Statistiques appliquées à la biologie : estimation.

Alexandre Popier

L3 BBTE, Université du Maine, Le Mans

Semestre 1

Plan du cours

1. Intervalle de confiance d'une proportion p
   - Résumer l'information, loi binomiale
   - Moyenne empirique
   - Intervalles de confiance

2. Estimation d'une moyenne
   - Si n est assez grand ($n \geq 30$)
   - Si n est petit...
   - Commentaires

Exemples

Exemple 1 : Contrôle qualité. Lors de la production

- Proportion inconnue d'objets défectueux $p \in [0, 1]$.
- Proportion évaluée par une fréquence d'apparition d'objets mal fabriqués dans des échantillons contrôlés au hasard.
  - Dépassement d'un premier seuil p_S : renforcement de la surveillance de la production (augmenter la taille des échantillons utilisés pour évaluer p).
  - Dépassement d'un autre seuil $p_C > p_S$: arrêt de la production pour réparer ou pour régler les machines.
- Tenir compte du coût respectif de chaque type d'opération pour définir les seuils précédents.

Exemples

Exemple 2 : fréquences cardiaques :

$$64; 67; 72; 58; 60; 65; 64; 57; 72; 66; 65; 59; 66;$$

$$63; 62; 64; 62; 66; 60; 61; 59; 62; 64; 61.$$

2/3 = 66% des données de cet échantillon sont inférieures (strictes) à 65.

Qu'en déduire pour la population ?

Exemples

Exemple 3 : Sondages sur l'élection présidentielle.

Pour 2012 :

- fait les 8 et 9 juillet 2011 par LH2 et publié par Yahoo et Le Monde,
- sur une population de 957 personnes âgées de plus de 18 ans, dont 827 inscrites sur les listes électorales.
- Résultat au second tour : Hollande 60 %, Sarkozy 40 %.

Pour 2007 :

- fait le 30 avril 2007 par LH2, sur 900 personnes,
- Résultat au second tour : Sarkozy 52 %, Royal 48 %.

Pour 2002 :

- fait le 19 avril 2002 par BVA, sur 1000 personnes,
- Résultat au premier tour : Chirac 19 %, Jospin 18 %, Le Pen 14 %.

Estimation

Question : est-ce que toutes ces proportions calculées sur un échantillon peuvent être extrapolées à la population entière ?

Définition

Chercher à connaître les valeurs de certaines grandeurs grâce à des observations réalisées sur un échantillon.

Deux types de réponses :

- Produire une valeur qui nous semble être la meilleure possible : estimation ponctuelle.
- Produire un intervalle de valeurs possibles, compatibles avec les observations : notion d'intervalle de confiance.

Intervalle de confiance

- Pas moyen d'être sûr de ce que vaut réellement la proportion dans la population !
- Au mieux calculer une étendue de valeurs qui inclut la vraie proportion de la population. Comment la choisir ?

Réponse excessive : $[0,00001\%; 99,99999\%]$, $\mathbb{R}$ ou $[0, +\infty[$.

Étendue plus étroite et plus utile :

- accepter la possibilité que l'étendue ne couvre pas la valeur vraie de la population !
- Marge d'erreur, ou encore probabilité que l'intervalle ne couvre pas la valeur vraie.
  - Exprimée en pourcentage, notée α. Classiquement $\alpha = 5\%$.
  - Intervalle appelé intervalle de confiance à $100 - \alpha\%$, noté $IC(100 - \alpha)$:
    « Certain à $100 - \alpha\%$ » que $IC(100 - \alpha)$ contient la vraie valeur.

Conditions techniques

Interprétation correcte si

1. Échantillon aléatoire simple ou représentatif.
2. Observations indépendantes.
3. Classification correcte.
4. Estimation de l'événement réellement intéressant.

Plan

1. Intervalle de confiance d'une proportion p
   - Résumer l'information, loi binomiale
   - Moyenne empirique
   - Intervalles de confiance
2. Estimation d'une moyenne
   - Si n est assez grand ($n \geq 30$)
   - Si n est petit...
   - Commentaires

RETOUR SUR LES EXEMPLES

- FRÉQUENCES CARDIAQUES : 1 si la fréquence est inférieure à 65, 0 sinon :

$$1; 0; 0; 1; 1; 0; 1; 1; 0; 0; 0; 1; 0; 1; 1; 1; 1; 0; 1; 1; 1; 1; 1; 1.$$

  La proportion sur l'échantillon est donc la somme de toutes ces valeurs divisée par la taille de l'échantillon :

$$\frac{16}{24} = \frac{2}{3}.$$

- CONTRÔLE QUALITÉ : l'opérateur comptera 1 si un objet est défectueux, 0 sinon.
- SONDAGE : le sondeur compte 1 si l'électeur se prononce pour Hollande, 0 si c'est pour Sarkozy.

Plan

1. Intervalle de confiance d'une proportion p
   - Résumer l'information, loi binomiale
   - Moyenne empirique
   - Intervalles de confiance
2. Estimation d'une moyenne
   - Si n est assez grand ($n \geq 30$)
   - Si n est petit...
   - Commentaires

Estimateur et loi binomiale

But : évaluer $p \in [0,1]$ à partir de l'observation $(x_1, \ldots, x_n)$, donc de n données.

Définition (Estimateur)

On appelle estimateur toute fonction de l'observation $h(x_1, \ldots, x_n)$.

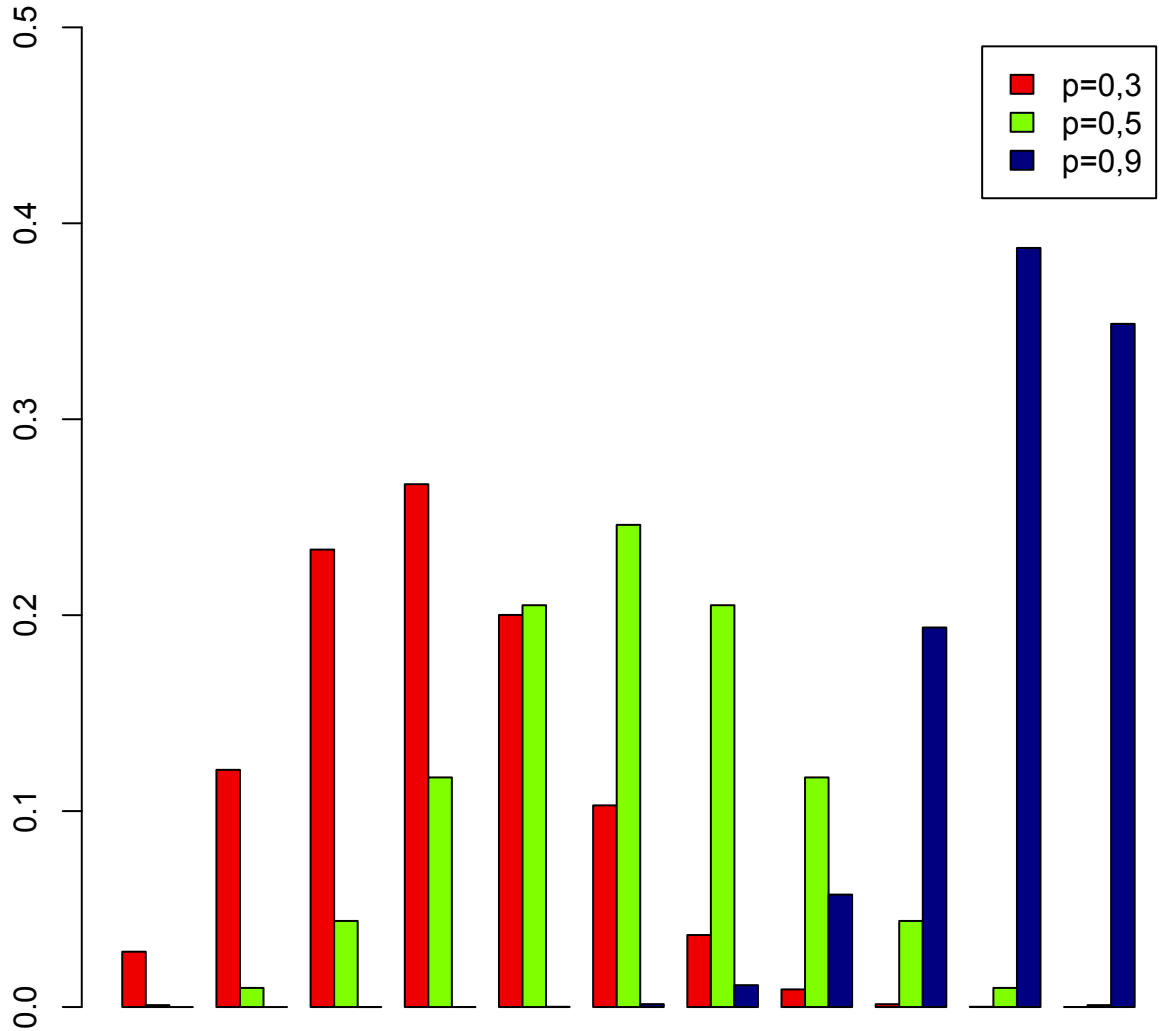
Ici x_i vaut 1 ou 0. Posons

$$\boxed{s_n = x_1 + \cdots + x_n.}$$

Loi théorique de s_n : binomiale de paramètres n et p :

$$\mathbb{P}(s_n = k) = C_n^k p^k (1-p)^{n-k}, \qquad k = 0, 1, \ldots, n.$$

ESTIMATEUR ET LOI BINOMIALE



Estimateur exhaustif

Calcul des probabilités (admis) :

- Si la somme s_n prend la valeur k, alors la configuration des x_i tels que
$$\sum_i x_i = s_n = k$$
n'apporte aucune information sur le paramètre inconnu p.
- On n'en apprendra pas plus de tout le résultat $(x_1, \ldots, x_n)$ de notre expérience aléatoire (qui peut prendre 2^n valeurs), que du résumé d'information constitué par s_n qui prend seulement $n+1$ valeurs.

s_n est appelé estimateur exhaustif de l'expérience $(x_1, \ldots, x_n)$: il rapporte toute l'information relative à p contenue dans notre expérience.

Plan

1. Intervalle de confiance d'une proportion p
   - Résumer l'information, loi binomiale
   - Moyenne empirique
   - Intervalles de confiance
2. Estimation d'une moyenne
   - Si n est assez grand ($n \geq 30$)
   - Si n est petit...
   - Commentaires

Moyenne empirique : propriétés

Fréquence (ou proportion ou moyenne) empirique :

$$\boxed{\widehat{p}_n = \frac{s_n}{n} = \frac{x_1 + \cdots + x_n}{n}.}$$

Cette valeur est calculée sur l'échantillon.

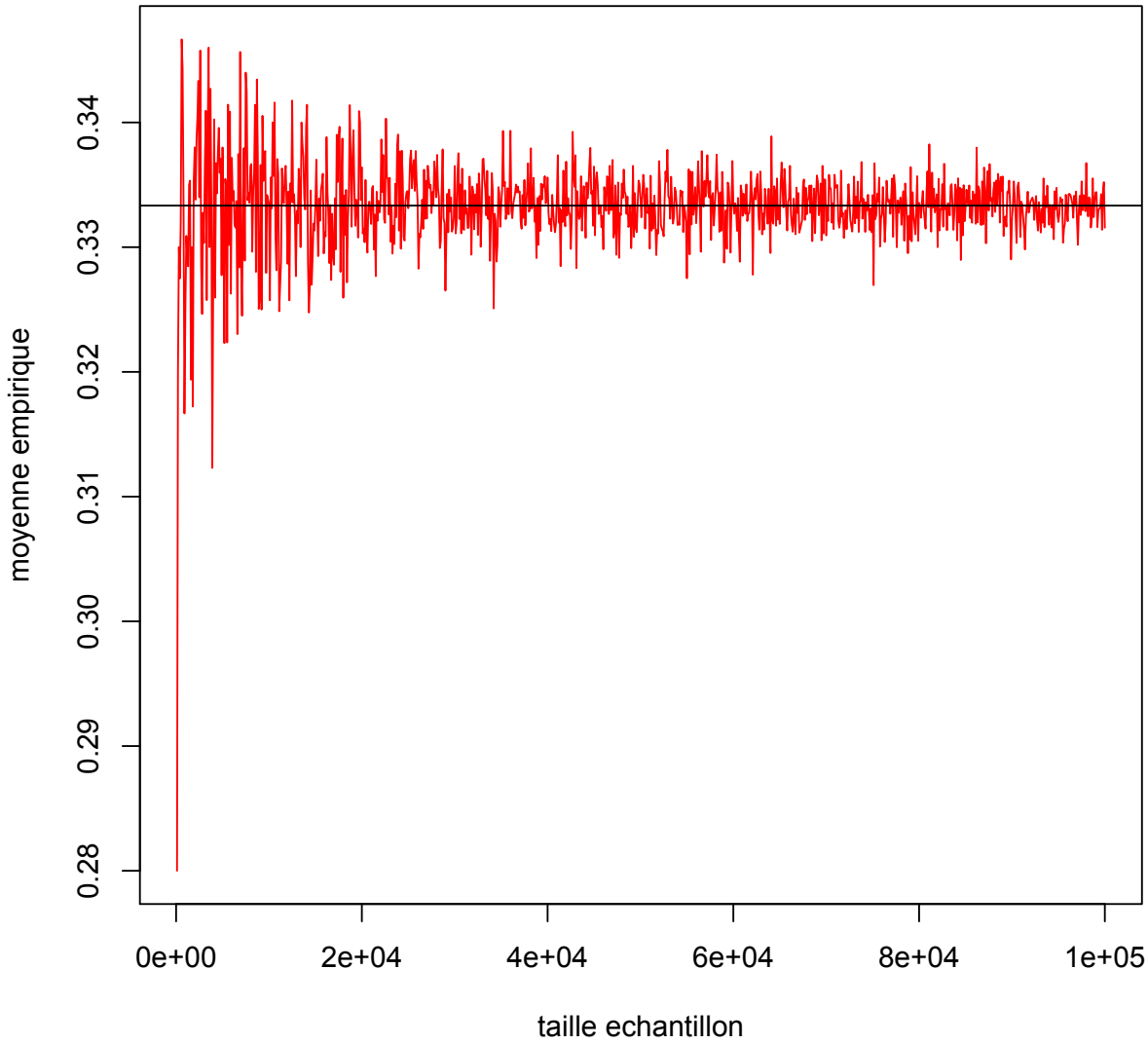
C'est un estimateur consistant :

Théorème (Loi des grands nombres (Bernoulli, 1690))

$$\lim_{n \to +\infty} \widehat{p}_n = p.$$

Remarque : pourquoi cet estimateur ? C'est le « meilleur » car il maximise la vraisemblance.

MOYENNE EMPIRIQUE : PROPRIÉTÉS



Plan

1. Intervalle de confiance d'une proportion p
   - Résumer l'information, loi binomiale
   - Moyenne empirique
   - Intervalles de confiance
2. Estimation d'une moyenne
   - Si n est assez grand ($n \geq 30$)
   - Si n est petit...
   - Commentaires

Contrôle de l'erreur d'estimation

- La moyenne empirique fluctue en permanence autour de la valeur asymptotique p. Même si ces fluctuations s'amoindrissent, elles continuent d'exister même pour $n = 10^5$.
- Il convient donc de quantifier l'erreur commise par l'estimation ponctuelle :

$$\widehat{p}_n - p.$$

- Impossible de donner une erreur absolue $\Rightarrow$ intervalle de confiance.

Calcul de l'intervalle de confiance :

1. de manière exacte via la loi biomiale ;
2. ou asymptotique avec la loi gaussienne.

Intervalles de confiance « exacts »

1. Soit α une marge d'erreur donnée, par exemple 5% = 0,05.
2. Chercher deux nombres a et b compris entre 0 et 1, tels que

$$1 - \alpha = \mathbb{P}(a \leq \widehat{p}_n \leq b) = \mathbb{P}(an \leq s_n \leq bn).$$

Intervalle de confiance au niveau de confiance $1 - \alpha$:

$$p \in IC(1 - \alpha) = [a, b].$$

Classiquement choisir a et b tels que :

$$\mathbb{P}(s_n \leq an) = \alpha/2, \qquad \mathbb{P}(bn \leq s_n) = \alpha/2,$$

en supposant que s_n suit une loi binomiale de paramètres n et $\widehat{p}_n$.

- Équation dure à résoudre.
- Utilisation de tables ou de la fonction `qbinom` sous R.

Intervalles de confiance « exacts »

Exemple des fréquences cardiaques.

- 24 données avec une estimation de p par $\widehat{p}_{24} = 66\%$.
- Marge d'erreur α de 5%.
- Code R permettant de calculer a et b.

```
a<-qbinom(0.025, 24, 2/3)/24
b<-qbinom(0.025, 24, 2/3,lower.tail=FALSE)/24
```

- $a = 0,46$ et $b = 0,83$: $p \in [0,46; 0,83]$. L'intervalle de confiance n'est pas symétrique par rapport à 66%.

Exemple des sondages.

- $n = 957$, p estimée par $\widehat{p}_{957} = 60\%$.
- Marge d'erreur de 5%.
- $a = 56,8\%$ et $b = 63,1\%$.

Intervalles de confiance asymptotiques

Avec données (0 ou 1) de l'échantillon :

- calcul de la fréquence empirique $\widehat{p}_n$,
- calcul de l'écart-type empirique : $\widehat{e}_n = \sqrt{\widehat{p}_n(1-\widehat{p}_n)}$.
  C'est une fonction de la fréquence empirique.

Théorème central limite :

$$\lim_{n\to+\infty} \mathbb{P}\left(a < \sqrt{n}\frac{(\widehat{p}_n - p)}{\widehat{e}_n} < b \right) = \mathbb{P}(a < Z < b),$$

où Z suit une distribution normale ou gaussienne.

- Z ne dépend pas de p !

INTERVALLE DE CONFIANCE ASYMPTOTIQUE : pour n « grand »

$$p \in IC(1-\alpha) \approx \left[\widehat{p}_n - \varphi_\alpha \sqrt{\frac{\widehat{p}_n(1-\widehat{p}_n)}{n}}, \widehat{p}_n + \varphi_\alpha \sqrt{\frac{\widehat{p}_n(1-\widehat{p}_n)}{n}}\right],$$

avec

$$\mathbb{P}(-\varphi_\alpha < Z < \varphi_\alpha) = 1-\alpha.$$

φ_α :

- Unique nombre réel tel que $\mathbb{P}(|Z| > \varphi_\alpha) = \alpha$.
- Pour $1-\alpha = 95\%$, $\varphi_\alpha = 1,96$.
- Pour obtenir cette valeur avec R : `qnorm`
  `qnorm(0.025,0,1,lower.tail=FALSE)`
- Sous Excel : `loi.normale.inverse(0,975;0;1)`

Remarques sur les intervalles de confiance

Taille de l'intervalle de confiance :

- Augmente avec le niveau de confiance $1 - \alpha$.
- Diminue avec le nombre de données n (en $\sqrt{n}$ pour être précis).

Exact ou approché ? Que signifie n « grand » ?

- En pratique, on se contente traditionnellement de supposer

$$n \geq 30, \quad np \geq 5, \quad n(1-p) \geq 5.$$

Illustration sur des exemples.

- Fréquences cardiaques : 24 données, moyenne empirique 66%.
  - IC(95%) exact : $[46\%, 83\%]$.
  - IC(80%) exact : $[54,2\%, 79,2\%]$.
  - IC(95%) approché : $[47\%; 85\%]$.

Exemple des sondages

- Pour 2012 : Hollande 60 %, Sarkozy 40 % (957 données).
  - IC(95%) exact : $[56,8\%, 63,1\%]$.
  - IC(95%) approché : $[56,8\%; 63,1\%]$.
  - IC(99%) $[56,4\%, 63,5\%]$.
- Pour 2007 : Sarkozy 52 %, Royal 48 %.
  - IC(95%) = $[48,74\%; 55,26\%]$.
- Pour 2002 : Chirac 19 %, Jospin 18 %, Le Pen 14 %.
  - IC(95%, Chirac) = $[16,57; 21,43]$.
  - IC(95%, Jospin) = $[15,62; 20,38]$.
  - IC(95%, Le Pen) = $[11,85; 16,15]$.

À RETENIR !

INTERVALLE DE CONFIANCE D'UNE PROPORTION

- de niveau de confiance $1-\alpha$,
- sur un échantillon de taille n.

1. n petit ($n \leq 30$) ou proportion petite ($np \leq 5$) :
   - Utilisation de la loi binomiale de paramètres n et $\widehat{p}_n$, via des tables ou sous R `qbinom`.
2. n grand, $np \geq 5$, $n(1-p) \geq 5$:
   - Utilisation d'intervalle de confiance asymptotique avec la loi normale (sous R `qnorm`) pour trouver φ_α.
   -

$$IC(1-\alpha) \approx \left[\widehat{p}_n - \varphi_\alpha \sqrt{\frac{\widehat{p}_n(1-\widehat{p}_n)}{n}}, \widehat{p}_n + \varphi_\alpha \sqrt{\frac{\widehat{p}_n(1-\widehat{p}_n)}{n}}\right].$$

TAILLE DE L'IC :

- Croît avec le niveau de confiance.
- Décroît avec la taille de l'échantillon.

Plan

1. Intervalle de confiance d'une proportion p
   - Résumer l'information, loi binomiale
   - Moyenne empirique
   - Intervalles de confiance
2. Estimation d'une moyenne
   - Si n est assez grand ($n \geq 30$)
   - Si n est petit...
   - Commentaires

Exemples

Exemple 1 : fréquences cardiaques :

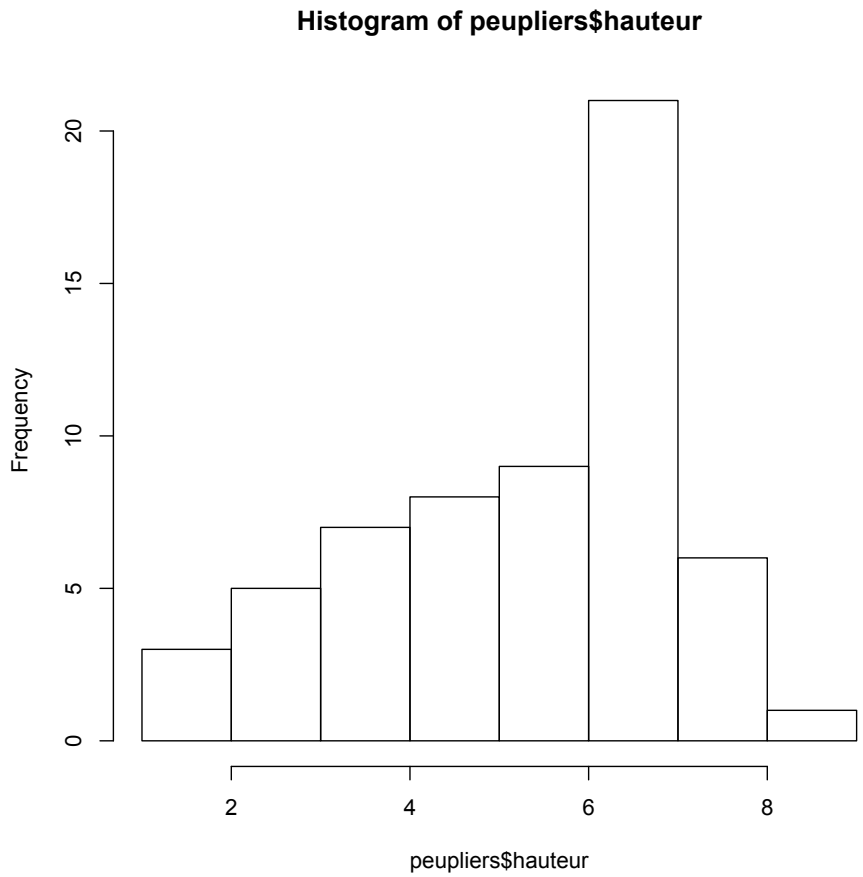
$$64; 67; 72; 58; 60; 65; 64; 57; 72; 66; 65; 59; 66;$$

$$63; 62; 64; 62; 66; 60; 61; 59; 62; 64; 61.$$

- Moyenne de l'échantillon : $\widehat{m} = 63,3$.
- Écart-type de l'échantillon : $\widehat{s} = 3,8$.

- Qu'en déduire pour la population ?

EXEMPLE 2 : La hauteur de 60 peupliers a été mesurée. Histogramme :



- Moyenne de l'échantillon : 5,3 mètres ; écart-type : 1,72 mètres.

Moyenne empirique et loi des grands nombres

But : estimer la valeur moyenne m de la population à partir des données de l'échantillon.

Estimateur : moyenne empirique

$$\boxed{\widehat{m}_n = \frac{1}{n}(x_1 + \ldots + x_n),}$$

avec

- n la taille de l'échantillon,
- les x_i étant les n mesures indépendantes effectuées.

Hypothèses : toutes ces mesures ont les mêmes caractéristiques (ou même loi) avec une moyenne m et une variance v.

Théorème (Loi des grands nombres)

La moyenne empirique a pour moyenne m et pour variance $\frac{v}{n}$. Et

$$\lim_{n \to +\infty} \widehat{m}_n = m.$$

QUANTIFIER L'ERREUR

QUESTION : comment maintenant quantifier l'erreur commise par cette estimation ponctuelle, c'est-à-dire : $|\widehat{m}_n - m|$?

DEUX APPROCHES :

1. Utilisation du théorème central limite, ce qui suppose que le nombre de données n est « grand ».
2. Si n est « petit », on suppose que les données sont distribuées suivant une loi gaussienne de paramètres m et v. Cette hypothèse est importante et devrait en principe être vérifiée ou testée avant de donner une estimation !

REMARQUE : ici il n'y a pas de lien a priori entre la moyenne et la variance. Dans le cas d'une proportion en revanche, si $m = p$, alors $v = p(1-p)$.

Plan

1. Intervalle de confiance d'une proportion p
   - Résumer l'information, loi binomiale
   - Moyenne empirique
   - Intervalles de confiance
2. Estimation d'une moyenne
   - Si n est assez grand ($n \geq 30$)
   - Si n est petit...
   - Commentaires

Deux estimateurs à calculer.

Estimateur de la moyenne :

$$\widehat{m}_n = \frac{1}{n}(x_1 + \ldots + x_n).$$

Estimateur de la variance :

$$\boxed{\widehat{v}_n = \frac{1}{n-1}\sum_{k=1}^{n}\left(x_k - \widehat{m}_n\right)^2.}$$

Théorème central limite :

$$\lim_{n\to+\infty} \mathbb{P}\left(a < \sqrt{n}\frac{(\widehat{m}_n - m)}{\sqrt{\widehat{v}_n}} < b\right) = \mathbb{P}(a < Z < b),$$

où Z suit une distribution normale ou gaussienne.

- Z ne dépend pas de p !

APPLICATION AUX EXEMPLES

INTERVALLE DE CONFIANCE ASYMPTOTIQUE : pour n « grand »

$$m \in IC(1-\alpha) \approx \left[\widehat{m}_n - \varphi_\alpha \sqrt{\frac{\widehat{v}_n}{n}}, \widehat{m}_n + \varphi_\alpha \sqrt{\frac{\widehat{v}_n}{n}} \right].$$

Pour $\alpha = 5\%$, $\varphi_\alpha = 1,96$.

HAUTEUR DES PEUPLIERS : $n = 60$, $\widehat{m} = 5,3$, $\sqrt{\widehat{v}} = 1,72$.

$$IC(95\%) = [4,864; 5,735].$$

Plan

1. Intervalle de confiance d'une proportion p
   - Résumer l'information, loi binomiale
   - Moyenne empirique
   - Intervalles de confiance
2. Estimation d'une moyenne
   - Si n est assez grand ($n \geq 30$)
   - Si n est petit...
   - Commentaires

Plus compliqué !

Théorème central limite : ne s'applique pas !

Hypothèse (comment la vérifier...)

Les données sont distribuées suivant une loi gaussienne de paramètres m et v. m n'est pas connue et on cherche à l'estimer.

Estimateurs :

- $\widehat{m}_n$: loi gaussienne.
- $\widehat{v}_n$: loi du chi-deux (χ^2).
- $T_n = \sqrt{n}\dfrac{\widehat{m}_n - m}{\sqrt{\widehat{v}_n}}$: loi de Student (ou t-distribution).

Intervalle de confiance.

$$m \in IC(1-\alpha) = \left[\widehat{m}_n - t_\alpha \sqrt{\frac{\widehat{v}_n}{n}}, \widehat{m}_n + t_\alpha \sqrt{\frac{\widehat{v}_n}{n}}\right]$$

t_α est telle que si Z suit une loi de Student de paramètre $n-1$

$$\mathbb{P}(|Z| > t_\alpha) = \alpha.$$

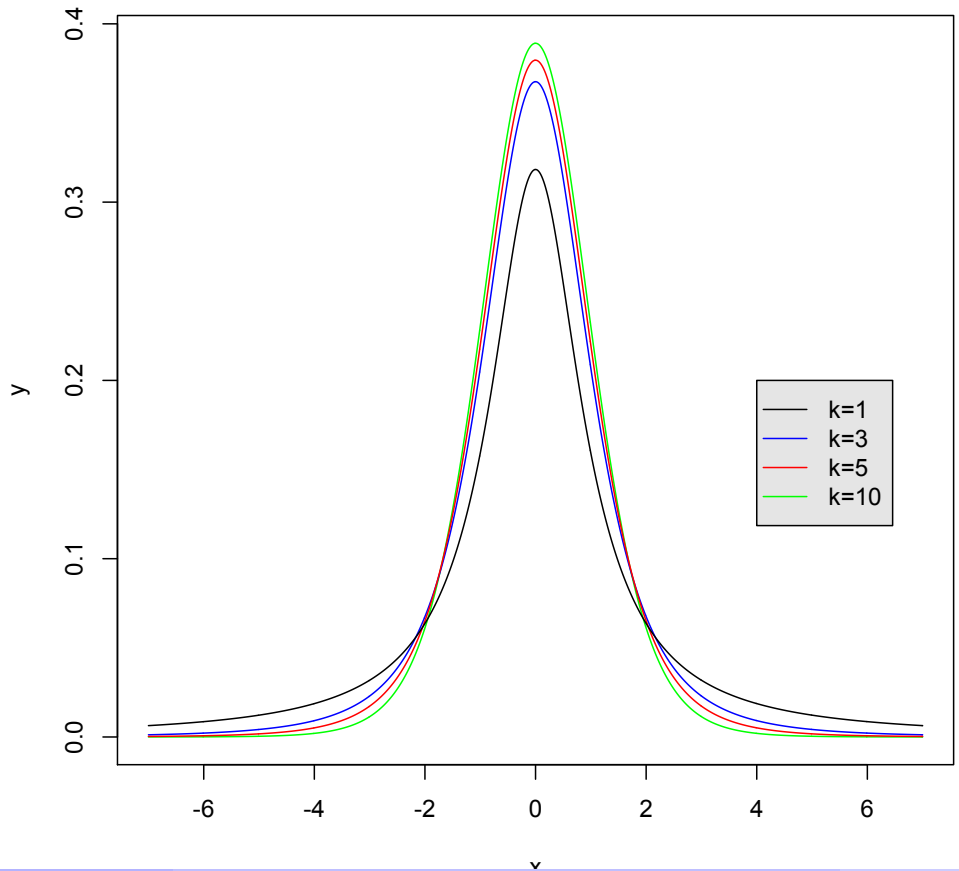
Calculs de t_α :

- Table.
- Sous R, fonction `qt`.
- Sous Excel, fonction `loi.student.inverse`.
- Exemple : pour $\alpha = 5\%$ et $n = 10$, on trouve $t_\alpha = 2,26$.

Loi de Student

Densité de la t-distribution de paramètre k :

$$f(x) = \frac{\Gamma((k+1)/2)}{\sqrt{\pi k}\Gamma(k/2)} \frac{1}{[(x^2/k)+1]^{(k+1)/2}}.$$



RETOUR SUR LES EXEMPLES

Tous les échantillons sont supposés gaussiens.

1. FRÉQUENCES CARDIAQUES : $n = 24$, $\widehat{m} = 63,3$, $\widehat{s} = 3,8$.

$$IC(95\%) = [61,7; 64,9] \ (t = 2,07),$$

$$IC(90\%) = [62; 64,6] \ (t = 1,71).$$

2. HAUTEUR DES PEUPLIERS : $n = 60$, $\widehat{m} = 5,3$, $\widehat{s} = 1,72$.

$$IC(95\%) = [4,86; 5,74] \ (t = 2).$$

$$IC_{asymp}(95\%) = [4,864; 5,735].$$

Plan

1. Intervalle de confiance d'une proportion p
   - Résumer l'information, loi binomiale
   - Moyenne empirique
   - Intervalles de confiance
2. Estimation d'une moyenne
   - Si n est assez grand ($n \geq 30$)
   - Si n est petit...
   - Commentaires

COMMENTAIRES

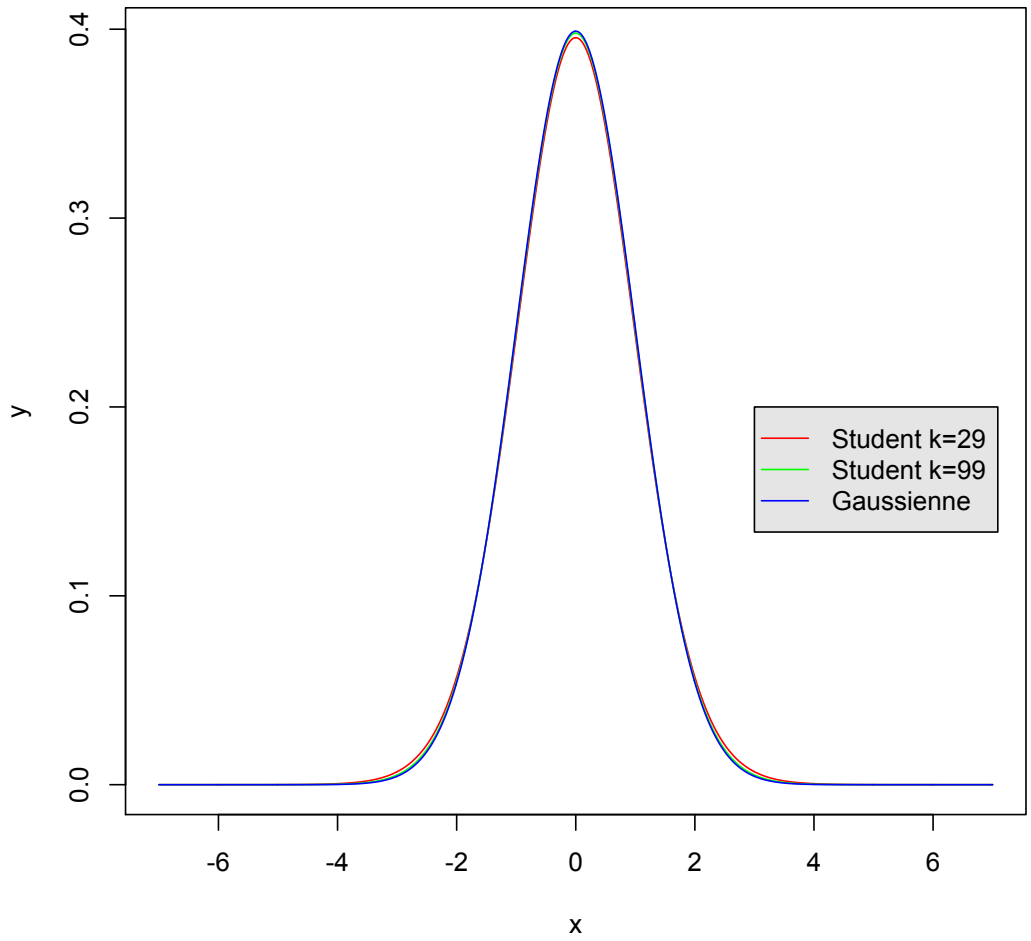
Dans tous les cas, l'intervalle de confiance dépend des quantités suivantes :

1. Le niveau de confiance : plus il est élevé, plus l'intervalle de confiance est grand.
2. Le nombre de données à travers la racine carrée de ce nombre : plus celui-ci est grand, plus l'intervalle de confiance est réduit.
3. L'écart-type : plus il est petit, plus l'intervalle de confiance est petit.

- Quand n petit, résultats reposent sur une hypothèse difficile à vérifier. Donc méfiance...

COMMENTAIRES

Loi de Student avec un degré de liberté grand ($n \geq 30$ en pratique)
= loi gaussienne.



À RETENIR !

INTERVALLE DE CONFIANCE D'UNE MOYENNE

- de niveau de confiance $1-\alpha$,
- sur un échantillon de taille n.

1. n grand :
   - Utilisation d'intervalle de confiance asymptotique avec la loi normale (sous R `qnorm`) pour trouver φ_α.
   - $$m \in IC(1-\alpha) \approx \left[\widehat{m}_n - \varphi_\alpha\sqrt{\frac{\widehat{v}_n}{n}}, \widehat{m}_n + \varphi_\alpha\sqrt{\frac{\widehat{v}_n}{n}}\right].$$
2. n petit ($n \leq 30$) :
   - Hypothèse sur la distribution des données : gaussienne.
   - Utilisation de la t-distribution de paramètres $n-1$ (sous R `qt`).
   - $$m \in IC(1-\alpha) = \left[\widehat{m}_n - t_\alpha\sqrt{\frac{\widehat{v}_n}{n}}, \widehat{m}_n + t_\alpha\sqrt{\frac{\widehat{v}_n}{n}}\right].$$