Statistiques

La statistique et les probabilités sont les deux aspects complémentaires de l'étude des phénomènes aléatoires. Ils sont cependant de natures bien différentes. Les probabilités peuvent être envisagées comme une branche des mathématiques pures, basée sur la théorie de la mesure. Les probabilités appliquées proposent des modèles probabilistes du comportement de phénomènes aléatoires concrets. **On peut alors, préalablement à toute expérience, faire des prévisions sur ce qui va se produire.**

Par exemple, il est usuel de modéliser la durée de bon fonctionnement d'un système par une variable aléatoire X de loi exponentielle de paramètre λ .

Ayant adopté ce modèle, on dira que la probabilité que le système ne soit pas encore tombé en panne à la date t est : $P(X> t)=e^{-\lambda t}$.

Dans la pratique, l'utilisateur d'un tel système est très intéressé par ces résultats. Il souhaite évidemment avoir une évaluation de la durée de bon fonctionnement de ce système, de la probabilité qu'il fonctionne correctement pendant plus d'un mois, un an, etc… Mais si l'on veut utiliser les résultats théoriques énoncés plus haut, il faut d'une part pouvoir s'assurer que la durée de vie de ce système est bien une variable aléatoire de loi exponentielle, et, d'autre part, pouvoir calculer d'une manière ou d'une autre la valeur du paramètre λ . **C'est la statistique qui va permettre de résoudre ces problèmes.**

I. Statistique descriptive

La statistique descriptive a pour but de résumer l'information contenue dans les données de façon à en dégager les caractéristiques essentielles sous une forme simple et intelligible. Les deux principaux outils de la statistique descriptive sont les représentations graphiques et les indicateurs statistiques.

1) Généralités

Les données dont nous disposons sont des mesures faites sur des individus (ou unités statistiques) issus d'une population. On s'intéresse à une ou plusieurs particularités des individus appelées variables ou caractères. L'ensemble des individus constitue l'échantillon étudié.
Une variable statistique peut être discrète ou continue, qualitative ou quantitative. Les méthodes de représentation des données diffèrent suivant la nature des variables étudiées.

On considère une population Ω d'individus $w_1,\dots,w_n$
Soit $X:\Omega \to \mathbb{R}$ une statistique simple quelconque.
Notons $X(\Omega)=\{x_i, i \in [\![1,p]\!]\}$ (quitte à les renuméroter, on les suppose rangés dans l'ordre croissant).

a) Série statistique discrète

L'effectif de la valeur x_i correspond au cardinal noté n_i de $X^{-1}(\{x_i,\})$
L'effectif cumulé en x_i correspond à $\sum_{j=1}^{i} x_j$,
L'effectif total est : $\sum_{i=1}^{p} n_i$

La fréquence de la valeur x_i correspond à $f_i = \frac{n_i}{\sum_{i=1}^{p} n_i}$

La fréquence cumulée de la valeur x_i correspond $\sum_{j=1}^{i} f_j$,

Remarque : Les familles (x_i,n_i) ou (x_i,f_i) sont des familles statistiques discrètes.

b) Série statistique groupée

Soit]a,b]$\subset \mathbb{R}$

L'effectif de la classe]a,b] est le cardinal de X^{-1}(]a,b])

L'effectif cumulé en b est le cardinal de X^{-1}(]-∞,b])

L'effectif total est card((Ω))

La fréquence de la classe]a,b] est $\frac{card(X^{-1}(]a,b]))}{\text{card}((\Omega))}$

La fréquence cumulée de la classe]a,b] est $\frac{card(X^{-1}(]-\infty,b]))}{\text{card}((\Omega))}$

2) Représentations graphiques

a) Série statistique discrète

Diagramme en bâtons ou colonnes pour les effectifs, fréquences qu'ils soient cumulés ou non

Polygone des effectifs ou fréquences, cumulés ou non.

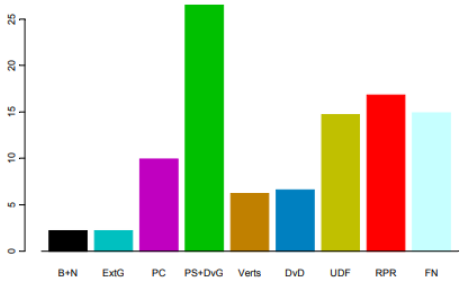

Figure 2.1. : élections législatives, diagramme en colonnes

b) Série statistique groupée

Histogramme dont les aires sont proportionnelles aux effectifs ou aux fréquences

Polygone des effectifs ou fréquences, cumulés ou non

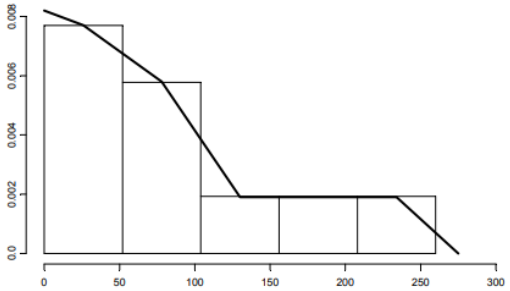

Figure 2.4. : trafic sur internet, histogramme à classes de même largeur et polygone des fréquences

3) Caractéristiques de position
   a) Moyenne empirique
   Si (x_i,n_i) avec $i \in [\![1,p]\!]$est une série discrète, alors la moyenne empirique de l'échantillon est la moyenne arithmétique des observations $\overline{x} = \frac{\sum_{i=1}^{p} n_i x_i}{\sum_{i=1}^{p} n_i}$

   Rappel : $\forall (a,b) \in \mathbb{R}^2$, E(aX+b) = aE(X)+b

   b) Valeurs extrêmes
   La plus petite valeur $x_1^* = \min x_i$ et la plus grande valeur $x_n^* = \max x_i$ d'un échantillon sont des indications intéressantes.

   Problème : Les deux indicateurs que l'on vient de définir sont très sensibles aux valeurs extrêmes. En particulier, il arrive parfois qu'une série statistique présente des valeurs aberrantes, c'est à dire des valeurs exagérément grandes ou petites par rapport aux autres valeurs de l'échantillon.
   Aussi est-il important de disposer d'indicateurs qui ne soient pas trop sensibles aux valeurs aberrantes. La médiane empirique est un indicateur de localisation construit pour être insensible aux valeurs aberrantes

   c) Médiane empirique
   On appelle médiane empirique d'une série statistique la valeur du caractère qui partage les fréquences et donc les effectifs en deux parties de même effectif

   Pour une série discrète :
   Soit X :$\Omega \to \mathbb{R}$ une statistique simple quelconque.
   Notons $X(\Omega)=\{x_i, i \in [\![1,p]\!]\}$ (quitte à les renuméroter, on les suppose rangés dans l'ordre croissant).
   Si p est impair, la médiane de cette série est le nombre $x_{\frac{p+1}{2}}$
   Si p est pair, la médiane de cette série est le nombre $\frac{x_{\frac{p}{2}}+x_{\frac{p}{2}+1}}{2}$

   Pour une série regroupée :
   La médiane est le réel μ correspondant à 50% des fréquences cumulées.

4) Caractéristiques de dispersion.
   Pour exprimer les caractéristiques d'un échantillon, il est nécessaire de compléter les indicateurs de localisation par des indicateurs de dispersion, qui mesureront la variabilité des données.

   Si (x_i,n_i) avec $i \in [\![1,p]\!]$est une série discrète, on appelle variance empirique de la série notée **V(X) la quantité : $E(X^2)-E(X)^2$**
   L'écart-type empirique : $\sigma(X) = \sqrt{V(X)}$

Etude d'un exemple :

	J	F	M	A	M	J	J	A	S	O	N	D
New-York	0	1	5	12	17	22	25	24	20	14	8	2
San Francisco	9	11	12	13	14	16	17	17	18	16	13	9

Tableau 2.9. : températures mensuelles moyennes à New-York et à San Francisco

La température annuelle moyenne est de 12.5° à New-York et de 13.7° à San Francisco. En se basant uniquement sur ces moyennes, on pourrait croire que les climats de ces deux villes sont similaires. Or il est clair que la différence de température entre l'hiver et l'été est beaucoup plus forte à New-York qu'à San Francisco. Ainsi, l'écart-type des températures annuelles est de 8.8° à New-York et de 3° à San Francisco, ce qui exprime bien la différence de variabilité des températures entre les deux villes.

II. Statistique descriptive à 2 variables

1) Généralités

Soient X et Y deux séries statistiques simples, on appelle statistique double l'application : (X,Y) :$\Omega \to \mathbb{R}^2$, notée $((x_i ; y_i), n_{i,j})$ avec $i \in [\![1,p]\!]$ et $j \in [\![1,q]\!]$

L'espérance de XY, notée E(XY) = $\frac{1}{n}\sum_{i\in[\![1,p]\!]}\sum_{j\in[\![1,q]\!]} n_{i,j}x_i y_j$

La covariance de X,Y notée cov(X,Y) = E(XY)-E(X)E(Y)

Exemple :

On dispose des données suivantes sur les variables X et Y :

X	Y
2	3
4	7
6	5
8	9
10	11

1. Calcule la moyenne de X et de Y.
2. Calcule la covariance entre X et Y.

La covariance entre X et Y, notée $\mathrm{Cov}(X,Y)$, est calculée comme suit :

$$\mathrm{Cov}(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$$

Pour chaque paire (X_i, Y_i), on calcule $(X_i - \bar{X})$ et $(Y_i - \bar{Y})$:

X_i	Y_i	$X_i - \bar{X}$	$Y_i - \bar{Y}$	$(X_i - \bar{X})(Y_i - \bar{Y})$
2	3	2 - 6 = -4	3 - 7 = -4	(-4) * (-4) = 16
4	7	4 - 6 = -2	7 - 7 = 0	(-2) * 0 = 0
6	5	6 - 6 = 0	5 - 7 = -2	0 * (-2) = 0
8	9	8 - 6 = 2	9 - 7 = 2	2 * 2 = 4
10	11	10 - 6 = 4	11 - 7 = 4	4 * 4 = 16

Ensuite, on fait la somme des produits $(X_i - \bar{X})(Y_i - \bar{Y})$ et on divise par n :

$$\mathrm{Cov}(X,Y) = \frac{1}{5}(16 + 0 + 0 + 4 + 16) = \frac{36}{5} = 7.2$$

2) Coefficient de corrélation linéaire :

On a $\rho = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$

Ainsi, si X et Y sont indépendantes, comme E(XY)=E(X)E(Y) alors cov(X,Y)=0, donc le coefficient de corrélation est nul donc les variables X et Y sont dites décorrélées.

La réciproque est cependant fausse, par exemple, si E(X)=0 et en prenant Y=|X|.

Etude d'un exemple :

On dispose des données suivantes sur les variables X et Y :

X	Y
1	2
2	3
3	5
4	4
5	6

1. Calcule la moyenne de X et de Y.
2. Calcule la covariance entre X et Y.
3. Calcule l'écart-type de X et de Y.
4. Calcule le coefficient de corrélation linéaire de Pearson entre X et Y.

On obtient : 1,8 pour la covariance et 0,906 de coefficient de corrélation !

En résumé (en bleu les commandes R permettant l'obtention des calculs)

- Mesures de tendance centrale (position)
  - Moyenne : $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ (`mean`)
  - Médiane : partage les valeurs en deux parties (`median`)
  - Quantiles : partagent les valeurs en k parties (`perctl`)
  - Quartiles ($k = 4$) : Q_1, Q_2 (médiane), Q_3 (`quart`)
  - Mode(s) : la (les) valeur(s) avec la plus grande fréquence
- Mesures de dispersion
  - Étendue : $x_{(n)} - x_{(1)}$ (`max` - `min`)
  - Intervalle interquartile (IQR) : $Q_3 - Q_1$ (`iqr`)
  - Variance de l'échantillon : (`variance`)
    $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{n\sum_{i=1}^{n}(x_i)^2 - \left(\sum_{i=1}^{n} x_i\right)^2}{n(n-1)}$ (attn. si $s/\bar{x} \ll 1$)
  - Écart-type de l'échantillon : s (`stdev`)
  - Écart absolu médian par rapport à la médiane (`mad`)
  - Coefficient de variation : $s/\bar{x}$

3) Méthode des moindres carrés.

i) Positionnement du problème

De nombreuses séries statistiques $(x_i; y_i)_{1\leq i\leq n}$sont reliées par des conditions du type y=ax+b. Ce peut être aussi le cas de grandeurs issues de la physique. En général, en raison des erreurs de mesure, les points $(x_i; y_i)_{1\leq i\leq n}$ ne sont pas alignés, mais sont "presque" sur une même droite. Il faut alors choisir a et b de sorte que la droite soit la meilleure possible.

Pour cela, il faut choisir une mesure de l'écart entre une droite y=ax+b et le nuage de points expérimentaux $(x_i ; y_i)_{1\leq i\leq n}$

On choisit en général le carré de la différence entre le point théorique et le point expérimental, c'est-à-dire $(y_i-(ax_i+b))^2$. L'écart total est donc :

$J(a,b)=\sum_{i=1}^{n}(y_i - ax_i - b)^2$

Effectuer une régression linéaire au sens des moindres carrés, c'est trouver la droite qui minimise l'écart précédent, c'est-à-dire la somme des carrés des différences : on parle de droite des moindres carrés.

ii) Théorème :

Si la variance Var(X) de la série statistique X=(xi) est non nulle, il existe une unique droite qui minimise la quantité J(a,b).
a=Cov(X,Y)/Var(X) et b=$\overline{Y} - a\overline{X}$,où Cov(X,Y) désigne la covariance de X et de Y $\overline{X}$ la moyenne de (xi) et $\overline{Y}$ la moyenne de (yi).

Démonstration :
Elle est faite complètement dans l'exercice n°1 de la feuille d'exercices.

iii) Exemple :
La tension U aux bornes d'une batterie de force électromotrice E et de résistance interne R est U=E−RI où I est l'intensité. On a procédé à différentes mesures :

Intensité mesurée (A) :	0	0,1	0,4	1
Tension mesurée (V) :	12	11	7	1

La régression linéaire donne E=11,9 et R=11,07.

III. Estimation

Dans ce chapitre, on suppose que les données $x_1, \dots, x_n$ sont n réalisations indépendantes d'une même variable aléatoire X , appelée variable parente. Il est équivalent de supposer que $x_1, \dots, x_n$ sont les réalisations de variables aléatoires $X_1, \dots, X_n$ indépendantes et de même loi. Nous adopterons ici la seconde formulation, qui est plus pratique à manipuler.

Les techniques de statistique descriptive, comme l'histogramme ou le graphe de probabilités, permettent de faire des hypothèses sur la nature de la loi de probabilité des X_i .

Des techniques statistiques plus sophistiquées, appelées tests d'adéquation, permettent de valider ou pas ces hypothèses. On supposera ici que ces techniques ont permis d'adopter une famille de lois de probabilité bien précises (par exemple, loi normale, loi binomiale, etc ...) pour la loi des X_i , mais que la valeur du ou des paramètres de cette loi est inconnue.

On notera θ le paramètre inconnu. A priori, θ peut-être un paramètre à plusieurs dimensions, mais on supposera ici que **θ est un réel**. On notera F(x;θ) la fonction de répartition des X_i . Pour les variables aléatoires discrètes on notera $P(X = x;\theta)$ les probabilités élémentaires, et pour les variables aléatoires continues, f(x;θ)la densité.

Le problème traité dans ce chapitre est celui de l'estimation du paramètre θ . Il s'agit de donner, au vu des observations $x_1, \dots, x_n$ une approximation de θ que l'on espère la plus proche possible de la vraie valeur inconnue.

On pourra proposer une unique valeur vraisemblable pour θ (**estimation ponctuelle**) ou un ensemble de valeurs vraisemblables (**intervalle de confiance**) à l'aide d'estimateurs (**sans biais, convergents** etc.)

1) Généralités

a) Définition
Un n-échantillon de X est un n-uplet $(X_1, X_2, \dots, X_n)$ tel que les X_k ont la même loi que X et sont indépendantes.
Une réalisation de l'échantillon est alors un n-uplet $(x_1, x_2, \dots, x_n)$ de valeurs prises par l'échantillon.

b) Définition
Une statistique « s » est une fonction des observations $x_1, x_2, \dots, x_n$. En ce qui concerne ce cours, elle sera à valeurs dans $\mathbb{R}$.

Exemple : $x^* = \min x_i$, $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ sont des statistiques

Remarque : $s(x_1, x_2, \dots, x_n)$ et $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ sont des réalisations respectivement de $s(X_1, X_2, \dots, X_n)$ et de $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

Un estimateur d'une grandeur θ est une statistique S_n à valeurs dans l'ensemble des valeurs possibles de θ.
Une estimation est la valeur de l'estimateur correspondant à une réalisation de l'échantillon.

Exemple : $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ est un estimateur de l'espérance mathématique.

Soit X une variable aléatoire dont la loi dépend d'un paramètre θ
Soit $(X_1,\dots,X_n)$ un n-échantillon de la variable X
On appelle estimateur de θ toute variable aléatoire T_n qui s'exprime en fonction des variables aléatoires $X_1,\dots,X_n$ et dont l'expression ne fait pas apparaître θ.
En d'autres termes, T_n est un estimateur de θ s'il existe une fonction φ de n variables telle que : $T_n = \varphi(X_1,\dots,X_n)$
Dans ce cas, on appelle estimation de θ tout réel de la forme $\varphi(x_1, \dots, x_n)$ avec $(x_1, \dots, x_n) \in X_1(\Omega) \times \dots \times X_n(\Omega)$

Ainsi : $X_1 + \dots + X_n$ et $\sqrt{X_1 + \cdots + X_n}$ sont des estimateurs de θ
Alors que : $X_1 + \dots + X_n - \theta$ n'est pas un estimateur de θ puisqu'il s'exprime avec θ !

c) Définition
A priori, n'importe quelle fonction des observations à valeurs dans l'ensemble des valeurs possibles de θ est un estimateur de θ . Mais un estimateur S_n de θ ne sera satisfaisant que si, pour n'importe quelle observation $x_1, x_2, \dots, x_n$, il est « proche », en un certain sens, de θ .
Pour cela, il faut d'abord que, si on répète plusieurs fois l'expérience, la moyenne des estimations obtenues soit très proche, et dans l'idéal égale à θ .
Cela revient à souhaiter que l'espérance de l'estimateur soit égale à θ .

Il est ainsi naturel d'évaluer l'écart entre l'espérance de l'estimateur et θ.
Le biais de l'estimateur T de θ est $E[T]-\theta$.
S'il est nul, on dit que T est un estimateur sans biais.
L'estimateur T_n est asymptotiquement sans biais si $\lim E[T_n] = \theta$.

Notation : On note souvent le biais $b_\theta(T)$.

d) Définition
L'estimateur est dit convergent si la suite (T_n) converge en probabilité vers θ_0 : $\forall \varepsilon > 0, \lim_{n\to+\infty} P(|T_n - \theta_0| > \varepsilon) = 0$

Remarque : D'après Bienaymé-Tchebychev pour qu'un estimateur asymptotiquement sans biais soit convergent il suffit que $V(T_n) \to 0$

e) Définition
Soit T un estimateur de θ.
Le risque quadratique est défini par $R(T, \theta) = E[(T - \theta)^2] = V(T) + b_\theta(T)^2$

On dit que T_1 est un meilleur estimateur que T_2 si $\forall \theta \in I$, $R(T_1, \theta) \leq R(T_2, \theta)$

De deux estimateurs sans biais, le meilleur est celui qui a la plus petite variance. C'est logique : il faut non seulement que la moyenne des estimations soit proche de θ , mais aussi que chaque estimation soit la plus proche possible de θ , donc que la variabilité de l'estimateur S_n soit faible.
Finalement, on considèrera que le meilleur estimateur possible de θ est un

estimateur sans biais et de variance minimum (ESBVM). Un tel estimateur n'existe pas forcément

f) Exemples classiques
Soit X une variable aléatoire telle que $E[X] = m$ et $Var(X) = \sigma^2$.
$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ est un estimateur sans biais et convergent de l'espérance mathématique m.

2) Estimateurs par la méthode des moments

C'est la méthode la plus naturelle. L'idée de base est d'estimer une espérance mathématique par une moyenne empirique, une variance par une variance empirique, etc.
Si le paramètre à estimer est l'espérance de la loi des X_i , alors on peut l'estimer par la moyenne empirique de l'échantillon. Autrement dit, si $\theta = E[X]$, alors l'estimateur de θ par la méthode des moments (EMM) est $\widetilde{\theta_n} = \overline{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i$

Plus généralement, pour $\theta \in \mathbb{R}$, si $E[X] = \phi(\theta)$, où ϕ est une fonction inversible, alors l'estimateur de θ par la méthode des moments est $\widetilde{\theta_n} = \phi^{-1}(\overline{X_n})$

De la même manière, on estime la variance de la loi des X_i par la variance empirique de l'échantillon $S_n{}^2 = \frac{1}{n}\sum_{i=1}^{n} X_i{}^2$ - $E(X_n)^2$

Exemple n°1 : Loi de Bernoulli
Si $X_1, \ldots, X_n$ sont indépendantes et de même loi de Bernoulli B(p), $E[X] = p$.
L'estimateur de p par la méthode des moments est $\widetilde{p_n} = \overline{X_n}$.
Cet estimateur n'est autre que la proportion de 1 dans l'échantillon. On retrouve donc le principe d'estimation d'une probabilité par une proportion.

Exemple n°2 : Loi exponentielle
Si $X_1, \ldots, X_n$ sont indépendantes et de même loi exponentielle exp(λ), $E[X] = 1/\lambda$.
Donc l'estimateur de λ par la méthode des moments est $\widetilde{\lambda_n} = \frac{1}{\overline{X_n}}$

Exemple n°3 : Loi normale
Si $X_1, \ldots, X_n$ sont indépendantes et de même loi normale N (m, σ^2), $E[X] = m$ et Var[X] = σ^2 , donc les estimateurs de m et σ^2 par la méthode des moments sont $\widetilde{m_n} = \overline{X_n}$.et $\widetilde{\sigma_n}{}^2 = = S_n{}^2$

Exemple n°4 :
Si $X_1, \ldots, X_n$ sont indépendantes et de même loi avec $E[X] = a/\lambda$ et $Var[X] = a/\lambda^2$.
On en déduit facilement que : $\lambda = E[X]/Var[X]$ et $a = E[X]^2/ Var[X]$

Donc la EMM donne $\widetilde{\lambda_n} = E[X_n]/ S_n{}^2$ et $\widetilde{a_n} = E[X_n]^2/ S_n{}^2$

3) Estimateurs par intervalle de confiance

Pour l'estimation ponctuelle, on considère un paramètre inconnu θ, un ensemble de valeurs observées $(x_1, \ldots, x_n)$, réalisations d'un échantillon aléatoire $(X_1, \ldots, X_n)$, et son estimation ponctuelle $\overline{x_n} = \frac{1}{n}\sum_{i=1}^{n} x_i$.

Les estimations ponctuelles n'apportent pas d'information sur la précision des résultats, c'est-à-dire qu'elles ne tiennent pas compte des erreurs dues aux fluctuations d'échantillonnage

Pour évaluer la confiance que l'on peut avoir en une valeur, il est nécessaire de déterminer un intervalle contenant, avec une certaine probabilité fixée au préalable, la vraie valeur du paramètre : c'est l'estimation par intervalle de confiance

a) Définition

Soit $(X_1, \dots, X_n)$ un n-échantillon aléatoire et θ un paramètre inconnu de la loi des X_i .

Soit $\alpha \in]0, 1[$. S'il existe des v.a.r. $\theta\min(X_1, \dots, X_n)$ et $\theta\max(X_1, \dots, X_n)$ telles que $P(\theta \in [\theta\min(X_1, \dots, X_n), \theta\max(X_1, \dots, X_n)]) = 1-\alpha$

On dit alors que $[\theta\min(X_1, \dots, X_n), \theta\max(X_1, \dots, X_n)]$ est un intervalle de confiance pour θ avec un coefficient de confiance $1-\alpha$ noté IC

Dans la pratique, on peut prendre par exemple $\alpha = 5\%$, ce qui nous donne un IC à 95%. Cela signifie qu'il y a 95% de chance que la valeur inconnue θ soit comprise entre $\theta\min(x1, \dots, xn)$ et $\theta\max(x1, \dots, xn)$.

b) Intervalle de confiance pour l'espérance et la variance dans un échantillon Gaussien.

Soit $(X_1, \dots, X_n)$ un n-échantillon de v.a.r. de loi $N(\mu, \sigma^2)$.

i) **Si la variance σ^2 est connue**

Pour estimer μ, on utilise la moyenne empirique $\overline{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i$ qui a pour loi $N(\mu, \sigma^2/n)$.

Ainsi, $\sqrt{n}\frac{\overline{X_n}-\mu}{\sigma} \sim N(0,1)$

On détermine $z_{1-\alpha/2}$, tel que $P(-z_{1-\alpha/2} \leq \sqrt{n}\frac{\overline{X_n}-\mu}{\sigma} \leq z_{1-\alpha/2}) = 1-\alpha$, on obtient un $IC = [\overline{X_n} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}};\ \overline{X_n} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}]$

Au fait, pourquoi, $z_{1-\alpha/2}$ et non $z_{1-\alpha}$?
Si $X \sim N(0,1)$, $P(-x \leq X \leq x) = P(X \leq x) - P(X \leq -x)$
Or par symétrie de la Gaussienne : $P(X \leq -x) = P(X \geq x)$
Et $P(-x \leq X \leq x) = P(X \leq x) - P(X \geq x) = P(X \leq x) - (1 - P(X \leq x))$
Donc : $P(-x \leq X \leq x) = 2P(X \leq x) - 1$

Ainsi $P(-x \leq X \leq x) = 1-\alpha \Leftrightarrow 2P(X \leq x) - 1 = 1-\alpha$
D'où l'utilisation de $IC = [\overline{x_n} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}};\ \overline{x_n} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}]$ où $\overline{x_n}$ est l'estimation ponctuelle de μ associée à la réalisation du n-échantillon $(X_1, \dots, X_n)$.

Remarque : Dans le cadre d'un intervalle à 95%, on $z_{1-\alpha/2} \approx 1.96$

	0	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,7580	0,7611	0,7642	0,7673	0,7703	0,7734	0,7764	0,7793	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,8849	0,8869	0,8888	0,8906	0,8925	0,8943	0,8962	0,8980	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
2	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,1	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2,3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,4	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,5	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2,6	0,9953	0,9955	0,9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964
2,7	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972	0,9973	0,9974
2,8	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,9980	0,9981
2,9	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986

ii) Si la variance σ^2 est inconnue

Lorsque la variance σ^2 est inconnue, il est alors nécessaire de remplacer dans les formules précédentes cette quantité par la variance empirique, qui en est un estimateur convergent. Il faut donc considérer non plus la quantité $\sqrt{n}\frac{\overline{X_n}-\mu}{\sigma}$ mais $\sqrt{n}\frac{\overline{X_n}-\mu}{S_n}$ qui ne suit pas une loi normale mais une loi de Student à n-1 degrés de liberté.

Ainsi, $\sqrt{n}\frac{\overline{X_n}-\mu}{S_n} \sim T_{n-1}$

A l'aide de tables de la loi de Student, on détermine $t_{1-\alpha/2}$, tel que $P(-t_{1-\alpha/2} \leq \sqrt{n}\frac{\overline{X_n}-\mu}{S_n} \leq t_{1-\alpha/2})$=1-$\alpha$, on obtient un IC=$[\overline{X_n}-t_{1-\alpha/2}\frac{S_n}{\sqrt{n}};\ \overline{X_n}+t_{1-\alpha/2}\frac{S_n}{\sqrt{n}}]$

D'où l'utilisation de IC=$[\overline{x_n}-t_{1-\alpha/2}\frac{S_n}{\sqrt{n}};\ \overline{x_n}+t_{1-\alpha/2}\frac{S_n}{\sqrt{n}}]$où $\overline{x_n}$ est l'estimation ponctuelle de μ associée à la réalisation du n-échantillon $(X_1, \ldots, X_n)$.

Remarque : Dans le cadre d'un intervalle à 95%, on $t_{3;0.975} = 3.182$.

ν \ P	0,60	0,70	0,80	0,90	0,95	0,975	0,990	0,995	0,999	0,9995
1	0,325	0,727	1,376	3,078	6,314	12,71	31,82	63,66	318,3	636,6
2	0,289	0,617	1,061	1,886	2,920	4,303	6,965	9,925	22,33	31,60
3	0,277	0,584	0,978	1,638	2,353	3,182	4,541	5,841	10,22	12,94
4	0,271	0,569	0,941	1,533	2,132	2,776	3,747	4,604	7,173	8,610
5	0,267	0,559	0,920	1,476	2,015	2,571	3,365	4,032	5,893	6,859
6	0,265	0,553	0,906	1,440	1,943	2,447	3,143	3,707	5,208	5,959
7	0,263	0,549	0,896	1,415	1,895	2,365	2,998	3,499	4,785	5,405
8	0,262	0,546	0,889	1,397	1,860	2,306	2,896	3,355	4,501	5,041
9	0,261	0,543	0,883	1,383	1,833	2,262	2,821	3,250	4,297	4,781
10	0,260	0,542	0,879	1,372	1,812	2,228	2,764	3,169	4,144	4,587
11	0,260	0,540	0,876	1,363	1,796	2,201	2,718	3,106	4,025	4,437
12	0,259	0,539	0,873	1,356	1,782	2,179	2,681	3,055	3,930	4,318
13	0,259	0,538	0,870	1,350	1,771	2,160	2,650	3,012	3,852	4,221
14	0,258	0,537	0,868	1,345	1,761	2,145	2,624	2,977	3,787	4,140
15	0,258	0,536	0,866	1,341	1,753	2,131	2,602	2,947	3,733	4,073
16	0,258	0,535	0,865	1,337	1,746	2,120	2,583	2,921	3,686	4,015
17	0,257	0,534	0,863	1,333	1,740	2,110	2,567	2,898	3,646	3,965
18	0,257	0,534	0,862	1,330	1,734	2,101	2,552	2,878	3,611	3,922
19	0,257	0,533	0,861	1,328	1,729	2,093	2,539	2,861	3,579	3,883
20	0,257	0,533	0,860	1,325	1,725	2,086	2,528	2,845	3,552	3,850
21	0,257	0,532	0,859	1,323	1,721	2,080	2,518	2,831	3,527	3,819
22	0,256	0,532	0,858	1,321	1,717	2,074	2,508	2,819	3,505	3,792
23	0,256	0,532	0,858	1,319	1,714	2,069	2,500	2,807	3,485	3,767
24	0,256	0,531	0,857	1,318	1,711	2,064	2,492	2,797	3,467	3,745
25	0,256	0,531	0,856	1,316	1,708	2,060	2,485	2,787	3,450	3,725
26	0,256	0,531	0,856	1,315	1,706	2,056	2,479	2,779	3,435	3,707
27	0,256	0,531	0,855	1,314	1,703	2,052	2,473	2,771	3,421	3,690
28	0,256	0,530	0,855	1,313	1,701	2,048	2,467	2,763	3,408	3,674
29	0,256	0,530	0,854	1,311	1,699	2,045	2,462	2,756	3,396	3,659
30	0,256	0,530	0,854	1,310	1,697	2,042	2,457	2,750	3,385	3,646
32	0,256	0,530	0,853	1,309	1,694	2,037	2,449	2,738	3,365	3,622
34	0,255	0,529	0,852	1,307	1,691	2,032	2,441	2,728	3,348	3,601
36	0,255	0,529	0,852	1,306	1,688	2,028	2,434	2,719	3,333	3,582
38	0,255	0,529	0,851	1,304	1,686	2,024	2,429	2,712	3,319	3,566
40	0,255	0,529	0,851	1,303	1,684	2,021	2,423	2,704	3,307	3,551
50	0,255	0,528	0,849	1,298	1,676	2,009	2,403	2,678	3,261	3,496
60	0,254	0,527	0,848	1,296	1,671	2,000	2,390	2,660	3,232	3,460
70	0,254	0,527	0,847	1,294	1,667	1,994	2,381	2,648	3,211	3,435
80	0,254	0,527	0,846	1,292	1,664	1,990	2,374	2,639	3,195	3,415
90	0,254	0,526	0,846	1,291	1,662	1,987	2,368	2,632	3,183	3,402
100	0,254	0,526	0,845	1,290	1,660	1,984	2,365	2,626	3,174	3,389
200	0,254	0,525	0,843	1,286	1,653	1,972	2,345	2,601	3,131	3,339
500	0,253	0,525	0,842	1,283	1,648	1,965	2,334	2,586	3,106	3,310
∞	0,253	0,524	0,842	1,282	1,645	1,960	2,326	2,576	3,090	3,291

c) Etude d'un exemple

Une entreprise chimique commercialise un polymère servant à la fabrication de microprocesseurs et stocké dans une cuve dont la caractéristique à contrôler est la viscosité ; celle-ci doit être comprise entre 75 et 95 pour pouvoir commercialiser le polymère.

Quatre extractions ont été réalisées dans des zones différentes de la cuve et ont conduit aux valeurs de l'échantillon :$x_1 = 78$, $x_2 = 85$, $x_3 = 91$, $x_4 = 76$, réalisation des variables aléatoires X_1, X_2, X_3, X_4.

L'entreprise a besoin d'estimer la viscosité et aussi de connaître la précision de cette estimation. Ayant choisi a priori un seuil de 5%, il s'agit de fournir aux clients des intervalles de confiances à 95% pour μ.

1) Résumer les estimations ponctuelles
2) Donner l'intervalle de confiance avec $\sigma = 5$
3) Donner l'intervalle de confiance sans connaître la variance

– Le modèle considère que les variables X_i sont indépendantes selon une loi N (μ, σ^2); μ représente la moyenne de la viscosité dans la cuve tandis que σ^2 prend en compte la variabilité de la viscosité au sein de la cuve et celle due à l'erreur de mesure.
– Les paramètres sont la moyenne μ et la variance σ^2 .
– Les estimateurs sont $\bar{X}$ de μ et S^2 de σ^2 .
– Les estimations ponctuelles sont $x = 82.5$ et $s = 6.86$.

Il est admis que la variabilité du processus de fabrication est constante et connue avec $\sigma = 5$. Dans ce cas, l'estimateur de μ est gaussien, $z_{1-\alpha/2} = 1.96$ et les formules précédentes

conduisent à l'estimation de l'intervalle de confiance de μ : $[82.5 - 1.96 \times 5/2; 82.5 + 1.96 \times 5/2] = [77.6; 87.4]$.
L'intervalle obtenu est bien à l'intérieur de la spécification ([75; 95]).

La variance n'est plus supposée constante et connue, elle doit être estimée. L'estimation de l'écart-type est $s = 6.86$. Celui-ci est certes plus important que la valeur théorique précédente mais surtout, l'estimateur de la moyenne μ suit maintenant une loi de Student à $n - 1 = 3$ degrés de liberté. La table de la loi en question fournit le $1 - \alpha/2$-quantile $t_{3;0.975} = 3.182$.
$82.5 - 3.182 \times 6.86/2; 82.5 + 3.182 \times 6.86/2] = [71.6; 93.4]$.
L'intervalle n'est pas contenu dans la spécification. On peut noter l'augmentation sensible de la taille de cet intervalle par le simple fait de devoir estimer la variance plutôt que de la supposer connue…

4) Estimation d'une proportion

On désire évaluer la probabilité p qu'un événement A se produise au cours d'une expérience donnée : $p = P(A)$. Pour cela, on fait n expériences identiques et indépendantes et on compte le nombre x de fois où A s'est produit. x est la réalisation d'une variable aléatoire X qu'on sait être de loi binomiale B(n, p) .

Exemple : 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 et 380 pour B. Estimer le résultat de l'élection, c'est estimer le pourcentage p de voix qu'obtiendra le candidat A. En supposant que les réponses des 800 personnes interrogées sont indépendantes, on est bien dans le cas de figure de l'estimation d'une proportion.

a) estimation ponctuelle
Remarquons que nous n'avons ici qu'une seule réalisation de X (ne pas confondre la variable binomiale X qui compte les succès avec les Bernoulli), c'est à dire un échantillon de taille 1.
Pour une fois, la notation n ne désigne pas la taille de l'échantillon. Il est naturel d'estimer la probabilité p que A se produise par le pourcentage $\frac{X}{n}$ de fois où A s'est produit au cours des n expériences.
Par la méthode des moments, on a $E(X) = np$, donc l'EMM de p est $\frac{X}{n}$

b) estimation par intervalle de confiance
On peut dans le cas où $np \geq 5$ et $n(1-p) \geq 5$, approcher la loi binomiale B(n,p) par la loi normale $N(np, \sqrt{np(1-p)}^2)$, ainsi $\frac{X-np}{\sqrt{np(1-p)}}$ suit approximativement la loi N(0,1).
On peut alors déterminer u_α tel que : $P\left(\left|\frac{X-np}{\sqrt{np(1-p)}}\right| \leq u_\alpha\right) = 1 - \alpha$
Il reste à construire l'IC à partir de : $\left|\frac{X-np}{\sqrt{np(1-p)}}\right| \leq u_\alpha$

Remarque : On peut utiliser une version approchée de l'intervalle de confiance d'une proportion avec la formule : **$IC = [p_n - z_{1-\alpha/2}\frac{1}{2\sqrt{n}};\ p_n + z_{1-\alpha/2}\frac{1}{2\sqrt{n}}]$ où p_n représente la proportion observée dans l'échantillon de taille n avec erreur à α%** (*Avec $z_{1-\alpha/2}$, tel que $P(-z_{1-\alpha/2} \leq X \leq z_{1-\alpha/2}) = 1-\alpha$ où $X \sim N(0,1)$, formule basée uniquement sur la majoration de $\sqrt{p(1-p)}$ par $\frac{1}{2}$*)