

Chapitre 3 – Estimations et tests d'hypothèse (1^{re} partie)

STT-1920 Méthodes statistiques

Département de mathématiques et de statistique



UNIVERSITÉ
LAVAL

Introduction à l'inférence statistique

Où en sommes-nous ?

- Au **chapitre 1**, nous avons appris à **décrire** un ensemble de données (moyenne, écart-type, médiane, histogrammes...).
- Au **chapitre 2**, nous avons étudié la **théorie des probabilités** en supposant les lois et leurs paramètres *parfaitement connus*.
 - *Exemple* : la masse des oranges suit une loi $\mathcal{N}(\mu = 204, \sigma^2 = 256)$.

Où en sommes-nous ?

- Au **chapitre 1**, nous avons appris à **décrire** un ensemble de données (moyenne, écart-type, médiane, histogrammes...).
- Au **chapitre 2**, nous avons étudié la **théorie des probabilités** en supposant les lois et leurs paramètres *parfaitement connus*.
 - *Exemple* : la masse des oranges suit une loi $\mathcal{N}(\mu = 204, \sigma^2 = 256)$.

Le problème en pratique

Dans la vraie vie, on ne connaît **jamaïs** la vraie distribution ni les vraies valeurs des paramètres μ , σ^2 ou p de la population !

⇒ C'est l'objet de l'**inférence statistique** : tirer des conclusions sur une **population** à partir d'un **échantillon**.

Démarche générale



1. On postule un **modèle probabiliste** adapté au phénomène (loi normale, binomiale, Poisson...).
2. Ce modèle comporte des **paramètres inconnus** (μ , σ^2 , p ...).
3. On recueille un **échantillon aléatoire** de taille n pour **estimer** ces paramètres.

Démarche générale



1. On postule un **modèle probabiliste** adapté au phénomène (loi normale, binomiale, Poisson...).
2. Ce modèle comporte des **paramètres inconnus** (μ , σ^2 , p ...).
3. On recueille un **échantillon aléatoire** de taille n pour **estimer** ces paramètres.

Deux grandes familles d'inférence

- **L'estimation** : ponctuelle (partie 1) et par intervalle de confiance (partie 2).
- **Les tests d'hypothèses** : décider face à une affirmation (partie 3).

Exemples motivants

- **Biologie cellulaire** : Le diamètre des cellules hépatiques suit une loi $\mathcal{N}(\mu, \sigma^2)$.
 - On mesure 20 cellules : comment estimer le diamètre moyen μ et la variabilité σ^2 de toutes les cellules de ce type ?
- **Épidémiologie animale** : Le nombre de veaux infectés parmi n veaux examinés suit une loi Binomiale(n, p).
 - Sur 80 veaux examinés, 12 présentent des symptômes. Quelle est la proportion réelle p d'infection dans le troupeau ?
- **Agronomie** : Deux engrais sont testés sur des parcelles. L'engrais A donne-t-il un rendement moyen supérieur à l'engrais B ?

Vocabulaire et définitions

Population vs Échantillon

Population (inaccessible)

- Ensemble complet de toutes les unités d'intérêt.
- Caractérisée par des **paramètres** fixes mais inconnus : μ , σ^2 , σ , p .

Échantillon (observé)

- Sous-ensemble de n unités tirées au hasard.
- Fournit des **statistiques** calculées sur les données : $\bar{x}$, s^2 , s , $\hat{p}$.



QUIZ!

La masse des souris adultes d'une lignée suit une loi $\mathcal{N}(\mu = 25 \text{ g}, \sigma^2 = 4 \text{ g}^2)$. Deux équipes de recherche pèsent chacune 16 souris choisies au hasard et calculent la moyenne de leur échantillon : $\bar{x}_A$ et $\bar{x}_B$.

Que peut-on dire des valeurs $\bar{x}_A$ et $\bar{x}_B$?

- (a) Elles sont toutes deux égales à 25 g.
- (b) Elles sont égales entre elles, mais pas nécessairement égales à 25 g.
- (c) Elles sont différentes l'une de l'autre, mais probablement « pas trop loin » de 25 g.

QUIZ!

La masse des souris adultes d'une lignée suit une loi $\mathcal{N}(\mu = 25 \text{ g}, \sigma^2 = 4 \text{ g}^2)$. Deux équipes de recherche pèsent chacune 16 souris choisies au hasard et calculent la moyenne de leur échantillon : $\bar{x}_A$ et $\bar{x}_B$.

Que peut-on dire des valeurs $\bar{x}_A$ et $\bar{x}_B$?

- (a) Elles sont toutes deux égales à 25 g.
- (b) Elles sont égales entre elles, mais pas nécessairement égales à 25 g.
- (c) Elles sont différentes l'une de l'autre, mais probablement « pas trop loin » de 25 g.

Réponse (c). La moyenne d'un échantillon **change d'un échantillon à l'autre** : c'est une **variable aléatoire**. Reste à préciser ce que veut dire « pas trop loin »...

Statistique et loi échantillonnale

Statistique

Quantité calculée **uniquement à partir de l'échantillon** (elle ne dépend d'aucun paramètre inconnu). Ex. : $\bar{X}$, S^2 , $\hat{P}$, la médiane, le maximum...

Loi échantillonnale

Une statistique dépend de l'échantillon tiré au hasard : c'est une **variable aléatoire**. Sa loi de probabilité s'appelle sa **loi échantillonnale**.

Statistique et loi échantillonnale

Statistique

Quantité calculée **uniquement à partir de l'échantillon** (elle ne dépend d'aucun paramètre inconnu). Ex. : $\bar{X}$, S^2 , $\hat{P}$, la médiane, le maximum...

Loi échantillonnale

Une statistique dépend de l'échantillon tiré au hasard : c'est une **variable aléatoire**. Sa loi de probabilité s'appelle sa **loi échantillonnale**.

⚠ $(\bar{X} - \mu)/(\sigma/\sqrt{n})$ n'est *pas* une statistique si μ et σ sont inconnus : on ne peut pas la calculer!

QUIZ!

On capture au hasard 50 truites dans un lac et on mesure leur longueur.
Chacune des quantités suivantes est-elle un **paramètre** ou une **statistique**?

- La longueur moyenne des 50 truites capturées (32,4 cm).
- La longueur moyenne de *toutes* les truites du lac.
- La proportion des 50 truites porteuses d'un parasite.
- L'écart-type des longueurs de toutes les truites du lac.

QUIZ!

On capture au hasard 50 truites dans un lac et on mesure leur longueur.
Chacune des quantités suivantes est-elle un **paramètre** ou une **statistique**?

- La longueur moyenne des 50 truites capturées (32,4 cm).
→ **Statistique** ($\bar{x}$) : elle est calculée sur l'échantillon.
- La longueur moyenne de *toutes* les truites du lac.
→ **Paramètre** (μ) : fixe, mais inconnu.
- La proportion des 50 truites porteuses d'un parasite.
→ **Statistique** ($\hat{p}$).
- L'écart-type des longueurs de toutes les truites du lac.
→ **Paramètre** (σ).

Variables aléatoires vs Observations

Distinction entre majuscules et minuscules

- **Avant l'expérience** : les mesures $X_1, X_2, \dots, X_n$ sont des **variables aléatoires**.
- **Après l'expérience** : les valeurs obtenues $x_1, x_2, \dots, x_n$ sont des **nombre réels** fixés.

Variables aléatoires vs Observations

Distinction entre majuscules et minuscules

- **Avant l'expérience** : les mesures $X_1, X_2, \dots, X_n$ sont des **variables aléatoires**.
- **Après l'expérience** : les valeurs obtenues $x_1, x_2, \dots, x_n$ sont des **nombre réels** fixés.

Estimateur vs Estimation ponctuelle

- **Estimateur** $T = g(X_1, \dots, X_n)$: **variable aléatoire** (formule générale).
- **Estimation** $t = g(x_1, \dots, x_n)$: **valeur numérique** sur l'échantillon observé.

⇒ Un estimateur a donc une **loi échantillonnale**, une **espérance** et une **variance** !

Tableau synthèse des notations

Paramètre (population)	Estimateur (v.a.)	Estimation (valeur)	Espérance $E[T]$	Erreur-type $\sigma_T = \sqrt{\text{Var}(T)}$
Moyenne μ	$\bar{X} = \frac{1}{n} \sum X_i$	$\bar{x} = \frac{1}{n} \sum x_i$	μ	$\frac{\sigma}{\sqrt{n}}$
Variance σ^2	$S^2 = \frac{1}{n-1} \sum (X_i - \bar{X})^2$	s^2	σ^2	$\sqrt{\frac{2\sigma^4}{n-1}} (*)$
Proportion p	$\hat{P} = \frac{Y}{n}$	$\hat{p} = \frac{y}{n}$	p	$\sqrt{\frac{p(1-p)}{n}}$

(*) Sous l'hypothèse de normalité de la population.

Erreur-type

L'**erreur-type** d'un estimateur est l'écart-type de sa loi échantillonnale : $ET(T) = \sqrt{\text{Var}(T)}$. Elle mesure la **précision** de l'estimateur.

Propriétés d'un bon estimateur

Qualités d'un bon estimateur

On souhaite qu'un estimateur T d'un paramètre θ possède deux qualités :

1. Être sans biais (juste)

En moyenne, il vise dans le mille :

$$E[T] = \theta \quad (\text{Biais} = 0)$$

2. Être précis (faible variance)

Les estimations varient très peu :

$\text{Var}(T)$ est petite

Qualités d'un bon estimateur

On souhaite qu'un estimateur T d'un paramètre θ possède deux qualités :

1. Être sans biais (juste)

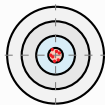
En moyenne, il vise dans le mille :

$$E[T] = \theta \quad (\text{Biais} = 0)$$

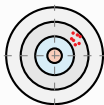
2. Être précis (faible variance)

Les estimations varient très peu :

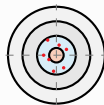
$\text{Var}(T)$ est petite



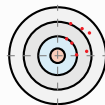
Sans biais et précis



Biaisé mais précis



Sans biais mais imprécis



Biaisé et imprécis

Chaque point = l'estimation obtenue avec un échantillon différent ; le centre = la vraie valeur θ .

Le biais : attention aux confusions !

Il faut distinguer plusieurs sens du mot « biais » :

- **Biais de sélection / d'échantillonnage :**

- L'échantillon n'est pas représentatif de la population cible (ex. : sondage téléphonique auprès des abonnés à un journal spécifique).
- C'est une erreur de **protocole** ou de collecte, impossible à corriger mathématiquement par la suite !

- **Biais mathématique de l'estimateur :**

- $\text{Biais}(T) = E[T] - \theta$.
- C'est une propriété de la formule mathématique elle-même.
- *Exemple* : $\frac{1}{n} \sum (X_i - \bar{X})^2$ est biaisé (sous-estime σ^2 en moyenne). On corrige ce biais en divisant par $n - 1$ (on y reviendra).

QUIZ!

Pour chaque situation : l'estimation est-elle biaisée? imprécise? De quel type de biais s'agit-il?

- Une balance mal étalonnée ajoute toujours environ 2 g à la masse des souris pesées, mais ses lectures sont très répétables.
- On estime la pression artérielle moyenne μ d'une population par $\bar{x}$ calculée sur $n = 3$ patients plutôt que $n = 300$.
- Pour estimer la prévalence du diabète au Québec, on ne recrute que des patients d'une clinique d'endocrinologie.

QUIZ!

Pour chaque situation : l'estimation est-elle biaisée? imprécise? De quel type de biais s'agit-il?

- Une balance mal étalonnée ajoute toujours environ 2 g à la masse des souris pesées, mais ses lectures sont très répétables.
→ **Biaisée mais précise** (erreur systématique de mesure).
- On estime la pression artérielle moyenne μ d'une population par $\bar{x}$ calculée sur $n = 3$ patients plutôt que $n = 300$.
→ **Sans biais mais imprécise** ($\bar{X}$ est sans biais, mais $\sigma / \sqrt{3}$ est grand).
- Pour estimer la prévalence du diabète au Québec, on ne recrute que des patients d'une clinique d'endocrinologie.
→ **Biais de sélection** : aucun calcul ne pourra le corriger!

Comment agir sur la précision ?

L'erreur-type mesure l'imprécision d'une estimation. Pour la réduire, on dispose de deux leviers principaux :

1. Choisir un estimateur de faible variance :

- Entre deux estimateurs sans biais, on préfère toujours celui qui a la plus petite variance (estimateur le plus efficace).

2. Augmenter la taille de l'échantillon n :

- Pour la moyenne, $ET(\bar{X}) = \frac{\sigma}{\sqrt{n}}$.
- Pour doubler la précision (diviser l'erreur-type par 2), il faut quadrupler la taille de l'échantillon ($4n$)!

Biais ou précision ? L'erreur quadratique moyenne

Que faire si un estimateur est **légèrement biaisé** mais **beaucoup plus précis** qu'un autre ?

Erreur quadratique moyenne (EQM)

$$\text{EQM}(T) = E[(T - \theta)^2] = \text{Var}(T) + [\text{Biais}(T)]^2$$

Biais ou précision ? L'erreur quadratique moyenne

Que faire si un estimateur est **légèrement biaisé** mais **beaucoup plus précis** qu'un autre ?

Erreur quadratique moyenne (EQM)

$$\text{EQM}(T) = E[(T - \theta)^2] = \text{Var}(T) + [\text{Biais}(T)]^2$$

- L'EQM mesure l'écart moyen (au carré) entre l'estimateur et la cible : elle **combine les deux qualités**.
- Pour un estimateur sans biais : $\text{EQM}(T) = \text{Var}(T)$.
- Entre deux estimateurs, on préfère celui qui a la **plus petite EQM**.
- Un exemple surprenant suivra avec l'estimation de σ^2 .

Estimation de la moyenne μ

Exemple 1 : Pression artérielle systolique

On suppose que la pression artérielle systolique (en mm Hg) d'une population d'adultes sains suit une loi normale $\mathcal{N}(\mu, \sigma^2)$.

On mesure la pression chez $n = 7$ individus choisis au hasard :

141 130 135 138 137 122 135

Exemple 1 : Pression artérielle systolique

On suppose que la pression artérielle systolique (en mm Hg) d'une population d'adultes sains suit une loi normale $\mathcal{N}(\mu, \sigma^2)$.

On mesure la pression chez $n = 7$ individus choisis au hasard :

141 130 135 138 137 122 135

Estimation ponctuelle de la moyenne

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{141 + 130 + 135 + 138 + 137 + 122 + 135}{7} = \frac{938}{7} = 134 \text{ mm Hg}$$

Exemple 1 : Pression artérielle systolique

On suppose que la pression artérielle systolique (en mm Hg) d'une population d'adultes sains suit une loi normale $\mathcal{N}(\mu, \sigma^2)$.

On mesure la pression chez $n = 7$ individus choisis au hasard :

141 130 135 138 137 122 135

Estimation ponctuelle de la moyenne

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{141 + 130 + 135 + 138 + 137 + 122 + 135}{7} = \frac{938}{7} = 134 \text{ mm Hg}$$

Question : Si on avait choisi 7 autres individus, aurait-on obtenu exactement 134? **Non!** La moyenne observée varie d'un échantillon à l'autre.

Variabilité d'échantillonnage : une simulation

Supposons que la vraie population suit $\mathcal{N}(\mu = 133, \sigma^2 = 62)$. Si on tirait 7 échantillons indépendants de taille $n = 7$:

Éch.	x_1	x_2	x_3	x_4	x_5	x_6	x_7	$\bar{x}$
1	141	130	135	138	137	122	135	134,0
2	132	128	139	133	132	127	138	132,7
3	140	125	122	126	136	128	131	129,7
4	131	137	136	130	140	134	124	133,1
5	143	125	131	121	139	150	131	134,3
6	132	140	128	136	141	135	140	136,0
7	135	143	134	133	138	123	134	134,3

Variabilité d'échantillonnage : une simulation

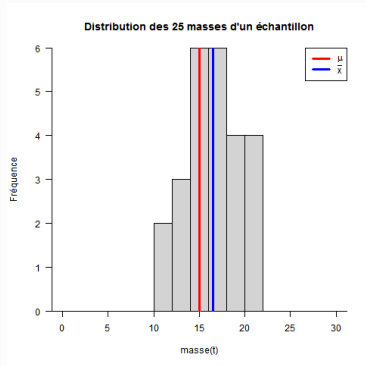
Supposons que la vraie population suit $\mathcal{N}(\mu = 133, \sigma^2 = 62)$. Si on tirait 7 échantillons indépendants de taille $n = 7$:

Éch.	x_1	x_2	x_3	x_4	x_5	x_6	x_7	$\bar{x}$
1	141	130	135	138	137	122	135	134,0
2	132	128	139	133	132	127	138	132,7
3	140	125	122	126	136	128	131	129,7
4	131	137	136	130	140	134	124	133,1
5	143	125	131	121	139	150	131	134,3
6	132	140	128	136	141	135	140	136,0
7	135	143	134	133	138	123	134	134,3

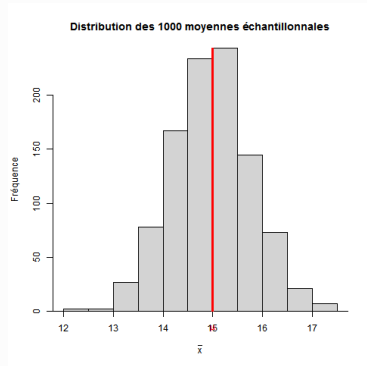
⇒ Les moyennes s'éparpillent autour de la vraie moyenne ($\mu = 133$), mais beaucoup moins que les observations individuelles.

Simulation : 1000 échantillons de taille 25

Population de moyenne $\mu = 15$ et de variance $\sigma^2 = 16$ ($\sigma = 4$).



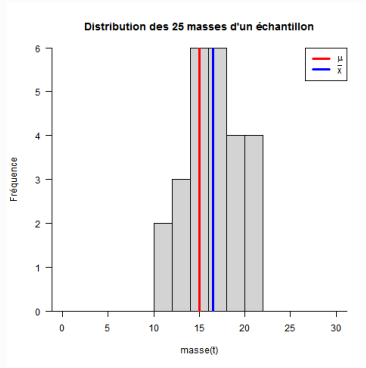
Un échantillon ($n = 25$) : $\bar{x}$ (bleu) $\neq \mu$ (rouge).



Les $\bar{x}$ de 1000 échantillons de taille 25.

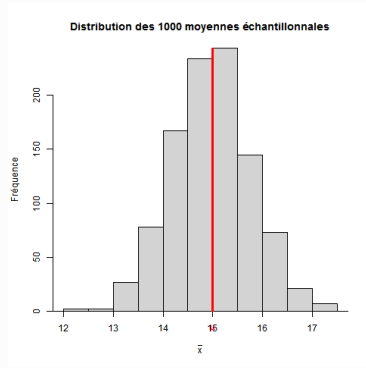
Simulation : 1000 échantillons de taille 25

Population de moyenne $\mu = 15$ et de variance $\sigma^2 = 16$ ($\sigma = 4$).



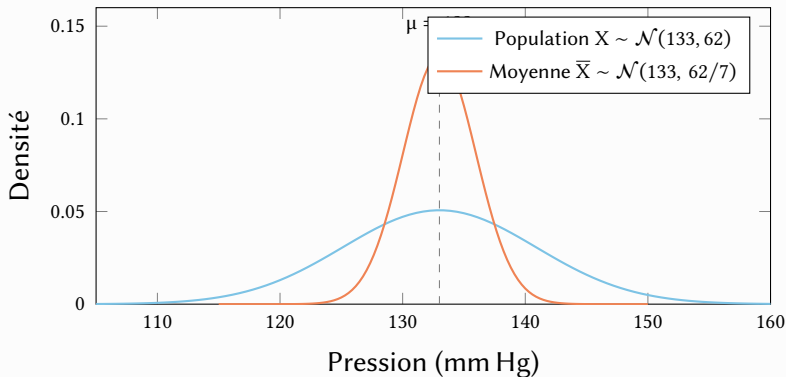
Un échantillon ($n = 25$) : $\bar{x}$ (bleu) $\neq \mu$ (rouge).

Les 1000 moyennes sont **centrées sur $\mu = 15$** et beaucoup moins dispersées que les données : leur écart-type est d'environ $\sigma / \sqrt{25} = 0,8$.



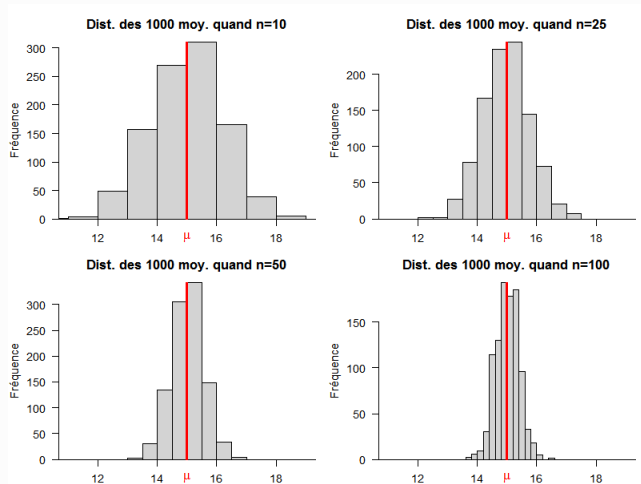
Les $\bar{x}$ de 1000 échantillons de taille 25.

Distribution de la population vs distribution de $\bar{X}$



Constat majeur : $\bar{X}$ a la même moyenne μ , mais une variance n fois plus petite!

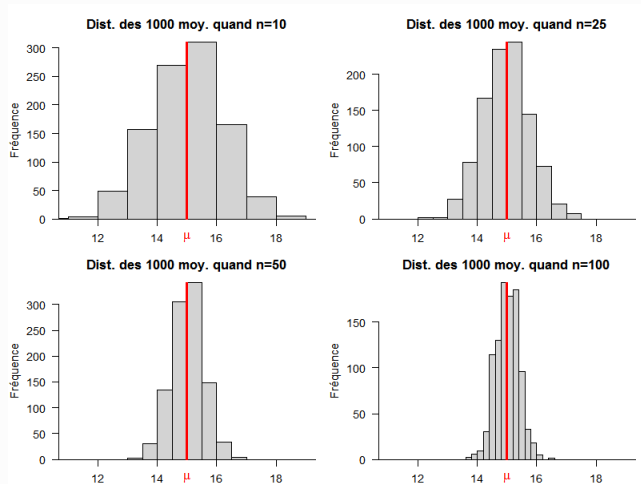
Effet de la taille d'échantillon n



Même population ($\mu = 15$, $\sigma = 4$) : 1000 échantillons pour chaque n .

n	$\sigma/\sqrt{n}$
10	1,26
25	0,80
50	0,57
100	0,40

Effet de la taille d'échantillon n



Même population ($\mu = 15$, $\sigma = 4$) : 1000 échantillons pour chaque n .

n	$\sigma/\sqrt{n}$
10	1,26
25	0,80
50	0,57
100	0,40

Plus n est grand, plus les $\bar{x}$ se **concentrent** autour de μ .

Propriétés théoriques de $\bar{X}$

Soit un échantillon aléatoire $X_1, \dots, X_n$ i.i.d. d'espérance μ et de variance σ^2 .

1. Espérance (sans biais)

$$E[\bar{X}] = E\left[\frac{1}{n} \sum_{i=1}^n X_i\right] = \frac{1}{n} \sum_{i=1}^n E[X_i] = \frac{1}{n}(n\mu) = \mu$$

Propriétés théoriques de $\bar{X}$

Soit un échantillon aléatoire $X_1, \dots, X_n$ i.i.d. d'espérance μ et de variance σ^2 .

1. Espérance (sans biais)

$$E[\bar{X}] = E\left[\frac{1}{n} \sum_{i=1}^n X_i\right] = \frac{1}{n} \sum_{i=1}^n E[X_i] = \frac{1}{n}(n\mu) = \mu$$

2. Variance et erreur-type (grâce à l'indépendance)

$$\text{Var}(\bar{X}) = \frac{1}{n^2} \sum_{i=1}^n \text{Var}(X_i) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n} \quad \implies \quad \text{ET}(\bar{X}) = \frac{\sigma}{\sqrt{n}}$$

QUIZ!

La glycémie à jeun d'adultes en santé a un écart-type $\sigma = 0,8$ mmol/L.

- Avec un échantillon de $n = 16$ patients, que vaut l'erreur-type de $\bar{X}$?
- Si on passe de $n = 16$ à $n = 64$ patients, l'erreur-type est : (a) divisée par 4; (b) divisée par 2; (c) inchangée.
- Vrai ou faux : augmenter n réduit aussi la variabilité des glycémies *individuelles*.

QUIZ!

La glycémie à jeun d'adultes en santé a un écart-type $\sigma = 0,8$ mmol/L.

- Avec un échantillon de $n = 16$ patients, que vaut l'erreur-type de $\bar{X}$?
→ $ET(\bar{X}) = 0,8 / \sqrt{16} = 0,2$ mmol/L.
- Si on passe de $n = 16$ à $n = 64$ patients, l'erreur-type est : (a) divisée par 4; (b) divisée par 2; (c) inchangée.
→ (b) : $0,8 / \sqrt{64} = 0,1$. Quatre fois plus de patients pour deux fois plus de précision.
- Vrai ou faux : augmenter n réduit aussi la variabilité des glycémies individuelles.
→ **Faux!** σ est une caractéristique de la population; seul l'écart-type de la *moyenne* diminue.

Estimation de la variance σ^2

Estimateur de la variance σ^2

On définit la variance échantillonnale par :

Variance échantillonnale S^2

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

Estimateur de la variance σ^2

On définit la variance échantillonnale par :

Variance échantillonnale S^2

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

Pourquoi diviser par $n-1$ plutôt que par n ?

- Les écarts se mesurent par rapport à $\bar{X}$ (estimé) et non par rapport au vrai μ .
- Puisque $\sum (X_i - \bar{X}) = 0$, dès qu'on connaît $n-1$ écarts, le dernier est imposé : il y a **$n-1$ degrés de liberté**.
- En divisant par $n-1$, l'estimateur devient **sans biais** : $E[S^2] = \sigma^2$.

Variance de S^2

Puisque S^2 est calculé à partir de l'échantillon, c'est aussi une variable aléatoire qui fluctue.

Propriétés de S^2 (sous l'hypothèse de normalité)

- **Espérance** : $E[S^2] = \sigma^2$ (sans biais)
- **Variance** : $\text{Var}(S^2) = \frac{2\sigma^4}{n-1}$
- **Erreur-type** : $\text{ET}(S^2) = \sqrt{\frac{2\sigma^4}{n-1}} = \frac{\sqrt{2}\sigma^2}{\sqrt{n-1}}$

Variance de S^2

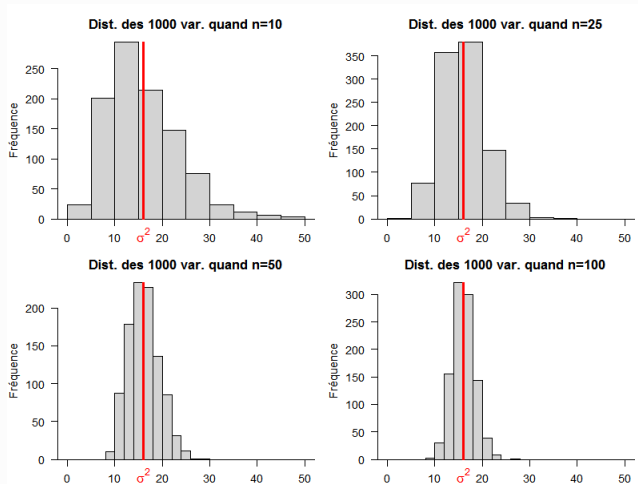
Puisque S^2 est calculé à partir de l'échantillon, c'est aussi une variable aléatoire qui fluctue.

Propriétés de S^2 (sous l'hypothèse de normalité)

- **Espérance** : $E[S^2] = \sigma^2$ (sans biais)
- **Variance** : $\text{Var}(S^2) = \frac{2\sigma^4}{n-1}$
- **Erreur-type** : $\text{ET}(S^2) = \sqrt{\frac{2\sigma^4}{n-1}} = \frac{\sqrt{2}\sigma^2}{\sqrt{n-1}}$

⇒ Plus n est grand, plus S^2 se rapproche précisément de la vraie variance σ^2 .

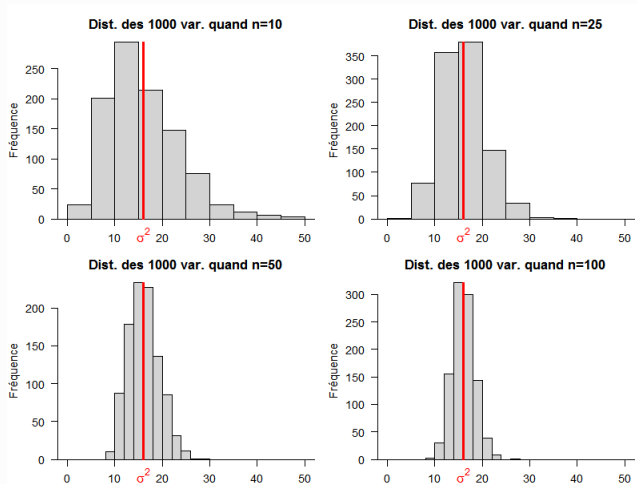
Simulation : distribution de S^2



Même population ($\sigma^2 = 16$) :
1000 valeurs de s^2 pour
chaque n .

- Centrées sur σ^2 : **sans biais.**
- Dispersion qui diminue avec n .

Simulation : distribution de S^2



Même population ($\sigma^2 = 16$) :
1000 valeurs de s^2 pour
chaque n .

- Centrées sur σ^2 : **sans biais.**
- Dispersion qui diminue avec n .
- **Asymétrique** (queue à droite), surtout si n est petit : **pas normale!**

⇒ Il faut une nouvelle loi : le **khi-deux.**

Retour sur l'exemple 1 : Calculs complets

Échantillon de pression systolique ($n = 7$) : 141, 130, 135, 138, 137, 122, 135

1. **Moyenne échantillonnale** : $\bar{x} = \frac{938}{7} = 134 \text{ mm Hg}$

2. **Variance échantillonnale** :

$$s^2 = \frac{1}{7-1} \left[(141 - 134)^2 + \dots + (135 - 134)^2 \right] = \frac{236}{6} = 39,33 \text{ (mm Hg)}^2$$

Retour sur l'exemple 1 : Calculs complets

Échantillon de pression systolique ($n = 7$) : 141, 130, 135, 138, 137, 122, 135

1. **Moyenne échantillonnale** : $\bar{x} = \frac{938}{7} = 134$ mm Hg

2. **Variance échantillonnale** :

$$s^2 = \frac{1}{7-1} [(141-134)^2 + \dots + (135-134)^2] = \frac{236}{6} = 39,33 \text{ (mm Hg)}^2$$

3. **Écart-type échantillonnal** : $s = \sqrt{39,33} = 6,27$ mm Hg

4. **Erreur-type estimée de la moyenne** : $\widehat{ET}(\bar{X}) = \frac{s}{\sqrt{n}} = \frac{6,27}{\sqrt{7}} = 2,37$ mm Hg

Dans R : `x <- c(141, 130, 135, 138, 137, 122, 135)`, puis
`mean(x)`, `var(x)`, `sd(x)` et `sd(x)/sqrt(7)`.

Diviser par n ou par $n - 1$? Comparaison par l'EQM

Certaines formules (et calculatrices) utilisent $\hat{\sigma}^2 = \frac{1}{n} \sum (X_i - \bar{X})^2 = \frac{n-1}{n} S^2$.

Comparons-les pour la population de pressions $\mathcal{N}(133, 62)$ avec $n = 7$:

	S^2	$\hat{\sigma}^2 = \frac{6}{7} S^2$
Espérance	62	$\frac{6}{7} \times 62 = 53,1$
Biais	0	-8,9
Variance	$\frac{2 \times 62^2}{6} = 1281,3$	$\left(\frac{6}{7}\right)^2 \times 1281,3 = 941,4$
EQM = Var + Biais ²	1281,3	$941,4 + 8,9^2 = 1019,8$

Diviser par n ou par $n - 1$? Comparaison par l'EQM

Certaines formules (et calculatrices) utilisent $\hat{\sigma}^2 = \frac{1}{n} \sum (X_i - \bar{X})^2 = \frac{n-1}{n} S^2$.

Comparons-les pour la population de pressions $\mathcal{N}(133, 62)$ avec $n = 7$:

	S^2	$\hat{\sigma}^2 = \frac{6}{7} S^2$
Espérance	62	$\frac{6}{7} \times 62 = 53,1$
Biais	0	-8,9
Variance	$\frac{2 \times 62^2}{6} = 1281,3$	$\left(\frac{6}{7}\right)^2 \times 1281,3 = 941,4$
EQM = Var + Biais ²	1281,3	941,4 + 8,9² = 1019,8

Surprise : $\hat{\sigma}^2$ est biaisé, mais son EQM est **plus petite** ! (C'est vrai pour tout n .)

Diviser par n ou par $n - 1$? Comparaison par l'EQM

Certaines formules (et calculatrices) utilisent $\hat{\sigma}^2 = \frac{1}{n} \sum (X_i - \bar{X})^2 = \frac{n-1}{n} S^2$.

Comparons-les pour la population de pressions $\mathcal{N}(133, 62)$ avec $n = 7$:

	S^2	$\hat{\sigma}^2 = \frac{6}{7} S^2$
Espérance	62	$\frac{6}{7} \times 62 = 53,1$
Biais	0	-8,9
Variance	$\frac{2 \times 62^2}{6} = 1281,3$	$\left(\frac{6}{7}\right)^2 \times 1281,3 = 941,4$
EQM = Var + Biais ²	1281,3	941,4 + 8,9² = 1019,8

Surprise : $\hat{\sigma}^2$ est biaisé, mais son EQM est **plus petite** ! (C'est vrai pour tout n .)

Pourquoi garder S^2 ? Il est sans biais, la différence devient négligeable quand n est grand, et c'est S^2 qui mène aux lois du khi-deux et de Student (IC et tests). C'est aussi ce que calcule `var()` dans R.

Lois échantillonnales

Pourquoi connaître la loi d'un estimateur ?

$E[T]$ et $ET(T)$ ne suffisent pas pour répondre à des questions comme :

- Quelle est la probabilité que $\bar{X}$ soit à moins de 5 mm Hg de μ ?
- Quelles valeurs de S^2 sont plausibles si $\sigma^2 = 62$?

⇒ Il faut la **loi échantillonnale** complète : c'est la base des intervalles de confiance (3b) et des tests (3c).

Pourquoi connaître la loi d'un estimateur ?

$E[T]$ et $ET(T)$ ne suffisent pas pour répondre à des questions comme :

- Quelle est la probabilité que $\bar{X}$ soit à moins de 5 mm Hg de μ ?
- Quelles valeurs de S^2 sont plausibles si $\sigma^2 = 62$?

⇒ Il faut la **loi échantillonnale** complète : c'est la base des intervalles de confiance (3b) et des tests (3c).

Quantité (population normale)	Loi	Où ?
$(\bar{X} - \mu)/(\sigma/\sqrt{n})$	Normale $\mathcal{N}(0, 1)$	ici
$(n - 1)S^2/\sigma^2$	Khi-deux χ_{n-1}^2	ici
$(\bar{X} - \mu)/(S/\sqrt{n})$	Student t_{n-1}	chapitre 3b
S_1^2/S_2^2 (deux échantillons)	Fisher	chapitre 4

Loi de $\bar{X}$: cas d'une population normale

Théorème. Si $X_1, \dots, X_n \stackrel{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$, alors, pour tout n ,

$$\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right) \quad \text{et donc} \quad Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1).$$

Loi de $\bar{X}$: cas d'une population normale

Théorème. Si $X_1, \dots, X_n \stackrel{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$, alors, **pour tout n** ,

$$\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right) \quad \text{et donc} \quad Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1).$$

- Une somme (ou une moyenne) de v.a. normales indépendantes est encore normale (chapitre 2).
- On connaissait déjà $E[\bar{X}] = \mu$ et $\text{Var}(\bar{X}) = \sigma^2/n$; la nouveauté, c'est la **forme** de la loi.
- On peut donc calculer des probabilités sur $\bar{X}$ avec la table de la loi normale, en **standardisant avec $\sigma/\sqrt{n}$** (et non σ).

Exemple 3 : Pression artérielle (suite)

Supposons que la pression systolique suit $\mathcal{N}(\mu = 133, \sigma^2 = 62)$.

- (a) Quelle est la probabilité qu'**un** adulte ait une pression supérieure à 138 mm Hg?
- (b) Quelle est la probabilité que la **moyenne de 7** adultes dépasse 138 mm Hg?
- (c) *À vous* : quelle est la probabilité que $\bar{X}$ soit à moins de 5 mm Hg de μ ?

Exemple 3 : Pression artérielle (suite)

Supposons que la pression systolique suit $\mathcal{N}(\mu = 133, \sigma^2 = 62)$.

- (a) Quelle est la probabilité qu'un adulte ait une pression supérieure à 138 mm Hg ?

$$P(X > 138) = P\left(Z > \frac{138 - 133}{\sqrt{62}}\right) = P(Z > 0,64) = \mathbf{0,2611}$$

- (b) Quelle est la probabilité que la **moyenne de 7** adultes dépasse 138 mm Hg ?

- (c) *À vous* : quelle est la probabilité que $\bar{X}$ soit à moins de 5 mm Hg de μ ?

Exemple 3 : Pression artérielle (suite)

Supposons que la pression systolique suit $\mathcal{N}(\mu = 133, \sigma^2 = 62)$.

- (a) Quelle est la probabilité qu'un adulte ait une pression supérieure à 138 mm Hg ?

$$P(X > 138) = P\left(Z > \frac{138 - 133}{\sqrt{62}}\right) = P(Z > 0,64) = \mathbf{0,2611}$$

- (b) Quelle est la probabilité que la **moyenne de 7** adultes dépasse 138 mm Hg ?

$$P(\bar{X} > 138) = P\left(Z > \frac{138 - 133}{\sqrt{62/7}}\right) = P\left(Z > \frac{5}{2,976}\right) = P(Z > 1,68) = \mathbf{0,0465}$$

- (c) *À vous* : quelle est la probabilité que $\bar{X}$ soit à moins de 5 mm Hg de μ ?

⇒ Une pression individuelle dépasse 138 dans 26 % des cas, mais une moyenne de 7 pressions, dans moins de 5 % des cas.

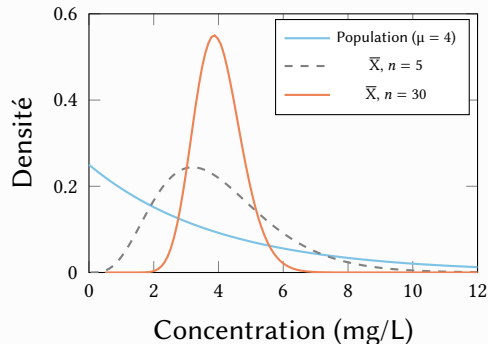
Et si la population n'est pas normale ?

Rappel : théorème limite central

$X_1, \dots, X_n$ i.i.d. de moyenne μ et de variance σ^2 (loi quelconque). Si n est assez grand :

$$\bar{X} \approx \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

- En pratique, $n \geq 30$ suffit souvent (chapitre 2).
- Population très asymétrique : il faut un n plus grand.



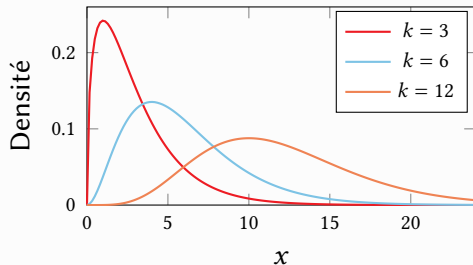
Concentration d'un polluant : population très asymétrique (loi exponentielle).

La loi du khi-deux χ_k^2

Définition (intuitive)

Si $Z_1, \dots, Z_k$ sont des v.a. $\mathcal{N}(0, 1)$ indépendantes, alors $K = Z_1^2 + Z_2^2 + \dots + Z_k^2$ suit une loi du khi-deux à k degrés de liberté, notée χ_k^2 .

- Somme de carrés : valeurs toujours positives.
- Loi asymétrique (queue à droite).

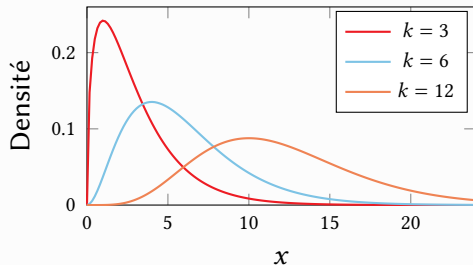


La loi du khi-deux χ_k^2

Définition (intuitive)

Si $Z_1, \dots, Z_k$ sont des v.a. $\mathcal{N}(0, 1)$ indépendantes, alors $K = Z_1^2 + Z_2^2 + \dots + Z_k^2$ suit une **loi du khi-deux à k degrés de liberté**, notée χ_k^2 .

- Somme de carrés : valeurs **toujours positives**.
- Loi **asymétrique** (queue à droite).
- $E[K] = k$ et $\text{Var}(K) = 2k$.
- Plus k est grand, plus elle ressemble à une loi normale.



Loi échantillonnale de S^2

Théorème. Si $X_1, \dots, X_n \stackrel{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$, alors

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi_{n-1}^2.$$

Loi échantillonnale de S^2

Théorème. Si $X_1, \dots, X_n \stackrel{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$, alors

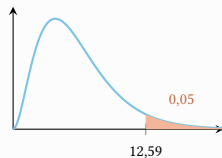
$$\frac{(n-1)S^2}{\sigma^2} \sim \chi_{n-1}^2.$$

- **Intuition :** $\sum \left(\frac{X_i - \mu}{\sigma}\right)^2$ est une somme de n carrés de $\mathcal{N}(0, 1)$, donc une χ_n^2 . En remplaçant μ par $\bar{X}$, on **perd un degré de liberté** (le même $n - 1$ que dans S^2).
- **Cohérence** avec $E[K] = k$ et $\text{Var}(K) = 2k$:
 - $E[(n-1)S^2/\sigma^2] = n-1 \Rightarrow E[S^2] = \sigma^2$;
 - $\text{Var}[(n-1)S^2/\sigma^2] = 2(n-1) \Rightarrow \text{Var}(S^2) = 2\sigma^4/(n-1)$.
- **⚠** Ce résultat exige une population **normale** : contrairement à $\bar{X}$, le TLC ne vient pas à la rescousse.

Table de la loi du khi-deux

On note $\chi^2_{k;\gamma}$ la valeur telle que $P(K > \chi^2_{k;\gamma}) = \gamma$ lorsque $K \sim \chi^2_k$ (aire à droite).

	γ									
k	0,995	0,975	0,950	0,900	0,500	0,100	0,050	0,025	0,005	
1	0,00	0,00	0,00	0,02	0,45	2,71	3,84	5,02	7,88	
2	0,01	0,05	0,10	0,21	1,39	4,61	5,99	7,38	10,60	
3	0,07	0,22	0,35	0,58	2,37	6,25	7,81	9,35	12,84	
4	0,21	0,48	0,71	1,06	3,36	7,78	9,49	11,14	14,86	
5	0,41	0,83	1,15	1,61	4,35	9,24	11,07	12,83	16,75	
6	0,68	1,24	1,64	2,20	5,35	10,64	12,59	14,45	18,55	
7	0,99	1,69	2,17	2,83	6,35	12,02	14,07	16,01	20,28	
8	1,34	2,18	2,73	3,49	7,34	13,36	15,51	17,53	21,95	
9	1,73	2,70	3,33	4,17	8,34	14,68	16,92	19,02	23,59	
10	2,16	3,25	3,94	4,87	9,34	15,99	18,31	20,48	25,19	
15	4,60	6,26	7,26	8,55	14,34	22,31	25,00	27,49	32,80	
20	7,43	9,59	10,85	12,44	19,34	28,41	31,41	34,17	40,00	
30	13,79	16,79	18,49	20,60	29,34	40,26	43,77	46,98	53,67	



$$\chi^2_{6;0,05} = 12,59$$

Dans R :

```
qchisq(0.05, 6,
lower.tail = FALSE)
```

Extrait ; la table complète du cours va
jusqu'à $k = 100$.

Exemple 4 : Variabilité de S^2 (pression artérielle)

Toujours $X \sim \mathcal{N}(133, 62)$ et $n = 7$, donc $\frac{6S^2}{62} \sim \chi_6^2$.

(a) Quelle est la probabilité que la variance échantillonnale dépasse 110 ?

(b) Entre quelles valeurs S^2 se trouvera-t-elle avec probabilité 0,90 ?

(c) *À vous* : encadrez $P(S^2 > 100)$ à l'aide de la table.

Exemple 4 : Variabilité de S^2 (pression artérielle)

Toujours $X \sim \mathcal{N}(133, 62)$ et $n = 7$, donc $\frac{6S^2}{62} \sim \chi_6^2$.

(a) Quelle est la probabilité que la variance échantillonnale dépasse 110 ?

$$P(S^2 > 110) = P\left(\chi_6^2 > \frac{6 \times 110}{62}\right) = P(\chi_6^2 > 10,65) \approx \mathbf{0,10} \quad (\text{table}).$$

(b) Entre quelles valeurs S^2 se trouvera-t-elle avec probabilité 0,90 ?

$$P(1,64 < \chi_6^2 < 12,59) = 0,90 \Rightarrow \frac{62 \times 1,64}{6} < S^2 < \frac{62 \times 12,59}{6}, \text{ soit} \\ \mathbf{16,9 < S^2 < 130,1.}$$

(c) *À vous* : encadrez $P(S^2 > 100)$ à l'aide de la table.

Exemple 4 : Variabilité de S^2 (pression artérielle)

Toujours $X \sim \mathcal{N}(133, 62)$ et $n = 7$, donc $\frac{6S^2}{62} \sim \chi_6^2$.

(a) Quelle est la probabilité que la variance échantillonnale dépasse 110 ?

$$P(S^2 > 110) = P\left(\chi_6^2 > \frac{6 \times 110}{62}\right) = P(\chi_6^2 > 10,65) \approx \mathbf{0,10} \quad (\text{table}).$$

(b) Entre quelles valeurs S^2 se trouvera-t-elle avec probabilité 0,90 ?

$$P(1,64 < \chi_6^2 < 12,59) = 0,90 \Rightarrow \frac{62 \times 1,64}{6} < S^2 < \frac{62 \times 12,59}{6}, \text{ soit} \\ \mathbf{16,9 < S^2 < 130,1}.$$

(c) *À vous* : encadrez $P(S^2 > 100)$ à l'aide de la table.

$\Rightarrow$ Avec $n = 7$, S^2 varie énormément (de $\approx \sigma^2/4$ à $\approx 2\sigma^2$) ! Notre $s^2 = 39,3$ est plausible. Pour estimer σ^2 , un **intervalle de confiance** (3b) sera bien utile.

QUIZ!

$X_1, \dots, X_n$ est un échantillon d'une population $\mathcal{N}(\mu, \sigma^2)$. Quelle est la loi de...

- $\dots \bar{X}$?
- $\dots (n-1)S^2/\sigma^2$?
- Vrai ou faux : la loi de S^2 est symétrique autour de σ^2 .
- $\dots \frac{\bar{X} - \mu}{S/\sqrt{n}}$ (on a remplacé σ par S)?

QUIZ!

$X_1, \dots, X_n$ est un échantillon d'une population $\mathcal{N}(\mu, \sigma^2)$. Quelle est la loi de...

- $\dots \bar{X}$? $\rightarrow \mathcal{N}(\mu, \sigma^2/n)$, exactement, pour tout n .
- $\dots (n-1)S^2/\sigma^2$? $\rightarrow \chi_{n-1}^2$.
- Vrai ou faux : la loi de S^2 est symétrique autour de σ^2 .
 $\rightarrow$ **Faux** : S^2 est un multiple d'un khi-deux, asymétrique à droite.
- $\dots \frac{\bar{X} - \mu}{S/\sqrt{n}}$ (on a remplacé σ par S)?
 $\rightarrow$ Plus tout à fait $\mathcal{N}(0, 1)$: c'est la **loi de Student** t_{n-1} , au chapitre 3b!

Estimation d'une proportion p

Exemple 2 : Un sondage d'opinion

Dans un sondage réalisé auprès de $n = 1028$ Québécois, on demande : « *Quel enjeu influencera le plus votre vote ?* »

Résultat : $y = 216$ personnes répondent l'environnement.

Exemple 2 : Un sondage d'opinion

Dans un sondage réalisé auprès de $n = 1028$ Québécois, on demande : « *Quel enjeu influencera le plus votre vote ?* »

Résultat : $y = 216$ personnes répondent l'environnement.

Modélisation

Soit Y le nombre de répondants choisissant l'environnement parmi n .

$$Y \sim \text{Binomiale}(n, p)$$

où p est la vraie proportion inconnue dans la population québécoise.

Comment estimer p et quelle est la précision de notre résultat ?

Propriétés de l'estimateur de proportion

Estimateur de la proportion

$$\hat{P} = \frac{Y}{n} \quad \text{où } Y \sim \text{Binomiale}(n, p), \quad E[Y] = np, \quad \text{Var}(Y) = np(1 - p).$$

Propriétés de l'estimateur de proportion

Estimateur de la proportion

$$\hat{P} = \frac{Y}{n} \quad \text{où } Y \sim \text{Binomiale}(n, p), \quad E[Y] = np, \quad \text{Var}(Y) = np(1 - p).$$

- **Espérance (sans biais) :** $E[\hat{P}] = \frac{1}{n}E[Y] = \frac{np}{n} = p$
- **Variance :** $\text{Var}(\hat{P}) = \frac{1}{n^2}\text{Var}(Y) = \frac{np(1-p)}{n^2} = \frac{p(1-p)}{n}$
- **Erreur-type :** $\text{ET}(\hat{P}) = \sqrt{\frac{p(1-p)}{n}}$

Retour sur l'exemple 2 : Application

Données du sondage : $n = 1028$, $y = 216$.

- **Estimation ponctuelle de la proportion** : $\hat{p} = \frac{216}{1028} \approx 0,2101$ (21,0 %)

Retour sur l'exemple 2 : Application

Données du sondage : $n = 1028$, $y = 216$.

- **Estimation ponctuelle de la proportion** : $\hat{p} = \frac{216}{1028} \approx 0,2101$ (21,0 %)
- **Erreur-type estimée** : Puisque p est inconnu, on remplace p par $\hat{p}$ dans la formule :

$$\widehat{\text{ET}}(\hat{P}) = \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} = \sqrt{\frac{0,2101 \times 0,7899}{1028}} \approx 0,0127 \quad (1,27 \%)$$

- **Conclusion** : L'estimation est de 21,0 % avec une incertitude d'échantillonnage d'environ 1,3 %.

Loi échantillonnale de $\hat{P}$

Loi approximative de $\hat{P}$

$\hat{P}$ est la moyenne de n variables valant 0 ou 1 (Bernoulli). Par le **TLC** (approximation normale de la binomiale, chapitre 2), si n est grand :

$$\hat{P} \approx \mathcal{N}\left(p, \frac{p(1-p)}{n}\right)$$

- Ce résultat servira à construire l'intervalle de confiance et le test pour une proportion.
- À vous (*exemple 2*) : si la vraie proportion était $p = 0,20$, quelle serait la probabilité qu'un sondage de 1028 personnes donne $\hat{p} \geq 0,21$?

Conclusion de la partie 1

Ce que nous savons faire à ce stade

- Calculer une **estimation ponctuelle** ($\bar{x}$, s^2 , $\hat{p}$) et son **erreur-type**.
- Juger un estimateur : **biais**, **variance**, **EQM**.
- Utiliser les **lois échantillonales** : $\bar{X}$ normale (exactement ou par le TLC); $(n-1)S^2/\sigma^2 \sim \chi_{n-1}^2$.

Conclusion de la partie 1

Ce que nous savons faire à ce stade

- Calculer une **estimation ponctuelle** ($\bar{x}$, s^2 , $\hat{p}$) et son **erreur-type**.
- Juger un estimateur : **biais**, **variance**, **EQM**.
- Utiliser les **lois échantillonales** : $\bar{X}$ normale (exactement ou par le TLC); $(n-1)S^2/\sigma^2 \sim \chi_{n-1}^2$.

Limite de l'estimation ponctuelle

Une valeur unique (ex. : $\bar{x} = 134$) n'indique pas à quel point la vraie valeur μ est proche.

La suite du chapitre 3

- **Partie 2 (chapitre 3b) : intervalles de confiance.**
 - Encadrer μ avec un niveau de confiance contrôlé, grâce à une nouvelle loi : la **loi de Student**.
 - La loi du khi-deux permettra d'encadrer σ^2 ; la loi approximative de $\hat{P}$, la proportion p .
- **Partie 3 (chapitre 3c) : tests d'hypothèses.**
 - Décider, par exemple, si la pression moyenne de la population dépasse 130 mm Hg.
- **Chapitre 4 : comparaison de deux échantillons** ; la **loi de Fisher** servira à comparer deux variances.

Réponses aux exemples

- **Quiz des souris :** (c). Avec $\bar{X} \sim \mathcal{N}(25, 4/16)$: $P(24 < \bar{X} < 26) = P(-2 < Z < 2) = 0,9544$.
- **Exemple 1 :** $\bar{x} = 134$ mm Hg; $s^2 = 39,33$ (mm Hg)²; $s = 6,27$ mm Hg; $\widehat{ET}(\bar{X}) = 2,37$ mm Hg.
- **EQM** ($n = 7, \sigma^2 = 62$) : $EQM(S^2) = 2\sigma^4/(n-1) = 1281,3$;
 $EQM(\hat{\sigma}^2) = (2n-1)\sigma^4/n^2 = 1019,8$.
- **Exemple 3 :** (a) $P(X > 138) = P(Z > 0,64) = 0,2611$.
(b) $P(\bar{X} > 138) = P(Z > 1,68) = 0,0465$.
(c) $P(128 < \bar{X} < 138) = P(-1,68 < Z < 1,68) = 0,9070$.
- **Exemple 4 :** (a) $P(\chi_6^2 > 10,65) \approx 0,10$ (R : 0,0999). (b) $16,9 < S^2 < 130,1$.
(c) $5,35 < 6 \times 100/62 = 9,68 < 10,64$, donc $0,10 < P(S^2 > 100) < 0,50$ (R : 0,139).
- **Exemple 2 :** $\hat{p} = 0,2101$; $\widehat{ET}(\hat{P}) = 0,0127$. Si $p = 0,20$: $ET(\hat{P}) = \sqrt{0,2 \times 0,8/1028} = 0,0125$ et
 $P(\hat{P} \geq 0,21) \approx P(Z \geq 0,80) = 0,2119$ (valeur exacte avec la binomiale : 0,219).