] = n\,\mathrm{Var}\left[X_i\right] = n\sigma_X^2$
- En conclusion, le résultat important :

La variance de la moyenne empirique $\overline{X}$ est

$$\mathrm{Var}\left[\overline{X}\right] = \frac{\sigma_X^2}{n}$$

La loi des grands nombres

- Le fait que la moyenne d'échantillon tende asymptotiquement vers l'espérance de la population résulte de la **loi des grands nombres**.
- En effet, la loi (faible) des grands nombres affirme :

$$\forall \varepsilon > 0, \quad \lim_{n\to+\infty} \Pr\left[\left|\overline{X} - E(X)\right| \geqslant \varepsilon\right] = 0$$

- C'est une conséquence de l'inégalité de Bienaymé-Chebychev qui appliquée à la variable aléatoire $\overline{X}$ s'écrit :

$$\Pr[|\overline{X} - \mathrm{E}\left(\overline{X}\right)| \geqslant \varepsilon] \leqslant \frac{\mathrm{Var}\left[\overline{X}\right]}{\varepsilon^2} = \frac{\sigma_X^2}{n\varepsilon^2}$$

- Par conséquent : $\Pr\left[\left|\overline{X} - \mu\right| \geqslant \varepsilon\right] \leqslant \frac{\sigma_X^2}{n\,\varepsilon^2}$.

La statistique de la moyenne $\overline{X}$

- Distribution asymptotique de $\overline{X}$ (pour n grand)

  Les X_i constitutifs de l'échantillon sont identiquement distribués.

- Le théorème de la limite centrale affirme que quelle que soit la distribution des X_i admettant des moments d'ordre 1 et 2 :

$$\sum_{i=1}^{n} X_i \xrightarrow{\mathcal{L}} \mathcal{N}(n\mu, n\sigma^2)$$

- La moyenne empirique converge donc vers une loi normale de moyenne μ et d'écart type $\sigma_X/\sqrt{n}$:

$$\overline{X} \xrightarrow{\mathcal{L}} \mathcal{N}(\mu, \sigma^2/n)$$

Théorème de la limite centrale

On en déduit que la variable centrée réduite $\left(\dfrac{\overline{X}-\mu}{\sigma/\sqrt{n}}\right)$ converge vers une loi normale $\mathcal{N}(0,1)$ quelle que soit la loi suivie par les X_i lorsque n est suffisamment grand (en pratique $n > 30$)

$$\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$$

Exemple : Comment générer une v.a. $Z \sim \mathcal{N}(0,1)$ à partir des tirrages de v.a. $X_i \sim \mathcal{U}(0,1)$?

Statistique de la fréquence empirique F

Considérons une population dont certains individus possèdent une caractéristique C avec une probabilité p. On cherche à estimer cette probabilité à partir d'un n-échantillon.

- Définissons la v.a. de Bernoulli X :

$$X = \begin{cases} 1 & \text{si la caractéristique } C \text{ est présente ;} \\ 0 & \text{sinon.} \end{cases}$$

  La somme $\sum_{i=1}^{n} X_i$ permet de comptabiliser le nombre d'éléments de l'échantillon possédant la caractéristique C.

- **Définition** La statistique F ou fréquence empirique, ou encore **proportion** est définie par :

$$F = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Statistique de la fréquence empirique F

Conséquence de la loi des grands nombres, F converge vers la probabilité p.

- En effet, $\sum_i X_i = nF$ suit une loi binomiale de paramètre (n,p). Donc :

$$\mathrm{E}[F] = \frac{1}{n}\mathrm{E}\left[\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}[X_i] = \frac{1}{n}np = p$$

- D'autre part, la variance de F tend vers 0 quand $n \to \infty$:

$$\mathrm{Var}[F] = \frac{1}{n^2}\mathrm{Var}\left[\sum_i X_i\right] = \frac{1}{n^2}\sum_i \mathrm{Var}[X_i] = \frac{1}{n^2}np(1-p) = \frac{p(1-p)}{n}$$

- donc : $\mathrm{Var}[F] = \dfrac{p(1-p)}{n} \xrightarrow[n\to\infty]{} 0$

Statistique de la fréquence empirique F

- La somme $\sum_i X_i$ suit une loi binomiale $\mathcal{B}(n, p)$.
- La loi binomiale tend asymptotiquement vers la loi normale pour $np > 5$:

$$\sum_i X_i \xrightarrow[np>5]{\mathcal{L}} \mathcal{N}\left(np, np(1-p)\right)$$

- Par conséquent, si $np > 5$, la v.a. $F = \frac{1}{n}\sum_i X_i$ est normalement distribuée, de moyenne p et de variance $\frac{p(1-p)}{n}$

$$\boxed{F \xrightarrow[np>5]{\mathcal{L}} \mathcal{N}\left(p, \frac{p(1-p)}{n}\right)}$$

Application : fonction de répartition empirique

L'estimation de la fonction de répartition d'une variable X issue d'un n-échantillon repose sur une évaluation de la fréquence empirique.

- À partir d'un n-échantillon $(x_1, \ldots, x_n)$, on évalue la proportion N_x/n d'éléments de l'échantillon qui sont inférieurs à une valeur x :

$$\widehat{F_X}(x) = \frac{1}{n}\sum_{i=1}^{n}\delta_i(x) = \frac{N_x}{n} \quad \text{où} \quad \delta_i(x) = \begin{cases} 1 & \text{si } x_i \leqslant x, \\ 0 & \text{sinon.} \end{cases}$$

- Pour chaque x_i, N_{x_i} suit une loi binomiale $\mathcal{B}(n, F_X(x_i))$. La fonction en escalier $\widehat{F_X}(x)$ est la fonction de répartition empirique de **l'échantillon**.
- $\widehat{F_X}(x)$ est une estimation de la fonction de répartition F_X de la **population** X.

Application : fonction de répartition empirique

Théorème : La fonction de répartition empirique d'un n-échantillon $\widehat{F_X}$ converge vers la fonction de répartition $F_X(x)$ de la population.

$$\forall x, \qquad \lim_{n\to\infty} \widehat{F_X}(x) = F_X(X)$$

- **Démonstration** : Pour x donné, soit N_x le nombre d'éléments x_k tels que $x_k \leqslant x$, $k = (1, \ldots, n)$. Par définition $\widehat{F_x}(x) = N_x/n$.

– La v.a. N_x est une somme de n variables de Bernoulli de paramètre $F_X(x)$:

$$N_x \sim \mathcal{B}(n, F_X(x))$$

– Par conséquent : $\mathrm{E}[N_x] \equiv \mathrm{E}\left[n\widehat{F_x}(x)\right] = nF_X(x)$

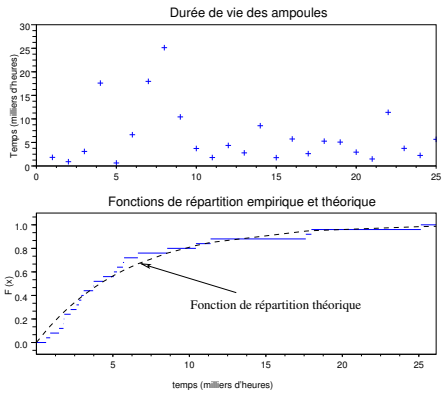
– et donc : $\mathrm{E}\left[\widehat{F_x}(x)\right] = F_x(x)$

Application : fonction de répartition empirique

Pratiquement on construit la fonction de répartition empirique comme suit :

- Trier l'échantillon par ordre croissant. Soient $(x_{(1)}, x_{(2)} \ldots, x_{(n)})$ les éléments de l'échantillon ainsi trié.
- Alors : $\widehat{F_X}(x) = \begin{cases} 0 & x < x_{(1)} \\ k/n & x_{(k)} \leqslant x < x_{(k+1)} \quad (1 \leqslant k \leqslant n-1) \\ 1 & x \geqslant x_{(n)} \end{cases}$

Exemple : on a observé la durée de vie de 25 lampes (figure du haut). La fonction de répartition empirique, $\widehat{F_X}(x)$, ainsi qu'une fonction de répartition théorique (loi exponentielle), $F_X(x)$, sont montrées sur la figure du bas. La fonction empirique $\widehat{F_X}(x)$ est une fonction en escalier, constante entre deux valeurs observées de la durée de vie.



La variance empirique S_n^2

On considère un n-échantillon issu d'une population de moyenne μ et de variance σ^2. On cherche à évaluer σ^2 à partir de la connaissance du n-échantillon.

- **Définition** : La statistique S_n^2 ou variance empirique est définie par :

$$S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

- Notons que l'espérance $\mathrm{E}[X]$ n'est pas connue mais seulement estimée à partir de la moyenne $\overline{X}$ de l'échantillon.
- On voit aisément que : $S_n^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}^2 = \overline{X^2} - \overline{X}^2$

La variance empirique S_n^2

- Calculons $\mathrm{E}[S_n^2]$.
  Pour cela, on décompose S_n^2 en une somme de deux termes en posant : $X_i - \overline{X} = (X_i - \mu) - (\overline{X} - \mu)$.
- Il vient : $S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 - (\overline{X} - \mu)^2$
- Par conséquent :

$$\mathrm{E}[S_n^2] = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}[(X_i-\mu)^2] - E[(\overline{X}-\mu)^2] = \frac{1}{n}\sum_{i=1}^{n}\mathrm{Var}[X] - \mathrm{Var}[\overline{X}]$$

- soit : $\mathrm{E}[S_n^2] = \sigma^2 - \frac{\sigma^2}{n} = \frac{n-1}{n}\sigma^2$ car $\mathrm{Var}[\overline{X}] = \frac{\sigma^2}{n}$.

La variance empirique S_n^2

L'espérance de S_n^2 est inférieure à la variance σ^2 de la population.

- La statistique S_n^2 est biaisée, le biais valant σ^2/n.
- Le biais provient du fait que, ne connaissant pas μ, la moyenne de la population μ est estimée par $\overline{X}$.
- L'estimateur non-biaisé de la variance est :

$$S_{n-1}^2 = \frac{n}{n-1}S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

- Pour n grand ($n > 30$), le biais est négligeable. Par contre pour de petits échantillons, le biais est sensible.

Distribution de S_{n-1}^2

- La somme $Z = \sum_{i=1}^{n}\frac{(X_i - \overline{X})^2}{\sigma^2}$ suit une loi du χ^2 à $(n-1)$ degrés de liberté. Or $\sum_{i=1}^{n}(X_i - \overline{X})^2 = (n-1)S_{n-1}^2$ Par conséquent :

$$Z = (n-1)\frac{S_{n-1}^2}{\sigma^2} \sim \chi_{n-1}^2$$

Soit $X_1, \ldots, X_n$ un n-échantillon de v.a. i.i.d.. On désigne par F_X et f_X la fonction de répartition et la densité de probabilité des X_i.

- Trions par ordre croissant les x_i réalisations de l'échantillon. Soit $x_{(1)} < x_{(2)} \cdots < x_{(n)}$ l'échantillon ordonné par ordre croissant.
- Les v.a. $X_{(i)}$ sont appelées **statistiques d'ordre.**
  - Par exemple, la v.a. $X_{(1)}$ décrit la plus petite valeur de l'échantillon, c'est à dire le minimum.
  - La v.a. $X_{(n)}$ décrit le maximum de l'échantillon.

La médiane constitue un exemple particulièrement utile de statistique d'ordre.

- La médiane $\text{med}(x)$ est la valeur qui permet de partager un échantillon ordonné en deux parties comprenant le même nombre d'éléments.
- Soit $(x_{(1)}, \ldots, x_{(n)})$ un n échantillon trié par ordre croissant. Deux cas sont à considérer.
  – Si l'effectif de l'échantillon est impair : $\boxed{\text{med}(x) = x_{(\frac{n+\mathbf{1}}{\mathbf{2}})}}$
  – Si l'effectif de l'échantillon est pair : $\boxed{\text{med}(x) = (x_{(\frac{n}{\mathbf{2}})} + x_{(\frac{n}{\mathbf{2}}+1)})/2}$
- Contrairement à la moyenne arithmétique, la médiane est un estimateur robuste, i.e. peu sensible aux valeurs aberrantes, permettant d'atténuer l'influence perturbatrice des valeurs extrêmes.
- **Exemple** : On préfère souvent caractériser le niveau des revenus d'une population par le revenu médian. Peu sensible aux quelques très grosses fortunes, le revenu médian partage la population en deux groupes «égaux» : ceux qui ont un revenu inférieur au revenu médian et ceux qui ont un revenu supérieur.

Estimation

L'estimation consiste à évaluer un paramètre caractérisant une population à partir d'un échantillon. Soit θ un paramètre à évaluer (par exemple la proportion de trèfles à quatre feuilles dans un champ de luzerne, ou le niveau moyen de pluie dans l'année à Paris).

- Si $(x_1, \ldots, x_n)$ est un échantillon de taille n de la population, on **estime** θ par une fonction des valeurs de l'échantillon :

$$\boxed{\widehat{\theta} = \phi(x_1, \ldots, x_n)}$$

- **Exemple** : On a vu au chapitre précédent que la moyenne de l'échantillon $\overline{X}$ ou sa variance S_n^2 sont des estimations de l'espérance μ et de la variance σ^2 d'une population.

  De même, la proportion $\widehat{F}$ d'un caractère C dans l'échantillon est une estimation de la proportion p de ce caractère au sein de la population.

Il est clair qu'un autre échantillon issu de la même population eut donné une autre estimation $\widehat{\theta}$ de θ. Par conséquent, $\widehat{\theta}$ est une variable aléatoire, fonction de l'échantillon aléatoire $(X_1, \ldots, X_n)$. Une telle variable aléatoire est une statistique.

Définition : On appelle estimateur une statistique $\widehat{\theta}$ donnant une évaluation d'un paramètre θ d'une population.
Une valeur obtenue de cette statistique (issue d'un échantillon) est une estimation du paramètre.

- **Exemple** : Les statistiques $\widehat{\mu} = \overline{X}$ et $\widehat{\sigma^2} = S_n^2$ sont des estimateurs de μ et σ^2.

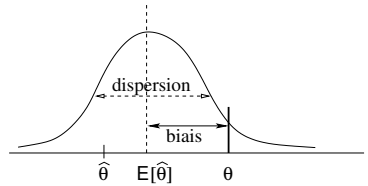
  Les valeurs particulières obtenues à partir d'un échantillon donné $x_1, \ldots, x_n$, $\overline{x}$ et s^2 sont des estimations.

- Il existe deux façons d'estimer un paramètre θ :
  - √ soit on estime une valeur $\widehat{\theta}$ par une statistique issue de l'échantillon. C'est le cas rencontré jusqu'ici. On parle alors d'estimation ponctuelle.
  - √ soit on estime un intervalle dans lequel θ se trouve avec une probabilité donnée. On parle alors d'estimation par intervalle.

Considérons une population caractérisée par une grandeur θ.

- Soit $\widehat{\theta}$ un estimateur de θ obtenu à partir du n-échantillon $X_1, \ldots, X_n$.
- La statistique $\widehat{\theta}$ est une v.a. de densité $f_{\widehat{\theta}}$, et d'espérance $E[\widehat{\theta}]$.

Ci-contre, la d.d.p. (quelconque) de l'estimateur $\widehat{\theta}$ ainsi que la valeur estimée θ (trait gras).



d.d.p. d'un estimateur $\widehat{\theta}$

La qualité d'un estimateur est définie selon les critères suivants : biais, précision, convergence (ou consistance) et robustesse.

Le biais b d'un estimateur $\widehat{\theta}$ est défini par : $\boxed{b = E[\widehat{\theta}] - \theta}$

- Le biais décrit la tendance d'un estimateur à une erreur systématique.
- Un estimateur est sans biais si $b = 0$. Alors $\boxed{E[\widehat{\theta}] = \theta}$.
- **Exemple** : Les estimateurs $\overline{X}$ et F sont des estimateurs sans biais de l'espérance μ et de la proportion p de la population.

  Par contre l'estimateur S_n^2 est un estimateur biaisé de la variance σ^2 puisque :

$$E[S_n^2] = \frac{n-1}{n}\sigma^2$$

La dispersion d'un estimateur $\widehat{\theta}$ est quantifiée par l'écart type $\sigma_{\widehat{\theta}}$:

$$\boxed{\sigma_{\widehat{\theta}}^2 = E\left[(\widehat{\theta} - E[\widehat{\theta}])^2\right]}$$

La précision résume les deux caractères précédents, prenant en compte à la fois le biais et la dispersion.

On mesure **la précision** d'un estimateur par l'erreur quadratique moyenne :

$$\text{EQM} = \mathrm{E}\left[(\widehat{\theta} - \theta)^2\right]$$

L'erreur quadratique moyenne s'exprime en fonction du biais b et de la dispersion $\sigma_{\widehat{\theta}}$:

$$\begin{aligned}\text{EQM} &= \mathrm{E}\left[(\widehat{\theta} - \theta)^2\right] = \mathrm{E}\left[\left(\widehat{\theta} - \mathrm{E}\left[\widehat{\theta}\right] + \mathrm{E}\left[\widehat{\theta}\right] - \theta\right)^2\right] \\ &= \mathrm{E}\left[\left(\widehat{\theta} - \mathrm{E}\left[\widehat{\theta}\right]\right)^2\right] + \mathrm{E}\left[\left(\mathrm{E}\left[\widehat{\theta}\right] - \theta\right)^2\right] = \sigma_{\widehat{\theta}}^2 + b^2\end{aligned}$$

$$\boxed{\text{EQM} == \sigma_{\widehat{\theta}}^2 + b^2}$$

Qualité de l'estimateur ponctuel : précision

- De deux estimateurs sans biais, le plus **précis** est celui dont **la variance est minimale**.

Si l'estimateur est biaisé : $\mathrm{E}\,[(\widehat{\theta}-\theta)^2] \neq \mathrm{E}\,[\widehat{\theta}^2] - \theta^2$.
Voyez vous pourquoi ?

- Si l'estimateur n'est pas biaisé, $b = 0 \iff \mathrm{E}\,[\widehat{\theta}] = \theta$.

  Par conséquent, l'erreur quadratique moyenne est simplement égale à la variance de l'estimateur :

$$EQM = \mathrm{E}\left[(\widehat{\theta} - \mathrm{E}\,[\widehat{\theta}])^2\right] = \sigma^2_{\widehat{\theta}}$$

- **Exemple** : L'erreur quadratique moyenne associée à S_n^2 est

$$\begin{aligned} \mathrm{EQM}(S_n^2) &= \mathrm{E}\,[\,(S_n^2 - \sigma^2)^2] = \mathrm{E}\,[(S_n^2)^2 - 2S_n^2\sigma^2 + \sigma^4] \\ &= \mathrm{E}\,[(S_n^2)^2] - \sigma^4 \frac{n-2}{n} \end{aligned}$$

  où l'on a utilisé le fait que $\mathrm{E}\,[S_n^2] = \dfrac{n-1}{n}\sigma^2$.

- **Exemple** : Si on connaît μ, l'estimation de σ^2 par $\widehat{\sigma^2} = \overline{(X-\mu)^2}$ est meilleure que S_n^2 car de variance moindre.

Qualité de l'estimateur ponctuel : convergence

Un estimateur $\widehat{\theta}$ est convergent si :

$$\widehat{\theta} \xrightarrow[n\to\infty]{} \theta$$

- **Exemple** : Les trois estimateurs, $\overline{X}$, F et S_n^2 sont convergents. Pour les estimateurs non biaisés, $\overline{X}$ et F, c'est une simple conséquence de la loi des grands nombres.
  - Par exemple, considérons la proportion F au sein d'un n-échantillon. La loi des grands nombres s'écrit :

$$\Pr[\,|F-p| > \alpha] \leqslant \frac{\sigma_F^2}{\alpha^2} = \frac{p(1-p)}{n\alpha^2}$$

    car $\mathrm{Var}\,[F] = \dfrac{np(1-p)}{n^2} = \dfrac{p(1-p)}{n}$.
    L'estimateur F étant sans biais, l'erreur quadratique moyenne EQM s'écrit :

$$EQM = \sqrt{\mathrm{Var}\,[F]} = \sqrt{\frac{p(1-p)}{n}} \xrightarrow[n\to\infty]{} 0$$

Qualité de l'estimateur ponctuel : convergence

- La variance de S_n^2 tend également vers 0 pour $n \to \infty$.
  - Cependant, il faut aussi tenir compte du biais b. Or :

$$b(S_n^2) = \frac{n-1}{n}\sigma^2 - \sigma^2 = -\frac{\sigma^2}{n} \xrightarrow[n\to\infty]{} 0$$

    L'erreur quadratique $EQM(S_n^2)$ tendant asymptotiquement vers 0, la variance d'échantillon est un estimateur convergent.

Qualité de l'estimateur ponctuel : Robustesse

Il arrive parfois que des valeurs extrêmes et rares apparaissent dans un échantillon. Ce peut être dû à une erreur d'échantillonnage, quelques individus étant issus d'une autre population (on parle alors d'intrus), ou encore à l'occurrence d'un ou de quelques individus hors-normes, voire à des erreurs de mesures.

- Un estimateur est robuste si la valeur de l'estimateur dépend peu des valeurs extrêmes.
- **Exemple :** La médiane est un estimateur robuste, très peu sensible à la présence de quelques valeurs aberrantes.

  En revanche, la moyenne n'est pas un estimateur robuste, surtout si l'échantillon est de petite taille.

Estimation d'un intervalle de confiance

L'estimation ponctuelle n'offre qu'un intérêt limité si l'on ne sait préciser la précision de l'estimation. Aussi, il conviendra d'évaluer une erreur standard (à partir de l'erreur quadratique moyenne).
Alternativement, on cherchera à préciser un intervalle de confiance dans lequel trouver «presque sûrement» θ. Cette dernière façon de procéder est appelée estimation par intervalle.

- On recherche donc les limites d'un l'intervalle $[\theta_1, \theta_2]$ dans lequel se situe θ avec le **risque à priori** α de se tromper.
- Il n'est possible de déterminer un tel intervalle que si l'on connaît (ou si l'on suppose connue) la loi de probabilité de l'estimateur $\widehat{\theta}$.
- On fixe à priori le niveau de confiance proche de l'unité, qu'on définit en fonction du risque α de se tromper.
- Les valeurs de risque couramment utilisées sont $\alpha = 0,05$ ou $\alpha = 0,01$, ou encore $\alpha = 0,001$.
- Le niveau de confiance est la probabilité :

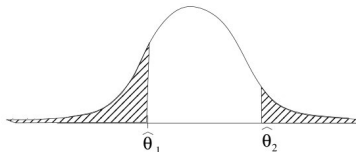
$$1 - \alpha = 95\% \quad \text{ou} \quad 99\% \quad \text{ou encore} \quad 99,9\%$$

Estimation d'un intervalle de confiance

Le niveau de confiance $1 - \alpha$ étant fixé, on cherche les limites de **l'intervalle de confiance** $[\theta_1, \theta_2]$ dans lequel placer $\widehat{\theta}$.

$$\Pr[\theta_1 < \widehat{\theta} < \theta_2] = 1 - \alpha$$

L'intervalle de confiance n'est pas complètement déterminé par le risque α. En effet, la probabilité α pour que θ soit en dehors de l'intervalle de confiance impose seulement :



$$\Pr\left[(\widehat{\theta} > \theta_2) \cup (\widehat{\theta} < \theta_1)\right] = \Pr[\widehat{\theta} > \theta_2] + \Pr[\widehat{\theta} < \theta_1] = \alpha$$

Il existe une infinité d'intervalles vérifiant la propriété ci-dessus. On choisit habituellement de diviser le risque en deux parts égales :

$$\Pr[\widehat{\theta} > \theta_2] = \Pr[\widehat{\theta} < \theta_1] = \frac{\alpha}{2}$$

On exprime les limites de l'intervalle en fonction θ (espérance de $\widehat{\theta}$) et de l'écart type $\sigma_{\widehat{\theta}}$:

$$\begin{cases} \theta_1 &= \theta - \zeta_{\alpha/2}\sigma_{\widehat{\theta}} \\ \theta_2 &= \theta + \zeta_{1-\alpha/2}\sigma_{\widehat{\theta}} \end{cases}$$

Estimation d'un intervalle de confiance

La probabilité $\Pr[\widehat{\theta} > \theta_2]$ s'écrit : $\Pr[\widehat{\theta} > \theta + \zeta_{1-\alpha/2}\sigma_{\widehat{\theta}}] = \dfrac{\alpha}{2}$

soit : $\Pr\left[\dfrac{\widehat{\theta} - \theta}{\sigma_{\widehat{\theta}}} > \zeta_{1-\alpha/2}\right] = \dfrac{\alpha}{2}$. De même : $\Pr\left[\dfrac{\widehat{\theta} - \theta}{\sigma_{\widehat{\theta}}} < -\zeta_{\alpha/2}\right] = \dfrac{\alpha}{2}$

Les deux équations précédentes sont résumées par :

$$\Pr\left[-\zeta_{\alpha/2} \leqslant \frac{\widehat{\theta} - \theta}{\sigma_{\widehat{\theta}}} \leqslant \zeta_{1-\alpha/2}\right] = 1 - \alpha$$

Si la distribution est symétrique, $\zeta_{\alpha/2} = \zeta_{1-\alpha/2}$. On en déduit l'intervalle de confiance pou θ :

$$\boxed{\Pr\left[\widehat{\theta} - \zeta_{1-\alpha/2}\sigma_{\widehat{\theta}} \leqslant \theta \leqslant \widehat{\theta} + \zeta_{1-\alpha/2}\sigma_{\widehat{\theta}}\right] = 1 - \alpha}$$

si la loi de la variable centrée-réduite est connue, l'estimation d'un intervalle de confiance consiste à déterminer la valeur critique $\zeta_{\alpha/2}$ correspondant au niveau de confiance $1 - \alpha$.

Intervalle de confiance pour la moyenne

Intervalle de confiance de la moyenne d'une population, σ **étant connu**.

- On cherche à estimer un intervalle de confiance pour la moyenne m d'une population X. Soit $\overline{X}$ la moyenne d'un n-échantillon. Supposons σ connu.
- Si l'estimateur $\overline{X}$ de m est distribué (au moins approximativement) suivant une loi normale $\mathcal{N}(m, \sigma^2/n)$, l'intervalle de confiance au niveau $1 - \alpha$ s'écrit :

$$\Pr\left[\overline{X} - z_{1-\alpha/2}\sigma_{\overline{X}} \leqslant m \leqslant \overline{X} + z_{1-\alpha/2}\sigma_{\overline{X}}\right] = 1 - \alpha$$

ou, de façon équivalente :

$$\boxed{\Pr\left[\left|\frac{m - \overline{X}}{\sigma_{\overline{X}}}\right| \leqslant z_{1-\alpha/2}\right] = 1 - \alpha}$$

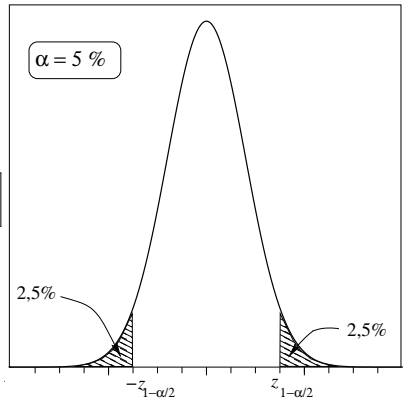
On sait que $\sigma_{\overline{X}} = \sigma/\sqrt{n}$.

Intervalle de confiance pour la moyenne

En fonction de la variable centrée-réduite $U = \dfrac{m - \overline{X}}{\sigma_{\overline{X}}}$:

$$\boxed{\Pr\left[-z_{1-\alpha/2} \leqslant U \leqslant z_{1-\alpha/2}\right] = 1 - \alpha}$$

Le choix du risque α détermine complètement les bornes de l'intervalle de confiance $[-z_{1-\alpha/2}, z_{1-\alpha/2}]$.



$$\begin{aligned} Pr[-z_{1-\alpha/2} \leqslant U \leqslant z_{1-\alpha/2}] &= Pr[U \leqslant z_{1-\alpha/2}] - Pr[U \leqslant -z_{1-\alpha/2}] \\ &= \Phi(z_{1-\alpha/2}) - \Phi(-z_{1-\alpha/2}) = 2\Phi(z_{1-\alpha/2}) - 1 \end{aligned}$$

où Φ est la fonction de répartition de la loi normale centrée-réduite ($U \sim \mathcal{N}(0,1)$).

Intervalle de confiance pour la moyenne

Il vient : $2\Phi(z_{1-\alpha/2}) - 1 = 1 - \alpha$ soit : $\Phi(z_{1-\alpha/2}) = 1 - \dfrac{\alpha}{2}$

Et donc :

$$\boxed{z_{1-\alpha/2} = \Phi^{-1}(1 - \alpha/2)}$$

La table de la loi normale indique :

- pour $1 - \alpha = 0.95$, $z_{1-\alpha/2} = 1,96$.
- pour $1 - \alpha = 0.99$, $z_{1-\alpha/2} = 2,58$.

Ainsi, on dira que l'intervalle de confiance au niveau $0,95$ pour la moyenne est

$$m \in [\overline{X} - 1,96\frac{\sigma}{\sqrt{n}}, \overline{X} + 1,96\frac{\sigma}{\sqrt{n}}]$$

Intervalle de confiance pour la moyenne

Le plus souvent, σ n'est pas connu mais estimé à partir de l'échantillon :

$$\widehat{\sigma_X^2} = S_{n-1}^2 = \frac{1}{n-1}\sum(X_i - \overline{X})^2$$

L'espérance μ est estimée par $\widehat{\mu} = \overline{X}$. On forme la variable centrée, réduite :

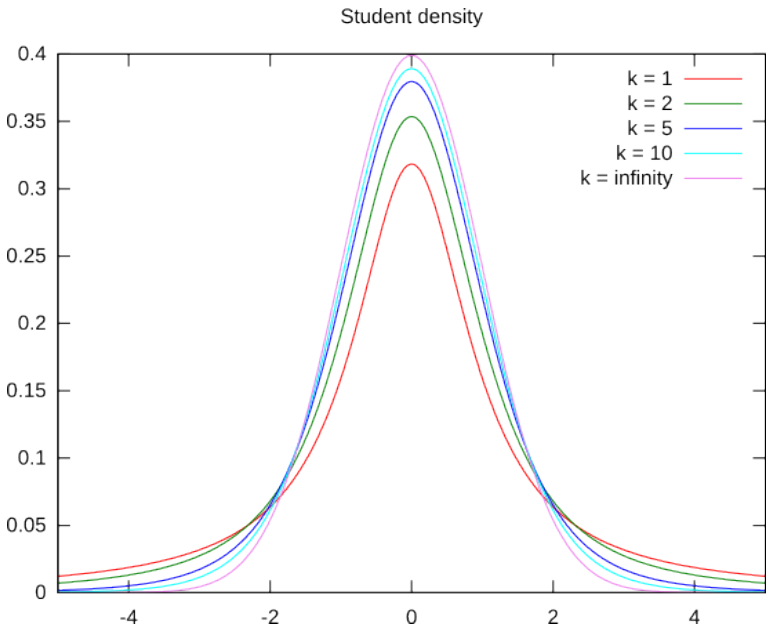
$$\boxed{T = \frac{\overline{X} - \mu}{\widehat{\sigma}_X/\sqrt{n}}}$$

La v.a. T suit une loi de Student à $(n-1)$ degrés de liberté.

Rappel :

Soit $U = (X - \mu)/\sigma$ une v.a. centrée réduite normalement distribuée.
Soit Z une v.a. indépendante de U suivant une loi du χ^2 à n degrés de liberté.
Alors la v.a. $T = U/\sqrt{Z/n}$ suit une loi de Student à n degrés de liberté.

Loi de Student



d.d.p. de lois de Student. Pour $k \to \infty$, $t(\infty) \to \mathcal{N}(0,1)$

Intervalle de confiance pour la moyenne

Montrons que T suit bien une loi de Student.

$$T = \frac{\overline{X} - \mu}{\frac{\widehat{\sigma}_X}{\sqrt{n}}\frac{\sigma_X}{\sigma_X}} = \frac{U}{\widehat{\sigma}_X/\sigma_X} = \frac{U}{\frac{\sqrt{n-1}\widehat{\sigma}_X}{\sqrt{n-1}\sigma_X}} = \frac{U}{\sqrt{Z/(n-1)}}$$

où $U = (\overline{X} - m)/(\sigma_X/\sqrt{n}) \sim \mathcal{N}(0,1)$ et $Z = (n-1)\widehat{\sigma}_X^2/\sigma_X^2 \sim \chi^2_{n-1}$.

Donc $T \sim t(n-1)$.

On déduit un intervalle de confiance pour μ au seuil de confiance α :

$$\boxed{\Pr\left[\mu \in \left\{\overline{X} - t^{n-1}_{1-\alpha/2}\frac{\widehat{\sigma}_X}{\sqrt{n}} \; ; \; \overline{X} + t^{n-1}_{1-\alpha/2}\frac{\widehat{\sigma}_X}{\sqrt{n}}\right\}\right] = 1-\alpha}$$

La loi de Student converge vers la loi normale pour n grand. Pratiquement, pour $n > 100$, on approchera la loi de Student par la loi normale.

Intervalle de confiance pour une proportion

Soit $F = k/n$ une fréquence empirique d'occurrence d'un caractère C au sein d'une population. F est une estimation de la proportion p au sein de la population.
On cherche un intervalle de confiance pour F

- Posons $Y = nF = k$. On sait que $Y \sim \mathcal{B}(n,p)$, et donc, $\mathrm{E}[Y] = np$, $\mathrm{Var}[Y] = np(1-p)$.
- **Rappel :** Si $np > 5$, la loi binomiale peut être approchée par une loi normale : $\mathcal{B}(n,p) \underset{np>5}{\longrightarrow} \mathcal{N}(np, np(1-p))$
- Sous cette condition : $F \sim \mathcal{N}\left(p, \frac{p(1-p)}{n}\right) = \mathcal{N}\left(p, \sigma_F^2\right)$
- On définit la v.a. centrée réduite $U = (F-p)/\sigma_F$, où $U \sim \mathcal{N}(0,1)$. L'intervalle de confiance au niveau de confiance $1-\alpha$ s'écrit :

$$\boxed{\Pr\left[F - z_{1-\alpha/2}\sigma_F \leqslant p \leqslant F + z_{1-\alpha/2}\sigma_F\right] = 1-\alpha}$$

où $z_{1-\alpha/2} = \Phi^{-1}(1-\alpha/2)$, Φ étant la fonction de répartition de la loi normale centrée-réduite.

Intervalle de confiance pour une proportion

L'écart type σ_F n'est pas connu. Une possibilité est de l'estimer par :

$$\widehat{\sigma}_F = \sqrt{\frac{F(1-F)}{n}}$$

Exemple : $n = 100$, $F = 0,6$. Quel intervalle donne une confiance de 0,9 ?

L'intervalle s'écrit :

$$p \in \left\{0,6 - z_{1-\alpha/2}\sqrt{\frac{0,6\times 0,4}{100}} \; ; \; 0,6 + z_{1-\alpha/2}\sqrt{\frac{0,6\times 0,4}{100}}\right\}$$

où $z_{1-\alpha/2} = \Phi^{-1}(1-\alpha/2) = \Phi^{-1}(0,95) = 1,641$, Φ étant la fonction de répartition de la loi normale centrée-réduite.

Il vient : $\boxed{0,5196 \leqslant p \leqslant 0,6804}$ au niveau de confiance 90%.

Intervalle de confiance pour un écart type

On sait que la variable $Z^2 = (n-1)\frac{\widehat{\sigma}^2}{\sigma^2}$ suit une loi du χ^2 de $(n-1)$ degrés de liberté.

- Pour estimer un intervalle de confiance au niveau $1-\alpha$ pour σ, il suffit de construire un intervalle de confiance pour Z.

$$\Pr[x^2_{\alpha/2} \leqslant Z \leqslant x^2_{1-\alpha/2}] = 1-\alpha$$

où $x^2_{\alpha/2}$ est donné par l'inverse de la fonction de répartition $F^{-1}_{\chi^2}(\alpha/2)$.

- Par conséquent, l'intervalle de confiance au niveau $1-\alpha$ pour σ^2 s'écrit :

$$(n-1)\frac{\widehat{\sigma}^2}{x^2_{1-\alpha/2}} \leqslant \sigma^2 \leqslant (n-1)\frac{\widehat{\sigma}^2}{x^2_{\alpha/2}}$$

- soit pour l'écart type σ : $\boxed{\sqrt{n-1}\frac{\widehat{\sigma}}{x_{1-\alpha/2}} \leqslant \sigma \leqslant \sqrt{n-1}\frac{\widehat{\sigma}}{x_{\alpha/2}}}$

Exemple : $n = 10$ et $\alpha = 5\%$:

$$3\frac{\widehat{\sigma}}{\sqrt{19}} \leqslant \sigma \leqslant 3\frac{\widehat{\sigma}}{\sqrt{2,7}} \quad \Longrightarrow \quad 0.69\,\widehat{\sigma} \leqslant \sigma \leqslant 1.83\,\widehat{\sigma}$$

Décision : tests d'hypothèse

Peut-on valider telle ou telle hypothèse sur une population à partir de la seule connaissance d'un échantillon ? Le test d'hypothèse permet de prendre une décision en quantifiant le risque.

- Introduisons le sujet de la décision par un exemple emprunté à l'excellent ouvrage de Saporta (2006).
- Des relevés effectués durant plusieurs années montrent que le niveau de pluie annuel dans la Beauce, exprimé en mm, suit une loi normale $\mathcal{N}(600, 100^2)$.
- Des entrepreneurs, surnommés faiseurs de pluie, affirmaient pouvoir augmenter le niveau moyen de pluie en inséminant des nuages par de l'iodure d'argent.
- Le procédé fut testé entre 1951 et 1959 et on releva les niveaux de pluie suivants :

Année	1951	1952	1953	1954	1955	1956	1957	1958	1959
mm	510	614	780	512	501	534	603	788	650

Comment conclure ? Deux hypothèses s'opposent : soit l'insémination augmente le niveau de pluie, soit elle est sans effet.

- Précisons les hypothèses. Soit X la variable aléatoire égale au niveau annuel de pluie. Les deux hypothèses sont notées H_0 et H_1. Si H_0 est vraie, le niveau moyen de pluie est 600 mm.
- Dans le cas contraire, le niveau moyen est 650 mm (au dire des faiseurs de pluie). Notons
  - √ $\mathrm{E}[X|H_0]$ l'espérance de X sous l'hypothèse H_0,
  - √ $\mathrm{E}[X|H_1]$ l'espérance sous l'hypothèse H_1.

$$\begin{cases} \mathrm{E}[X|H_0] = \mu_0 = 600 \text{ mm} \\ \mathrm{E}[X|H_1] = \mu_1 = 650 \text{ mm} \end{cases}$$

- Les données relevées indiquent $\overline{X} = 610$ mm. Peut-on conclure que H_1 est vraie ?
- Nous devons choisir l'une ou l'autre hypothèse comme vraie, au risque, bien sûr, de se tromper (le risque zéro n'existe pas).

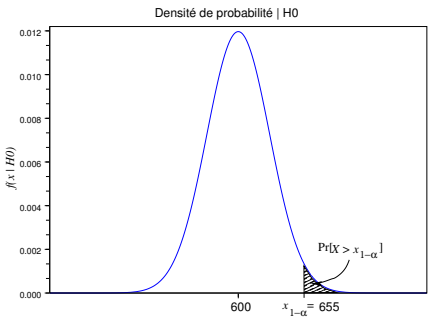
- Fixons le risque d'erreur à $\alpha = 0.05$. La probabilité α est le risque de choisir H_1 alors que H_0 est vraie. Ce faisant, on assume le risque de se tromper dans 5% des cas.

Le test porte sur le niveau moyen de pluie $\overline{X}$ sur les 9 années d'observations. On connaît la distribution de $(\overline{X}|H_0)$ (courbe ci-contre) :

$$(\overline{X}|H_0) \sim \mathcal{N}(\mu_0, \sigma^2/9)$$

Compte tenu du risque assumé (5%), le niveau de pluie critique, c-à-d le niveau au delà duquel on ne retient plus l'hypothèse H_0 (on choisit donc H_1) est $x_{1-\alpha}$ tel que :

$$\Pr[\overline{X} \geqslant x_{1-\alpha}] \leqslant 0,05$$



d.d.p. X conditionné à H_0

Test d'hypothèse : un exemple

- La variable du test est $\xi = \dfrac{\overline{X} - \mu_0}{\sigma/3} = 0.3$.
- Exprimons la probabilité que la variable du test soit supérieure à ξ :

$$\Pr\left[\frac{\overline{X} - \mu_0}{\sigma/3} > \xi\right] = 0{,}38 \gg 0{,}05$$

  La probabilité obtenue $(0, 38)$ s'appelle la valeur p.
- On conclut donc qu'on ne peut rejeter H_0 au risque 5%, c'est à dire que l'ensemencement est sans effet notable.
- On dira qu'on conserve l'hypothèse H_0 au niveau de signification de 5%.
- À partir de la table de la loi normale on trouve que le seuil permettant de rejeter H_0 est :

$$\xi_{1-\alpha} = \Phi^{-1}(0,95) = 1,65 \implies \overline{x}_{0.95} = \mu_0 + \xi_{1-\alpha}\frac{\sigma}{\sqrt{9}} = 600 + 1,65 \times \frac{100}{3} = 655 \text{ mm}$$

Test d'hypothèse : un exemple

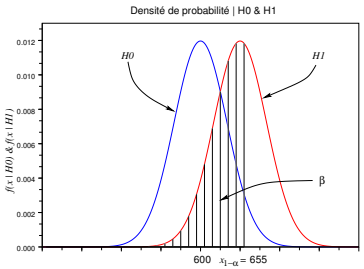
Cependant, on ne peut affirmer que les faiseurs de pluie ont tort.

1. Croire les faiseurs de pluie alors qu'ils ont tort $\iff$ rejeter H_0 à tort. La probabilité de commettre cette erreur est $\alpha = 5\%$
2. Ne pas croire les faiseurs de pluie alors qu'ils ont raison $\iff$ rejeter H_1 à tort.

Supposons que les faiseurs de pluie aient raison (hypothèse H_1 vraie, $\mu_1 = 650$ mm). Alors $\overline{X} \sim \mathcal{N}(\mu_1, \sigma^2/9)$.
On commet donc une erreur chaque fois que $\overline{X} < 655$. Or cette erreur survient avec une probabilité β :

$$\beta = \Pr[\overline{X} < 655] = \Pr\left[\frac{\overline{X} - \mu_1}{\sigma/3} < 0,15\right]$$

Il vient $\beta = 0.56$, ce qui est considérable !



d.d.p. de X conditionné à H_0 et H_1

Il existe donc deux sortes d'erreurs associées à la décision :

- α est l'erreur de première espèce (indépendante de H_1).
- β est l'erreur de seconde espèce.

La valeur p

Définition : Le niveau de signification (en anglais p-value) est la probabilité d'obtenir une valeur plus extrême (supérieure ou inférieure) que la valeur obtenue pour la variable du test si l'hypothèse H_0 était vraie.

Si la valeur-p est inférieure au niveau de signification α préalablement défini, i.e. $\alpha = 5\%$ ou 1%, on rejette l'hypothèse H_0.

En d'autres termes, la valeur p est la probabilité de commettre une erreur de première espèce, c'est-à-dire de rejeter à tort l'hypothèse H_0.

Exemple : Pour l'exemple des faiseurs de pluie, $\overline{x} = 610$, donc

$$z = \frac{\overline{x} - 600}{\sigma/3} = 0.3$$

- La valeur p est donnée par $p = \Pr[U > z = 0.3] = 0,38$.
- Comme p est beaucoup plus grand que $\alpha = 0.05$, on retient H_0.
- Pour tout seuil $p > \alpha$, on retiendra H_0.

Tests d'hypothèses : généralités

Définition : Un test d'hypothèse est une démarche probabiliste permettant de faire faire un choix concernant une hypothèse H_0 au vu des résultats d'un échantillon.

Soient H_0 et H_1 deux hypothèses qui s'excluent l'une l'autre, c-à-d que l'une seulement est vraie. La décision consiste à faire un choix, ce qui conduit aux quatre possibilités suivantes :

	Vérité	
Décision	H_0	H_1
H_0	$1-\alpha$	β
H_1	α	$1-\beta$

α et β sont les probabilités d'erreur.

- α est la probabilité de choisir H_1 alors que H_0 est vrai (erreur de première espèce).
- β est la probabilité de choisir H_0 alors que H_1 est vrai (erreur de deuxième espèce).

Dans la pratique, il est de règle de fixer α, la distribution de la v.a. étant supposée connue sous H_0.

Les valeurs couramment choisies pour α sont 0.1, 0.05 ou 0.01.

Puissance du test

- L'hypothèse H_0 joue un rôle prééminent puisque c'est celle que l'on rejette ou accepte. Le choix de H_0 résulte souvent du fait que l'on ne sait exprimer la loi de la variable X que sous cette seule hypothèse.
- α et H_0 choisis, β sera calculé si on sait exprimer la loi de probabilité de X sous l'hypothèse H_1.
- Si on peut formuler une hypothese H_1 (ce qui n'est pas toujours possible), on pourra estimer β par des simulations numériques (méthodes de Monte-Carlo).
- Il est crucial de remarquer que α et β varient en sens contraires.
- Le seul moyen de réduire conjointement les deux erreurs est d'augmenter la taille de l'échantillon (la dispersion des moyennes diminue en $1/\sqrt{n}$).

Définition : La probabilité $1-\beta$ de choisir H_1 quand H_1 est effectivement réalisée définit la puissance du test.

Le test fournissant l'erreur β la plus petite pour une même valeur de α est, par définition, le plus puissant.

Région d'acceptation et région critique

- Lors de l'élaboration d'un test, la première étape consiste à définir une variable de décision.
- C'est une statistique qui doit permettre de faire un choix entre les deux hypothèses.
- La loi de probabilité de cette statistique doit être connue, au moins dans l'hypothèse H_0.
- On appelle région critique et l'on note W, l'ensemble des valeurs de la variable de décision conduisant au rejet de l'hypothèse H_0.
- La région complémentaire, appelée région d'acceptation et notée $\overline{W}$ conduit donc à l'acceptation de H_0.

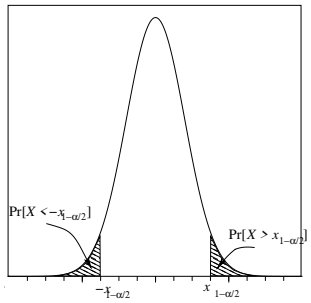
Test bilatéral

On appelle test bilatéral un test pour lequel la région critique est de part et d'autre de la région d'acceptation.

- Dans ce cas, les hypothèses sont du type :

$$\begin{aligned} H_0 &: \theta = \theta_0 \\ H_1 &: \theta \neq \theta_0 \end{aligned}$$

- Peu importe que le paramètre soit plus grand ou plus petit que θ_0, ce qui importe, c'est qu'il diffère peu de la valeur supposée.
- Un test bilatéral s'applique quand on cherche une différence entre deux estimations, ou entre une estimation et une valeur supposée, sans se préoccuper du signe de la différence.



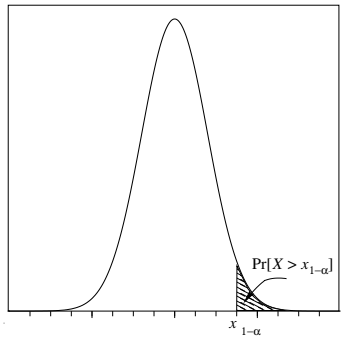
Test unilatéral

Un test unilatéral est un test pour lequel la région critique est d'un côté seulement de la région d'acceptation.

- Dans ce cas, les hypothèses sont de la forme :

$$\begin{aligned} H_0 &: \theta \leqslant \theta_0 \\ H_1 &: \theta > \theta_0 \end{aligned}$$

- Peu importe que le paramètre soit proche ou non de θ_0, ce qui compte, c'est qu'il dépasse un seuil critique.
- Un test unilatéral s'applique quand on cherche à déterminer si une estimation est supérieure à une autre, ou à une valeur supposée à priori.



Il existe un nombre considérable de tests, de conception très différentes. Voyons en quelques exemples.

Tests d'hypothèse simple : comparaison à un standard

Cas correspondant à l'exemple introductif : le test permet de décider, au niveau de signification α, si une statistique est égale ou non une valeur supposée :

$$\begin{cases} H_0 : \theta = \theta_0 \\ H_1 : \theta = \theta_1 \end{cases}$$

On compare une statistique (issue d'un échantillon) à une valeur supposée à priori.

Le test vise à déterminer si la valeur issue de l'échantillon est ou non compatible avec celle fixée à priori.

Exemple : test d'une moyenne d'une population

Soit X une v.a. de moyenne m et d'écart type σ, dont on connaît un n-échantillon. On souhaite savoir si la moyenne m de l'échantillon est ou non égale à la moyenne attendue m_0 (moyenne de la population).

$$\begin{cases} H_0 : m = m_0 \\ H_1 : m \neq m_0 \end{cases}$$

La moyenne de l'échantillon est simplement : $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.

On suppose que $\overline{X}$ est normalement distribué : $\overline{X} \sim \mathcal{N}(m, \sigma^2/n)$

> Même si la population X n'est pas normalement distribuée, cette dernière hypothèse est justifiée pour $\overline{X}$ par le théorème de la limite centrale pour n grand.

Test d'une moyenne d'une population

Cas où σ est connu

- La statistique du test est la variable centrée-réduite :

$$\boxed{Z = \frac{\overline{X} - m_0}{\sigma/\sqrt{n}}} \text{ où } Z \sim \mathcal{N}(0,1)$$

- Le test est bilatéral, et symétrique, car m peut différer de m_0 par défaut ou par excès.
- On calcule la valeur-p associée à $|Z|$ (table de la loi normale)
- Reste à comparer la valeur-p associée à $|Z|$ au seuil $\alpha/2$.
  - √ Si $p \geqslant \alpha/2$, on ne peut rejeter H_0, donc on retient H_0,
  - √ Si $p < \alpha/2$, on rejette H_0.

Test d'une moyenne d'une population

Cas où σ n'est pas connu :

- Dans ce cas, la variance $\widehat{\sigma^2}$ est estimée par la statistique S^2_{n-1} de l'échantillon :

$$\widehat{\sigma^2} \equiv S^2_{n-1} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

- La statistique du test est : $\boxed{T_{n-1} = \frac{\overline{X} - m_0}{\widehat{\sigma}/\sqrt{n}}}$
- T_{n-1} suit une **loi de Student** à $(n-1)$ degrés de liberté.
- Reste à comparer la valeur-p associée à $|T_{n-1}|$ au seuil $\alpha/2$:
  - √ Si $p \geqslant \alpha/2$, on ne peut pas rejeter H_0,
  - √ Si $p < \alpha/2$, on rejette H_0.

  Le seuil $t_{1-\alpha/2}$ est donné par la table de la loi de Student à $n-1$ degrés de liberté.

Test d'une proportion

On souhaite tester la fréquence d'occurrence d'un caractère au sein d'une population à partir d'une proportion dans un n-échantillon.

- L'hypothèse H_0 postule que l'échantillon est représentatif de la population :
  - le nombre d'occurrences N dans l'échantillon suit alors une loi binomiale de paramètre (n, p), p étant la probabilité d'occurrence.
  - la fréquence $F = N/n$ de l'échantillon suit une loi binomiale $\mathcal{B}(n, p)$. Si $np > 5$, $F \sim \mathcal{N}(p, p(1-p)/n)$.
- On se ramène alors au cas d'un test sur une variable centrée-réduite normalement distribuée :

$$Z = \frac{F - p}{\sqrt{p(1-p)/n}}$$

- Reste à comparer la valeur-p associée à $|Z|$ au seuil $\alpha/2$:
  - √ Si $p \geqslant \alpha/2$, on ne peut rejeter H_0,
  - √ Si $p < \alpha/2$, on rejette H_0.

Test d'une proportion

Exemple : on prélève $n = 100$ pièces produites par une machine. On en trouve $N = 8$ défectueuses. Peut-on admettre la proportion annoncée de 6% de pièces défectueuses au niveau de confiance 95% ?

- La fréquence observée, $F = 0,08$, est à comparer à la fréquence attendue $f_0 = 0,06$.
- Définissions les deux hypothèses : $\begin{cases} H_0 : \mathrm{E}[F] = f_0 \\ H_1 : \mathrm{E}[F] = f_1 \neq f_0 \end{cases}$
- Sous l'hypothèse H_0, le nombre de pièces défectueuses de l'échantillon suit une loi binomiale : $N \sim \mathcal{B}(100, f_0)$.
- L'écart type de N est $\sigma_F = \sqrt{\frac{f_0(1-f_0)}{n}}$. l'échantillon étant de 100 pièces, il est légitime de faire l'approximation de la loi normale. La statistique du test est :

$$Z = \frac{F - f_0}{\sigma_F} = 0.84$$

- La valeur-p, $\Pr[Z > 0.84]$, est 0,20. Elle est supérieure à $\alpha/2 = 0.025$. On ne peut rejeter H_0.

Comparaison de deux échantillons

On se trouve devant deux échantillons dont on ignore s'ils proviennent ou non de la même population.

On cherche à savoir si les statistiques X_1 et X_2 observées dans les deux échantillons sont significativement différentes car issues de populations différentes.

- On teste : $\begin{cases} H_0 : X_2 = X_1 = 0 \\ H_1 : X_2 \neq X_1 \end{cases}$
- La statistique du test est la différence $\Delta X = X_1 - X_2$.
- Sous l'hypothèse H_0 :
  - $\mathrm{E}[\Delta X] = 0$
  - $\mathrm{Var}[\Delta X] = \mathrm{Var}[X_1] + \mathrm{Var}[X_2]$ pourvu que les échantillons soient indépendants.

Test des proportions dans deux échantillons

On veut savoir si les proportions P_1 et P_2 issues de deux échantillons différents sont ou non issus d'une même population. Sous l'hypothèse H_0 les proportions observées sont représentatives d'une même population.

- Sous l'hypothèse H_0, P_1 et P_2 suivent des lois binomiales $\mathcal{B}(n_i, p)$, $i = \{1, 2\}$.
- Pourvu que $n_i p > 5$, ces lois binomiales peuvent être approchées par des loi normales $\mathcal{N}(p, p(1-p)/n_i)$.
- Sous H_0, la différence ΔP suit donc une loi normale de moyenne nulle.

$$\Delta P_{H_0} = (P_2 - P_1)_{|_{H_0}} \sim \mathcal{N}\left(0, p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right)$$

- Sous H_1, la différence ΔP suit également une loi normale, mais d'espérance $\neq 0$:

$$\Delta P_{H_1} = (P_2 - P_1)_{|_{H_1}} \sim \mathcal{N}\left(p_2 - p_1, \frac{p_2(1-p_2)}{n_2} + \frac{p_1(1-p_1)}{n_1}\right)$$

Test des proportions dans deux populations

- Ne connaissant pas p, on l'estime (sous H_0) par :

$$\widehat{P} = \frac{n_1 P_1 + n_2 P_2}{n_1 + n_2}$$

- La variance de ΔP est estimée par :

$$\widehat{\sigma^2_{\Delta P}} = \widehat{P}(1-\widehat{P})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$$

- La statistique du test est la v.a. de Student centrée réduite de $n_1 + n_2 - 2$ degrés de liberté : $T_{n-2} = \dfrac{\Delta P}{\sigma_{\Delta P}}$.
- Reste à comparer la valeur-p associée à $|T_{n-2}|$ au seuil $\alpha/2$:
  - √ Si $p \geqslant \alpha/2$, on ne peut pas rejeter H_0,
  - √ Si $p < \alpha/2$, on rejette H_0.

Test des proportions dans deux populations

Exemple : Il y a deux sortes de shadoks, les shadoks d'en haut et les shadoks d'en bas. Les individus des deux groupes de shadoks possèdent un caractère C en proportion p et q. On souhaite savoir si la proportion d'individus possédant le caractère C est la même dans les deux groupes. Les deux hypothèses sont :

$$\begin{cases} H_0 : p = q \\ H_1 : p \neq q \end{cases}$$

- On prélève un échantillon de taille $n_1 = 200$ dans la population de shadoks du bas, et un échantillon de taille $n_2 = 50$ dans la population des shadoks du haut (ils sont beaucoup plus rares).
- Sur ces échantillons, on observe les fréquences $P_1 = 0.34$ et $P_2 = 0.4$.

Test des proportions dans deux populations

- sous H_0 :

$$\widehat{P} = \frac{n_1 P_1 + n_2 P_2}{n_1 + n_2} = 0.352$$

$$\widehat{\sigma^2_{\Delta P}} = \widehat{P}(1-\widehat{P})\left(\frac{1}{n_1} + \frac{1}{n_2}\right) = 0{,}0057, \text{ soit } \widehat{\sigma_{\Delta P}} = 0{,}076.$$

- La variable du test : $T_{248} = \dfrac{P_2 - P_1}{\sigma_{\Delta P}} = 0{,}79$.
- La p-value associée ($= 1 - \Phi^{-1}(0{,}79)$) est égale à 0,21. Elle est bien supérieure à $\alpha/2$. On ne peut rejeter H_0, les deux populations ne sont pas différentiables.

Test de l'égalité de deux variances : test de Fisher

Les estimations non-biaisées de la variance de deux échantillons X et Y sont $\widehat{\sigma^2_X}$ et $\widehat{\sigma^2_Y}$. On veut savoir si les variances sont bien égales pour les deux échantillons :

- $H_0 : \widehat{\sigma^2_X} = \widehat{\sigma^2_Y}$;
  $H_1 : \widehat{\sigma^2_X} \neq \widehat{\sigma^2_Y}$;
- Le test porte sur la statistique F définie par : $\boxed{F = \dfrac{\widehat{\sigma^2_X}}{\widehat{\sigma^2_Y}}}$ la variance la plus grande étant au numérateur.
- F suit une loi de Fisher $\mathcal{F}(n_x - 1, n_y - 1)$.
- Il s'agit d'un test bilatéral. L'hypothèse H_0 est acceptée au niveau de confiance $1 - \alpha$ si la valeur-p associée à F est supérieure à $\alpha/2$.

Les tests d'ajustement

Les tests d'ajustement ont pour objet de déterminer si un échantillon suit ou non une loi de probabilité.

- Soit $\mathcal{L}$ la loi de probabilité supposée de la population X.
- À partir d'un vecteur d'observation $(x_1, \ldots, x_n)$, il s'agit de tester l'hypothèse $H_0 : X \sim \mathcal{L}$ contre $H_1 : X \nsim \mathcal{L}$.

Nous présenterons ici deux des tests les plus couramment utilisés, tous deux basés sur une «distance» entre l'échantillon et la distribution théorique :

– distance quadratique moyenne entre les histogrammes empirique et théorique, le test du χ^2 ;

– distance maximale entre fonctions de répartition empirique et théorique, le test de Kolmogorov-Smirnov.

Le test du χ^2

Il existe plusieurs tests du χ^2 : ce sont des tests pour lesquels la variable de décision suit une distribution du χ^2 sous l'hypothèse H_0

Un exemple particulièrement important est le test du χ^2 de Pearson (ou test de Pearson).

Le test du χ^2 de Pearson

Le but du test du χ^2 de Pearson est de comparer la distribution issue d'un n-échantillon d'une v.a. X à une distribution de probabilité théorique.

- On réalise un histogramme du n-échantillon.
- Les valeurs de X sont réparties en k classes $C_1, \ldots, C_k$. La probabilité théorique (sous H_0) pour X de tomber dans chacune des classes est notée $(p_1, \ldots, p_k)$.
- À partir du n-échantillon, on obtient un effectif N_i pour la classe C_i. Bien sûr, $N_1 + \cdots + N_k = n$.
- Sous l'hypothèse H_0, le nombre d'éléments N_1 de l'échantillon tombant dans la classe C_1 est décrit par une loi binomiale de paramètre (n, p_1) (soit l'élément est dans C_1, soit il n'y est pas).
- Par conséquent, l'espérance de N_1 (sous H_0) est $\mathrm{E}[N_1] = np_1$. Pour la classe i, $(1 \leqslant i \leqslant k)$:

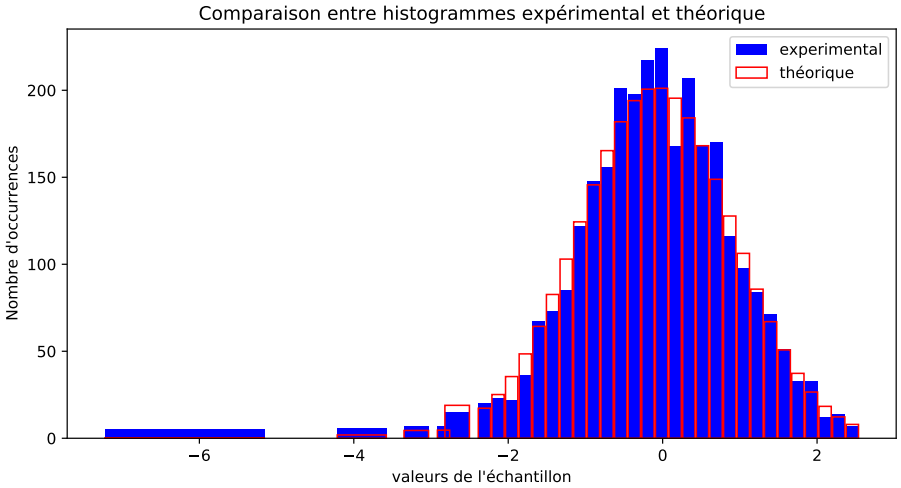
$$\mathrm{E}[N_i] = np_i$$

Le test du χ^2 de Pearson

- La statistique du test est la somme des carrés des écarts $(N_i - \mathrm{E}[N_i])$ normalisés par $\mathrm{E}[N_i]$:

$$D^2 = \sum_{j=1}^{k} \frac{(N_j - \mathrm{E}[N_j])^2}{\mathrm{E}[N_j]} = \sum_{j=1}^{k} \frac{(N_j - np_j)^2}{np_j}$$

- Si l'effectif de chaque classe est suffisamment grand, la v.a. D^2 suit une loi du χ^2 à $k-1$ degrés de liberté car $\sum N_i = n$. Pratiquement, cette condition est vérifiée si $N_i > 5$, $\forall i = 1, \ldots, k$.
- Hypothèse H_0 : "La distribution observée est conforme à la distribution théorique",
- Au seuil de signification α, on retient H_0 si $D^2 < x_\alpha^2$.
- Alternativement, on retient H_0 si la valeur-p associée à D^2 est supérieure à α.



Superposition des histogrammes

Déroulement du test du χ^2 de Pearson :

1. On comptabilise le nombre d'éléments d'un n-échantillon dans k classes : $N_i = \text{card}\{X_j; X_j \in C_i,\ i = (1, \ldots, k)\}$.
2. Compte tenu de la loi $\mathcal{L}$ supposée, on calcule les effectifs théoriques de chacune des classes C_i : $n \times p_i$.
3. On vérifie les conditions d'application du test : il faut que l'effectif de chacune des classes soit supérieur ou égal à 5. Sinon, on redéfinit les classes. Retour au 1.
4. Calcul de D^2.
5. Recherche de la valeur-p χ^2_α dans la table de la loi du χ^2 à $k-1$ degrés de liberté.
6. Le test est unilatéral. La valeur-p est ici $\Pr[\chi^2(k-1) > D^2]$.
7. Si la valeur-p est supérieure à α, on retient H_0. Sinon on rejette l'hypothèse H_0.

Exemple : On souhaite tester un dé à six faces. On jette le dé 1000 fois (on est patient). On obtient les nombres d'occurrences suivants :

1	2	3	4	5	6
188	153	159	155	150	195

Le dé est-il juste ? Soit H_0 **: le dé est juste**. Sous cette hypothèse :

- le nombre d'occurrences de chaque face suit une loi binomiale de paramètre $(1000, 1/6)$.
- l'effectif théorique de chaque classe est $n \times 1/6 \approx 166,7$.

On calcule la statistique $D^2 = \sum_{i=1}^{6} \frac{(N_i - 166,7)^2}{166,7} = 11,5$

Sous l'hypothèse H_0, la statistique D^2 suit une loi du χ^2_5. Il n'y a que 5 degrés de liberté car, si la somme définissant D^2 comprend six termes, seulement 5 d'entre eux sont indépendants, le sixième étant spécifié par les cinq premiers puisque $\sum N_i = n$).
La valeur-p, $\Pr[D^2 \geqslant 11{,}5]$, est inférieure à $0,05$ ($x^2_{0,05}(5) = 11{,}07$). On est conduit à rejeter l'hypothèse H_0 avec un risque de $\alpha = 5\%$.

- Dans le cas d'une v.a. discrète, le choix des classes est évident.
- Pour des variables continue, le choix des classes est arbitraire.
- Le fait que le résultat du test dépende du choix des classes (de la résolution de l'histogramme) est une faiblesse de ce test.
- Le test de Kolmogorov-Smirnov évite cet écueil.

Le test de Kolmogorov-Smirnov est un test d'ajustement utilisé pour déterminer si un échantillon suit ou non une loi de probabilité donnée $\mathcal{L}$.

- La fonction de répartition de la loi $\mathcal{L}$ est comparée à la fonction de répartition empirique, issue du n-échantillon $(x_1, \ldots, x_n)$.
- Ce test peut aussi être appliqué pour déterminer si deux échantillons suivent la même loi.
- La fonction de répartition empirique $\widehat{F_n}(x)$ est une estimation de la probabilité $\Pr[X \leqslant x]$. Elle est évaluée par la proportion d'individus de l'échantillon qui sont inférieurs ou égaux à x :

$$\widehat{F_n}(x_i) = \frac{1}{n}\sum_{j=1}^{n} \delta_j(x_i) \quad \text{où} \quad \delta_j(x) = \begin{cases} 1 & \text{si } x_i \leqslant x, \\ 0 & \text{sinon.} \end{cases}$$

- Sous l'hypothèse H_0, $\widehat{F_n} \to F$. L'échantillon est issu d'une population X suivant la loi $\mathcal{L}$. Alors, le sup de la différence $|\widehat{F_n}(x) - F(x)| \underset{n\to\infty}{\longrightarrow} 0$.

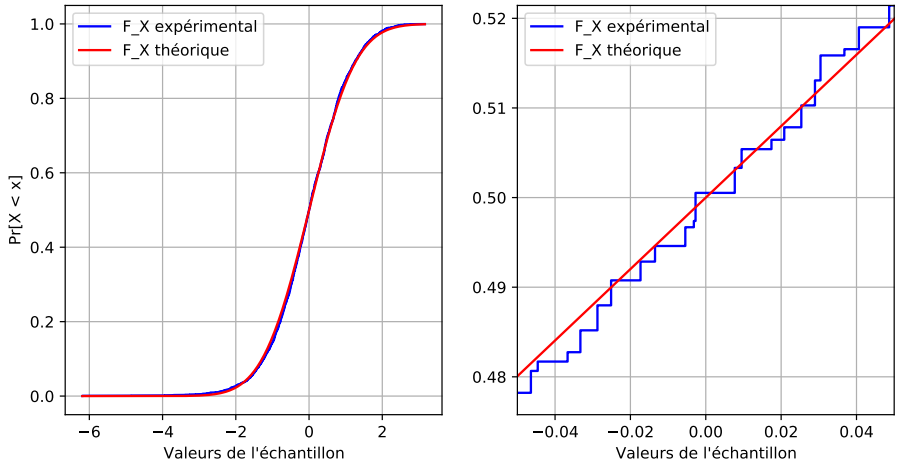
On définit une distance D_{KS} entre $\widehat{F_n}$ et F.

- Pour ce faire, on trie en ordre croissant l'échantillon : soit $x_{(i)}$ la $i$ème valeur triée.
- On en déduit la fonction de répartition expérimentale $\widehat{F_n}$ qui est une fonction en escalier discontinue aux points x_i (de $(i-1)/n$ à i/n).
- Sous H_0, la fonction de répartition théorique $F(x_{(i)})$ est telle que :

$$\frac{i-1}{n} \leqslant F(x_{(i)}) \leqslant \frac{i}{n}$$

.

Comparaison des fonctions de répartition expérimentale et théorique



Superposition des fonctions de répartition

- La distance D_{KS} est définie par :

$$D_{KS}(F, \widehat{F_n}) = \max_{i=1,\ldots,n} \left\{ \left| F(x_{(i)}) - \frac{i-1}{n} \right|, \left| F(x_{(i)}) - \frac{i}{n} \right| \right\}$$

- La statistique du test est $\sqrt{n}D_{KS}$. Le test est unilatéral : si la valeur p associée à $\sqrt{n}D_{KS}$ est inférieure à α, l'hypothèse H_0 est rejetée.
- Les valeurs-p sont tabulées. Pour n grand ($n > 35$), le seuil critique $k_\alpha \approx C(\alpha)/\sqrt{n}$.

α	0,2	0,1	0,05	0,02	0,01
C	1,07	1,22	1,36	1,52	1,63

Régression linéaire

Cette section traite de couples de variables aléatoires X et Y **dépendantes** entre elles. C'est à dire que la connaissance de X diminue l'incertitude sur Y.

On suppose qu'existe une relation de dépendance linéaire entre deux variables aléatoires X et Y (lien de cause à effet, ou cause commune). Même si elle n'est pas comprise, cette relation doit permettre de prédire Y à partir de la connaissance de X. On recherche donc une fonction f de X :

$$\widehat{Y} = f(X)$$

exprimant cette relation. $\widehat{Y}$ est donc une estimation de Y résultante de la seule connaissance de X.
On cherche une fonction f sans biais, i.e. $\mathrm{E}[Y - f(X)] = 0$, minimisant l'erreur de prévision. L'erreur est mesurée par la variance de :

$$\epsilon = Y - \widehat{Y} = Y - f(X)$$

Le modèle le plus simple, et le plus important, est celui où f est une **fonction affine** (une droite). On parle alors de **régression linéaire**.

Le problème consiste, connaissant X, à déduire la valeur la plus «vraisemblable» pour Y. Autrement dit, on cherche l'espérance de Y conditionnée à X : $\mathrm{E}[Y|X]$.

Si on se limite au modèle linéaire, on cherche Y tel que :

$$\mathrm{E}[Y|X] = \alpha X + \beta \tag{1}$$

où α et β sont deux nombres réels.

Si l'on suppose (1), alors Y est de la forme :

$$Y = \alpha X + \beta + \epsilon \tag{2}$$

où ϵ est une variable aléatoire, indépendante de X et d'espérance nulle. L'erreur résiduelle ϵ peut s'interpréter comme la part de Y qui n'est pas expliquée par la relation linéaire. L'espérance de Y est :

$$\mathrm{E}[Y] = \alpha\,\mathrm{E}[X] + \beta \tag{3}$$

La droite de régression passe donc par le point $(\mathrm{E}[X], \mathrm{E}[Y])$.
Soustrayant (3) de (2), on obtient :

$$Y - \mathrm{E}[Y] = \alpha(X - \mathrm{E}[X]) + \epsilon \tag{4}$$

Multipliant les deux membres de (4) par $X - \mathrm{E}[X]$ il vient :

$$(X - E[X])(Y - \mathrm{E}[Y]) = \alpha(X - \mathrm{E}[X])^2 + \epsilon(X - E[X])$$

Prenons l'espérance des deux membres de cette dernière équation :

$$\left.\begin{array}{lcl} \mathrm{E}[(X - E[X])(Y - \mathrm{E}[Y])] & = & \mathrm{Cov}(X, Y) \\ \mathrm{E}\left[(X - E[X])^2\right] & = & \mathrm{Var}[X] \\ \mathrm{E}[\epsilon(X - E[X])] & = & \mathrm{Cov}(\epsilon, X) \end{array}\right\} \mathrm{Cov}(X, Y) = \alpha\,\mathrm{Var}[X] + \mathrm{Cov}(\epsilon, X)$$

Comme ϵ n'est pas corrélé à X : $\alpha = \dfrac{\mathrm{Cov}(X, Y)}{\sigma_X^2} = \rho_{XY}\dfrac{\sigma_Y}{\sigma_X}$ (5)

où ρ_{XY} est le coefficient de corrélation entre X et Y Connaissant α, on déduit β de la relation (3).

L'équation théorique de la droite de régression est donc :

$$\widehat{Y} = \mathrm{E}[Y] + \rho_{XY}\frac{\sigma_Y}{\sigma_X}(X - \mathrm{E}[X]) \qquad (6)$$

Prenant les variances des deux membres de l'équation (2) et compte tenu de (5) :

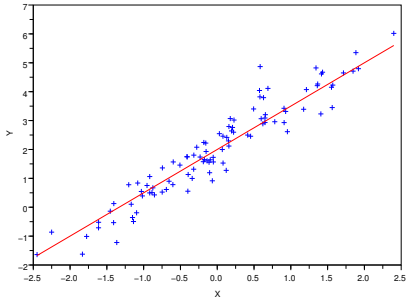
$$\mathrm{Var}[Y] = \rho_{XY}^2\frac{\sigma_Y^2}{\sigma_X^2}\mathrm{Var}[X] + \mathrm{Var}[\epsilon] = \underbrace{\rho_{XY}^2\sigma_Y^2}_{\substack{\text{variance expliquée}\\ \text{par le modèle linéaire}}} + \underbrace{\mathrm{Var}[\epsilon]}_{\substack{\text{variance}\\ \text{inexpliquée}}}$$

À priori, il existe une infinité de modèles permettant d'exprimer une relation entre les variables X et Y. Cependant, des modèles non-linéaires peuvent parfois se ramener à un modèle linéaire.

Exemple : Considérons le modèle $Y = \beta X^\alpha$. Le changement de variable $Z = \ln Y = \alpha \ln X + \ln \beta$ permet de linéariser le modèle.

Pratiquement, on ne dispose que d'observations (x_i, y_i) du couple de variables (X, Y). Chaque observation peut être représentée par un point de coordonnées (x_i, y_i) dans le plan (x, y). Un échantillon est donc représenté par un «nuage» de points (x_i, y_i), $i = 1, \ldots, n$ (figure ci-contre).

Le plus souvent, puisqu'on recherche la façon dont Y dépend de X, on considère X comme un paramètre contrôlé, c'est à dire sans incertitude.



Tracé d'un nuage de points (x_i, x_i). Une droite de régression – en rouge – est superposée.

On a alors le modèle : $\boxed{y_i = \alpha x_i + \beta + \epsilon_i\,,} \qquad \forall i \in (1, \ldots, n)$

où les erreurs ϵ_i sont des réalisations indépendantes d'une variable aléatoire ϵ, d'espérance nulle et de variance σ^2.

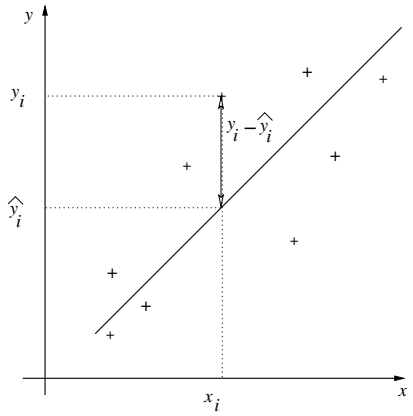
La méthode des moindres carrés est due à **Johann Carl Friedrich Gauss**.

On cherche à ajuster au nuage de points (x_i, y_i) une droite $\widehat{y} = ax + b$ telle que la variance des erreurs

$$e^2 = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 \quad (7)$$

soit minimale. La somme e^2 est la somme des carrés des écarts résiduels.

Les paramètres a et b sont des estimations de α et β.



La méthode pour déterminer a et b est la suivante.

- Explicitons l'expression de e en fonction des coefficients a et b :

$$e^2(a,b) = \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 = \sum_{i=1}^{n}(y_i - (ax_i + b_i))^2$$

- La fonction $e(a, b)$ est minimale pour les valeurs a et b qui annulent les dérivées :

$$\frac{\partial e}{\partial a} = 0 \quad \text{et} \quad \frac{\partial e}{\partial b} = 0$$

ce qui conduit aux deux équations : $\begin{cases} \sum_{i=1}^{n} x_i(y_i - ax_i - b) = 0 \\ \sum_{i=1}^{n}(y_i - ax_i - b) = 0 \end{cases}$

- équations que nous pouvons réécrire : $\begin{cases} \overline{xy} = a\overline{x^2} + b\overline{x} \\ \overline{y} = a\overline{x} + b \end{cases} \qquad (8)$

- La solution du système (8) s'obtient aisément en multipliant la seconde équation par $\overline{x}$ puis en soustrayant la seconde équation à la première. Il vient :

$$a = \frac{\overline{xy} - \overline{x}\,\overline{y}}{\overline{x^2} - \overline{x}^2} \qquad (9)$$

  Le numérateur de a n'est autre que l'estimation de la covariance de (X, Y)

- Notons $\mathrm{cov}(x, y)$ l'estimateur de la covariance $\mathrm{Cov}(X, Y)$:

$$\mathrm{cov}(x, y) = \overline{(x - \overline{x})(y - \overline{y})} = \overline{xy} - \overline{x}\,\overline{y} \quad \text{où} \quad \overline{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i$$

- Le dénominateur de l'équation (9) est la variance empirique de X, notée s_x^2 :

$$s_x^2 = \overline{(x - \overline{x})^2} = \overline{x^2} - \overline{x}^2$$

Définissons la corrélation empirique r_{xy} de l'échantillon :

$$r_{xy} = \frac{\mathrm{cov}(x, y)}{s_x s_y}$$

Le coefficient de la droite de régression a s'écrit :

$$a = \frac{\mathrm{cov}(x, y)}{s_x^2} = r_{xy}\frac{s_y}{s_x} \qquad (10)$$

Cette dernière expression est à comparer à l'expression théorique (5). On conclut que a et b sont des estimations des coefficients théoriques α et β, tels que définis par les équations (5) et (3).

Les statistiques a, b et $\widehat{y}$ sont des estimateurs sans biais de p, β et $\mathrm{E}[Y|X] = \alpha X + \beta$.

Finalement, la droite des moindres carrés s'écrit :

$$\widehat{y} = \overline{y} + r_{xy}\frac{s_y}{s_x}(x - \overline{x}) \qquad (11)$$

Nous avons défini l'écart résiduel par $e_i = y_i - \widehat{y}$. Or :

$$\sum_i e_i = \sum_i y_i - \widehat{y} = \sum_i (y_i - \overline{y}) - \alpha \sum_i (x_i - \overline{x})$$

Ces deux sommes sont identiquement nulles. On en déduit :
Théorème : Les écarts résiduels e_i sont de moyenne nulle.

$$\sum_{i=1}^{n} e_i = 0$$

Par conséquent, e^2 tel que définie en (7) est la variance empirique des e_i (sans biais, car $\mathrm{E}[e_i] = 0$) : $e^2 = \sum_{i=1}^{n} e_i^2$.

On alors le résultat : $\boxed{e^2 = (1 - r_{xy}^2)s_y^2}$

En effet, on montre aisément que : $e^2 = s_y^2 + a^2 s_x^2 - 2a\mathrm{cov(x,y)}$ soit, en remplaçant a par $r_{xy}s_y/s_x$: $e^2 = s_y^2 + r_{xy}^2 s_y^2 - 2r_{xy}^2 s_y^2$. □