

Chapitre 3 – Estimations et tests d'hypothèse (2^e partie)

STT-1920 Méthodes statistiques

Département de mathématiques et de statistique



UNIVERSITÉ
LAVAL

Notion d'intervalle de confiance

Limite de l'estimation ponctuelle

Dans la première partie (chapitre 3a), nous avons vu que :

- L'estimateur ponctuel $\bar{X}$ fournit **une seule valeur** pour estimer le paramètre μ .
- Parce que $\bar{X}$ est une variable aléatoire, sa valeur change d'un échantillon à l'autre.
- Pour l'exemple 1 de la pression ($n = 7$), nous avons obtenu $\bar{x} = 134$ mm Hg.

Limite de l'estimation ponctuelle

Dans la première partie (chapitre 3a), nous avons vu que :

- L'estimateur ponctuel $\bar{X}$ fournit **une seule valeur** pour estimer le paramètre μ .
- Parce que $\bar{X}$ est une variable aléatoire, sa valeur change d'un échantillon à l'autre.
- Pour l'exemple 1 de la pression ($n = 7$), nous avons obtenu $\bar{x} = 134$ mm Hg.

Une question fondamentale

Quelle confiance accorder à cette unique valeur 134? À quelle distance de la vraie moyenne inconnue μ a-t-on des chances de se trouver?

Limite de l'estimation ponctuelle

Dans la première partie (chapitre 3a), nous avons vu que :

- L'estimateur ponctuel $\bar{X}$ fournit **une seule valeur** pour estimer le paramètre μ .
- Parce que $\bar{X}$ est une variable aléatoire, sa valeur change d'un échantillon à l'autre.
- Pour l'exemple 1 de la pression ($n = 7$), nous avons obtenu $\bar{x} = 134$ mm Hg.

Une question fondamentale

Quelle confiance accorder à cette unique valeur 134 ? À quelle distance de la vraie moyenne inconnue μ a-t-on des chances de se trouver ?

⇒ On souhaite encadrer μ par un intervalle $[L_1, L_2]$ associé à un **degré de certitude contrôlé**.

Définition d'un intervalle de confiance

Intervalle de confiance de niveau $1 - \alpha$

Deux statistiques L_1 et L_2 calculées à partir de l'échantillon telles que

$$P(L_1 \leq \theta \leq L_2) = 1 - \alpha.$$

- θ est le **paramètre inconnu** (valeur fixe de la population).

Définition d'un intervalle de confiance

Intervalle de confiance de niveau $1 - \alpha$

Deux statistiques L_1 et L_2 calculées à partir de l'échantillon telles que

$$P(L_1 \leq \theta \leq L_2) = 1 - \alpha.$$

- θ est le **paramètre inconnu** (valeur fixe de la population).
- $1 - \alpha$ est le **niveau de confiance** (typiquement 90%, 95% ou 99%).

Définition d'un intervalle de confiance

Intervalle de confiance de niveau $1 - \alpha$

Deux statistiques L_1 et L_2 calculées à partir de l'échantillon telles que

$$P(L_1 \leq \theta \leq L_2) = 1 - \alpha.$$

- θ est le **paramètre inconnu** (valeur fixe de la population).
- $1 - \alpha$ est le **niveau de confiance** (typiquement 90%, 95% ou 99%).
- α est le **seuil de risque** (probabilité que l'intervalle rate la vraie valeur).

Définition d'un intervalle de confiance

Intervalle de confiance de niveau $1 - \alpha$

Deux statistiques L_1 et L_2 calculées à partir de l'échantillon telles que

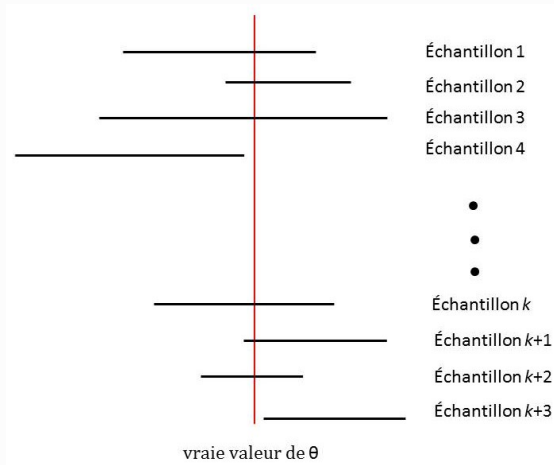
$$P(L_1 \leq \theta \leq L_2) = 1 - \alpha.$$

- θ est le **paramètre inconnu** (valeur fixe de la population).
- $1 - \alpha$ est le **niveau de confiance** (typiquement 90%, 95% ou 99%).
- α est le **seuil de risque** (probabilité que l'intervalle rate la vraie valeur).
- Lorsque l'intervalle est symétrique autour de l'estimation, on écrit

Estimation $\pm$ Marge d'erreur.

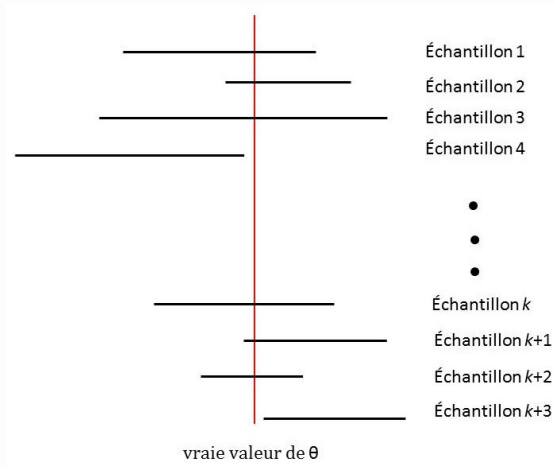
La **marge d'erreur** (ME) est la demi-longueur de l'intervalle.

Que veut dire « confiance à 95% » ?



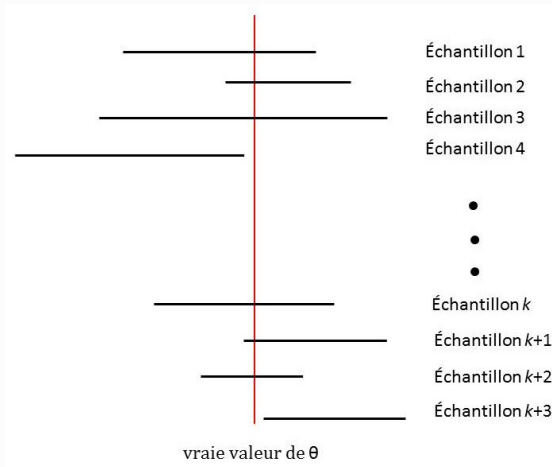
- Chaque échantillon donne un intervalle **différent** (centre et longueur).
- La vraie valeur θ (ligne rouge) est **fixe**.

Que veut dire « confiance à 95% » ?



- Chaque échantillon donne un intervalle **différent** (centre et longueur).
- La vraie valeur θ (ligne rouge) est **fixe**.
- Certains intervalles la contiennent, d'autres la ratent (ex. : échantillons 4 et $k + 3$).

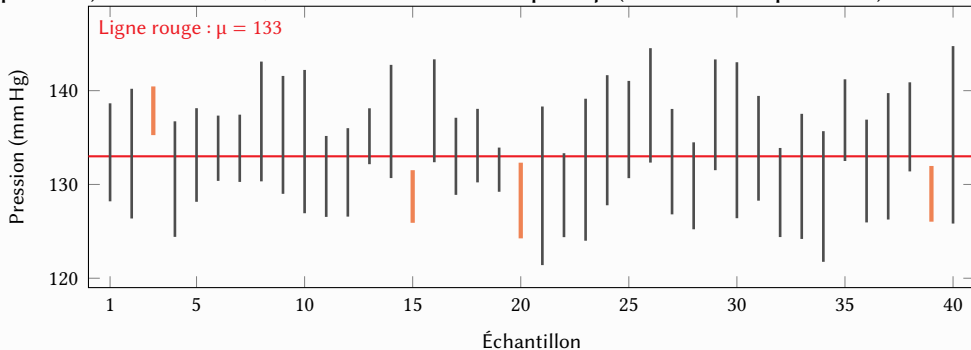
Que veut dire « confiance à 95% » ?



- Chaque échantillon donne un intervalle **différent** (centre et longueur).
- La vraie valeur θ (ligne rouge) est **fixe**.
- Certains intervalles la contiennent, d'autres la ratent (ex. : échantillons 4 et $k + 3$).
- Le niveau de confiance est la proportion des intervalles qui contiennent θ **à long terme**.

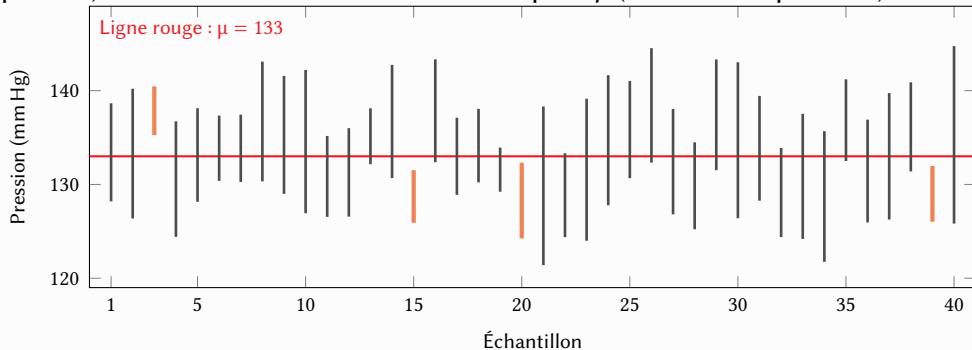
Simulation : 40 intervalles de confiance à 90%

40 échantillons de taille $n = 7$ tirés d'une population $\mathcal{N}(\mu = 133, \sigma^2 = 62)$ (comme au chapitre 3a). Pour chacun, on calcule l'IC à 90% pour μ (formule vue plus loin).



Simulation : 40 intervalles de confiance à 90%

40 échantillons de taille $n = 7$ tirés d'une population $\mathcal{N}(\mu = 133, \sigma^2 = 62)$ (comme au chapitre 3a). Pour chacun, on calcule l'IC à 90% pour μ (formule vue plus loin).



36 intervalles sur 40 (90%) contiennent μ ; les 4 intervalles orange la ratent.

Interprétation rigoureuse

✓ Interprétation correcte

Si on prélevait un grand nombre d'échantillons de même taille n et que pour chacun on construisait l'intervalle à 95%, alors **environ 95% de ces intervalles contiendraient la vraie valeur θ .**

Interprétation rigoureuse

✓ Interprétation correcte

Si on prélevait un grand nombre d'échantillons de même taille n et que pour chacun on construisait l'intervalle à 95%, alors **environ 95% de ces intervalles contiendraient la vraie valeur θ .**

✗ Attention aux interprétations incorrectes!

- **Faux** : « Il y a 90% de chances que μ appartienne à $[129,39 ; 138,61]$. » $\rightarrow \mu$ n'est pas aléatoire : c'est l'intervalle $[L_1, L_2]$ qui l'est!
- **Faux** : « 90% des individus ont une valeur entre 129,39 et 138,61. » $\rightarrow$ L'IC vise la *moyenne* μ , pas les individus!

QUIZ!

Dans la simulation précédente (40 intervalles à 90%) :

- Combien d'intervalles s'attendait-on à voir rater μ ?

QUIZ!

Dans la simulation précédente (40 intervalles à 90%) :

- Combien d'intervalles s'attendait-on à voir rater μ ?
- Si on avait plutôt calculé des IC à 95%, combien en moyenne rateraient μ ?
Les intervalles seraient-ils plus longs ou plus courts ?

QUIZ!

Dans la simulation précédente (40 intervalles à 90%) :

- Combien d'intervalles s'attendait-on à voir rater μ ?
- Si on avait plutôt calculé des IC à 95%, combien en moyenne rateraient μ ?
Les intervalles seraient-ils plus longs ou plus courts?
- Pourquoi les 40 intervalles n'ont-ils pas tous la même longueur?

Construction de l'intervalle pour la moyenne

Cas théorique : si σ était connue

Supposons $X_1, \dots, X_n$ i.i.d. de loi $\mathcal{N}(\mu, \sigma^2)$, avec σ connue.

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

Cas théorique : si σ était connue

Supposons $X_1, \dots, X_n$ i.i.d. de loi $\mathcal{N}(\mu, \sigma^2)$, avec σ connue.

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

Pour un niveau $1 - \alpha$, on prend le quantile $z_{\alpha/2}$ tel que $P(Z > z_{\alpha/2}) = \alpha/2$:

$$P\left(-z_{\alpha/2} \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}\right) = 1 - \alpha$$

Cas théorique : si σ était connue

Supposons $X_1, \dots, X_n$ i.i.d. de loi $\mathcal{N}(\mu, \sigma^2)$, avec σ connue.

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

Pour un niveau $1 - \alpha$, on prend le quantile $z_{\alpha/2}$ tel que $P(Z > z_{\alpha/2}) = \alpha/2$:

$$P\left(-z_{\alpha/2} \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}\right) = 1 - \alpha$$

IC pour μ (σ connue) en isolant μ au centre

$$\left[\bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right] \quad (z_{0,05} = 1,645, z_{0,025} = 1,96, z_{0,005} = 2,576)$$

Cas théorique : si σ était connue

Supposons $X_1, \dots, X_n$ i.i.d. de loi $\mathcal{N}(\mu, \sigma^2)$, avec σ connue.

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

Pour un niveau $1 - \alpha$, on prend le quantile $z_{\alpha/2}$ tel que $P(Z > z_{\alpha/2}) = \alpha/2$:

$$P\left(-z_{\alpha/2} \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}\right) = 1 - \alpha$$

IC pour μ (σ connue) en isolant μ au centre

$$\left[\bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right] \quad (z_{0,05} = 1,645, z_{0,025} = 1,96, z_{0,005} = 2,576)$$

Le problème pratique : en réalité, on ne connaît presque jamais σ !

Que faire quand σ est inconnue ?

Puisque σ est inconnue, l'idée naturelle est de la remplacer par l'écart-type échantillonnal S :

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

Que faire quand σ est inconnue ?

Puisque σ est inconnue, l'idée naturelle est de la remplacer par l'écart-type échantillonnal S :

$$T = \frac{\bar{X} - \mu}{S / \sqrt{n}}$$

- Au dénominateur, S n'est plus une constante, mais une **variable aléatoire**.

Que faire quand σ est inconnue ?

Puisque σ est inconnue, l'idée naturelle est de la remplacer par l'écart-type échantillonnal S :

$$T = \frac{\bar{X} - \mu}{S / \sqrt{n}}$$

- Au dénominateur, S n'est plus une constante, mais une **variable aléatoire**.
- Cette incertitude supplémentaire rend T **plus dispersée** que la loi normale centrée réduite Z .

Que faire quand σ est inconnue ?

Puisque σ est inconnue, l'idée naturelle est de la remplacer par l'écart-type échantillonnal S :

$$T = \frac{\bar{X} - \mu}{S / \sqrt{n}}$$

- Au dénominateur, S n'est plus une constante, mais une **variable aléatoire**.
- Cette incertitude supplémentaire rend T **plus dispersée** que la loi normale centrée réduite Z .
- T ne suit plus une loi normale $\mathcal{N}(0, 1)$!

Que faire quand σ est inconnue ?

Puisque σ est inconnue, l'idée naturelle est de la remplacer par l'écart-type échantillonnal S :

$$T = \frac{\bar{X} - \mu}{S / \sqrt{n}}$$

- Au dénominateur, S n'est plus une constante, mais une **variable aléatoire**.
- Cette incertitude supplémentaire rend T **plus dispersée** que la loi normale centrée réduite Z .
- T ne suit plus une loi normale $\mathcal{N}(0, 1)$!

Quelle est la loi exacte de T ?

La réponse a été découverte en 1908 par William Sealy Gosset : la **loi de Student**.

William Sealy Gosset (1876–1937)

- Chimiste et statisticien en chef à la brasserie **Guinness** à Dublin.
- Travaillait avec de **très petits échantillons** (variétés d'orge, température de brassage, levures).
- Découvre la distribution exacte de $\frac{\bar{X}-\mu}{S/\sqrt{n}}$.
- Guinness interdisait à ses employés de publier des secrets d'entreprise : il publia en 1908 sous le pseudonyme « **Student** ».



William Sealy Gosset

La loi de Student t_k

Théorème de Student

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

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1},$$

une loi de Student à $k = n - 1$ degrés de liberté.

La loi de Student t_k

Théorème de Student

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

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1},$$

une loi de Student à $k = n - 1$ degrés de liberté.

- Symétrique en 0, en forme de cloche, tout comme la loi normale.

La loi de Student t_k

Théorème de Student

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

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1},$$

une loi de Student à $k = n - 1$ degrés de liberté.

- Symétrique en 0, en forme de cloche, tout comme la loi normale.
- Ses queues sont plus lourdes : les valeurs extrêmes sont plus probables en raison de la variabilité de S.

La loi de Student t_k

Théorème de Student

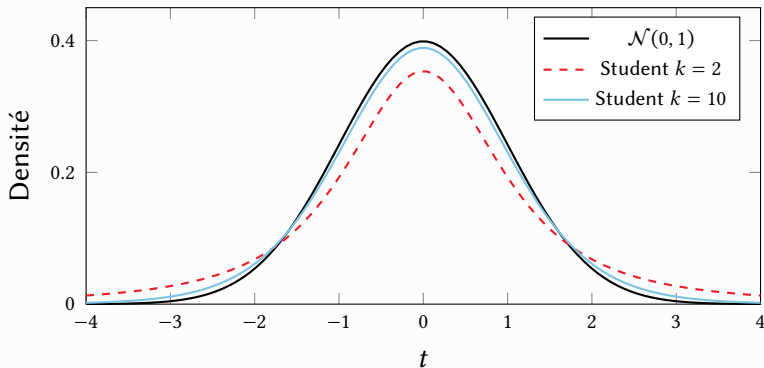
Si $X_1, \dots, X_n \stackrel{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$, alors

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1},$$

une loi de Student à $k = n - 1$ degrés de liberté.

- Symétrique en 0, en forme de cloche, tout comme la loi normale.
- Ses queues sont plus lourdes : les valeurs extrêmes sont plus probables en raison de la variabilité de S.
- Quand $k \rightarrow \infty$, la loi de Student converge vers la loi normale $\mathcal{N}(0, 1)$.

Comparaison de la loi normale et des lois de Student



Pour k petit, les quantiles sont plus grands que 1,96 pour compenser l'incertitude sur σ (ex. : $t_{6,0,025} = 2,447$; $t_{29,0,025} = 2,045$).

Formule de l'IC pour μ (σ inconnue)

Intervalle de confiance pour la moyenne μ

Soit un échantillon aléatoire d'une population normale :

$$\left[\bar{X} - t_{n-1, \alpha/2} \frac{S}{\sqrt{n}}, \bar{X} + t_{n-1, \alpha/2} \frac{S}{\sqrt{n}} \right]$$

Formule de l'IC pour μ (σ inconnue)

Intervalle de confiance pour la moyenne μ

Soit un échantillon aléatoire d'une population normale :

$$\left[\bar{X} - t_{n-1, \alpha/2} \frac{S}{\sqrt{n}}, \bar{X} + t_{n-1, \alpha/2} \frac{S}{\sqrt{n}} \right]$$

- $\bar{X}$: estimation ponctuelle de la moyenne.
- $\frac{S}{\sqrt{n}}$: erreur-type estimée (ÉT).
- $t_{n-1, \alpha/2}$: valeur critique de la table de Student à $n - 1$ d.l.
(dans R : `qt(1-alpha/2, n-1)`).
- $ME = t_{n-1, \alpha/2} \frac{S}{\sqrt{n}}$: la **marge d'erreur**.

Exemple 1 (suite) : pression artérielle

Mesures ($n = 7$) : 141, 130, 135, 138, 137, 122, 135; $\bar{x} = 134$ mm Hg,
 $s = 6,27$ mm Hg.

On cherche un intervalle de confiance à 90%, donc $\alpha = 0,10$ et $\alpha/2 = 0,05$.

Exemple 1 (suite) : pression artérielle

Mesures ($n = 7$) : 141, 130, 135, 138, 137, 122, 135; $\bar{x} = 134$ mm Hg,
 $s = 6,27$ mm Hg.

On cherche un intervalle de confiance à 90%, donc $\alpha = 0,10$ et $\alpha/2 = 0,05$.

1. **Degrés de liberté** : $k = n - 1 = 7 - 1 = 6$.
2. **Quantile critique** : dans la table de Student, $t_{6,0,05} = 1,943$.
3. **Erreur-type estimée** : $\frac{s}{\sqrt{n}} = \frac{6,27}{\sqrt{7}} = 2,37$.
4. **Marge d'erreur** : $ME = 1,943 \times 2,37 = 4,61$ mm Hg.
5. **Intervalle de confiance à 90%** :
 $[134 - 4,61; 134 + 4,61] = [129,39; 138,61]$ mm Hg

Exemple 1 (suite) : pression artérielle

Mesures ($n = 7$) : 141, 130, 135, 138, 137, 122, 135; $\bar{x} = 134$ mm Hg,
 $s = 6,27$ mm Hg.

On cherche un intervalle de confiance à 90%, donc $\alpha = 0,10$ et $\alpha/2 = 0,05$.

1. **Degrés de liberté** : $k = n - 1 = 7 - 1 = 6$.
2. **Quantile critique** : dans la table de Student, $t_{6,0,05} = 1,943$.
3. **Erreur-type estimée** : $\frac{s}{\sqrt{n}} = \frac{6,27}{\sqrt{7}} = 2,37$.
4. **Marge d'erreur** : $ME = 1,943 \times 2,37 = 4,61$ mm Hg.
5. **Intervalle de confiance à 90%** :
 $[134 - 4,61; 134 + 4,61] = [129,39; 138,61]$ mm Hg

Remarque : on garde toutes les décimales dans les calculs intermédiaires ($s/\sqrt{n} = 2,3705$), puis on arrondit à la fin.

Avec R

```
> x <- c(141, 130, 135, 138, 137, 122, 135)
> t.test(x, conf.level = 0.90)$conf.int
[1] 129.3938 138.6062
attr(,"conf.level")
[1] 0.9
> t.test(x)$conf.int                # niveau 95 % par défaut
[1] 128.1997 139.8003
```

- La fonction `t.test()` affiche aussi un **test d'hypothèses** (t , p -value) : nous verrons au chapitre 3c comment l'interpréter.
- Quantile à la main : `qt(0.95, df = 6)` donne 1,943.

Quiz conceptuel : Vrai ou Faux ?

Question 1

Énoncé 1

« 90% des gens de cette population ont une pression artérielle comprise entre 129,39 et 138,61 mm Hg. »

Question 1

Énoncé 1

« 90% des gens de cette population ont une pression artérielle comprise entre 129,39 et 138,61 mm Hg. »

FAUX!

L'intervalle de confiance est construit pour estimer le paramètre moyen μ de la population, et non pas pour contenir 90% des observations individuelles. La variabilité des individus est régie par σ (environ 6,3), alors que l'intervalle est basé sur $\sigma/\sqrt{n}$ (environ 2,4).

Question 2

Énoncé 2

« 90% des gens de cet échantillon ont une pression artérielle comprise entre 129,39 et 138,61 mm Hg. »

Question 2

Énoncé 2

« 90% des gens de cet échantillon ont une pression artérielle comprise entre 129,39 et 138,61 mm Hg. »

FAUX!

Pour le vérifier, regardons notre échantillon : {141, 130, 135, 138, 137, 122, 135}.

- 122 et 141 sont à l'extérieur de l'intervalle.
- Seules 5 des 7 valeurs (71,4%) sont à l'intérieur!

L'intervalle ne prétend absolument pas décrire l'échantillon, il vise la **vraie moyenne** μ .

Question 3

Énoncé 3

« Si on prenait des mesures sur 7 autres individus, il y a 90% de chances que la moyenne $\bar{x}_{\text{nouveau}}$ de ce nouvel échantillon soit comprise entre 129,39 et 138,61. »

Question 3

Énoncé 3

« Si on prenait des mesures sur 7 autres individus, il y a 90% de chances que la moyenne $\bar{x}_{\text{nouveau}}$ de ce nouvel échantillon soit comprise entre 129,39 et 138,61. »

FAUX!

Une nouvelle moyenne $\bar{X}_{\text{nouv}}$ a sa propre variabilité autour de μ . L'écart entre deux moyennes indépendantes a pour écart-type $\sqrt{\frac{\sigma^2}{n} + \frac{\sigma^2}{n}} = \sqrt{2} \frac{\sigma}{\sqrt{n}}$, qui est plus grand que $\frac{\sigma}{\sqrt{n}}$. Pour prédire une nouvelle moyenne ou observation, on utilise un **intervalle de prédiction**.

Question 4

Énoncé 4

« Si on tirait 1000 échantillons de taille 7 dans cette population et qu'on calculait pour chacun l'IC à 90%, environ 900 d'entre eux contiendraient la vraie moyenne μ . »

Question 4

Énoncé 4

« Si on tirait 1000 échantillons de taille 7 dans cette population et qu'on calculait pour chacun l'IC à 90%, environ 900 d'entre eux contiendraient la vraie moyenne μ . »

VRAI!

C'est **exactement** la signification probabiliste du niveau de confiance de 90% (voir la simulation de 40 intervalles)! C'est le mécanisme de construction de l'intervalle qui est fiable à 90%.

Question 5

Énoncé 5

« Avec un échantillon de taille $n \geq 30$, on peut utiliser les quantiles de la loi normale au lieu de la loi de Student. »

Question 5

Énoncé 5

« Avec un échantillon de taille $n \geq 30$, on peut utiliser les quantiles de la loi normale au lieu de la loi de Student. »

VRAI (en pratique)!

- Par le théorème limite central (TLC), la loi de $\bar{X}$ est approximativement normale même si la population ne l'est pas.
- De plus, pour $n \geq 30$, la loi de Student est presque identique à la normale : $t_{29, 0,025} = 2,045 \approx 1,96$.
- Avec un logiciel comme R, on utilise toutefois toujours la loi de Student.

Questions 6 et 7

Énoncé 6

« Avec un échantillon plus grand, on obtiendrait un intervalle de confiance plus long. »

Questions 6 et 7

Énoncé 6

« Avec un échantillon plus grand, on obtiendrait un intervalle de confiance plus long. »

FAUX : la longueur vaut $2 t_{n-1, \alpha/2} \frac{s}{\sqrt{n}}$. Puisque $\sqrt{n}$ est au dénominateur, **plus n augmente, plus l'intervalle est étroit.**

Questions 6 et 7

Énoncé 6

« Avec un échantillon plus grand, on obtiendrait un intervalle de confiance plus long. »

FAUX : la longueur vaut $2 t_{n-1, \alpha/2} \frac{s}{\sqrt{n}}$. Puisque $\sqrt{n}$ est au dénominateur, **plus n augmente, plus l'intervalle est étroit.**

Énoncé 7

« Si la vraie moyenne est de 133, l'estimation obtenue $\bar{x} = 134$ est biaisée. »

Questions 6 et 7

Énoncé 6

« Avec un échantillon plus grand, on obtiendrait un intervalle de confiance plus long. »

FAUX : la longueur vaut $2 t_{n-1, \alpha/2} \frac{s}{\sqrt{n}}$. Puisque $\sqrt{n}$ est au dénominateur, **plus n augmente, plus l'intervalle est étroit.**

Énoncé 7

« Si la vraie moyenne est de 133, l'estimation obtenue $\bar{x} = 134$ est biaisée. »

FAUX : le biais est une propriété de l'estimateur ($E[\bar{X}] = \mu$). Une réalisation $\bar{x}$ n'est jamais exactement égale à μ en raison de l'aléa, sans que l'estimateur soit biaisé pour autant.

Cas n grand, marge d'erreur et taille d'échantillon

Et si la population n'est pas normale ?

Plusieurs variables biologiques sont **asymétriques** (concentrations de contaminants, charges parasitaires...), alors que la formule de Student suppose la normalité.

Et si la population n'est pas normale ?

Plusieurs variables biologiques sont **asymétriques** (concentrations de contaminants, charges parasitaires...), alors que la formule de Student suppose la normalité.

IC pour μ : loi inconnue, n grand (en pratique $n \geq 30$)

Par le TLC (chapitre 2), $\bar{X} \approx \mathcal{N}(\mu, \sigma^2/n)$ et $S \approx \sigma$. Un IC de niveau **approximativement** $1 - \alpha$ est

$$\left[\bar{X} - z_{\alpha/2} \frac{S}{\sqrt{n}}, \bar{X} + z_{\alpha/2} \frac{S}{\sqrt{n}} \right].$$

Et si la population n'est pas normale ?

Plusieurs variables biologiques sont **asymétriques** (concentrations de contaminants, charges parasitaires...), alors que la formule de Student suppose la normalité.

IC pour μ : loi inconnue, n grand (en pratique $n \geq 30$)

Par le TLC (chapitre 2), $\bar{X} \approx \mathcal{N}(\mu, \sigma^2/n)$ et $S \approx \sigma$. Un IC de niveau **approximativement** $1 - \alpha$ est

$$\left[\bar{X} - z_{\alpha/2} \frac{S}{\sqrt{n}}, \bar{X} + z_{\alpha/2} \frac{S}{\sqrt{n}} \right].$$

- Si σ est connue, on l'utilise au lieu de S . Plus la population est asymétrique, plus n doit être grand.
- n petit et population très asymétrique : aucune formule fiable !

Exemple 3 : mercure dans la chair des dorés

On mesure la concentration de mercure ($\mu\text{g/g}$) dans la chair de $n = 50$ dorés jaunes pêchés dans un réservoir. La distribution est asymétrique à droite. On obtient $\bar{x} = 0,42 \mu\text{g/g}$ et $s = 0,18 \mu\text{g/g}$.

Question : donnez un IC à 95% pour la concentration moyenne μ .

Exemple 3 : mercure dans la chair des dorés

On mesure la concentration de mercure ($\mu\text{g/g}$) dans la chair de $n = 50$ dorés jaunes pêchés dans un réservoir. La distribution est asymétrique à droite. On obtient $\bar{x} = 0,42 \mu\text{g/g}$ et $s = 0,18 \mu\text{g/g}$.

Question : donnez un IC à 95% pour la concentration moyenne μ .

1. $n = 50$ est grand : on utilise le TLC et $z_{0,025} = 1,96$.
2. Erreur-type estimée : $s/\sqrt{n} = 0,18/\sqrt{50} = 0,0255$.
3. Marge d'erreur : $ME = 1,96 \times 0,0255 = 0,050 \mu\text{g/g}$.
4. IC à 95% : $0,42 \pm 0,050 = [0,370 ; 0,470] \mu\text{g/g}$.

Exemple 3 : mercure dans la chair des dorés

On mesure la concentration de mercure ($\mu\text{g/g}$) dans la chair de $n = 50$ dorés jaunes pêchés dans un réservoir. La distribution est asymétrique à droite. On obtient $\bar{x} = 0,42 \mu\text{g/g}$ et $s = 0,18 \mu\text{g/g}$.

Question : donnez un IC à 95% pour la concentration moyenne μ .

1. $n = 50$ est grand : on utilise le TLC et $z_{0,025} = 1,96$.
2. Erreur-type estimée : $s / \sqrt{n} = 0,18 / \sqrt{50} = 0,0255$.
3. Marge d'erreur : $ME = 1,96 \times 0,0255 = 0,050 \mu\text{g/g}$.
4. IC à 95% : $0,42 \pm 0,050 = [0,370 ; 0,470] \mu\text{g/g}$.

La ligne directrice de Santé Canada pour le poisson est de $0,5 \mu\text{g/g}$: l'intervalle entier est sous cette valeur. Nous verrons au chapitre 3c comment formaliser cette conclusion par un **test**.

Marge d'erreur et erreur-type : une structure commune

Les trois intervalles pour μ ont la même forme :

Forme générale

$$\text{Estimation} \pm \underbrace{\text{quantile} \times \text{erreur-type}}_{\text{marge d'erreur}} \quad \text{Longueur} = 2 \times \text{ME}$$

Situation	Estimation	Quantile	Erreur-type
Normale, σ connue	$\bar{x}$	$z_{\alpha/2}$	$\sigma / \sqrt{n}$
Normale, σ inconnue	$\bar{x}$	$t_{n-1, \alpha/2}$	$s / \sqrt{n}$
Loi inconnue, n grand	$\bar{x}$	$z_{\alpha/2}$	$s / \sqrt{n}$

Marge d'erreur et erreur-type : une structure commune

Les trois intervalles pour μ ont la même forme :

Forme générale

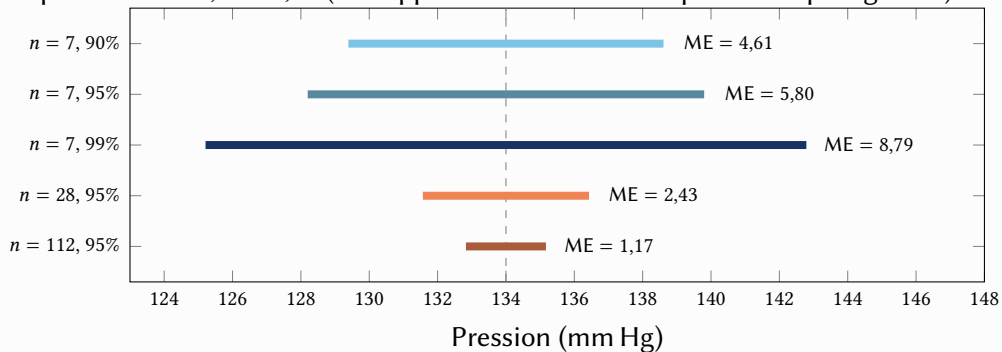
$$\text{Estimation} \pm \underbrace{\text{quantile} \times \text{erreur-type}}_{\text{marge d'erreur}} \quad \text{Longueur} = 2 \times \text{ME}$$

Situation	Estimation	Quantile	Erreur-type
Normale, σ connue	$\bar{x}$	$z_{\alpha/2}$	$\sigma / \sqrt{n}$
Normale, σ inconnue	$\bar{x}$	$t_{n-1, \alpha/2}$	$s / \sqrt{n}$
Loi inconnue, n grand	$\bar{x}$	$z_{\alpha/2}$	$s / \sqrt{n}$

Trois facteurs déterminent la marge d'erreur : la **variabilité** (σ), la **taille d'échantillon** (n) et le **niveau de confiance** (quantile).

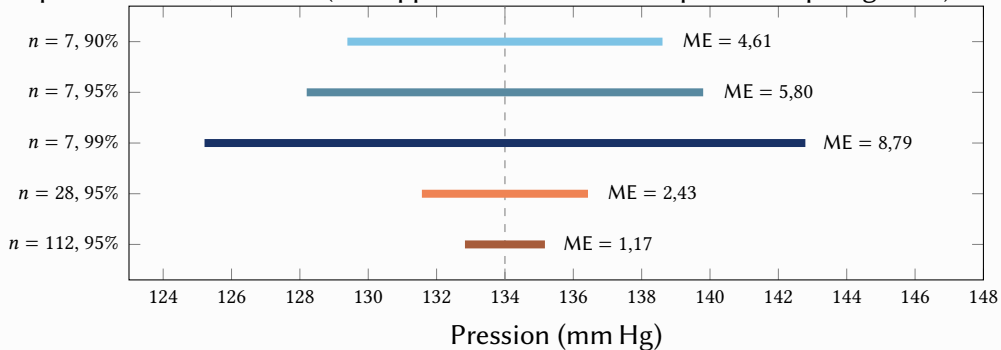
Effet du niveau de confiance et de la taille d'échantillon

Exemple 1 : $\bar{x} = 134$, $s = 6,27$ (on suppose les mêmes $\bar{x}$ et s pour les n plus grands).



Effet du niveau de confiance et de la taille d'échantillon

Exemple 1 : $\bar{x} = 134$, $s = 6,27$ (on suppose les mêmes $\bar{x}$ et s pour les n plus grands).



Plus de confiance $\Rightarrow$ intervalle **plus long**. n multiplié par 4 $\Rightarrow$ ME environ **divisée par 2**.

QUIZ!

Vrai ou faux ? La marge d'erreur d'un IC pour μ augmente (donc la précision diminue) si :

- la variabilité des mesures augmente;

QUIZ!

Vrai ou faux ? La marge d'erreur d'un IC pour μ augmente (donc la précision diminue) si :

- la variabilité des mesures augmente;
- la taille d'échantillon augmente;

QUIZ!

Vrai ou faux ? La marge d'erreur d'un IC pour μ augmente (donc la précision diminue) si :

- la variabilité des mesures augmente;
- la taille d'échantillon augmente;
- le niveau de confiance augmente;

QUIZ!

Vrai ou faux ? La marge d'erreur d'un IC pour μ augmente (donc la précision diminue) si :

- la variabilité des mesures augmente;
- la taille d'échantillon augmente;
- le niveau de confiance augmente;
- la population étudiée est plus grande (ex. : tout le Québec plutôt qu'une seule ville).

Planifier une étude : quelle taille d'échantillon ?

Avant de collecter les données, on veut que la marge d'erreur ne dépasse pas une valeur fixée E :

$$z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq E \quad \Longleftrightarrow \quad \sqrt{n} \geq \frac{z_{\alpha/2} \sigma}{E}$$

Planifier une étude : quelle taille d'échantillon ?

Avant de collecter les données, on veut que la marge d'erreur ne dépasse pas une valeur fixée E :

$$z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq E \quad \Longleftrightarrow \quad \sqrt{n} \geq \frac{z_{\alpha/2} \sigma}{E}$$

Taille d'échantillon pour estimer μ avec une marge d'erreur E

$$n \geq \left(\frac{z_{\alpha/2} \sigma}{E} \right)^2 \quad (\text{on arrondit toujours vers le haut})$$

Planifier une étude : quelle taille d'échantillon ?

Avant de collecter les données, on veut que la marge d'erreur ne dépasse pas une valeur fixée E :

$$z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq E \quad \Longleftrightarrow \quad \sqrt{n} \geq \frac{z_{\alpha/2} \sigma}{E}$$

Taille d'échantillon pour estimer μ avec une marge d'erreur E

$$n \geq \left(\frac{z_{\alpha/2} \sigma}{E} \right)^2 \quad (\text{on arrondit toujours vers le haut})$$

- On utilise $z_{\alpha/2}$ car n (donc le d.l. de Student) est encore inconnu.
- Il faut une **valeur plausible de σ** : étude pilote, littérature, données antérieures.

Exemple 1 (suite) : taille d'échantillon

Une étude antérieure indique que l'écart-type de la pression systolique est d'environ $\sigma = 8$ mm Hg. Combien d'adultes faut-il mesurer pour estimer μ avec une marge d'erreur d'au plus 2 mm Hg, à un niveau de confiance de 95%?

Exemple 1 (suite) : taille d'échantillon

Une étude antérieure indique que l'écart-type de la pression systolique est d'environ $\sigma = 8$ mm Hg. Combien d'adultes faut-il mesurer pour estimer μ avec une marge d'erreur d'au plus 2 mm Hg, à un niveau de confiance de 95% ?

$$n \geq \left(\frac{1,96 \times 8}{2} \right)^2 = 7,84^2 = 61,47 \quad \Rightarrow \quad \mathbf{n = 62.}$$

Exemple 1 (suite) : taille d'échantillon

Une étude antérieure indique que l'écart-type de la pression systolique est d'environ $\sigma = 8$ mm Hg. Combien d'adultes faut-il mesurer pour estimer μ avec une marge d'erreur d'au plus **2 mm Hg**, à un niveau de confiance de 95%?

$$n \geq \left(\frac{1,96 \times 8}{2} \right)^2 = 7,84^2 = 61,47 \quad \Rightarrow \quad \mathbf{n = 62.}$$

Et pour une marge d'erreur d'au plus 1 mm Hg?

$$n \geq \left(\frac{1,96 \times 8}{1} \right)^2 = 245,86 \quad \Rightarrow \quad \mathbf{n = 246.}$$

Exemple 1 (suite) : taille d'échantillon

Une étude antérieure indique que l'écart-type de la pression systolique est d'environ $\sigma = 8$ mm Hg. Combien d'adultes faut-il mesurer pour estimer μ avec une marge d'erreur d'au plus **2 mm Hg**, à un niveau de confiance de 95%?

$$n \geq \left(\frac{1,96 \times 8}{2} \right)^2 = 7,84^2 = 61,47 \quad \Rightarrow \quad \mathbf{n = 62.}$$

Et pour une marge d'erreur d'au plus 1 mm Hg?

$$n \geq \left(\frac{1,96 \times 8}{1} \right)^2 = 245,86 \quad \Rightarrow \quad \mathbf{n = 246.}$$

⚠ Diviser la marge d'erreur par 2 exige environ **4 fois plus** de sujets : la précision coûte cher!

Intervalle de confiance pour une proportion

IC pour une proportion p

Rappel (chapitre 3a) : $\hat{P} = Y/n$ avec $Y \sim \text{Binomiale}(n, p)$, $E[\hat{P}] = p$ et $ET(\hat{P}) = \sqrt{p(1-p)/n}$.

IC pour une proportion p

Rappel (chapitre 3a) : $\hat{P} = Y/n$ avec $Y \sim \text{Binomiale}(n, p)$, $E[\hat{P}] = p$ et $ET(\hat{P}) = \sqrt{p(1-p)/n}$.

Par le TLC (approximation normale de la binomiale), si n est grand :
 $\hat{P} \approx \mathcal{N}\left(p, \frac{p(1-p)}{n}\right)$.

IC pour une proportion p

Rappel (chapitre 3a) : $\hat{P} = Y/n$ avec $Y \sim \text{Binomiale}(n, p)$, $E[\hat{P}] = p$ et $ET(\hat{P}) = \sqrt{p(1-p)/n}$.

Par le TLC (approximation normale de la binomiale), si n est grand :
 $\hat{P} \approx \mathcal{N}\left(p, \frac{p(1-p)}{n}\right)$.

IC de niveau approximatif $1 - \alpha$ pour p

$$\left[\hat{p} - z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right]$$

Conditions d'utilisation : $n\hat{p} \geq 5$ et $n(1-\hat{p}) \geq 5$.

IC pour une proportion p

Rappel (chapitre 3a) : $\hat{P} = Y/n$ avec $Y \sim \text{Binomiale}(n, p)$, $E[\hat{P}] = p$ et $ET(\hat{P}) = \sqrt{p(1-p)/n}$.

Par le TLC (approximation normale de la binomiale), si n est grand : $\hat{P} \approx \mathcal{N}\left(p, \frac{p(1-p)}{n}\right)$.

IC de niveau approximatif $1 - \alpha$ pour p

$$\left[\hat{p} - z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right]$$

Conditions d'utilisation : $n\hat{p} \geq 5$ et $n(1-\hat{p}) \geq 5$.

Comme p est inconnue, on la remplace par $\hat{p}$ dans l'erreur-type.

Exemple 2 (suite) : le sondage

Sur $n = 1028$ Québécois, $y = 216$ citent l'environnement comme principal enjeu.

Exemple 2 (suite) : le sondage

Sur $n = 1028$ Québécois, $y = 216$ citent l'environnement comme principal enjeu.

1. Estimation : $\hat{p} = 216/1028 = 0,2101$. Conditions : $216 \geq 5$ et $812 \geq 5$ ✓.
2. Erreur-type : $\sqrt{0,2101 \times 0,7899/1028} = 0,0127$.
3. Marge d'erreur à 95% : $1,96 \times 0,0127 = 0,0249$.
4. IC à 95% : $0,2101 \pm 0,0249 = [0,185 ; 0,235]$, soit de 18,5% à 23,5%.

Exemple 2 (suite) : le sondage

Sur $n = 1028$ Québécois, $y = 216$ citent l'environnement comme principal enjeu.

1. Estimation : $\hat{p} = 216/1028 = 0,2101$. Conditions : $216 \geq 5$ et $812 \geq 5$ ✓.
2. Erreur-type : $\sqrt{0,2101 \times 0,7899/1028} = 0,0127$.
3. Marge d'erreur à 95% : $1,96 \times 0,0127 = 0,0249$.
4. IC à 95% : $0,2101 \pm 0,0249 = [0,185 ; 0,235]$, soit de 18,5% à 23,5%.

« Marge d'erreur de 3,1%, 19 fois sur 20 »

Les sondeurs annoncent la marge d'erreur **maximale**, obtenue avec $\hat{p} = 0,5$: $1,96\sqrt{0,5 \times 0,5/1028} = 0,031$. « 19 fois sur 20 » signifie un niveau de confiance de 95%.

Exemple 4 : prévalence d'une infection chez des veaux

Sur 80 veaux examinés dans un troupeau, 12 présentent des symptômes d'infection (chapitre 3a).

1. Estimez la prévalence p de l'infection dans le troupeau.
2. Les conditions d'utilisation de l'IC approximatif sont-elles respectées ?
3. Construisez un IC à 95% pour p et interprétez-le.
4. Le vétérinaire affirme que « moins de 5% des veaux sont infectés ». Cette affirmation est-elle plausible ?

Exemple 4 : prévalence d'une infection chez des veaux

Sur 80 veaux examinés dans un troupeau, 12 présentent des symptômes d'infection (chapitre 3a).

1. Estimez la prévalence p de l'infection dans le troupeau.
2. Les conditions d'utilisation de l'IC approximatif sont-elles respectées ?
3. Construisez un IC à 95% pour p et interprétez-le.
4. Le vétérinaire affirme que « moins de 5% des veaux sont infectés ». Cette affirmation est-elle plausible ?

(Réponses à la fin du chapitre.)

Taille d'échantillon pour une proportion

On veut que $z_{\alpha/2} \sqrt{p(1-p)/n} \leq E$, d'où

Taille d'échantillon pour estimer p avec une marge d'erreur E

$$n \geq \frac{z_{\alpha/2}^2 p^*(1-p^*)}{E^2}$$

où p^* est une valeur plausible de p . **Sans information**, on prend $p^* = 0,5$ (cas le plus défavorable, car $p(1-p) \leq 0,25$).

Taille d'échantillon pour une proportion

On veut que $z_{\alpha/2} \sqrt{p(1-p)/n} \leq E$, d'où

Taille d'échantillon pour estimer p avec une marge d'erreur E

$$n \geq \frac{z_{\alpha/2}^2 p^*(1-p^*)}{E^2}$$

où p^* est une valeur plausible de p . **Sans information**, on prend $p^* = 0,5$ (cas le plus défavorable, car $p(1-p) \leq 0,25$).

Exemple 5 : on veut estimer la proportion de tiques porteuses de la bactérie de la maladie de Lyme dans un parc, à $\pm 5\%$ près (95%).

Taille d'échantillon pour une proportion

On veut que $z_{\alpha/2} \sqrt{p(1-p)/n} \leq E$, d'où

Taille d'échantillon pour estimer p avec une marge d'erreur E

$$n \geq \frac{z_{\alpha/2}^2 p^*(1-p^*)}{E^2}$$

où p^* est une valeur plausible de p . **Sans information**, on prend $p^* = 0,5$ (cas le plus défavorable, car $p(1-p) \leq 0,25$).

Exemple 5 : on veut estimer la proportion de tiques porteuses de la bactérie de la maladie de Lyme dans un parc, à $\pm 5\%$ près (95%).

- Sans information : $n \geq 1,96^2(0,5)(0,5)/0,05^2 = 384,2 \Rightarrow n = 385$.

Taille d'échantillon pour une proportion

On veut que $z_{\alpha/2} \sqrt{p(1-p)/n} \leq E$, d'où

Taille d'échantillon pour estimer p avec une marge d'erreur E

$$n \geq \frac{z_{\alpha/2}^2 p^*(1-p^*)}{E^2}$$

où p^* est une valeur plausible de p . **Sans information**, on prend $p^* = 0,5$ (cas le plus défavorable, car $p(1-p) \leq 0,25$).

Exemple 5 : on veut estimer la proportion de tiques porteuses de la bactérie de la maladie de Lyme dans un parc, à $\pm 5\%$ près (95%).

- Sans information : $n \geq 1,96^2(0,5)(0,5)/0,05^2 = 384,2 \Rightarrow \mathbf{n = 385}$.
- Si une étude voisine suggère $p \approx 0,20$:
 $n \geq 1,96^2(0,2)(0,8)/0,05^2 = 245,9 \Rightarrow \mathbf{n = 246}$.

Avec R : `prop.test()` et `binom.test()`

```
> prop.test(216, 1028)$conf.int      # Wilson  
[1] 0.1858517 0.2365822  
> binom.test(216, 1028)$conf.int    # exact  
[1] 0.1855853 0.2363200
```

- R n'utilise pas exactement la formule vue en classe, mais des méthodes plus précises lorsque n est petit ou $\hat{p}$ proche de 0 ou 1 (Wilson avec correction de continuité; Clopper-Pearson, dit « exact »).
- Avec $n = 1028$, les trois méthodes donnent pratiquement le même intervalle : $[0,185 ; 0,235]$ (formule), $[0,186 ; 0,237]$ (Wilson), $[0,186 ; 0,236]$ (exact).
- Pour un petit échantillon, préférer `binom.test()`.

QUIZ!

Vrai ou faux?

- Un sondage auprès de 1000 Québécois (8,9 millions d'habitants) est beaucoup moins précis qu'un sondage auprès de 1000 Madelinots (12 000 habitants).

QUIZ!

Vrai ou faux?

- Un sondage auprès de 1000 Québécois (8,9 millions d'habitants) est beaucoup moins précis qu'un sondage auprès de 1000 Madelinots (12 000 habitants).
- Si un sondage donne $\hat{p} = 0,10$ avec $n = 1000$, la marge d'erreur est plus petite qu'avec $\hat{p} = 0,50$.

QUIZ!

Vrai ou faux?

- Un sondage auprès de 1000 Québécois (8,9 millions d'habitants) est beaucoup moins précis qu'un sondage auprès de 1000 Madelinots (12 000 habitants).
- Si un sondage donne $\hat{p} = 0,10$ avec $n = 1000$, la marge d'erreur est plus petite qu'avec $\hat{p} = 0,50$.
- Sur 20 souris traitées, 1 développe une tumeur. L'IC à 95% « $\hat{p} \pm 1,96\sqrt{\hat{p}(1 - \hat{p})/n}$ » est approprié.

Intervalle de confiance pour σ^2 et σ

Pourquoi estimer une variance ?

En biologie, la **variabilité** est souvent l'objet d'intérêt :

- uniformité de la dose de principe actif dans des comprimés;
- homogénéité d'une lignée de souris de laboratoire;
- stabilité d'un appareil de mesure (précision d'un tensiomètre);
- variabilité du rendement d'un cultivar d'une parcelle à l'autre.

Pourquoi estimer une variance ?

En biologie, la **variabilité** est souvent l'objet d'intérêt :

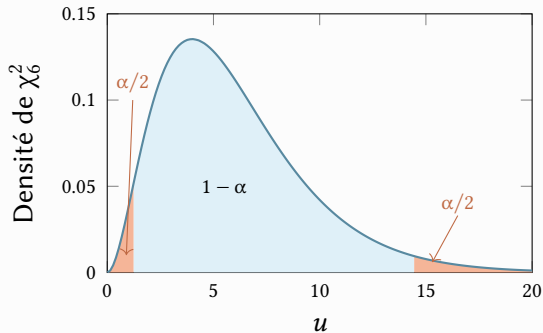
- uniformité de la dose de principe actif dans des comprimés;
- homogénéité d'une lignée de souris de laboratoire;
- stabilité d'un appareil de mesure (précision d'un tensiomètre);
- variabilité du rendement d'un cultivar d'une parcelle à l'autre.

Rappel : la loi du khi-deux (chapitre 3a)

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

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

Quantiles de la loi du khi-deux



- Loi **asymétrique**, à valeurs positives, d'espérance $n - 1$.
- On note $\chi_{n-1, \alpha/2}^2$ le quantile qui laisse $\alpha/2$ **à droite**.
- Il faut **deux** quantiles distincts : $\chi_{n-1, 1-\alpha/2}^2$ et $\chi_{n-1, \alpha/2}^2$.
- Ex. ($n = 7, 95\%$) : $\chi_{6; 0,975}^2 = 1,237$ et $\chi_{6; 0,025}^2 = 14,449$.

Construction de l'IC pour σ^2

On part de

$$P \left(\chi_{n-1, 1-\alpha/2}^2 \leq \frac{(n-1)S^2}{\sigma^2} \leq \chi_{n-1, \alpha/2}^2 \right) = 1 - \alpha$$

Construction de l'IC pour σ^2

On part de

$$P \left(\chi_{n-1, 1-\alpha/2}^2 \leq \frac{(n-1)S^2}{\sigma^2} \leq \chi_{n-1, \alpha/2}^2 \right) = 1 - \alpha$$

et on isole σ^2 (attention : inverser une inégalité change son sens!).

Construction de l'IC pour σ^2

On part de

$$P \left(\chi_{n-1, 1-\alpha/2}^2 \leq \frac{(n-1)S^2}{\sigma^2} \leq \chi_{n-1, \alpha/2}^2 \right) = 1 - \alpha$$

et on isole σ^2 (attention : inverser une inégalité change son sens!).

IC de niveau $1 - \alpha$ pour σ^2 et pour σ (population normale)

$$\sigma^2 \in \left[\frac{(n-1)S^2}{\chi_{n-1, \alpha/2}^2}, \frac{(n-1)S^2}{\chi_{n-1, 1-\alpha/2}^2} \right] \quad \sigma \in \left[\sqrt{\frac{(n-1)S^2}{\chi_{n-1, \alpha/2}^2}}, \sqrt{\frac{(n-1)S^2}{\chi_{n-1, 1-\alpha/2}^2}} \right]$$

Construction de l'IC pour σ^2

On part de

$$P \left(\chi_{n-1, 1-\alpha/2}^2 \leq \frac{(n-1)S^2}{\sigma^2} \leq \chi_{n-1, \alpha/2}^2 \right) = 1 - \alpha$$

et on isole σ^2 (attention : inverser une inégalité change son sens!).

IC de niveau $1 - \alpha$ pour σ^2 et pour σ (population normale)

$$\sigma^2 \in \left[\frac{(n-1)S^2}{\chi_{n-1, \alpha/2}^2}, \frac{(n-1)S^2}{\chi_{n-1, 1-\alpha/2}^2} \right] \quad \sigma \in \left[\sqrt{\frac{(n-1)S^2}{\chi_{n-1, \alpha/2}^2}}, \sqrt{\frac{(n-1)S^2}{\chi_{n-1, 1-\alpha/2}^2}} \right]$$

Le **grand** quantile donne la borne **inférieure**. Pour σ , on prend la racine carrée des bornes.

Exemple 1 (suite) : variabilité de la pression

$n = 7$, $s^2 = 39,33$, donc $(n - 1)s^2 = 236$. IC à 95% pour σ^2 et σ ?

Exemple 1 (suite) : variabilité de la pression

$n = 7$, $s^2 = 39,33$, donc $(n - 1)s^2 = 236$. IC à 95% pour σ^2 et σ ?

1. Quantiles (6 d.l.) : $\chi^2_{6;0,025} = 14,449$ et $\chi^2_{6;0,975} = 1,237$.

2. IC pour σ^2 : $\left[\frac{236}{14,449} ; \frac{236}{1,237} \right] = [16,33 ; 190,73] \text{ (mm Hg)}^2$.

3. IC pour σ : $\left[\sqrt{16,33} ; \sqrt{190,73} \right] = [4,04 ; 13,81] \text{ mm Hg}$.

Exemple 1 (suite) : variabilité de la pression

$n = 7$, $s^2 = 39,33$, donc $(n - 1)s^2 = 236$. IC à 95% pour σ^2 et σ ?

1. Quantiles (6 d.l.) : $\chi^2_{6;0,025} = 14,449$ et $\chi^2_{6;0,975} = 1,237$.

2. IC pour σ^2 : $\left[\frac{236}{14,449} ; \frac{236}{1,237} \right] = [16,33 ; 190,73] \text{ (mm Hg)}^2$.

3. IC pour σ : $\left[\sqrt{16,33} ; \sqrt{190,73} \right] = [4,04 ; 13,81] \text{ mm Hg}$.

Attention!

- L'intervalle **n'est pas symétrique** autour de $s^2 = 39,33$ (ni de $s = 6,27$).
- Avec $n = 7$, il est **très large** : estimer une variance demande beaucoup de données.
- Cette méthode est **sensible à la normalité**, même quand n est grand (pas de TLC ici).

Exemple 6 : homogénéité d'une lignée de souris

Une animalerie produit une lignée consanguine de souris, censée être très homogène. On pèse $n = 20$ souris de 8 semaines et on obtient un écart-type $s = 1,6$ g. On suppose que les masses suivent une loi normale.

1. Donnez un IC à 95% pour la variance σ^2 des masses.
2. Déduisez-en un IC à 95% pour l'écart-type σ .
3. Le fournisseur affirme que $\sigma = 1$ g. Est-ce plausible ?

Indices : $\chi^2_{19;0,025} = 32,852$ et $\chi^2_{19;0,975} = 8,907$.

Dans R : `qchisq(c(0.975, 0.025), df = 19)`.

QUIZ!

Vrai ou faux?

- L'IC pour σ^2 est de la forme $s^2 \pm \text{marge d'erreur}$.

QUIZ!

Vrai ou faux?

- L'IC pour σ^2 est de la forme $s^2 \pm \text{marge d'erreur}$.
- Si $[16,33 ; 190,73]$ est un IC à 95% pour σ^2 , alors $[4,04 ; 13,81]$ est un IC à 95% pour σ .

QUIZ!

Vrai ou faux ?

- L'IC pour σ^2 est de la forme $s^2 \pm \text{marge d'erreur}$.
- Si $[16,33 ; 190,73]$ est un IC à 95% pour σ^2 , alors $[4,04 ; 13,81]$ est un IC à 95% pour σ .
- Avec $n = 200$ mesures, l'IC pour σ^2 basé sur le khi-deux reste valide même si la population est très asymétrique.

Synthèse

Tableau récapitulatif des intervalles de confiance

Paramètre et situation	IC de niveau $1 - \alpha$	R
μ ; normale, σ connue	$\bar{x} \pm z_{\alpha/2} \sigma / \sqrt{n}$	<code>qnorm()</code>
μ ; normale, σ inconnue	$\bar{x} \pm t_{n-1, \alpha/2} s / \sqrt{n}$	<code>t.test()</code>
μ ; loi inconnue, n grand	$\bar{x} \pm z_{\alpha/2} s / \sqrt{n}$	<code>t.test()</code>
p ; $n\hat{p} \geq 5$, $n(1 - \hat{p}) \geq 5$	$\hat{p} \pm z_{\alpha/2} \sqrt{\hat{p}(1 - \hat{p})/n}$	<code>prop.test()</code> , <code>binom.test()</code>
σ^2 ; normale	$\left[\frac{(n-1)s^2}{\chi_{n-1, \alpha/2}^2}; \frac{(n-1)s^2}{\chi_{n-1, 1-\alpha/2}^2} \right]$	<code>qchisq()</code>

Tailles d'échantillon : $n \geq (z_{\alpha/2} \sigma / E)^2$ pour μ ; $n \geq z_{\alpha/2}^2 p^*(1 - p^*)/E^2$ pour p .

Prochaine étape (chapitre 3c) : utiliser ces mêmes lois pour **tester des hypothèses**.

Réponses aux exemples

- **Quiz (simulation)** : on s'attend à $40 \times 0,10 = 4$ intervalles qui ratent μ ; à 95%, 2 en moyenne, avec des intervalles plus longs; les longueurs varient car s change d'un échantillon à l'autre.
- **Quiz vrai/faux (énoncés 1 à 7)** : 1 F; 2 F; 3 F; 4 V; 5 V (en pratique); 6 F; 7 F.
- **Exemple 1** : IC à 90% : [129,39; 138,61]; à 95% : [128,20; 139,80]; à 99% : [125,21; 142,79] mm Hg. Taille : $n = 62$ ($ME \leq 2$) et $n = 246$ ($ME \leq 1$). IC à 95% pour σ : [4,04; 13,81] mm Hg.
- **Exemple 3 (mercure)** : [0,370; 0,470] $\mu\text{g/g}$.
- **Quiz (marge d'erreur)** : V; F; V; F (la taille de la population n'intervient pas, tant qu'elle est grande devant n).
- **Exemple 2 (sondage)** : [0,185; 0,235]; marge maximale 3,1%.
- **Exemple 4 (veaux)** : $\hat{p} = 0,15$; $n\hat{p} = 12$ et $n(1 - \hat{p}) = 68$: conditions respectées; $ET = 0,0399$; IC à 95% : $0,15 \pm 0,078 = [0,072; 0,228]$. La valeur 5% est sous la borne inférieure : l'affirmation est peu plausible.
- **Exemple 5 (tiques)** : $n = 385$ sans information; $n = 246$ si $p \approx 0,20$.
- **Quiz (proportions)** : F (la précision dépend de n , presque pas de la taille de la population); V ($0,1 \times 0,9 < 0,5 \times 0,5$); F ($n\hat{p} = 1 < 5$: utiliser `binom.test()`).
- **Exemple 6 (souris)** : $(n - 1)s^2 = 48,64$; IC pour σ^2 : [1,48; 5,46] g^2 ; IC pour σ : [1,22; 2,34] g. La valeur $\sigma = 1$ g est hors de l'intervalle : peu plausible.
- **Quiz (variance)** : F (intervalle non symétrique); V; F (la méthode exige la normalité).