Statistique descriptive et inférentielle

Ségolen Geffray

IUT Carquefou

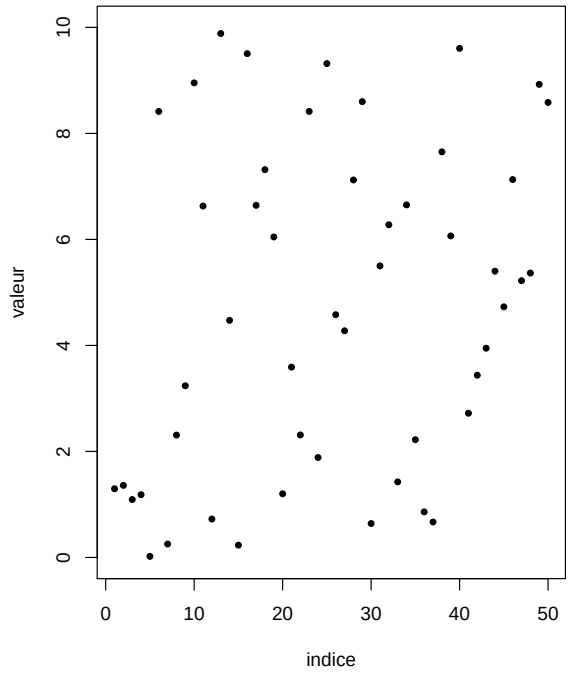
Année 2008-2009

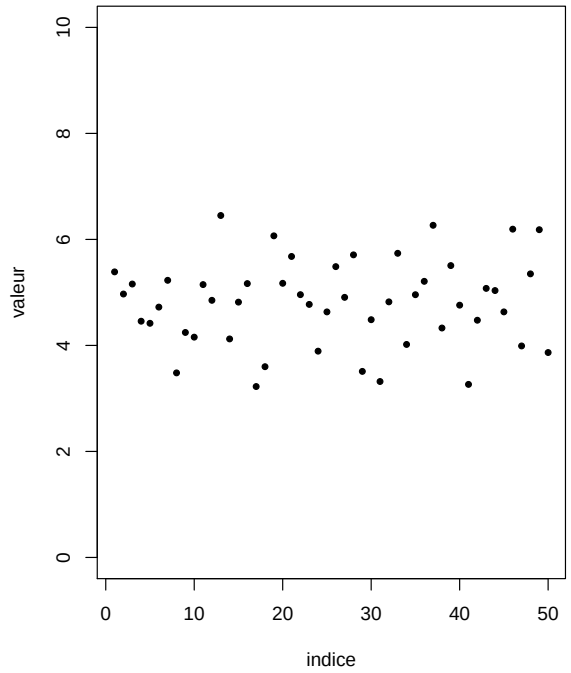
1ère partie

Statistique descriptive

Regarder ses données ! ! !

- Considérons un jeu de données $(x_1, ..., x_n)$ = échantillon.
- Première chose à faire : tracer le nuage de points. Y a-t-il des points d'accumulation ou non ? Y-a-t-il des tendances ou non ? Y-a-t-il des valeurs extrêmes ?
- ex : les deux nuages de points ci-dessous comportent $n = 50$ observations.





Caractéristiques de position

- **moyenne empirique** (=moyenne de l'échantillon) :

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- Cas des données groupées : lorsque les données sont présentées sous la forme (x_i, n_i) pour $i = 1, ..., k$ où x_i représente la valeur obtenue avec un effectif n_i, on calcule $\overline{X}$ au moyen de la formule :

$$\overline{x} = \frac{\sum_{i=1}^{k} n_i x_i}{\sum_{i=1}^{k} n_i}$$

- Si les données sont fournies en classes $[a_i, a_{i+1}[$ pour $i = 1, ...k$, on approxime $\overline{x}$ par $\frac{\sum_{i=1}^{k} n_i c_i}{\sum_{i=1}^{k} n_i}$ où n_i=effectif de la classe n°i et c_i=centre de la classe n°i : $c_i = (a_i + a_{i+1})/2$.
- La moyenne empirique est très sensible aux valeurs extrêmes (peu robuste).

Caractéristiques de position (suite)

- **médiane empirique** (=médiane de l'échantillon) : m=valeur qui partage l'effectif total rangé par ordre croissant en deux classes de même taille. Notons $x_{(1)} \leq x_{(2)} \leq ... \leq x_{(n)}$ l'échantillon ordonné. La médiane empirique est donnée par

$$m = \begin{cases} x_{((n+1)/2)} & \text{si } n \text{ est impair} \\ \frac{x_{(n/2)}+x_{(n/2+1)}}{2} & \text{si } n \text{ est pair} \end{cases}$$

  La médiane empirique est un indicateur robuste.
- Pour une distribution parfaitement symétrique, on a : moyenne=médiane. C'est utile en particulier pour vérifier rapidement la plausibilité d'une hypothèse de normalité des données : pour la loi $\mathcal{N}(0,1)$, moyenne=médiane.

Caractéristiques de dispersion

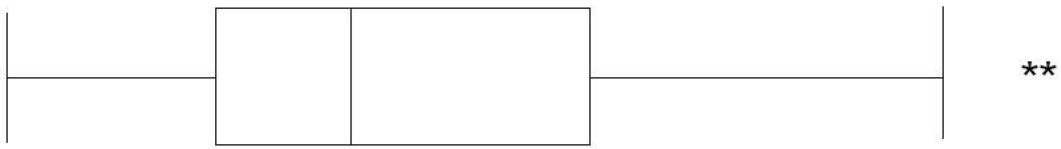
- **variance empirique** :

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

  La variance d'un jeu de données exprime à quel point les valeurs sont dispersées autour de la valeur moyenne. Plus la variance est grande, plus les données sont dispersées.
- **écart-type empirique** : $s = \sqrt{s^2}$.
- **étendue** : $r = x_{(n)} - x_{(1)}$=value maximale-valeur minimale. C'est un indicateur instable car il ne dépend que des valeurs extrêmes.
- **intervalle interquartile** : les quartiles q_1, q_2, q_3 sont les 3 valeurs partageant l'effectif total ordonné en 4 parties égales ($q_2 = m$=médiane). L'intervalle interquartile est plus robuste que l'étendue.

Caractéristiques de dispersion (suite)

- **boxplot (boîte à moustaches)** : ce diagramme représente schématiquement les principales caractéristiques d'un jeu de données en utilisant les quartiles. La partie centrale de la distribution est représentée par une boîte dont la longueur correspond à l'intervalle interquartile. On trace à l'intérieur la position de la médiane. On complète par les moustaches correspondant aux valeurs adjacentes :
  - adjacente supérieure : plus grande valeur inférieure à $q_3 + 1.5(q_3 - q_1)$
  - adjacente inférieure : plus petite valeur supérieure à $q_1 - 1.5(q_3 - q_1)$
- Les valeurs extérieures représentées par des étoiles sont celles qui sortent des moustaches.



Caractéristiques de forme

- Le **coefficient de symétrie empirique** (skewness) est défini par :

$$\widehat{\nu}_1 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^3}{(s^2)^{3/2}}.$$

- Le **coefficient d'aplatissement empirique** (kurtosis) est défini par :

$$\widehat{\nu}_2 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^4}{(s^2)^{2}}.$$

- Utile en particulier pour vérifier rapidement la plausibilité d'une hypothèse de normalité des données : pour la loi $\mathcal{N}(0,1)$, $\nu_1 = 0$ et $\nu_2 = 3$.

$2^{\text{ème}}$ partie

Statistique inférentielle

Notion d'échantillon

- On appelle **échantillon** de taille n d'une variable X une succession de n variables aléatoires $(X_1, ..., X_n)$ indépendantes et toutes de même loi.
- Cela correspond aux conditions suivantes :
  1. tous les individus sont sélectionnés dans la même population et sont donc identiques à quelques variations près
  2. les individus sont sélectionnés de manière indépendante
- Si on note P la loi de probabilité commune des X_i, parfois on dit que P est la loi parente de l'échantillon. Parfois on introduit une variable X de même loi que les X_i et on dit que X est la **variable parente** de l'échantillon.
- Après expérience, on recueille un jeu de données constitué des observations $(x_1, ..., x_n)$. C'est une réalisation de l'échantillon aléatoire $(X_1, ..., X_n)$: x_1 est la **réalisation** de la variable aléatoire X_1=1$^{\text{ère}}$ valeur obtenue en tirant au sort n sujets, etc...

Notion de statistique, d'estimateur

Soit $(X_1, ..., X_n)$ un échantillon aléatoire.

- Une **statistique** est une variable aléatoire qui est une fonction de l'échantillon $(X_1, ..., X_n)$ soit $\widehat{\theta} = \varphi(X_1, ..., X_n)$, par exemple : $\widehat{\theta} = \frac{1}{n}\sum_{i=1}^{n} X_i$.
- Soit θ un paramètre inconnu à estimer. Pour estimer un paramètre d'une distribution, on utilise forcément l'échantillon ! L'**estimateur** du paramètre est une variable aléatoire qui est une fonction $\widehat{\theta}$ des variables de l'échantillon, soit $\widehat{\theta} = \varphi(X_1, ..., X_n)$, et qui doit "approcher" $\theta \Rightarrow$ est-ce que $\widehat{\theta} = \frac{1}{n}\sum_{i=1}^{n} X_i$ approche $\theta = \mathbb{E}[X]$?
- La réalisation $\widehat{\theta} = \varphi(x_1, ..., x_n)$ s'appelle une **estimation** de θ et fournit une approximation de la valeur de θ.
- A chaque fois qu'on réalise une expérience, on obtient une nouvelle réalisation de l'échantillon et donc vraisemblablement une nouvelle valeur de $\widehat{\theta}$. Ainsi $\widehat{\theta}$ est une variable aléatoire et donc suit une loi de probabilité, a une moyenne, une variance...

Qualités d'un estimateur

- La $1^{\text{ère}}$ qualité d'un estimateur est d'être **convergent** : quand la taille de l'échantillon n augmente, $\widehat{\theta}$ doit avoir tendance à se rapprocher de θ puisque la quantité d'information augmente.
- La $2^{\text{ème}}$ qualité d'un estimateur est d'être précis. La précision d'un estimateur peut se mesurer au moyen du biais et de la variance.
  - Le **biais** est l'écart moyen entre $\widehat{\theta}$ et θ i.e. biais $= \mathbb{E}[\widehat{\theta}] - \theta$. Un estimateur est **sans biais** lorsque $\mathbb{E}[\widehat{\theta}] = \theta$ i.e. si en utilisant $\widehat{\theta}$ un grand nombre de fois, il donne en moyenne la valeur du paramètre recherché. A l'inverse, un estimateur est biaisé si en l'utilisant un grand nombre de fois, il ne donne pas en moyenne la valeur du paramètre recherché. On souhaite qu'un estimateur soit non biasé !
  - On préfère qu'un estimateur sans biais ait une **variance minimale** de sorte que ses réalisations oscillent autour de θ sans jamais s'en éloigner trop.

L'estimateur "moyenne empirique"

- Soit $(X_1, ..., X_n)$ un échantillon de variable parente X. On veut estimer le paramètre $\theta = \mathbb{E}[X]$.
- Un estimateur de $\mathbb{E}[X]$ est $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. C'est une variable aléatoire donc $\overline{X}$ une moyenne, une variance, une loi de probabilité...
- Cet estimateur est convergent : $\overline{X} \longrightarrow \theta$ quand $n \to \infty$.
- Cet estimateur est sans biais : $\mathbb{E}[\overline{X}] = \theta$.
- Sa variance satisfait : $\mathrm{Var}(\overline{X}) = \frac{\mathrm{Var}(X)}{n}$ donc la variance de $\overline{X}$ décroît vers 0 quand n tend vers $+\infty$.
- Quelque soit la loi de X, la loi de $\overline{X}$ converge toujours vers la loi normale. En pratique, pour n assez grand ($n \geq 30$), la loi de $\overline{X}$ peut être approximée par une loi normale de moyenne θ et de variance $\frac{\mathrm{Var}(X)}{n}$ soit $\mathcal{N}(\theta, \mathrm{Var}(X)/n)$.
- La loi d'échantillonnage révèle la façon dont les réalisations de $\overline{X}$ oscillent autour de θ. Cette loi sert à
  - contrôler la marge d'erreur
  - construire une estimation par intervalle

Estimation par intervalle de confiance (IC) : fluctuations prévisibles de l'estimation

- Une estimation ponctuelle ne nous renseigne pas ni sur le niveau de confiance que l'on peut avoir en l'estimation, ni sur la marge d'erreur :
  - le niveau de confiance nous dit dans quelle mesure la méthode est fiable en usage répétée,
  - la marge d'erreur nous dit dans quelle mesure la méthode est sensible, i.e. avec quelle précision l'intervalle localise le paramètre en train d'être estimé.
- Lorsqu'on est intéressé non seulement par l'estimation en elle-même mais aussi par le niveau de confiance et la marge d'erreur, on effectue une estimation par intervalle.
- A taille d'échantillon fixé à n, lorsqu'on augmente le niveau de confiance $1 - \alpha$, la largeur de l'IC augmente.
- A niveau de confiance fixé à $(1 - \alpha)$, lorsqu'on augmente la taille de l'échantillon n, la largeur de l'IC diminue.

Construction d'un IC bilatéral pour $\theta = \mathbb{E}[X]$ pour un grand échantillon

- Pour n assez grand, la loi de $\overline{X}$ peut être approchée par une loi normale de moyenne θ et de variance $\mathrm{Var}(X)/n$ et donc d'écart-type $\sigma(X)/\sqrt{n}$ (d'après le théorème central limite).
- La loi de $\sqrt{n}\frac{\overline{X}-\theta}{\sigma(X)}$ peut donc être approchée par une loi $\mathcal{N}(0,1)$ pour n assez grand.
- En pratique, $\mathrm{Var}(X)$ et $\sigma(X)$ sont inconnus : il faut les estimer !

L'estimateur "variance empirique" et "écart-type empirique"

- Soit $(X_1, ..., X_n)$ un échantillon de variable parente X. On cherche un estimateur ponctuel de $\mathrm{Var}(X)$.
- $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 - n\overline{X}^2\right)$ est un estimateur de $\mathrm{Var}(X)$.
- Cet estimateur est convergent : $S^2 \to \mathrm{Var}(X)$ quand $n \to \infty$.
- S^2 est un estimateur sans biais : $\mathbb{E}[S^2] = \mathrm{Var}(X)$.
- La variance de S^2 satisfait $\mathrm{Var}(S^2) \to 0$ quand $n \to \infty$.
- Pour n assez grand ($n \geq 30$), S^2 suit approximativement une loi normale de moyenne $\mathbb{E}[S^2] = \mathrm{Var}(X)$ et de variance un peu compliquée dans le cas général.
- $S = \sqrt{S^2}$ est un estimateur de $\sigma(X)$.

Estimation par IC de $\theta = \mathbb{E}[X]$ pour un grand échantillon

- La loi de $\sqrt{n}\frac{\overline{X}-\theta}{S}$ peut être approchée par une loi $\mathcal{N}(0,1)$.
- L'intervalle de probabilité au niveau de confiance $(1-\alpha)$ s'approxime par :

$$\mathbb{P}\left[-F_N^{-1}(1-\alpha/2) \leq \sqrt{n}\frac{\overline{X}-\theta}{S} \leq F_N^{-1}(1-\alpha/2)\right] = 1-\alpha$$

où $F_N^{-1}(1-\alpha/2)$ est le fractile d'ordre $(1-\alpha/2)$ de la loi $\mathcal{N}(0,1)$ défini comme étant la valeur x telle que $\mathbb{P}[X \leq x] = 1-\alpha/2$ pour une variable X de loi $\mathcal{N}(0,1)$.

- L'**intervalle de confiance** bilatéral pour θ au niveau de confiance $(1-\alpha)$ (valable pour n grand ! !) peut être approximé par :

$$\mathbb{P}\left[\overline{X} - F_N^{-1}(1-\alpha/2)\frac{S}{\sqrt{n}} \leq \theta \leq \overline{X} + F_N^{-1}(1-\alpha/2)\frac{S}{\sqrt{n}}\right] = 1-\alpha$$

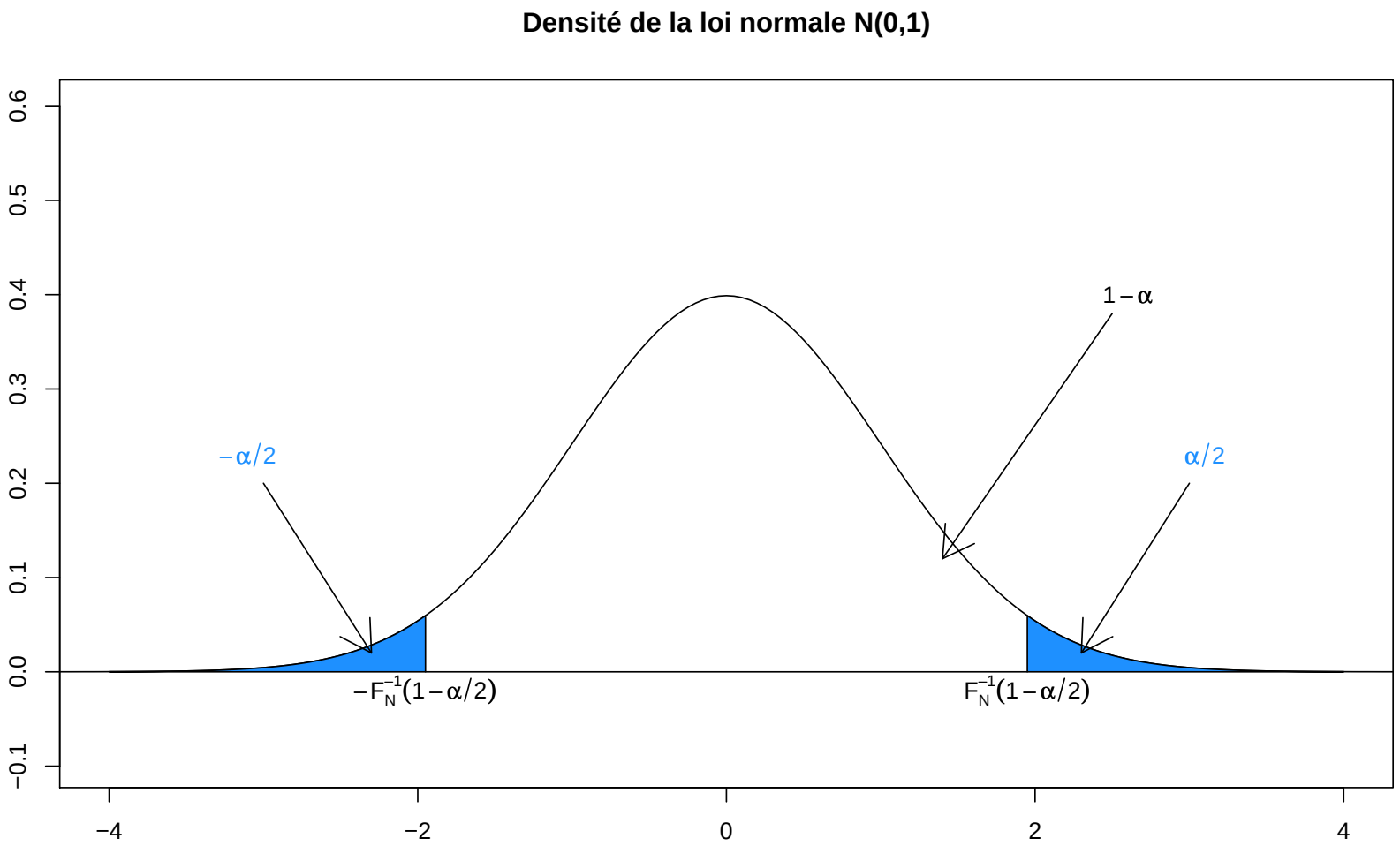
Il y a une probabilité d'environ $(1-\alpha)$ pour que l'IC contienne la moyenne de la population.

- La **marge d'erreur** est définie comme l'erreur absolue $|\overline{X}-\theta|$. Elle est inférieure à $F_N^{-1}(1-\alpha/2)\frac{S}{\sqrt{n}}$ avec une probabilité d'environ $(1-\alpha)$:

$$\mathbb{P}\left[-F_N^{-1}(1-\alpha/2)\frac{S}{\sqrt{n}} \leq \overline{X}-\theta \leq F_N^{-1}(1-\alpha/2)\frac{S}{\sqrt{n}}\right] = 1-\alpha$$

Fractile de la loi $\mathcal{N}(0,1)$ pour IC bilatéral

Le fractile d'ordre $1-\alpha/2$ de la loi $\mathcal{N}(0,1)$ est noté $F_N^{-1}(1-\alpha/2)$ et est défini comme étant le réel x tel que $F_N(x) = 1-\alpha/2$ ou encore $\mathbb{P}[X \leq x] = 1-\alpha/2$ pour X de loi $\mathcal{N}(0,1)$. On a alors $\mathbb{P}[-x \leq X \leq x] = 1-\alpha$.



Exercice : fiabilité d'un dispositif électronique

La firme Kopak vient de développer un nouveau dispositif électronique qui entre dans la fabrication du système de guidage des avions. Avant de mettre en production ce nouveau dispositif, Kopak veut effectuer des essais préliminaires sur un échantillon pour être en mesure d'en estimer la fiabilité en termes de durée de bon fonctionnement. D'après le bureau de Recherche et Développement de l'entreprise, l'écart-type de la durée de bon fonctionnement de ce nouveau dispositif électronique serait de l'ordre de 115h.

1. Déterminer la taille d'échantillon requise pour estimer, avec un niveau de confiance de 99% la durée de bon fonctionnement de sorte que la marge d'erreur dans l'estimation de la durée moyenne de bon fonctionnement n'excède pas 50h.
2. Pour le même niveau de confiance, quelle doit être la taille d'échantillon pour que la marge d'erreur dans l'estimation de la durée moyenne de bon fonctionnement n'excède pas 20h ?

Estimation par IC de Var(X) pour un grand échantillon

- Soit $(X_1, ..., X_n)$ un échantillon de variable parente X.
- Pour n assez grand, la loi de $\frac{\sqrt{n-1}}{\sqrt{2}}\left(\frac{S^2}{\text{Var}(X)} - 1\right)$ peut être approchée par une loi $\mathcal{N}(0,1)$.
- L'intervalle de probabilité au niveau de confiance $(1-\alpha)$ s'approxime par :

$$\mathbb{P}\left[-F_N^{-1}(1-\alpha/2) \leq \frac{\sqrt{n-1}}{2}\left(\frac{S^2}{\text{Var}(X)} - 1\right) \leq F_N^{-1}(1-\alpha/2)\right] = 1-\alpha$$

où $F_N^{-1}(1-\alpha/2)$ est le fractile d'ordre $(1-\alpha/2)$ de la loi $\mathcal{N}(0,1)$ défini comme étant la valeur x telle que $\mathbb{P}[X \leq x] = 1-\alpha/2$ pour une variable X de loi $\mathcal{N}(0,1)$.

- L'**intervalle de confiance** bilatéral pour Var(X) au niveau de confiance $(1-\alpha)$ peut être approché par : (valable pour n grand ! !)

$$\mathbb{P}\left[\frac{S^2}{1 + F_N^{-1}(1-\alpha/2)\frac{\sqrt{2}}{\sqrt{n-1}}} \leq \text{Var}(X) \leq \frac{S^2}{1 - F_N^{-1}(1-\alpha/2)\frac{\sqrt{2}}{\sqrt{n-1}}}\right] = 1-\alpha$$

- **La probabilité que l'IC contienne la variance de la population est d'environ $(1-\alpha)$.**

Estimation du paramètre m d'une loi normale

- Soit $(X_1, ..., X_n)$ un échantillon de loi $\mathcal{N}(m, \sigma^2)$.
- Deux estimateurs ponctuels de m sont $\widehat{m} = \overline{X}$ et $\widehat{m}$=médiane empirique.
- L'IC bilatéral pour m au niveau de confiance $(1 - \alpha)$ est donné par :

$$\mathbb{P}[\widehat{m}_{\text{inf}} \leq m \leq \widehat{m}_{\text{sup}}] = 1 - \alpha$$

avec

$$\widehat{m}_{\text{inf}} = \overline{X} - F^{-1}_{T(n-1)}(1 - \alpha/2)\frac{S}{\sqrt{n}}$$

$$\widehat{m}_{\text{sup}} = \overline{X} + F^{-1}_{T(n-1)}(1 - \alpha/2)\frac{S}{\sqrt{n}}$$

où $F^{-1}_{T(n-1)}(1 - \alpha/2)$ est le fractile d'ordre $(1 - \alpha/2)$ de $T(n-1)$, la loi de Student à $(n-1)$ degrés de libertés et est défini comme étant la valeur x telle que $\mathbb{P}[X \leq x] = 1 - \alpha/2$ pour une variable X de loi $T(n-1)$.

- La loi de Student à d degrés de liberté est tabulée pour certaines valeurs de d. Pour $d > 50$ et $\beta > 0$, on peut utiliser l'approximation :

$$F^{-1}_{T(d)}(\beta) \approx F_N(\beta)^{-1}\sqrt{\frac{d}{d-2}}$$

- **Il y a une probabilité $(1 - \alpha)$ pour que l'IC contienne le paramètre m.**

Estimation du paramètre σ d'une loi normale

- Soit $(X_1, ..., X_n)$ un échantillon de loi $\mathcal{N}(m, \sigma^2)$.
- Deux estimateurs de σ sont $\widehat{\sigma} = \frac{S}{K_S(n)}$ et $\widehat{\sigma} = \frac{X_{(n)} - X_{(1)}}{K_R(n)}$ où $K_S(n)$ et $K_R(n)$ sont tabulées. Pour $n > 30$, $K_S(n)$ s'approxime par $K_S(n) \approx 1 - \frac{1}{4(n-1)}$.
- L'IC bilatéral pour σ au niveau de confiance $(1 - \alpha)$ est donné par $\mathbb{P}[\widehat{\sigma}_{\text{inf}} \leq \sigma \leq \widehat{\sigma}_{\text{sup}}] = 1 - \alpha$ avec

$$\widehat{\sigma}_{\text{inf}} = S\sqrt{\frac{n-1}{F^{-1}_{\chi^2(n-1)}(1-\alpha/2)}} \quad \text{et} \quad \widehat{\sigma}_{\text{sup}} = S\sqrt{\frac{n-1}{F^{-1}_{\chi^2(n-1)}(\alpha/2)}}$$

  où le fractile d'ordre $(1 - \alpha/2)$ de la loi $\chi^2(n-1)$ est noté $F^{-1}_{\chi^2(n-1)}(1-\alpha/2)$ et est défini comme étant la valeur x telle que $\mathbb{P}[X \leq x] = 1 - \alpha/2$ pour X de loi $\chi^2(n-1)$. Le fractile d'ordre $(\alpha/2)$ de la loi $\chi^2(n-1)$ est noté $F^{-1}_{\chi^2(n-1)}(\alpha/2)$ et est défini comme étant la valeur x telle que $\mathbb{P}[X \leq x] = \alpha/2$ pour X de loi $\chi^2(n-1)$.
- La loi du χ^2 à d degrés de liberté est tabulée pour certaines valeurs de d. Pour $d > 50$ et $\beta > 0$, on peut utiliser l'approximation :

$$F^{-1}_{\chi^2(d)}(\beta) \approx \frac{1}{2}\left(F_N(\beta)^{-1} + \sqrt{2d-1}\right)^2$$

- **La probabilité que l'IC contienne le paramètre σ est de $(1 - \alpha)$.**

Tabulations des fonctions $K_S(n)$ et $K_R(n)$

n	2	3	4	5	6	7	8
$K_S(n)$	0.7979	0.8862	0.9213	0.9400	0.9515	0.9594	0.9650
$K_R(n)$	1.1284	1.6926	2.0588	2.3259	2.5344	2.7044	2.8472
n	9	10	11	12	13	14	15
$K_S(n)$	0.9693	0.9727	0.9754	0.9776	0.9764	0.9810	0.9823
$K_R(n)$	2.9700	3.0775	3.1729	3.2585	3.3360	3.4068	3.4718
n	16	17	18	19	20	21	22
$K_S(n)$	0.9835	0.9845	0.9854	0.9862	0.9869	0.9876	0.9882
$K_R(n)$	3.5320	3.5879	3.6401	3.6890	3.7350	3.7783	3.8194

23	24	25	26	27	28	29	30
0.9887	0.9892	0.9896	0.9901	0.9904	0.9908	0.9911	0.9914
3.8583	3.8953	3.9306	3.9643	3.9965	4.0274	4.0570	4.0855

Exercice : tiges tournées

Dans un atelier mécanique, l'ingénieur vérifie le diamètre de tiges tournées sur un tour automatique. Le diamètre des tiges peut fluctuer selon le réglage du tour. Vingt tiges prélévées au hasard ont été mesurées avec un micromètre de précision. Les résultats sont exposés ci-dessous (en mm) :
39.5 40.6 38.4 37.8 39.4 39.9 41.5 40.0 38.5 41.2 39.7 39.1 42.6 40.0 38.4 40.0 39.4 41.1 41.3 40.8
On admet que le diamètre des tiges est distribué selon une loi normale.

1. Déterminer la moyenne empirique, la variance empirique, l'étendue de l'échantillon ainsi que les quartiles empiriques.
2. Proposer une estimation des paramètres de la loi normale.
3. Estimer le diamètre moyen des tiges par intervalle de confiance avec le niveau de confiance 95%.
4. Recommencer en prenant un niveau de confiance de 99%. Que constatez-vous ?
5. Estimer l'écart-type du diamètre des tiges par intervalle de confiance avec le niveau de confiance 95%.

Exercice : résistance à l'éclatement

Un laboratoire indépendant est chargé par l'Office de protection des consommateurs de vérifier la résistance à l'éclatement (en kg/cm^2) d'un réservoir de carburant d'un certain fabricant. On admet que la résistance à l'éclatement est distribuée selon une loi normale.

1. Des essais sont effectués sur un échantillon de 10 réservoirs conduisant à une résistance moyenne à l'éclatement de 219 kg/cm^2 avec un écart-type empirique de $10kg/cm^2$. Proposer un intervalle de confiance pour la résistance moyenne à l'éclatement de ce type de réservoir avec un niveau de confiance de 95%.
2. Des essais sont effectués sur un échantillon de 100 réservoirs conduisant à une résistance moyenne à l'éclatement de 219 kg/cm^2 avec un écart-type empirique de $10kg/cm^2$. Proposer un intervalle de confiance pour la résistance moyenne à l'éclatement de ce type de réservoir avec un niveau de confiance de 95%. Que constatez-vous ?

Estimation du paramètre p d'une loi de Bernoulli

- Soit $(X_1, ..., X_n)$ un échantillon de loi $\mathcal{B}(p)$.
- Un estimateur ponctuel de p est $\widehat{p} = \overline{X}$.
- L'IC bilatéral pour p au niveau de confiance $(1-\alpha)$ est approché par $\mathbb{P}[\widehat{p}_{\text{inf}} \leq p \leq \widehat{p}_{\text{sup}}] = 1-\alpha$ avec

$$\widehat{p}_{\text{inf}} = \widehat{p} - F_N^{-1}(1-\alpha/2)\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}$$

$$\widehat{p}_{\text{sup}} = \widehat{p} + F_N^{-1}(1-\alpha/2)\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}$$

ou par $\mathbb{P}[\widehat{p}'_{\text{inf}} \leq p \leq \widehat{p}'_{\text{sup}}] = 1-\alpha$ avec

$$\widehat{p}'_{\text{inf}} = \frac{\widehat{p}}{\widehat{p} + (1-\widehat{p}) \times \exp\left(\frac{F_N^{-1}(1-\alpha/2)}{\sqrt{n\widehat{p}(1-\widehat{p})}}\right)}$$

$$\widehat{p}'_{\text{sup}} = \frac{\widehat{p}}{\widehat{p} + (1-\widehat{p}) \times \exp\left(-\frac{F_N^{-1}(1-\alpha/2)}{\sqrt{n\widehat{p}(1-\widehat{p})}}\right)}$$

- L'IC $[\widehat{p}_{\text{inf}}, \widehat{p}_{\text{inf}}]$ est plus simple à calculer mais l'IC $[\widehat{p}'_{\text{inf}}, \widehat{p}'_{\text{inf}}]$ est meilleur.
- **La probabilité que l'IC contienne le paramètre p est d'environ $(1-\alpha)$.**

Exercice : tubes de verre

L'entreprise Simtech produit des tubes de verres de haute qualité qu'elle vend à l'entreprise Gescom sous forme de lots de 100 tubes de verre. L'ingénieur d'usine de l'entreprise Simtech s'intéresse à la proportion de tubes défectueux issus de cette production. Il dispose des données suivantes obtenues sur 200 lots :

nombre de tubes défectueux	0	1	2	3	4	5
fréquence	98	60	22	16	2	2

1. Estimer la proportion du nombre de tubes défectueux dans l'ensemble de la production.
2. Estimer par intervalle la proportion de tubes défectueux dans l'ensemble de la production au niveau de confiance 95%.
3. Le qualiticien de l'entreprise Gescom utilise le plan de contrôle suivant à la réception de chaque lot. Prélever au hasard 5 tubes de verres. S'il y a, dans cet échantillon, 1 tube de verre (ou plus) défectueux, le lot est refusé et retourné à l'entreprise Simtech sans plus d'inspection. Estimer la probabilité qu'un lot soit refusé avec ce plan de contrôle.

Estimation du paramètre λ d'une loi de Poisson

- Soit $(X_1, ..., X_n)$ un échantillon de loi $\mathcal{P}(\lambda)$.
- Un estimateur ponctuel de λ est $\widehat{\lambda} = \overline{X}$.
- L'IC bilatéral pour λ au niveau de confiance $(1 - \alpha)$ est donné par :

$$\mathbb{P}[\widehat{\lambda}'_{\text{inf}} \leq \lambda \leq \widehat{\lambda}'_{\text{sup}}] = 1 - \alpha$$

avec

$$\widehat{\lambda}'_{\text{inf}} = \frac{1}{2n} F^{-1}_{\chi^2(2\sum_{i=1}^{n} x_i)}(\alpha/2)$$

$$\widehat{\lambda}'_{\text{sup}} = \frac{1}{2n} F^{-1}_{\chi^2(2(1+\sum_{i=1}^{n} x_i))}(1 - \alpha/2)$$

- L'IC bilatéral pour λ au niveau de confiance $(1 - \alpha)$ peut également être approché par :

$$\mathbb{P}[\widehat{\lambda}_{\text{inf}} \leq \lambda \leq \widehat{\lambda}_{\text{sup}}] = 1 - \alpha$$

avec

$$\widehat{\lambda}_{\text{inf}} = \widehat{\lambda} - F_N^{-1}(1 - \alpha/2) \frac{\sqrt{\widehat{\lambda}}}{\sqrt{n}}$$

$$\widehat{\lambda}_{\text{sup}} = \widehat{\lambda} + F_N^{-1}(1 - \alpha/2) \frac{\sqrt{\widehat{\lambda}}}{\sqrt{n}}$$

Exercice : urgences hospitalières

Le gérant d'un hôpital étudie l'arrivée des patients dans le service d'urgence en pédiatrie ouvert 24h sur 24h. Il compte le nombre de patients se présentant aux urgences en pédiatrie par heure, et ce, pendant 24h. Il obtient les résultats suivants :
9 6 3 5 4 5 4 6 3 2 2 10 7 8 6 4 4 6 8 6 3 4 1 9
On admet que le nombre de patients se présentant aux urgences en pédiatrie par heure suit une loi de Poisson.

1. Proposer une estimation ponctuelle du paramètre de la loi de Poisson.
2. Proposer une estimation par intervalle du paramètre de la loi de Poisson au niveau de confiance 95%.
3. Estimer la probabilité qu'au cours d'une heure plus de 6 patients se présentent aux urgences de pédiatrie.
4. Estimer la probabilité qu'au cours d'une heure un nombre inférieur ou égal à 4 de patients se présentent aux urgences de pédiatrie.

Estimation du paramètre λ d'une loi exponentielle

- Soit $(X_1, ..., X_n)$ un échantillon de loi $\mathcal{E}(\lambda)$.
- Un estimateur ponctuel de λ est $\widehat{\lambda} = \frac{n-1}{n\overline{X}}$.
- L'IC bilatéral pour λ au niveau de confiance $(1-\alpha)$ est donné par :

$$\mathbb{P}[\widehat{\lambda}_{\text{inf}} \leq \lambda \leq \widehat{\lambda}_{\text{sup}}] = 1 - \alpha$$

avec

$$\widehat{\lambda}_{\text{inf}} = \frac{\widehat{\lambda}}{2(n-1)} F^{-1}_{\chi^2(2n)}(\alpha/2)$$

$$\widehat{\lambda}_{\text{sup}} = \frac{\widehat{\lambda}}{2(n-1)} F^{-1}_{\chi^2(2n)}(1-\alpha/2)$$

- **Il y a une probabilité $(1-\alpha)$ pour que l'IC contienne le paramètre λ.**

Exercice : fiabilité

Dans le cadre du suivi de production, l'ingénieur d'usine de l'entreprise Electrotek étudie la durée de vie d'un composant électronique. Le service de maintenance lui a fourni les données suivantes (en jours) :
493 353 6208 2234 4108 205 1442 2322 1939 8308 587 1182 3987 5496 391
On admet que la durée de vie (en jours) des composants électroniques est distribuée selon une loi exponentielle.

1. Estimer la durée de vie moyenne du composant électronique.
2. Proposer une estimation ponctuelle du paramètre de la loi exponentielle.
3. Proposer une estimation par intervalle du paramètre de la loi exponentielle au niveau de confiance 95%.
4. Estimer la probabilité que la durée de vie du composant soit inférieure à 1000 jours puis estimer la probabilité que la durée de vie du composant soit supérieure à 10000 jours.