Université d'Evry-Val d'Essonne 2019-2020
L2 Mathématiques / DL2 Mathématiques Economie
L2 Informatique / DL2 Biologie Informatique
Statistiques

# Corrigé de la feuille de TD 4 : Estimation par intervalle de confiance

**Exercice 1.**

Pour déterminer la teneur en potassium d'une solution, on effectue des dosages à l'aide d'une technique expérimentale donnée.
On admet que le résultat d'un dosage est une variable aléatoire suivant une loi gaussienne $N(\mu, \sigma^2)$ dont l'espérance $\mu$ est la valeur que l'on cherche à déterminer, et dont l'écart-type $\sigma$ est de 1 mg/litre si l'on suppose que le protocole expérimental a été suivi scrupuleusement.
Les résultats pour cinq dosages indépendants réalisés en suivant rigoureusement le protocole expérimental sont les suivants (en mg/litre) : 74.0, 71.6, 73.4, 74.3, 72.2.

1. Déterminer à partir de ces mesures un intervalle de confiance pour $\mu$ de niveau de confiance 95% et calculer l'intervalle observé.
**Outils :** CM Chap.IV, §1 "Intervalle de confiance pour l'espérance d'un $n$-échantillon gaussien", §§ 1.1."Cas où la variance est connue.
**Corrigé :** Soit $X$ le résultat d'un dosage en mg/litre. Alors $X \sim \mathcal{N}(\mu, 1)$.
Soient $X_1, \ldots, X_n$ i.i.d. de loi de $X$. L'espérance $\mu$ est naturellement estimée par la moyenne empirique

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i \ ,$$

estimateur sans biais et consistant. Comme moyenne de $n$ variables aléatoires indépendantes de loi $\mathcal{N}(\mu, 1)$, la moyenne empirique suit la loi gaussienne. Plus précisément :

$$\bar{X}_n \sim \mathcal{N}(\mu, \frac{1}{n}) \quad \text{et donc} \quad \sqrt{n}(\bar{X}_n - \mu) \sim \mathcal{N}(0, 1). \tag{1}$$

On sait que si $Z \sim \mathcal{N}(0, 1)$, alors, lu dans les tables de $\mathcal{N}(0, 1)$

$$\mathbb{P}(Z \leq 1,96) = 0,975$$

et donc, par symétrie de la densité de $\mathcal{N}(0, 1)$ par rapport à l'axe des ordonnées,

$$\mathbb{P}(-1,96 \leq Z \leq 1,96) = 0,95.$$

Ce qui donne, en utilisant (1)

$$\mathbb{P}(-1,96 \leq \sqrt{n}(\bar{X}_n - \mu) \leq 1,96) = 0,95.$$

En résolvant la double inégalité on obtient :

$$\mathbb{P}(\bar{X}_n - \frac{1,96}{\sqrt{n}} \leq \mu \leq \bar{X}_n + \frac{1,96}{\sqrt{n}}) = 0,95$$

et donc l'intervalle de confiance pour $\mu$ de niveau de confiance 95% est

$$IC_{95\%}(\mu) = [\bar{X}_n - \frac{1,96}{\sqrt{n}}; \bar{X}_n + \frac{1,96}{\sqrt{n}}].$$

Pour trouver l'intervalle de confiance observée on calcule la valeur de la moyenne empirique sur l'échantillon donnée :

$$\bar{x} = \frac{1}{5}(74.0 + 71.6 + 73.4 + 74.3 + 72.2 + 73,1)$$

où on a utilisé $n = 5$. On obtient :

$$IC_{95\%,obs}(\mu) = [72,22346; 73,97654].$$

2. Quelle taille d'échantillon est nécessaire pour avoir au même niveau de confiance un intervalle de longueur inférieure à 0.1 mg/litre ?
**Corrigé :** Notons $l$ la longueur de l'intervalle de confiance. On a

$$l = \frac{2 \times 1,96}{\sqrt{n}}.$$

On cherche $n$ tel que $l \leq 0,1$, i.e. $n$ tel que

$$\frac{2 \times 1,96}{\sqrt{n}} \leq 0,1.$$

D'ou

$$n \geq (2 \times 1,96/0,1)^2 = 1536,64.$$

C'est donc à partir de $n = 1537$ que la longueur $l$ sera plus petite que $0,1$.

**Exercice 2.**

**Suite de l'exercice 1 du TD 3.**

Lors d'un sondage effectué en Ile de France, auprès de 550 personnes, il est apparu que 42 avaient de l'asthme. On se propose d'estimer par intervalle de confiance la probabilité $p$ d'avoir de l'asthme en Ile de France. On note $Z_i$ la variable aléatoire qui vaut 1 si la $i$-ème personne de l'échantillon sondé est atteinte et 0 sinon. On admet que les variables $Z_1, \cdots, Z_n$ sont indépendantes et de même loi. On note

$\overline{Z}_n = \frac{1}{n}\sum_{i=1}^{n} Z_i$. Calculer des intervalles de confiance pour $p$, basés sur l'approximation gaussienne, de coefficients de sécurité asymptotiques 90%, 95% puis 99%. Calculer les intervalles observés.

**Outils :** CM Chap. IV, §3 "Intervalle de confiance pour une probabilité dans le cas d'un grand échantillon."

**Corrigé :** Les v.a. $Z_1, \cdots, Z_n$ sont i.i.d. de loi de Bernoulli de paramètre $p$. On a aussi

$$\mathbb{E}Z_i = p \ , \quad Var(Z_i) = p(1-p).$$

L'estimation de $p$ par intervalle de confiance c'est donc un cas particulier de l'estimation de l'espérance d'un grand échantillon par intervalle de confiance. Par le Théorème Central Limite , on a la convergence en loi suivante :

$$\sqrt{n}\frac{\bar{Z}_n - p}{\sqrt{p(1-p)}} \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0,1).$$

La variance $Var(Z_i) = p(1-p)$ est inconnue. En utilisant le fait que $\bar{Z}_n$ est un estimateur consistant de $p$, on obtient que $\bar{Z}_n(1-\bar{Z}_n)$ est un estimateur consistant de la variance , d'où on peut déduire que

$$\sqrt{n}\frac{\bar{Z}_n - p}{\sqrt{\bar{Z}_n(1-\bar{Z}n)}} \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0,1).$$

Cette convergence en loi signifie précisément que $\forall a \leq b$,

$$\mathbb{P}(a \leq \sqrt{n}\frac{\bar{Z}_n - p}{\sqrt{\bar{Z}_n(1-\bar{Z}_n)}} \leq b) \xrightarrow[n\to\infty]{} \mathbb{P}(a \leq Z \leq b),$$

où $Z \sim \mathcal{N}(0,1)$.

Pour $n$ grand, et à partir de $n = 30$ on peut utiliser l'approximation :

$$\mathbb{P}\left(a \leq \sqrt{n}\frac{\bar{Z}_n - p}{\sqrt{\bar{Z}_n(1-\bar{Z}_n)}} \leq b\right) \approx \mathbb{P}(a \leq Z \leq b).$$

Dans les tables de la loi $\mathcal{N}(0,1)$ on choisit la valeur $t_\alpha$ telle que

$$\mathbb{P}(-t_\alpha \leq Z \leq t_\alpha) = 1-\alpha.$$

C'est à dire, $\mathbb{P}(Z \leq t_\alpha) = 1 - \frac{\alpha}{2}$. $t_\alpha$ est le quantile d'ordre $1-\frac{\alpha}{2}$ de la loi $\mathcal{N}(0,1)$. On obtient, en résolvant la double inégalité par rapport à $p$ :

$$\mathbb{P}\left(\bar{Z}_n - t_\alpha\sqrt{\frac{\bar{Z}_n(1-\bar{Z}_n)}{n}} \leq p \leq \bar{Z}_n + t_\alpha\sqrt{\frac{\bar{Z}_n(1-\bar{Z}_n)}{n}}\right) \approx 1-\alpha$$

1. $\alpha = 10\%, \quad t_\alpha = 1,645 \quad IC_{10\%,obs}(p) = [0,0577; 0,09499]$
2. $\alpha = 5\%, \quad t_\alpha = 1,96, \quad IC_{5\%,obs}(p) = [0,0542; 0,0985]$
3. $\alpha = 1\%, \quad t_\alpha = 2,575, \quad IC_{1\%,obs}(p) = [0,0472; 0,1055]$

**Exercice 3.**

En conduite, on sait que le temps de réaction (t.r.) est aléatoire et est lié à l'état du conducteur. On suppose que pour un conducteur dans un état normal (non atteint de maladie pouvant modifier son t.r., sous l'emprise d'aucun produit de type alcool, drogue, médicaments, ...), le temps de réaction mesuré en secondes est une variable aléatoire $X$ d'espérance $\mu$ et de variance $\sigma^2 > 0$. On considère un test de conduite où on met le conducteur en situation de danger imprévisible et on observe son temps de réaction. On fait passer le test à $n = 307$ conducteurs choisis aléatoirement et dans un état normal et le temps de réaction moyen pour l'échantillon choisi est $\bar{x}_n = 1.05\, s$.

1. On suppose tout d'abord que $X$ est une v.a. de loi gaussienne de variance $\sigma^2 = 0.2 : X \sim \mathcal{N}(\mu, 0, 2)$.
   (a) Donner un intervalle de confiance pour le t.r. moyen $\mu$, de niveau de confiance 95% et calculer l'intervalle observé.
   **Outils** CM Chap IV, §1, §§1.1 Intervalle de confiance pour l'espérance d'une loi gaussiènne, cas ou la variance est connue.
   **Corrigé** Soit $X_1, \ldots, X_n$ un $n$-échantillon de loi de $X$. Puisque $X \sim \mathcal{N}(\mu, 0, 2)$, on a $\bar{X}_n \sim \mathcal{N}(\mu; \frac{0,2}{n})$ et

$$\sqrt{n}\frac{(\bar{X}_n - \mu)}{\sqrt{0,2}} \sim \mathcal{N}(0,1).$$

   Dans les tables de la loi $\mathcal{N}(0,1)$ on choisit la valeur $t_\alpha$ telle que

$$\mathbb{P}(-t_\alpha \leq Z \leq t_\alpha) = 1 - \alpha.$$

   C'est à dire, $\mathbb{P}(Z \leq t_\alpha) = 1 - \frac{\alpha}{2}$. $t_\alpha$ est le quantile d'ordre $1 - \frac{\alpha}{2}$ de la loi $\mathcal{N}(0,1)$. Ici $\alpha = 0,05$ et $t_\alpha = 1,96$. On obtient,

$$\mathbb{P}(-1,96 \leq \sqrt{n}\frac{(\bar{X}_n - \mu)}{\sqrt{0,2}} \leq 1,96) = 0,95$$

   et en résolvant la double inégalité par rapport à $\mu$ :

$$IC_{95\%}(\mu) = [\bar{X}_n - 1,96\sqrt{\frac{0,2}{n}}; \bar{X}_n + 1,96\sqrt{\frac{0,2}{n}}].$$

   Pour calculer l'intervalle de confiance observée, on utilise $\bar{x}_{obs} = 1,05$ et $n = 307$.

$$IC_{95\%,obs} = [0,99998; 1,10003].$$

(b) Je conduis par beau temps sur une autoroute à $130\,km/h$. La distance parcourue pendant le temps de réaction est appelée distance de réaction. Donner un intervalle de confiance (et l'intervalle observé) pour la distance de réaction moyenne, de niveau de confiance 95%.
**Corrigé :** Notons $D$ distance de réaction. On a

$$D = X \times 130/3600, \quad E(D) = \mu \times 130/3600.$$

A partir de l'intervalle de confiance pour $\mu$ on trouvera celui pour $D$ :

$$\mathbb{P}\left(\bar{X}_n - 1,96\sqrt{\frac{0,2}{n}} \leq \mu \leq \bar{X}_n + 1,96\sqrt{\frac{0,2}{n}}\right);$$

et donc

$$\mathbb{P}\left(13/360 \times (\bar{X}_n - 1,96\sqrt{\frac{0,2}{n}}) \leq \mathbb{E}(D) \leq 13/360 \times (\bar{X}_n + 1,96\sqrt{\frac{0,2}{n}})\right).$$

Finalement l'intervalle de confiance observée pour $\mathbb{E}D$ est en km : $[0,03611; 0,03972]$ et en m : $[36,11; 39,72]$.

2. On ne suppose plus à présent que $X$ suit une loi gaussienne et que $\sigma^2$ est connu.
(a) Comment peut-on estimer $\sigma^2$ ? Quelle sont les propriétés de cet estimateur ?
**Corrigé :** On estime $\sigma^2$ par la variance empirique :

$$\hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2.$$

Remarque : C'est un estimateur consistant mais biaisé de $\sigma^2$. Pour avoir un estimateur non-biaisé, il faut normaliser par $n-1$. Mais comme dans cette exercice par la suite on va faire tendre $n \to \infty$ et utiliser le TCL, cela est équivalent. Aussi on a

$$\hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i)^2 - (\bar{X}_n)^2.$$

(b) On a estimé $\sigma^2$ à partir de l'échantillon considéré et on trouve comme estimation $\hat{\sigma}^2_{n,obs} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2 = 0.23$. Peut-on calculer des intervalles de confiance comme dans la question 1 ? Si oui, faites le.
**Outils :** CM Chap IV, § 2. Intervalle de confiance pour l'espérance d'une loi quelconque. Grand échantillon. §§2.2. Cas ou la variance est inconnue.
**Corrigé :** Par le Théorème Central Limite , on a la convergence en loi suivante :

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

En utilisant le fait que $\hat{\sigma}_n^2$ est un estimateur consistant de $\sigma^2$, on peut montrer que

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

On aura donc

$$\mathbb{P}\left(\bar{X}_n - 1,96\sqrt{\frac{\hat{\sigma}_n}{n}} \leq \mu \leq \bar{X}_n + 1,96\sqrt{\frac{\hat{\sigma}_n}{n}}\right) \approx 0,95.$$

$$IC_{95\%,obs}(\mu) = [0,99635; 110365].$$

On a aussi l'intervalle observé pour la moyenne de la distance $\mathbb{E}D$

$$\mathbb{P}\left(13/360 \times (\bar{X}_n - 1,96\sqrt{\frac{\hat{\sigma}_n}{n}}) \leq \mathbb{E}D \leq 13/360 \times (\bar{X}_n + 1,96\sqrt{\frac{\hat{\sigma}_n}{n}})\right) \approx 0,95.$$

$$IC_{95\%,obs}(\mathbb{E}(D)) = [35,98; 39,85].$$

(c) Peut-on affirmer que les niveaux de confiance sont exactement de 95% ?
**Corrigé :** Non, car il s'agit des intervalles asymptotiques : par le TCL on a seulement

$$\mathbb{P}(\bar{X}_n - 1,96\sqrt{\frac{\hat{\sigma}_n}{n}} \leq \mu \leq \bar{X}_n + 1,96\sqrt{\frac{\hat{\sigma}_n}{n}}) \xrightarrow[n\to\infty]{} 0,95.$$

ce qui donne $\approx 0,95$ et non $= 0,95$ pour les deux intervalles de confiance.

**Exercice 4 Suite de l'exercice 2 du TD 3.**

On suppose que le nombre $X$ de clients téléphonant à un central téléphonique chaque jour est une variable aléatoire de loi de Poisson de paramètre $\lambda$, avec $\lambda > 0$. Pendant $n = 100$ jours, on a compté le nombre $x_i$ de clients ayant appelé le jour $i$. Sur ces 100 jours, le nombre moyen d'appels par jour obtenu est de 2,89. On considère que chaque $x_i$ est la réalisation d'une variable aléatoire $X_i$ de loi de Poisson de paramètre $\lambda$, les $X_i$ étant supposées indépendantes et identiquement distribuées.

1. Déterminer un intervalle de confiance pour $\lambda$ de niveau de confiance approximatif $1 - \alpha$.
**Outil :** CM Chap IV §2, §§2.2 Intervalle de confiance pour l'espérance d'une loi quelconque. Grand échantillon. Variance inconnue.
**Corrigé :** On note $X_i$ nombre de clients téléphonant à le centrale le jour $i$. On a $X_1, \ldots, X_n$ i.i.d. de loi $\mathcal{P}(\lambda)$. $n = 100$ On a $\mathbb{E}X_i = \lambda$, $Var(X_i) = \lambda$. Par le TCL

$$\sqrt{n}\frac{(\bar{X}_n - \lambda)}{\lambda} \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0,1).$$

Comme $\bar{X}_n$ est un estimateur consistant de $\lambda$, on a aussi

$$\sqrt{n}\frac{(\bar{X}_n - \lambda)}{\bar{X}_n} \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0,1).$$

On cherche dans les tables de $\mathcal{N}(0,1)$ la valeur $t_\alpha$ telle que

$$\mathbb{P}(-t_\alpha \leq Z \leq t_\alpha) = 1 - \alpha$$

ou $Z \sim \mathcal{N}(0,1)$. On a

$$\mathbb{P}(-t_\alpha \leq \sqrt{n}\frac{(\bar{X}_n - \lambda)}{\bar{X}_n} \leq t_\alpha) \xrightarrow[n\to\infty]{\mathcal{L}} \mathbb{P}(-t_\alpha \leq Z \leq t_\alpha)$$

et donc pour $n$ grand

$$\mathbb{P}(-t_\alpha \leq \sqrt{n}\frac{(\bar{X}_n - \lambda)}{\bar{X}_n} \leq t_\alpha) \approx 1 - \alpha.$$

D'ou

$$\mathbb{P}(\bar{X}_n - t_\alpha\sqrt{\frac{\bar{X}_n}{n}} \leq \lambda \leq \bar{X}_n + t_\alpha\sqrt{\frac{\bar{X}_n}{n}}) \approx 1 - \alpha.$$

2. Donner les intervalles de confiance observés pour $\lambda$ calculés aux niveaux de confiance approximatifs 90%, 95% et 99%. Commenter.
   Pour calculer l'IC observé on trouve :
   (a) $\alpha = 10\%, \quad t_\alpha = 1,645 \quad IC_{10\%,obs}(\lambda) = [2,61035; 3,16465]$
   (b) $\alpha = 5\%, \quad t_\alpha = 1,96, \quad IC_{5\%,obs}(\lambda) = [2,5568; 3,2232]$
   (c) $\alpha = 1\%, \quad t_\alpha = 2,575, \quad IC_{1\%,obs}(\lambda) = [2,45225; 3,32775]$
3. Comment peut-on obtenir des intervalles de confiance de longueurs plus petites en gardant les mêmes niveaux de confiance ?
   Augmenter le $n$.

**Exercice 5** Pour déterminer la concentration en glucose d'un échantillon sanguin, on effectue des dosages à l'aide d'une technique expérimentale donnée. On considère que le résultat de chaque dosage est une variable aléatoire de loi gaussienne. On effectue 10 dosages indépendants, qui donnent les résultats suivants (en g/l) :

$$0.96, 1.04, 1.08, 0.92, 1.04, 1.18, 0.99, 0.99, 1.25, 1.08$$

1. Calculer une estimation de la concentration en glucose de cet échantillon sanguin.
   **Corrigé** Soit $X_i$ résultat du dosage $i$. $X_i \sim \mathcal{N}(m, \sigma^2)$, $X_1, \ldots, X_n$ i.i.d. et $n = 10$. On estime $m$ par$\bar{X}_n$ et $\bar{x}_{obs} = 1,053$.

2. Calculer un intervalle de confiance de cette concentration de niveau de confiance 95%.
**Outil** CM Chap IV § 1, §§1.2 : Intervalle de confiance pour l'espérance de la loi gaussienne, $n$ quelconque, variance inconnue.
**Corrigé** On estime la variance inconnue $\sigma^2$ par l'estimateur sans biais

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

On considère la statistique

$$\sqrt{n}\frac{\bar{X}_n - m}{S_n}.$$

Cette statistique ne suit pas la loi gaussienne. Sa loi est appelée la loi de Student à $n-1$ degré de libertés. On note

$$\sqrt{n}\frac{\bar{X}_n - m}{S_n} \sim St(n-1).$$

On cherche dans les tables de $St(n-1)$ la valeur de $t_\alpha$ telle que

$$\mathbb{P}(-t_\alpha \leq T \leq t_\alpha) = 1 - \alpha$$

où $T \sim St(n-1)$.
On remplace $T$ par $\sqrt{n}\frac{\bar{X}_n - m}{S_n}$ et on en déduit

$$\mathbb{P}(\bar{X}_n - t_\alpha\frac{S}{\sqrt{n}} \leq m \leq \bar{X}_n - t_\alpha\frac{S}{\sqrt{n}}) = 1 - \alpha$$

Comme on a utilisé une loi exacte (loi de Student à $n-1$ degré de liberté, c'est une égalité exacte, et pas une égalité approximative, puisque la loi de la statistique $\sqrt{n}\frac{\bar{X}_n - m}{S_n}$ c'est la loi $St(n-1)$.
Pour $\alpha = 0,05$ et $n = 10$ on trouve $t_\alpha = 2,262$. On calcule $\sum_{i=1}^{10} x_i^2 = 11,1791$. Enfin en utilisant

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 = \frac{1}{n}\sum_{i=1}^{n}X_i^2 - (\bar{X}_n)^2,$$

et

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

on calcule

$$s_{obs} = \sqrt{1,12 - 1,053^2} = 0,095.$$