Chapitre 8. Statistiques descriptives

14/05/2025

8.1 Introduction

La **statistique descriptive ou analyse des données** a pour but de résumer l'information contenue dans les données de façon synthétique et efficace. Elle utilise pour cela des représentations de données sous forme de graphiques, de tableaux et d'indicateurs numériques (par exemple des moyennes). Elle permet de dégager les caractéristiques essentielles du phénomène étudié et de suggérer des hypothèses pour une étude ultérieure plu sophistiquée. Les probabilités n'ont ici qu'un rôle minieur.

La **statistique inférentielle** va au dela de la simple description des données. Elle a pour but de faire des prévisions et de prendre des décisions au vu des observations. En général, il faut pour cela proposer des modèles probabilistes du phénomène aléatoire étudié et savoir gérer les risques d'erreurs. Les probaibiltés jouent un rôle fondamental.

Définitions

La statistique utilise un vocabulaire spécifique, tiré pour l'essentiel de la première activité : la démographie.

- population : c'est l'ensemble étudié ;
- individus : ce sont les éléments de l'ensemble ;
- échantillon : c'est un sous-ensemble de la population ;
- effectif total : le nombre d'éléments d'une popolation ;
- caractère : la caractéristique que l'on étudie ;
- variable discrète : elle ne peut prendre que des valeurs ponctuelle ;
- variable continue : elle peut prendre toutes les valeurs numérique d'un intervalle donné.

Une population peut être une population humaine, mais aussi un ensemble d'objets.

Vocabulaire

- population : en extension ou en compréhension ;
- échantillon : l'effectif est le nombre d'éléments de l'échantillon ;
- caractère : qualitatif (couleur, profession,..) ou quantitatif (numérique) ;
- modalité : une modalité est une des situations possibles pour un caratère qualitatif ou un caratère quantitatif discret.
- mode : le mode est la valeur du caratère correspondant au plus grand effectif (ou fréquence relative).
- collecte de données : la population étant ciblée, on effectue des relevés des valeurs du caratère étudié : les N données peuvent se présenter sous trois formes : individuelle ; individuelles groupées (on renseigne que la valeur x_i a été observée n_i fois) ; groupées par classes (dans le cas où les données sont trop nombreuses).

8.2. Présentation des données

Considérons une population staistique de n individus décrite suivant un caractère C dont les k modalités sont $C_1, C_2, \ldots, C_k$. Désignons par n_i le nombre d'individus présentant la modalité C_i : n_i est l'effectif de la modalité C_i et f_i est sa proportion :

$$f_i = \frac{n_i}{n}$$

On a donc $\sum_{i=1}^{k} n_i = n$ et $\sum_{i=1}^{k} f_i = 1$.
Un tableau statistique revient à associer modalités et effectifs, voire modalités et fréquences.
Généralement, un tableau statistique est traduit en graphe pour réaliser une synthèse visuelle, donc à la fois rapide et simple. Les méthodes utilisées, pour présenter les caractères de classes dans une population ou échantillon, dépendent de la nature du caractère en question (qualitatif ou quantitatif).

Tableaux statistiques

Considérons les notes (de 0 à 20) obtenues par 9 élèves dans 4 disciplines (mathématiques, physique, français, anglais) :

	MATH	PHYS	FRAN	ANGL
jean	6.00	6.00	5.00	5.50
alan	8.00	8.00	8.00	8.00
anni	6.00	7.00	11.00	9.50
moni	14.50	14.50	15.50	15.00
didi	14.00	14.00	12.00	12.50
andr	11.00	10.00	5.50	7.00
pier	5.50	7.00	14.00	11.50
brig	13.00	12.50	8.50	9.50
evel	9.00	9.50	12.50	12.00

Elèves	
MP	42
PC	16
PSI	38
PT	14
TSI	2

Classification suivant les filières des étudiants de première année d'une école d'ingé.

Caratère qualitatif : Le principe de la représentation des C.Q. est la proportionnalité des aires aux effectifs. On utilisera couramment des tuyaux d'orgue ou des secteurs. Les secteurs circulaires (camenberts) ont un angle au centre proportionnel à l'effectif correspondant. Les tuyaux d'orgue ont une base constante et une hauteur proportionnelle à l'effectif. Dans les deux cas, l'aire est proportionnelle à l'effectif. Si chaque modalité est représentée comme un disque, il convient de respecter cette règle de proportionnalité de la surface.
Caratère quantitatifs : Si le caractère est quantitatif, on utilise deux sortes de représentations :

- **diagramme différentiel** : diagramme en bâtons qui représentent en fonction des valeurs de x_i et les fréquences f_i correspondentes ; histogramme ;
- **digramme intégral** : courbe cummulatif

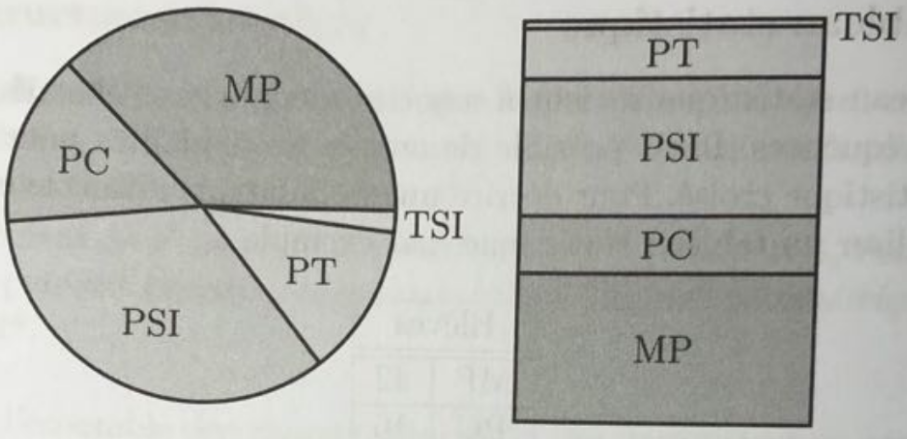


FIGURE 6.1 – Étudiants en première année d'une école d'ingénieurs

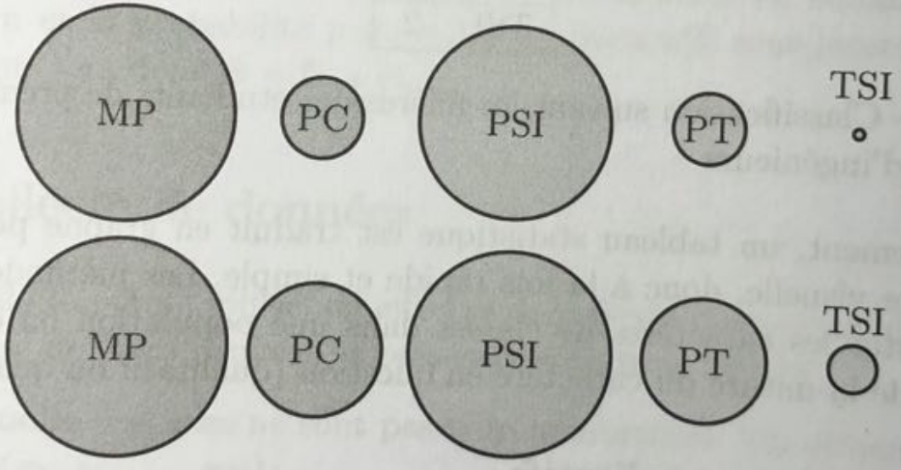


FIGURE 6.2 – Rayon (haut) ou surface (bas) proportionnel à l'effectif

Diagramme différentiel

Les variables discrètes ou les données qui sont groupées par classe, sont souvent présentées en diagramme en bâtons ou en histogramme.
Pour faire l'histogramme, on utilise les notions suivantes :

- limites de classe ;
- bornes de classes ;
- centres de classes ;
- amplitudes de classes ;
- effectifs de classes ;
- fréquences de classes.

L'objectif reste de faire une représentation visuelle fidèle : l'aire d'un objet géométrique est proportionnel à l'effectif de la classe d'il représente.
SI les classes n'ont pas toutes la même amplitude, il convient de ne pas faire des barres d'histogramme de hauteur proportionnelle à l'effectif, mais de hauteur proportionnelle à l'effectif divisé par l'amplitude.

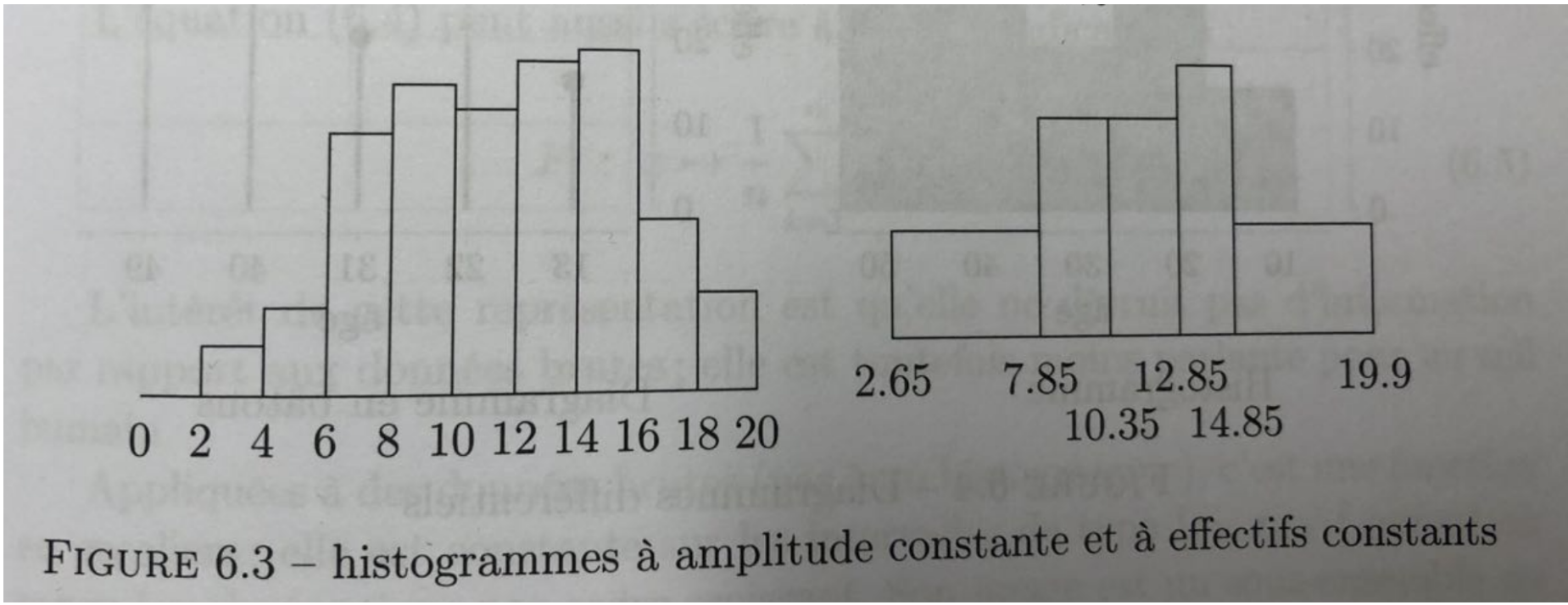


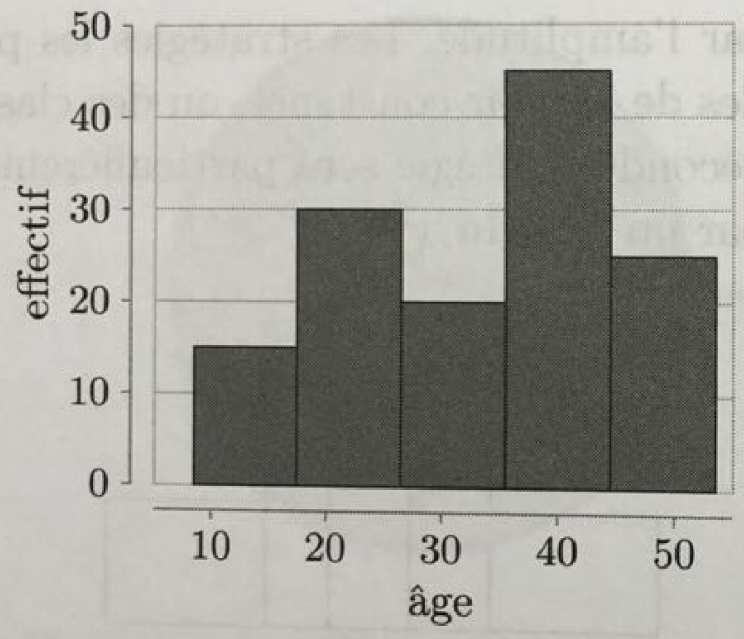
FIGURE 6.3 – histogrammes à amplitude constante et à effectifs constants

Exemple

Le tableau 6.2 décrit un échantillon de 135 individus selon le caractère « âge ».

Classes	Limites	Bornes	Centre	Effec.	Effec. Cum	Fréq. en %	Fréq. Cum. en %
9-17	9 et 17	8.5 et 17.5	$\frac{9+17}{2}$	15	15	$\frac{15\times100}{135}$	11
18-26	18 et 26	17.5 et 26.5	22	30	45	22	33
27-35	27 et 35	26.5 et 35.5	31	20	65	15	48
36-44	36 et 44	35.5 et 44.5	40	45	110	34	82
45-53	45 et 53	44.5 et 53.5	49	25	135	18	100

TABLE 6.2 – Un tableau statistique



Histogramme

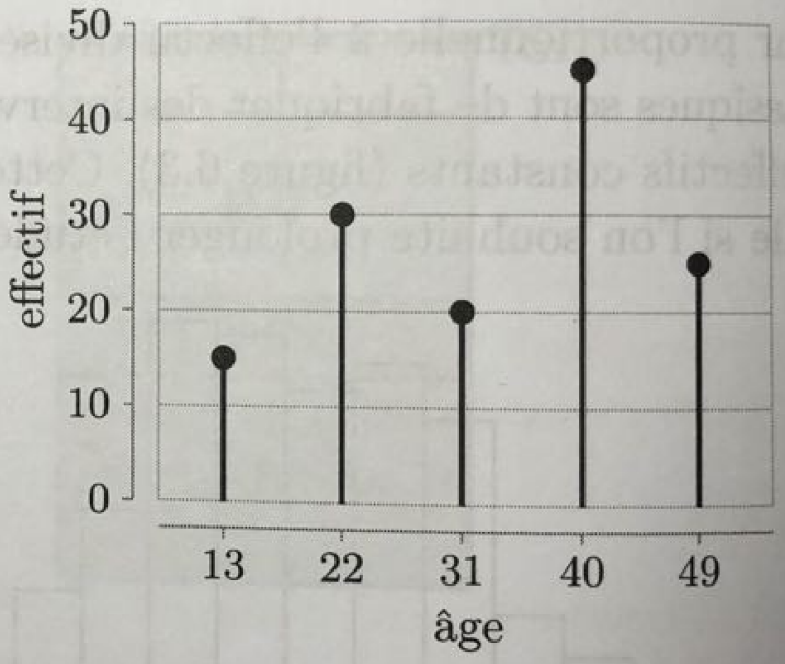


Diagramme en bâtons

FIGURE 6.4 – Diagrammes différentiels

Courbe cumulative

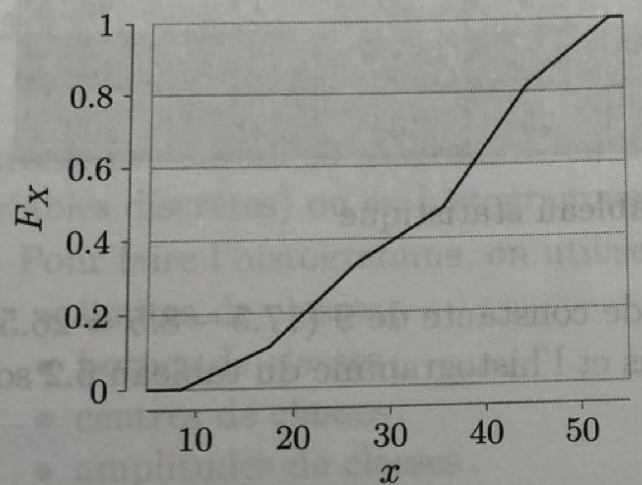
Soit $F(x)$ la proportion des individus de la population dont le caractère est inférieur à x. Cette fonction appelée *fonction cumulative* ou *fonction de répartition* est définie pour toute valeur x réelle. Elle est constante sur chaque intervalle séparant deux valeurs possibles consécutives. Si les valeurs x_i sont ordonnées de façon croissante, et que $x_i \leq x < x_{i+1}$, on a

$$F(x) = \sum_{j=1}^{i} f_j$$

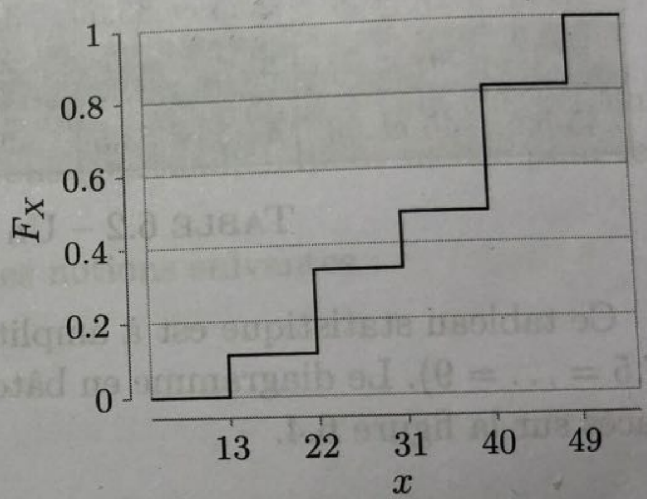
Par convention, on pose $F(-\infty) = 0$ et $F(+\infty) = 1$.
Si $\mathcal{E} = \{x_1, \ldots, x_n\}$, on a

$$F : x \to \frac{1}{n} \sum_{k=1}^{n} \mathbb{1}_{]-\infty, x]}(x_k)$$

La courbe cumulative ou courbe des fréquences cumulées est la courbe représentative de $F(x)$ en fonction de x, voir figure 6.5.



Classes homogènes



Classes réduites à leurs centres

FIGURE 6.5 – Courbes cumulatives

Résumé des données

Qaund les données sont regroupées par classes, les classes sont résumées en leur centre x_l de façon à se ramener en données pondérées :

$$(x_1, n_1),\ (x_2, n_2),\ \ldots, (x_k, n_k),$$

les n_i sont les effectifs de x_i, avec $n = n_1 + \cdots + n_k$, et $f_i = n_i/n$ sont leurs fréquences.
Pour résumer les données statistiques, un certain nombre de paramètres sont définis.

Définition

La moyenne, ou moyenne arithmétique,d'une variable statistique est la somme pondérée des valeurs possibles par les fréquences :

$$\bar{x} = \sum_{i=1}^{k} f_i x_i = \frac{1}{n} \sum_{i=1}^{k} n_i x_i$$

Médiane et autres quantiles

Définition

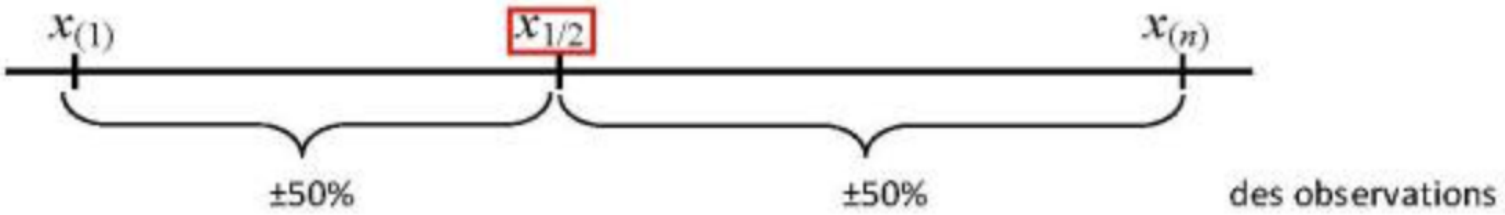
Une médiane d'une variable statistique est telle que la part de la population inférieure ou égale à la médiane soit d'au moins la moitié de la population, ainsi que la part de la population supérieure ou égale à la médiane.
Le quantile d'ordre α, avec $0 \leq \alpha \leq 1$, noté x_α, est la solution de l'équation :

$$F(x_\alpha) = \alpha.$$

Par exemple, si on considère une population de 5 personnes décritèes suivant la taille, et rangés par ordre de taille, la taille médiane est celle de la troisième personne.
En général, la médiane M est la valeur de la variable statistique telle que l'ordonnée de la courbe cummulative soit égal à $\frac{1}{2}$: $F(M) = \frac{1}{2}$.

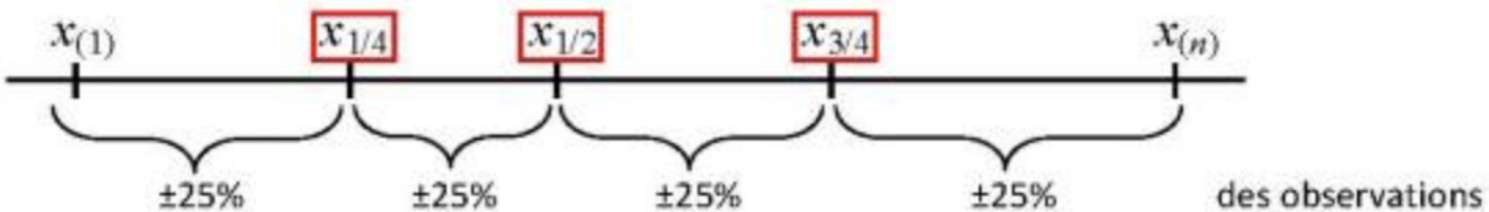
Quantiles les plus utilisés

1) la médiane : $\alpha = 1/2$



La médiane

La médiane partage la série statistique ordonnée en deux sous-ensembles qui contiennent chacun (environ) la moitié des observations.
2) le quartiles $\alpha = 1/4$ (1^{er} quartile), $\alpha = 1/2$ (2^{e} quartile=médiane), $\alpha = 3/4$ (3^{e} quartile)



Les quartiles

Quantiles les plus utilisés

3) les déciles : $\alpha = i/10,\ i = 1, 2, \ldots, 9$

Les déciles

Les 9 déciles partagent la série statistique ordonnée en 10 sous-ensembles qui contiennent chacun (environ) un dixième (10%) des observations.

4) les percentiles : $\alpha = i/100,\ i = 1, 2, \ldots, 99$
Les 99 percentiles partagent la série statistique ordonnée en 100 sous-ensembles qui contiennent chacun (environ) un centième (1%) des observations.

Variances

Définition

La variance S_n^2 est définie par

$$S_n^2 = \frac{1}{n}\sum_{i=1}^{k} n_i(x_i - \bar{x})^2 = \sum_{i=1}^{k} f_i(x_i - \bar{x})^2$$

La variance dite sans biais est définie par

$$S_n'^2 = \frac{1}{n-1}\sum_{i=1}^{k} n_i(x_i - \bar{x})^2$$

d'où la relation suivante : $nS_n^2 = (n-1)S_n'^2$.

On pourrait penser que S_n^2 est un bon estimateur de $Var(X)$. Cependant des calculs prouvent que cet estimateur est biaisé, l'espérance de S_n^2 est toujours inférieure à $Var(X)$.

Estimation d'une variance

On définit :

$$V_n = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 = \frac{1}{n}\sum X_i^2 - \mu^2$$

Si μ est connue, alors V_n est un estimateur **sans biais** de $\mathbb{V}[X]$

Preuve :

$$\begin{aligned}\mathbb{E}[\mathbb{V}] &= \frac{1}{n}\mathbb{E}\left[\sum_{i=1}^{n} X_i^2\right] - \mathbb{E}[\mu^2] \\ &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i^2] - \mu^2 \\ &= \mathbb{E}[X_i^2] - \mu^2 \equiv \mathbb{V}[X]\end{aligned}$$

Estimation d'une variance

On définit :

$$s_n^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \overline{x}^2$$

Si μ est inconnue, alors S_n^2 est un estimateur **biaisé** de $\mathbb{V}(X)$

Preuve :

$$\begin{aligned}
\mathbb{E}[S_n^2] &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i^2] - \mathbb{E}[\bar{X}^2] \\
&= \frac{1}{n}\sum_{i=1}^{n}(\mathbb{V}[X_i] + \mathbb{E}[X_i]^2) - \mathbb{V}[\bar{X}] - \mathbb{E}[\bar{X}]^2 \\
&= \frac{1}{n}(n\mathbb{V}[X] + n\mathbb{E}[X_i]^2) - \frac{1}{n}\mathbb{V}[X] - \mathbb{E}[X_i]^2 \\
&= \mathbb{V}[X] - \frac{1}{n}\mathbb{V}[X] = \frac{n-1}{n}\mathbb{V}[X]
\end{aligned}$$

Estimation d'une variance

On définit :

$$S_n'^2 = \frac{n}{n-1} S_n^2$$

Si μ est inconnue, alors $S_n'^2$ est un estimateur **sans biais** de $\mathbb{V}(X)$

Preuve :

$$\begin{aligned}
\mathbb{E}[S_n'^2] &= \frac{n}{n-1}\mathbb{E}[S_n^2] \\
&= \frac{n}{n-1}\frac{n-1}{n}\mathbb{V}[X] \\
&= \mathbb{V}[X]
\end{aligned}$$

Moments

Définition

Pour $r \in \mathbb{N}^{*r}$, le moment d'ordre r, m_r de X, est donné par :

$$m_r := m_r(X) = \sum_{i=1}^{k} f_i x_i^r$$

Le moment centré d'ordre r est défini par :

$$\mu_r := \mu_r(X) = \sum_{i=1}^{k} f_i (x - \bar{x})^r$$

Le moment d'ordre r est un estimateur sans biais du moment d'ordre r de la v.a. sous-jacente : $\mathrm{E}(m_r) = \mathrm{E}(X^r)$.

Covariances

Définition

La covariance de deux variables statistiques X et Y est définie par

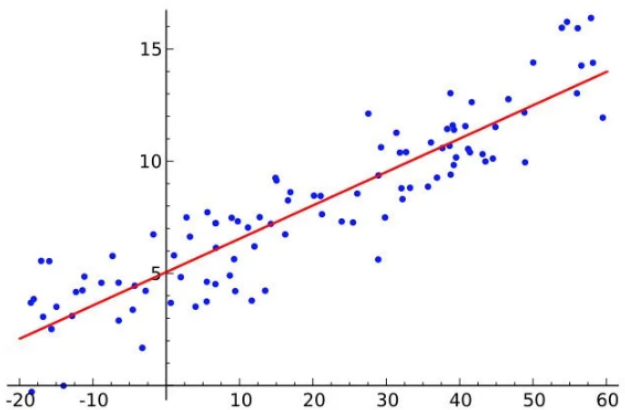
$$\begin{aligned} Cov(X,Y) &= \frac{1}{n}\sum_i\sum_j n_{ij}(x_i-\bar{x})(y_j-\bar{y}) \\ &= \frac{1}{n}\sum_i\sum_j n_{ij}(x_i-\bar{x})y_j \\ &= \frac{1}{n}\sum_i\sum_j n_{ij}x_iy_j - \bar{x}\bar{y} \end{aligned}$$

où $n = \sum_i \sum_j n_{ij}$.
Le coefficient de corrélation empirique est défini par

$$\rho_{X,Y} = \frac{Cov(X,Y)}{S_{X,n}S_{Y,n}}$$

Régression linéaire : une définition

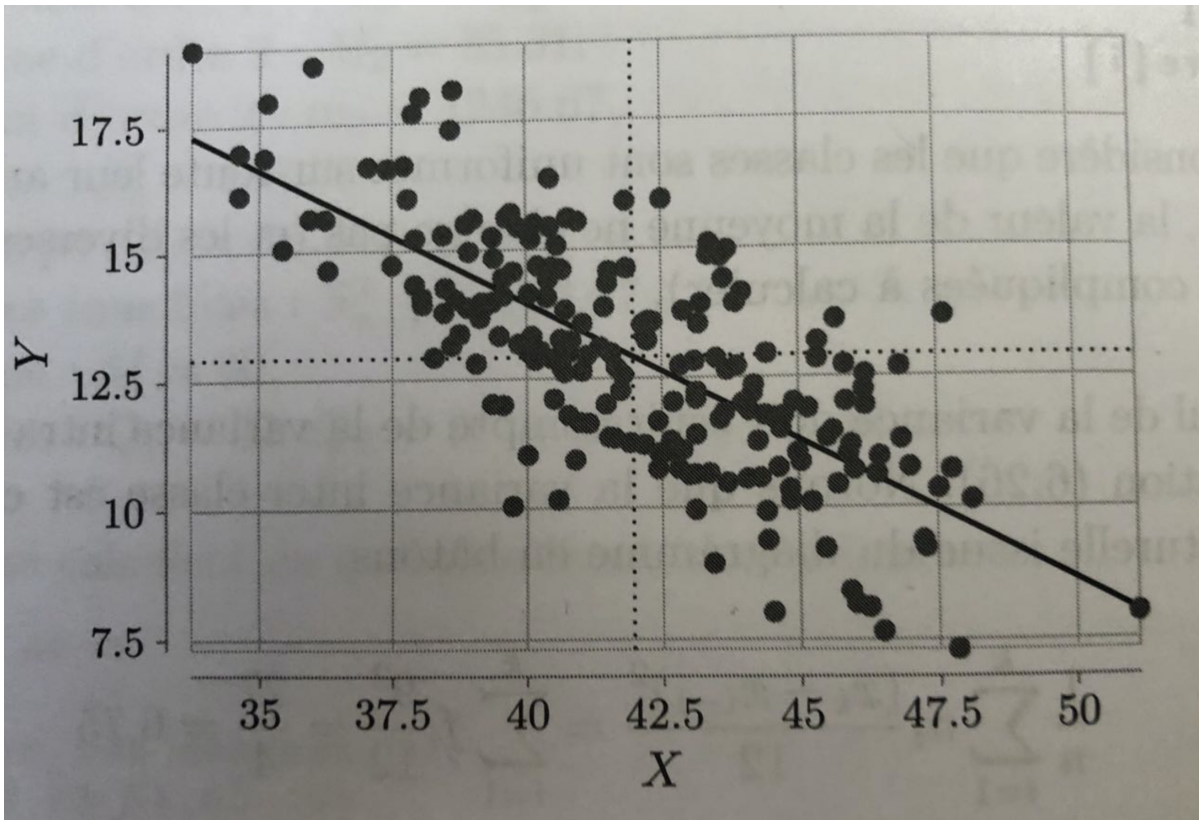
La régression linéaire est une technique statistique de **modélisation des relations entre différentes variables** (dépendantes et indépendantes). Utilisée pour décrire et analyser les valeurs ou données, la régression linéaire a pour objectif de réaliser des prédictions ou des prévisions.



Exemple de régression linéaire

Régression linéaire

Soient X et Y deux caractères mésurés sur une mêm population. Lorsque le nuage de points (x_i, y_i) $(i = 1, \ldots, n)$ se forme autour d'droite, il est raisonnable d'approcher la relation qui les lie par une fonction linéaire (affine).



Théorème

La droite de régression linéaire de Y par rapport à X est la droite d'équation $y = a(x - \bar{x}) + b$, où

$$a = \frac{Cov(X, Y)}{\mu_2(X)}, \quad b = \bar{y}.$$

Dém. On cherche a et b pour minimiser $\frac{1}{n}\sum_{i=1}^{n}[y_i - a(x_i - \bar{x}) - b]^2$, ce qui revient à minimiser

$$\begin{aligned} Q(a, b) &= \sum_{i=1}^{n}(y_i - a(x_i - \bar{x}) - b)^2 \\ &= \sum_i y_i^2 + a^2 \sum_i (x_i - \bar{x})^2 + nb^2 - 2bn\bar{y} - 2a\sum_i y_i(x_i - \bar{x}) \end{aligned}$$

Le minimum de Q est atteint lorsque sa dérivée s'annule :

$$\frac{\partial Q(a,b)}{\partial a} = 0 \quad \frac{\partial Q(a,b)}{\partial b} = 0$$

c'est-à-dire

$$0 = 2an\mu_2(X) - 2\sum_i y_i(x_i - \bar{x})$$

$$0 = 2nb - 2n\bar{y}$$

On a donc $b = \bar{y}$ et

$$a = \frac{\frac{1}{n}\sum_{i=1}^n y_i(x_i - \bar{x})}{\mu_2(X)} = \frac{\frac{1}{n}\sum_{i=1}^n y_i x_i - \bar{x}\bar{y}}{\mu_2(X)}$$

$$= \frac{Cov(X,Y)}{\mu_2(X)}$$

Intervalles de confiance

Soit X un caractère (ou variable) étudié sur une population, de moyenne m et de variance σ^2. On chercher ici à donner une estimation de la moyenne m de ce caractère, calculée à partir de valeurs observées sur un échantillon $(X_1, \ldots, X_n)$.
La fonction de l'échantillon qui estimera un paramètre est appelée **estimateur**. Son écart-type est appelé **erreur standard** et est noté SE.
L'estimateur de la moyenne m est la moyenne empirique

$$\bar{X}_n = \frac{X_1 + \cdots + X_n}{n}.$$

Que savons nous sur $\bar{X}_n$?

1. $E(\bar{X}_n] = m$,
2. $\mathrm{VAR}(\bar{X}_n) = \sigma^2/n$, donc $\mathrm{SE}(\bar{X}_n) = \sigma/\sqrt{n}$,
3. $\bar{X}_n \to m$ quand $n \to \infty$,
4. quand n est grand, $\bar{X}_n$ suit approximativement une loi normale $\mathcal{N}(m, \sigma^2/n)$.

D'après les propriétés de la loi normale, quand n est grand (mettons supérieur à 20), on sait que

$$\mathbb{P}[m - 2\sigma/\sqrt{n} \leq \overline{X}_n \leq m + 2\sigma/\sqrt{n}] = 0.954$$

ou, de manière équivalente,

$$P[\overline{X}_n - 2\sigma/\sqrt{n} \leq m \leq \overline{X}_n + 2\sigma/\sqrt{n}] = 0.954$$

Ce qui peut se traduire ainsi : quand on estime m par $\overline{X}_n$, l'erreur faite est inférieure à $2\sigma/\sqrt{n}$, pour 95,4% des échantillons. Ou avec une probabilité de 95,4%, la moyenne inconnue m est dans l'intervalle $[\overline{X}_n - 2\sigma/\sqrt{n}, \overline{X}_n + 2\sigma/\sqrt{n}]$.

Définition 56 *On peut associer à chaque incertitude α, un intervalle, appelé intervalle de confiance de niveau de confiance $1 - \alpha$, qui contient la vraie moyenne m avec une probabilité égale à $1 - \alpha$.*

Définition 57 *Soit Z une v.a.. Le fractile supérieur d'ordre α de la loi de Z est le réel z qui vérifie*

$$P[Z \geq z] = \alpha$$

Le fractile inférieur d'ordre α de la loi de Z est le réel z qui vérifie

$$P[Z \leq z] = \alpha$$

Proposition 58 *Un intervalle de confiance pour la moyenne, de niveau de confiance $1-\alpha$, est*

$$[\overline{X}_n - z_{\alpha/2}\sigma/\sqrt{n},\ \overline{X}_n + z_{\alpha/2}\sigma/\sqrt{n}]$$

où $z_{\alpha/2}$ est le **fractile** *supérieur d'ordre $\alpha/2$ de la loi normale $\mathcal{N}(0,1)$.*

preuve :

$$\begin{aligned} P[\overline{X}_n - z_{\alpha/2}\sigma/\sqrt{n} \leq m \leq \overline{X}_n + z_{\alpha/2}\sigma/\sqrt{n}] &= P\left[-z_{\alpha/2} \leq \frac{\overline{X}_n - m}{\sigma/\sqrt{n}} \leq z_{\alpha/2}\right] \\ &= P[-z_{\alpha/2} \leq Z \leq z_{\alpha/2}] \end{aligned}$$

où Z suit une loi normale centrée réduite. Par définition de $z_{\alpha/2}$, cette probabilité vaut bien $1-\alpha$. □

Remarque : soit Z une v.a. de loi $\mathcal{N}(0,1)$, $z_{\alpha/2}$ vérifie

$$P[Z \leq -z_{\alpha/2}] = P[Z \geq z_{\alpha/2}] = \frac{\alpha}{2}, \qquad P[-z_{\alpha/2} \leq Z \leq z_{\alpha/2}] = 1 - \alpha$$

On appelle aussi intervalle de confiance la réalisation de l'intervalle précédent

$$[\bar{x}_n - z_{\alpha/2}\sigma/\sqrt{n},\ \bar{x}_n + z_{\alpha/2}\sigma/\sqrt{n}]$$

Seules quelques valeurs de α sont utilisées habituellement. Les trois valeurs communes sont :

- $\alpha = 0.01$, et $z_{0.005} = 2.58$,
- $\alpha = 0.05$, et $z_{0.025} = 1.96$,
- $\alpha = 0.1$, et $z_{0.05} = 1.645$.

Voici deux exemples, l'un théorique et l'autre pratique.

Exemple 60 *Voici 30 mesures (en décibels) du bruit occasionné par le trafic routier le long de la nationale 7 :*

57	43	55	59	52	57	50	52	60	49	56	56	52	58	55
58	54	52	56	53	59	50	55	51	54	53	53	56	55	58

Regroupons les différentes modalités

x_i	n_i	f_i	F_i
43	1	0.033	0.033
49	1	0.033	0.066
50	2	0.066	0.133
51	1	0.033	0.167
52	4	0.133	0.3
53	3	0.1	0.4
54	2	0.066	0.466
55	4	0.133	0.6
56	4	0.133	0.733
57	2	0.066	0.8
58	3	0.1	0.9
59	2	0.066	0.966
60	1	0.033	1

Tout d'abord, observons une valeur extrême (43) trop petite : les données n'ont pas l'air

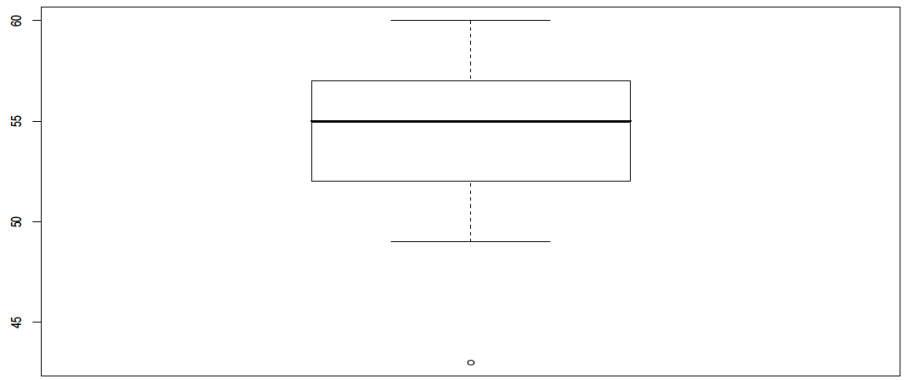


FIGURE 4.2 – Exemple 60

homogène (mesure le dimanche, ou appareil déréglé). Je préfère l'enlever avant l'étude, car elle risque de fausser la moyenne empirique et la variance. Il reste donc 29 observations. On cherche un intervalle de confiance pour le bruit moyen occasionné par le trafic routier. Mais il nous manque l'écart-type de ce bruit. Le calcul est donc impossible.

Quand l'écart-type théorique de la loi du caractère X étudié n'est pas connu, on l'estime par l'écart-type empirique s_{n-1}. Comme on dispose d'un grand échantillon (de taille supérieure à 20), l'erreur commise est petite. L'intervalle de confiance, de niveau de confiance $1-\alpha$ devient :

$$\left[\overline{x}_n - z_{\alpha/2}\frac{s_{n-1}}{\sqrt{n}},\ \overline{x}_n + z_{\alpha/2}\frac{s_{n-1}}{\sqrt{n}}\right]$$

où

$$s_{n-1}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

Exemple 60 (suite) On peut maintenant donner un intervalle de confiance, de niveau de confiance 95%, pour le volume sonore moyen le long de la RN 7. Calculons :

$$\overline{x} = 54.6552, \qquad s_{n-1} = 2.9553$$

d'où l'intervalle

$$\left[54.6552 - \frac{1.96 * 2.9553}{\sqrt{29}}, 54.6552 + \frac{1.96 * 2.9553}{\sqrt{29}}\right] = \left[53.5796, 55.7308\right]$$

Le bruit moyen occasionné par le trafic routier de la nationale 7 est compris entre 53,58 et 55,73, avec un niveau de confiance de 95%.

Estimation par intervalle de confiance. La "fourchette"

Considérons un vote avec un assez grand nombre d'électeurs. Quand le scrutin est clot, on commence à dépouiller les bulletins. Assez vite on est en mesure de donner une estimation du résultat final. En pratique, on ne donne pas une estimation numérique (telle liste obtient 18% des votes) mais une fourchette, c'est-à-dire un petit intervalle dans lequel on estime que le pourcentage exact figure.

- La taille de la fourchette dépend de la confiance qu'on souhaite avoir dans l'estimation.
- On peut vouloir que la probabilité que le pourcentage exact d'une liste soit bien dans la fourchette dépasse 0.95 (le niveau de confiance).
- Plus on exige un haut niveau de confiance, plus la fourchette sera large.

Estimation par intervalle de confiance

Notion d'intervalle de confiance Il est souvent plus réaliste et plus intéressant de fournir un renseignement de type $a < \theta < b$ plutôt que de calculer $\hat{\theta}$.
On cherche à determiner l'intervalle $[a; b]$, centré sur la valeur numérique estimée du paramètre inconnu θ, contenant la valeur vraie avec une probabilité $1 - \alpha$ $(0 < \alpha < 1)$:

$$\mathbb{P}(a < \theta < b) = 1 - \alpha$$

- L'intervalle $[a; b]$ est appelé intervalle de confiance, α le risque et $1 - \alpha$ le niveau de confiance.
- Données de départ : l'échantillon et la connaissance de la loi de probabilité du paramètre à estimer.

Estimation par intervalle de confiance d'une proportion

Soit une population dont les individus possèdent un caractère A avec une probabilité p. On dispose d'un échantillon de taille n, dont x individus possèdent le caractère A.

- On sait maintenant que la proportion $f_n = x/n$ est une estimation de la valeur vraie p ...
- Mais avec quelle confiance ?
- On cherche donc à construire un intervalle de confiance de l'estimateur.

Estimation par intervalle de confiance d'une proportion

Soit une population dont les individus possèdent un caractère A avec une probabilité p. La propotion $f_n = x/n$ est une estimation de la valeur vraie p.

Principe

- Soit $F_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. F_n est une v.a. construite comme somme de n v.a. indépendantes de type Bernoulli et de paramètre p, i.e., $X_i \sim \mathcal{B}(p)$.
- La loi de $T_n = nF_n$ suit une loi binomiale $\mathcal{B}(n, p)$.
- La loi de T_n tend vers une loi normale de moyenne np et de variance $np(1-p)$ (si $np > 10$ et $n(1-p) > 10$)
- La variable renormalisée approche une loi normale contrée réduite :

$$\frac{T_n - np}{\sqrt{np(1-p)}} \sim_\infty \mathcal{N}(0, 1)$$

Estimation par intervalle de confiance d'une proportion

On écrit alors

$$\mathbb{P}\left(\left|\frac{\mathbf{T_n - np}}{\sqrt{\mathbf{np(1-p)}}}\right| \leq \mathbf{u}_\alpha\right) \approx \mathbf{1} - \alpha\ ,$$

où u_α est une valeur (se lisant dans la table de la loi normale $\mathcal{N}(0,1)$) qui vérifie :

$$\mathbb{P}(|U| > u_\alpha) = \alpha, \quad U \sim \mathcal{N}(0,1)\ .$$

Pour en déduire un intervalle de confiance, il suffit d'écrire $\left|\frac{T_n - np}{\sqrt{np(1-p)}}\right| \leq u_\alpha$ sous la forme $Z_1 \leq p \leq Z_2$:

$$\left|\frac{T_n - np}{\sqrt{np(1-p)}}\right| \leq u_\alpha \iff \frac{(T_n - np)^2}{np(1-p)} \leq u_\alpha^2$$

$$\iff p^2(n + u_\alpha^2) - p(2T_n + u_\alpha^2) + \frac{T_n^2}{n} \leq 0$$

Estimation par intervalle de confiance d'une proportion

Intervalle de confiance asymptotique

Le trinôme $p^2(n+u_\alpha^2) - p(2T_n + u_\alpha^2) + \frac{T_n^2}{n}$ est toujours positif sauf entre ses racines. Donc ses racines sont les bornes de l'intervalle de confiance recherché :

$$\left[\frac{\frac{T_n}{n} + \frac{u_\alpha^2}{2n} - u_\alpha \sqrt{\frac{u_\alpha^2}{4n^2} + \frac{T_n(n-T_n)}{n^3}}}{1+\frac{u_\alpha^2}{n}}, \frac{\frac{T_n}{n} + \frac{u_\alpha^2}{2n} + u_\alpha \sqrt{\frac{u_\alpha^2}{4n^2} + \frac{T_n(n-T_n)}{n^3}}}{1+\frac{u_\alpha^2}{n}} \right]$$

Pour les valeurs usuelles de α et pour n grand, on peut négliger u_α^2 par rapport à n. D'où, avec $F_n = \frac{T_n}{n}$ et une réalisation f_n de F_n, on obtient **l'intervalle de confiance asymptotique** suivant :

$$\left[f_n - u_\alpha \sqrt{\frac{f_n(1-f_n)}{n}}, f_n + u_\alpha \sqrt{\frac{f_n(1-f_n)}{n}} \right]$$

Estimation par intervalle de confiance d'une proportion

Fourchette du sondage

Une élection oppose deux candidats A et B. Un institut de sondage interroge **800 personnes** sur leurs intentions de vote :

- 420 déclarent voter pour A
- 380 déclarent voter pour B

Estimer le résultat de l'élection, c'est estimer le pourcentage p de voix qu'obtiendra A le jour de l'élection, en inférant sur l'ensemble de la population. L'estimateur de p est la proportion $f_n = \frac{420}{800} = 52.5\%$. L'institut de sondage estime donc que le candidat A va gagner l'élection. Mais pour évaluer l'incertitude, on a besoin d'un intervalle de confiance de seuil disons 5% pour p. On obtient alors l'intervalle de confiance asymptotique suivant

$$[0.4904, 0.5596]$$

Conclusion : on a une confiance de 95% dans le fait que le pourcentage de voix qu'obtiendra le candidat A sera compris entre 49% et 56%.

Estimation par intervalle de confiance d'une proportion

Obtenir une prédiction plus précise

À quelle condition l'intervalle de confiance pour p sera entièrement situé au dessus de 50% ?

$\Rightarrow$ Il s'agit donc de réduire l'intervalle de confiance, de largeur :

$$\ell = 2u_\alpha \sqrt{\frac{f_n(1-f_n)}{n}}$$

Pour diminuer cette largeur ℓ, on peut :

- Diminuer u_α, c'est-à-dire augmenter α, donc augmenter la probabilité de se tromper en affirmant que le candidat est élu ;
- Augmenter n, c'est-à-dire augmenter le nombre de personnes interrogées.

Estimation par intervalle de confiance d'une proportion

Taille de l'échantillon minimum

À un seuil de confiance $\alpha = 5\%$ fixé, combien de personnes n doit-on interroger pour que l'intervalle de confiance n'excède pas une largeur ℓ ?

On sait que $\forall p \in [0,1], p(1-p) \leq \frac{1}{4}$, donc

$$2u_\alpha \sqrt{\frac{f_n(1-f_n)}{n}} \leq \frac{u_\alpha}{\sqrt{n}}$$

Ainsi il suffit de déterminer n tel que

$$\frac{\mathbf{u_\alpha}}{\sqrt{\mathbf{n}}} < \ell \iff \mathbf{n} > \frac{\mathbf{u_\alpha^2}}{\ell^2}$$

Pour $n = 800$, on a $\frac{u_\alpha}{\sqrt{n}} = \frac{1.96}{\sqrt{800}} \approx 7.5\%$ $\Rightarrow$ la précision sur l'estimation de p est donc avec une confiance de 95% de plus ou moins 3.5%, ce qu'on a constaté avec l'intervalle $[49\%, 56\%]$.

Estimation par intervalle de confiance d'une proportion

Taille de l'échantillon minimum

À un seuil de confiance $\alpha = 5\%$ fixé, combien de personnes n doit-on interroger pour que l'intervalle de confiance n'excède pas une largeur ℓ ?

On sait que $\forall p \in [0,1], p(1-p) \leq \frac{1}{4}$, donc

$$2u_\alpha \sqrt{\frac{f_n(1-f_n)}{n}} \leq \frac{u_\alpha}{\sqrt{n}}$$

Ainsi il suffit de déterminer n tel que

$$\frac{\mathbf{u_\alpha}}{\sqrt{\mathbf{n}}} < \ell \iff \mathbf{n} > \frac{\mathbf{u_\alpha^2}}{\ell^2}$$

Si on veut, avec le même niveau de confiance, avoir une précision $<$ à 1%, il faudra interroger au moins :

$$n = \frac{u_\alpha^2}{\ell^2} = \frac{1.96^2}{0.01^2} = 38\,416 \text{ personnes}$$

Introduction aux tests d'hypothèse : Jeu de Pile ou Face et triche

Karl et Ronald jouent à Pile ou Face. Karl parie systématiquement sur Pile et Ronald sur Face.

- Au bout de 6 lancers :
  - Karl obtient 1 fois Pile
  - Ronald obtient 5 fois Face

  $\Longrightarrow$ Cela vous semble-t-il suspect ?
- Ils continuent. Au bout de 18 lancers :
  - Karl obtient 4 fois Pile
  - Ronald obtient 14 fois Face

  $\Longrightarrow$ Cela vous semble-t-il suspect ?

Introduction : Jeu de Pile ou Face et triche

- Au bout de 6 lancers :
  - Karl obtient 1 fois Pile
  - Ronald obtient 5 fois Face

  $\Rightarrow$ Cela vous semble-t-il suspect ? Si $X \sim \mathcal{B}(6, \frac{1}{2})$ alors

$$\mathbb{P}(\mathbf{X \geq 5}) = \mathbf{1 - pbinom(4, 6, 0.5) \approx 0.109}$$

- Ils continuent. Au bout de 18 lancers :
  - Karl obtient 4 fois Pile
  - Ronald obtient 14 fois Face

  $\Rightarrow$ Cela vous semble-t-il suspect ? Si $X \sim \mathcal{B}(18, \frac{1}{2})$ alors

$$\mathbb{P}(\mathbf{X \geq 14}) = \mathbf{1 - pbinom(13, 18, 0.5) \approx 0.015}$$

Karl a 985 chances sur 1000 de ne pas se tromper en refusant d'attribuer au hasard seul sa perte au jeu.

Formulation des hypothèses

Jeu de Pile ou Face et triche

On souhaite déterminer si Ronald est un tricheur ou est honnête. On confronte alors ces deux hypothèses :

- H_0 : Ronald est honnête, chaque lancer a une chance sur deux de faire Face. (Hypothèse nulle)
- H_1 : Ronald est un tricheur, il utilise une pièce qui a plus de chances de faire Face. (Hypothèse alternative)

Ces deux hypothèses ne jouent pas des rôles symétriques :

- la première suppose que Ronald n'a pas d'effet sur le jeu, que seul le hasard intervient ;
- tandis que la second considère qu'un processus supplémentaire (par exemple la triche, utilisation d'une pièce truquée) modifie les résultats par rapport au premier cas de figure.

Formulation mathématiques des hypothèses

Jeu de Pile ou Face et triche

Le modèle probabiliste doit permettre de voir si l'échantillon observé est une « exception » ou s'il ne diffère pas significativement de la **majorité** des autres échantillons choisis au hasard.

Soit X la v.a comptant le nombre de Face obtenues après n lancers (ici $n = 18$), les hypothèses H_0 et H_1 peuvent se réécrire comme des hypothèses sur la loi de X, ce qui se traduit par un **test paramétrique** :

- $H_0 : X \sim \mathcal{B}(n, p)$ où $p = \frac{1}{2}$. (Hypothèse nulle)
- $H_1 : X \sim \mathcal{B}(n, p)$ où $p > \frac{1}{2}$. (Hypothèse alternative)

$\Rightarrow$ p n'est connue que dans le cas H_0 où Ronald est honnête, on ne peut donc calculer explicitement de probabilité que dans le cas de l'hypothèse H_0. On dit que H_0 est **testable** et que H_1 ne l'est pas directement.

Introduction : Jeu de Pile ou Face et triche

Principe du test d'hypothèse

Analogie avec un procès en justice

- On se place sous l'hypothèse H_0 (présomption d'innocence) pour voir s'il est raisonnable de maintenir cette hypothèse au vu des données observées (éléments de l'enquête).
- À l'issue du test statistique (après enquête), on pourra prendre la décision de **rejeter l'hypothèse** H_0 (de condamner Ronald) si l'on considère les résultats de l'expérience comme incompatibles avec cette l'hypothèse, jugée fortement improbable au vu des données.
- Si au contraire les résultats sont compatibles avec l'hypothèse, on dira que l'on ne rejette pas H_0 (Ronald est acquitté).

$\Rightarrow$ Cela ne signifie pas que l'on ait la certitude que H_0 soit vrai (être acquitté est différent que d'être innocent), mais que l'**on ne dispose pas d'assez de preuves pour la rejeter** (c'est-à-dire ici pour accuser Ronald).

Introduction : Jeu de Pile ou Face et triche

Risques d'erreur

Quand on prend des décisions en se basant sur des tests statistiques, on n'est pas à l'abri de commettre des erreurs. Elles sont de deux types :

- **Erreur de type I** : Rejeter à tord H_0, cela revient à accuser un innocent (erreur judiciaire).
- **Erreur de type II** : Accepter à tord H_0, cela revient à innocenter un coupable.

Etat \ Décision	Accepter H_0	Rejeter H_0
H_0 vraie	Pas d'erreur	Erreur de type I
H_1 vraie	Erreur de type II	Pas d'erreur

Introduction : Jeu de Pile ou Face et triche

Risques d'erreur

- **Erreur de type I** : Probabilité de rejeter H_0 alors que H_0 est vraie :

$$\alpha = \mathbb{P}(\text{rejeter } H_0 \mid H_0 \text{ est vraie})$$

Proposition contraire : accepter H_0 alors que H_0 est vraie (vraisemblance) :$1 - \alpha = \mathbb{P}(\text{accepter } H_0 \mid H_0 \text{ est vraie})$

- **Erreur de type II** : Probabilité d'accepter H_0 alors que H_1 est vraie :

$$\beta = \mathbb{P}(\text{accepter } H_0 \mid H_1 \text{ est vraie})$$

Proposition contraire : rejeter H_0 lorsque que H_1 est vraie (puissance) :$1 - \beta = \mathbb{P}(\text{rejeter } H_0 \mid H_1 \text{ est vraie})$

Etat \ Décision	Accepter H_0	Rejeter H_0
H_0 vraie	Pas d'erreur (proba $1-\alpha$)	Erreur (proba α)
H_1 vraie	Erreur (proba β)	Pas d'erreur (proba $1-\beta$)

Introduction : Jeu de Pile ou Face et triche

Région de rejet

La prise de décision se fera en fonction de l'appartenance des données observées à une certaine région de valeurs. Ici, on a envie :

- D'accuser Ronald de tricherie si le nombre de Faces est très élevé.
- De ne pas l'accuser si le nombre de Faces est raisonnable.

On cherche donc une région, que l'on notera W_α appelée **région critique**, composée de valeurs élevées, dans laquelle on a peu de chances de tomber si jamais H_0 est vraie :

$$\mathbb{P}(X \in W_\alpha \,|\, H_0) \leq \alpha$$

On choisit de **rejeter** H_0 dans cette région.

Introduction : Jeu de Pile ou Face et triche

Région de rejet

Dans notre exemple, on prend par exemple $\alpha = 0.05$ et on rejette H_0 si le nombre de Faces observé est trop grand au niveau α, c'est-à-dire s'il est plus grand qu'une valeur seuil k_α qui dépend du risque d'erreur que l'on est prêt à accepter.

Pour trouver cette région W_α la plus grande possible, on doit chercher tous les k tels que

$$\mathbb{P}(X \geq k) \leq \alpha$$

et prendre la plus petite parmi elles. Par exemple pour $X \sim \mathcal{B}(6, \frac{1}{2})$ on a

$$\mathbb{P}(X \geq 6) = 0.0156, \quad \mathbb{P}(X \geq 5) = 0.109$$

donc $W_\alpha = \{6\}$, tandis que pour $X \sim \mathcal{B}(18, \frac{1}{2})$ on a

$$W_\alpha = \{13, 14, 15, 16, 17, 18\}$$

Introduction : Jeu de Pile ou Face et triche

Notion de p-valeur

Si on prend un risque $\alpha = 0.01$ alors la région critique sera

$$W_\alpha = \{15, 16, 17, 18\}$$

qui ne contient pas la valeur observée 14. Donc **entre les deux niveaux de risques il y a une valeur α où on change de décision, cette valeur s'appelle la p-valeur.**

Dans notre exemple, pour quel niveau α a-t-on

$$W_\alpha = \{14, 15, 16, 17, 18\} \ ?$$

On obtient

$$\boxed{\alpha = \mathbb{P}(X \geq 14) = 0.015}$$

La p-valeur est donc par définition la probabilité sous l'hypothèse nulle d'observer des données au moins aussi grandes que la donnée observée.