Partie III : Estimation paramétrique

1. Introduction

2. Estimation ponctuelle
   - Estimateur statistique
   - Qualité d'un estimateur
   - Estimateur de l'espérance, d'une proportion et de la variance

3. Estimation par intervalle de confiance
   - Notion d'intervalle de confiance
   - Estimation par intervalle de confiance d'une proportion

1 Introduction

2 Estimation ponctuelle
- Estimateur statistique
- Qualité d'un estimateur
- Estimateur de l'espérance, d'une proportion et de la variance

3 Estimation par intervalle de confiance
- Notion d'intervalle de confiance
- Estimation par intervalle de confiance d'une proportion

Introduction

Questions abordées dans cette partie

On sait calculer des indicateurs numériques à partir d'un échantillon de données ...

- Mais comment généraliser à la population entière ?
- Quelles informations sur la population obtient-on en étudiant l'échantillon ?
- Quelle confiance peut-on accorder à ces informations ?

Introduction

Statistique inférentielle

L'idée : à partir d'échantillons représentatifs, on va introduire des résultats sur la population.

On étudie une variable X, dont on observe des réalisations. On suppose que X suit une loi connue, i.e. on choisit parmi les modèles existants la loi la plus appropriée au phénomène observé. Seule la valeur numérique du paramètre θ intervenant dans cette loi de probabilité est inconnue.

$$X \sim \mathcal{P}(\theta), \quad \theta \text{ inconnu}$$

Exemple : soit X la taille des habitants de Grenoble. On suppose que X suit une loi normale, de moyenne inconnue θ et de variance connue. On va donc chercher à estimer θ à partir d'un échantillon de données.

Introduction

Deux types d'estimation

On considère généralement deux types d'estimation :

- **L'estimation ponctuelle** : on cherche à calculer une *unique* valeur $\hat{\theta}$ estimant au mieux θ.

- **L'estimation par intervalle de confiance** : on estime la probabilité que la valeur vraie d'un paramètre appartienne à un intervalle donné, on a ainsi un ensemble de valeurs vraisemblables donc une estimation *ensembliste* ou région de confiance. Typiquement, on cherche *a* et *b* tel que, par exemple

$$\mathbb{P}(\mathbf{a} \leq \theta \leq \mathbf{b}) = \mathbf{0.95}$$

Introduction

Hypothèses effectuées

On note les données $x_1, x_2, \ldots x_n$.

- On regardera x_i comme le i-ème tirage d'une variable aléatoire X :

$$X \sim \mathcal{P}(\theta)$$

- Ou de façon équivalente comme une réalisation d'une variable X_i de même loi que X. De plus, les X_i sont supposées indépendantes :

$$X_i \sim \mathcal{P}(\theta)$$

1 Introduction

2 Estimation ponctuelle
- Estimateur statistique
- Qualité d'un estimateur
- Estimateur de l'espérance, d'une proportion et de la variance

3 Estimation par intervalle de confiance
- Notion d'intervalle de confiance
- Estimation par intervalle de confiance d'une proportion

Estimation ponctuelle

Soient $x_1, x_2, \ldots, x_n$ les n valeurs prises par la v.a. X dans un échantillon de taille n prélevé dans la population-mère.

- Une statistique t est une fonction des observations $x_1, x_2, \ldots, x_n$:

$$\begin{array}{rccc} t: & \mathbb{R}^n & \to & \mathbb{R}^m \\ & (x_1, \ldots, x_n) & \mapsto & t(x_1, \ldots, x_n) \end{array}$$

- Un estimateur de θ est une fonction construite à l'aide des $\{X_i\}$:

$$T_n = t(X_1, \ldots, X_n)$$

- Une estimation est une réalisation $t_n = t(x_1, \ldots, x_n)$ de l'estimateur T_n. On note la valeur numérique de cette estimation par

$$\hat{\theta} = t(x_1, \ldots, x_n)$$

Estimation ponctuelle

Exemple

1. On observe un phénomène de production de pièces manufacturées. Chaque pièce est associée à une mesure (un indicateur de qualité par exemple). Comme on ne peut pas vérifier chaque mesure, on procède à un échantillonnage qui nous fournit donc un échantillon.
2. Supposons que la connaissance de la nature de cet indicateur nous permet de faire l'hypothèse qu'il obéit à une loi de probabilité normale.
3. Le problème est maintenant, au vue de l'échantillon $\{x_i\}$, de proposer une valeur pour la moyenne de cette loi normale. Il faut procéder à une estimation du paramètre vrai θ qui se traduit par la valeur $\hat{\theta}$. Il y a une infinité de manière possible parmi lesquelles :
   - $\hat{\theta} =$ la moyenne,
   - $\hat{\theta} =$ la médiane,
   - $\hat{\theta} =$ le mode,
   - $\hat{\theta} = x_{29}$, ...

   $\Rightarrow$ Quel est le meilleur estimateur de la moyenne ? Existe-t-il ? Qu'est ce que cela veut dire **meilleur** ?

Qualité d'un estimateur

Il n'existe pas de "meilleur estimateur" ! mais il existe des critères de comparaison :

Biais	On souhaite que l'estimation ne soit pas systématiquement décalée par rapport à la valeur vraie.
Précision	Si l'on répète l'estimation sur un autre échantillon, on souhaite obtenir une estimation cohérente, donc peu de variation d'un échantillon à l'autre. On parlera aussi d'efficacité.
Convergence	Si l'on peut estimer la valeur du paramètre sur toute la population-mère, la valeur de l'estimation obtenue doit être la valeur vraie du paramètre.
Complexité	Toute estimation nécessite un calcul donc un temps. On s'attachera donc à évaluer la complexité du calcul en fonction de la taille des données.
Robustesse	Dans tout cas concret, il existe des sources de perturbations. On souhaite que l'estimation ne soit pas sensible à la présence de valeurs abérantes.

Estimation ponctuelle

- **Estimateur sans biais**

  Un estimateur T_n est dit sans biais lorsque son espérance mathématique est égale à la valeur vraie du paramètre.

$$\text{Biais} = \mathbb{E}(T_n) - \theta = 0$$

- **Précision d'un estimateur**

  On mesure généralement la précision d'un estimateur par l'erreur quadratique moyenne :

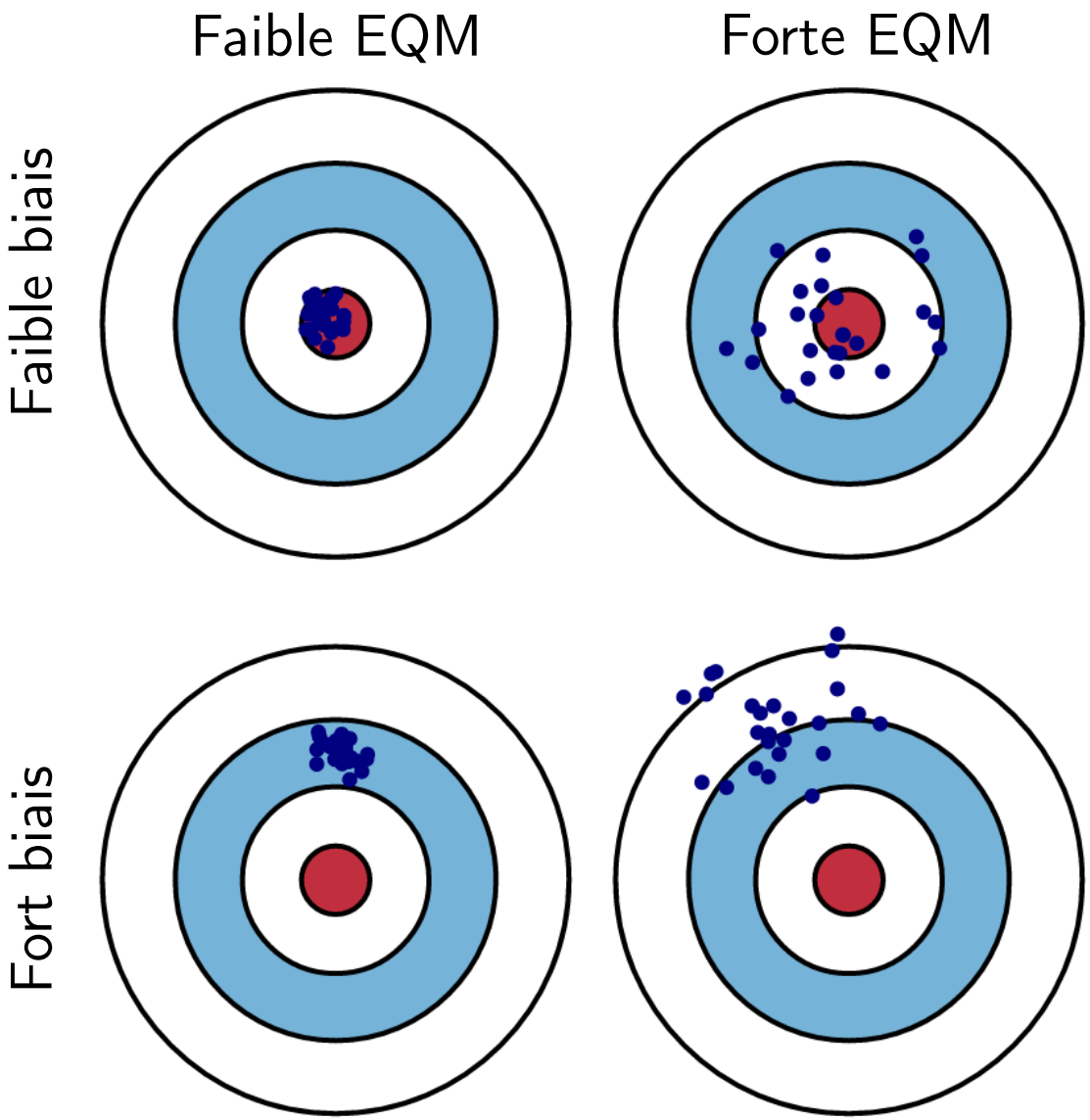
$$\text{EQM}(T_n) = \mathbb{E}((T_n - \theta)^2) = \mathbb{V}(T_n) + [\mathbb{E}(T_n) - \theta]^2$$

- **Estimateur convergent**

  Un estimateur T_n de θ est convergent en moyenne quadratique si

$$\mathbb{E}((T_n - \theta)^2) \to 0 \textbf{ quand } n \to \infty$$

Illustration



Crédits : http://www.cs.cornell.edu/courses/cs4780/2015fa/web/lecturenotes/lecturenote12.html

Estimation d'une espérance

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

est un estimateur sans biais et convergent de l'espérance

$$\theta = \mathbb{E}[X]$$

- La moyenne est un estimateur sans biais :

$$\mathbb{E}[\bar{X}_n] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\mathbb{E}\left[\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X] = \theta$$

- La moyenne est un estimateur convergent :

$$\mathbb{V}[\bar{X}_n] = \mathbb{V}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}\mathbb{V}\left[\sum_{i=1}^{n} X_i\right] \overset{(*)}{=} \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{V}[X_i] = \frac{1}{n}\mathbb{V}[X] \xrightarrow[n\to\infty]{} 0$$

$(*)$ car les X_i sont indépendants.

Estimation d'une proportion

Un estimateur fournit par la loi des grands nombres

La fréquence empirique d'un évènement
est une estimation de sa probabilité

Preuve : Soit $f_n(i)$ la fréquence de la valeur x_i dans l'échantillon de taille n $(X_1, \ldots, X_n)$, $B_k = \mathbb{1}_{(X_k = x_i)}$ et $p_i = \mathbb{P}(X = x_i)$. Ainsi la suite (B_k) est constituée de loi de Bernoulli indépendantes de paramètre p_i, de variance finie et d'espérance commune $\mathbb{E}(B_k) = p_i$, d'où d'après la loi des grands nombres :

$$f_n(i) = \frac{B_1 + \cdots + B_n}{n} \xrightarrow[n \to +\infty]{} p_i$$

$\Rightarrow$ L'estimation d'une proportion peut être vu comme un problème d'estimation de moyenne.

Estimation d'une proportion

Un estimateur sans biais convergent

La fréquence empirique d'un évènement
est une estimation de sa probabilité

- f_n est un estimateur sans biais :

$$\mathbb{E}[f_n] = \mathbb{E}\left[\frac{1}{n}\sum_{k=1}^{n} B_k\right] = \frac{1}{n}\mathbb{E}\left[\sum_{k=1}^{n} B_k\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[B_k] = \frac{1}{n}\sum_{i=1}^{n} p_i = p_i$$

- f_n est un estimateur convergent :

$$\mathbb{V}[f_n] = \mathbb{V}\left[\frac{1}{n}\sum_{k=1}^{n} B_k\right] = \frac{1}{n^2}\mathbb{V}\left[\sum_{k=1}^{n} B_k\right] \overset{(*)}{=} \frac{1}{n^2}\sum_{k=1}^{n}\mathbb{V}[B_k] = \frac{p_i(1-p_i)}{n} \to 0$$

$(*)$ car les X_i sont indépendants.

Estimation d'une variance

On définit :

$$\mathbf{V_n = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 = \frac{1}{n}\sum X_i^2 - \mu^2}$$

Si μ est connue, alors V_n est un estimateur **sans biais** de $\mathbb{V}[X]$

Preuve :

$$\begin{aligned}
\mathbb{E}[\mathbb{V}] &= \frac{1}{n}\mathbb{E}\left[\sum_{i=1}^{n} X_i^2\right] - \mathbb{E}[\mu^2] \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i^2] - \mu^2 \\
&= \mathbb{E}[X_i^2] - \mu^2 \equiv \mathbb{V}[X]
\end{aligned}$$

Estimation d'une variance

On définit :

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

Si μ est inconnue, alors S_n^2 est un estimateur **biaisé** de $\mathbb{V}(X)$

Preuve :

$$\begin{aligned}
\mathbb{E}[S_n^2] &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i^2] - \mathbb{E}[\bar{X}^2] \\
&= \frac{1}{n}\sum_{i=1}^{n}(\mathbb{V}[X_i] + \mathbb{E}[X_i]^2) - \mathbb{V}[\bar{X}] - \mathbb{E}[\bar{X}]^2 \\
&= \frac{1}{n}(n\mathbb{V}[X] + n\mathbb{E}[X_i]^2) - \frac{1}{n}\mathbb{V}[X] - \mathbb{E}[X_i]^2 \\
&= \mathbb{V}[X] - \frac{1}{n}\mathbb{V}[X] = \frac{n-1}{n}\mathbb{V}[X]
\end{aligned}$$

Estimation d'une variance

On définit :

$$\mathbf{S_n'^2 = \frac{n}{n-1} S_n^2}$$

Si μ est inconnue, alors $S_n'^2$ est un estimateur **sans biais** de $\mathbb{V}(X)$

Preuve :

$$\begin{aligned}\mathbb{E}[S_n'^2] &= \frac{n}{n-1}\mathbb{E}[S_n^2] \\ &= \frac{n}{n-1}\frac{n-1}{n}\mathbb{V}[X] \\ &= \mathbb{V}[X]\end{aligned}$$

Estimation d'une variance

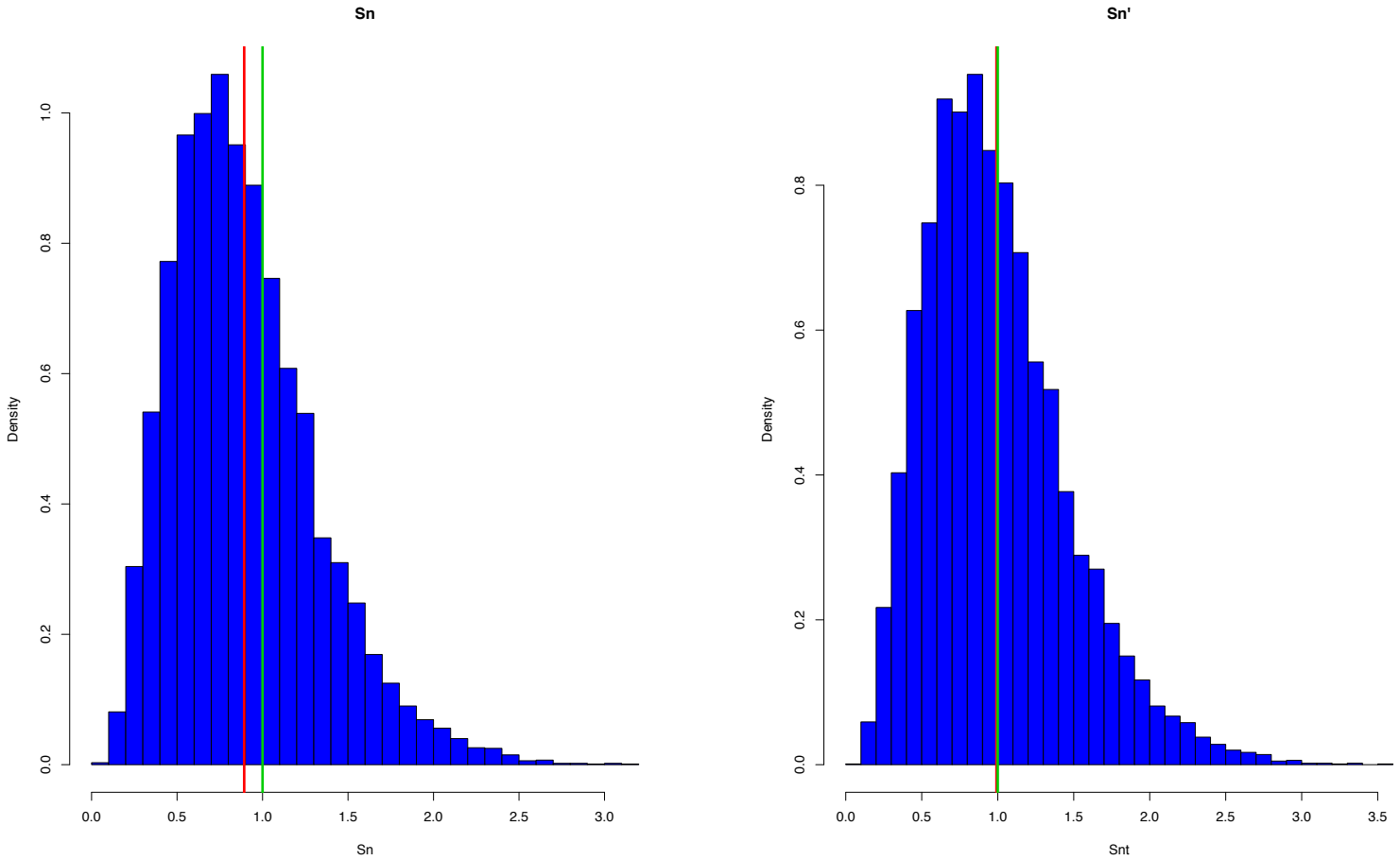
$$S_n'^2 = \frac{n}{n-1} S_n^2$$

```
K <- 10000
n <- 10
mu=0
sigma2=1
# On considère K échantillons composés de n valeurs issues de N(mu,sigma2)
# réparties sur chaque ligne de la matrice Xkn suivante
Xkn <- matrix(rnorm(n*K, mean=mu, sd=sqrt(sigma2)), ncol=n)
# On applique la variance selon les lignes, donnant K variances
Snt <- apply(Xkn,FUN=var,MARGIN=1)
Sn <- (n-1)/n*Snt
# S est la version biaisée (car var utilise la variance débiaisée)
hist(Sn, nclass=30, probability=TRUE, main="Sn", col="blue")
# la variance est décalée à n/(n-1)*sigma2=0.9
abline(v=mean(Sn), col="red") ; abline(v=sigma2, col=3)
hist(Snt, nclass=30, probability=TRUE, main="Sn'", col="blue")
abline(v=mean(Snt), col="red") ; abline(v=sigma2, col=3)
```

Estimation d'une variance

Illustration du biais

$$S_n'^2 = \frac{n}{n-1} S_n^2, \quad \mathbb{E}[S_n^2] = \frac{n-1}{n}\mathbb{V}[X], \quad \mathbb{E}[S_n'^2] = \mathbb{V}[X]$$



Exercice

On a mesuré le poids de raisins produits par souche sur 10 souches prises au hasard dans la vigne. On a obtenu les résultats suivants exprimés en kilogrammes :

$$2.4 \quad 3.4 \quad 3.6 \quad 4.1 \quad 4.3 \quad 4.7 \quad 5.4 \quad 5.9 \quad 6.5 \quad 6.9$$

1. Calculer la moyenne et la variance de l'échantillon.
2. En déduire les estimation ponctuelles non biaisées de la moyenne et de la variance de la population dont sont extraites les souches.

1 Introduction

2 Estimation ponctuelle
- Estimateur statistique
- Qualité d'un estimateur
- Estimateur de l'espérance, d'une proportion et de la variance

3 Estimation par intervalle de confiance
- Notion d'intervalle de confiance
- Estimation par intervalle de confiance d'une proportion

Estimation par intervalle de confiance

La "fourchette"

Considérons un vote avec un assez grand nombre d'électeurs. Quand le scrutin est clot, on commence à dépouiller les bulletins. Assez vite on est en mesure de donner une estimation du résultat final. En pratique, on ne donne pas une estimation numérique (telle liste obtient 18% des votes) mais une fourchette, c'est-à-dire un petit intervalle dans lequel on estime que le pourcentage exact figure.

- La **taille de la fourchette** dépend de la confiance qu'on souhaite avoir dans l'estimation.
- On peut vouloir que la probabilité que le pourcentage exact d'une liste soit bien dans la fourchette dépasse 0.95 (**le niveau de confiance**).
- Plus on exige un haut niveau de confiance, plus la fourchette sera large.

Estimation par intervalle de confiance

Notion d'intervalle de confiance

Il est souvent plus réaliste et plus intéressant de fournir un renseignement de type $a < \theta < b$ plutôt que de calculer $\hat{\theta}$.

On cherche à determiner l'intervalle $[a, b]$, centré sur la valeur numérique estimée du paramètre inconnu θ, contenant la valeur vraie avec une probabilité $1 - \alpha$ $(0 \leq \alpha \leq 1)$:

$$\mathbb{P}(a < \theta < b) = 1 - \alpha$$

- L'intervalle $[\mathbf{a}, \mathbf{b}]$ est appelé intervalle de confiance, α le risque et $\mathbf{1} - \alpha$ le niveau de confiance.
- Données de depart : **l'échantillon** et la connaissance de **la loi de probabilité du paramètre** à estimer.

Estimation par intervalle de confiance d'une proportion

Soit une population dont les individus possèdent un caractère A avec une probabilité p. On dispose d'un échantillon de taille n, dont x individus possèdent le caractère A.

- On sait maintenant que la proportion $\mathbf{f_n = x/n}$ est une estimation de la valeur vraie $\mathbf{p}$ ...
- Mais avec quelle confiance ?
- On cherche donc à construire un intervalle de confiance de l'estimateur.

Crédits : O. Gaudoin

Estimation par intervalle de confiance d'une proportion

Soit une population dont les individus possèdent un caractère A avec une probabilité $\mathbf{p}$. On dispose d'un échantillon de taille $\mathbf{n}$, dont $\mathbf{x}$ individus possèdent le caractère A. La proportion $\mathbf{f_n = x/n}$ est une estimation de la valeur vraie $\mathbf{p}$.

Principe

- Soit $\mathbf{F_n = \frac{1}{n}\sum_{i=1}^{n} X_i}$. $\mathbf{F_n}$ est une v.a. construite comme somme de n v.a indépendantes de type Bernoulli et de paramètre $\mathbf{p}$, i.e $\mathbf{X_i} \sim \mathcal{B}(\mathbf{p})$.
- La loi de $\mathbf{T_n = nF_n}$ suit une loi binomiale $\mathcal{B}(\mathbf{n}, \mathbf{p})$ ...
- La loi de $\mathbf{T_n}$ tend vers une loi normale de moyenne $\mathbf{np}$ et de variance $\mathbf{np(1-p)}$ (*cf.* TP 2, approximation valide si $np > 10$ et $n(1-p) > 10$)
- La variable renormalisée approche une loi normale centrée réduite :

$$\frac{\mathbf{T_n - np}}{\sqrt{\mathbf{np(1-p)}}} \sim_\infty \mathcal{N}(\mathbf{0}, \mathbf{1})$$

Crédits : O. Gaudoin

Estimation par intervalle de confiance d'une proportion

On écrit alors

$$\mathbb{P}\left(\left|\frac{\mathbf{T_n - np}}{\sqrt{\mathbf{np(1-p)}}}\right| \leq \mathbf{u}_\alpha\right) \approx \mathbf{1} - \alpha \ ,$$

où u_α est une valeur (se lisant dans la table de la loi normale $\mathcal{N}(0,1)$) qui vérifie :

$$\mathbb{P}(|U| > u_\alpha) = \alpha, \quad U \sim \mathcal{N}(0,1) \ .$$

Pour en déduire un intervalle de confiance, il suffit d'écrire $\left|\frac{T_n - np}{\sqrt{np(1-p)}}\right| \leq u_\alpha$ sous la forme $Z_1 \leq p \leq Z_2$:

$$\left|\frac{T_n - np}{\sqrt{np(1-p)}}\right| \leq u_\alpha \iff \frac{(T_n - np)^2}{np(1-p)} \leq u_\alpha^2$$

$$\iff p^2(n + u_\alpha^2) - p(2T_n + u_\alpha^2) + \frac{T_n^2}{n} \leq 0$$

Crédits : O. Gaudoin

Estimation par intervalle de confiance d'une proportion

Intervalle de confiance asymptotique

Le trinôme $p^2(n+u_\alpha^2) - p(2T_n + u_\alpha^2) + \frac{T_n^2}{n}$ est toujours positif sauf entre ses racines. Donc ses racines sont les bornes de l'intervalle de confiance recherché :

$$\left[\frac{\frac{T_n}{n} + \frac{u_\alpha^2}{2n} - u_\alpha\sqrt{\frac{u_\alpha^2}{4n^2} + \frac{T_n(n-T_n)}{n^3}}}{1+\frac{u_\alpha^2}{n}}, \frac{\frac{T_n}{n} + \frac{u_\alpha^2}{2n} + u_\alpha\sqrt{\frac{u_\alpha^2}{4n^2} + \frac{T_n(n-T_n)}{n^3}}}{1+\frac{u_\alpha^2}{n}}\right]$$

Pour les valeurs usuelles de α et pour n grand, on peut négliger u_α^2 par rapport à n. D'où, avec $F_n = \frac{T_n}{n}$ et une réalisation f_n de F_n, on obtient **l'intervalle de confiance asymptotique** suivant :

$$\left[f_n - u_\alpha\sqrt{\frac{f_n(1-f_n)}{n}}, f_n + u_\alpha\sqrt{\frac{f_n(1-f_n)}{n}}\right]$$

Crédits : O. Gaudoin

Estimation par intervalle de confiance d'une proportion

Fourchette du sondage

Une élection oppose deux candidats A et B. Un institut de sondage interroge **800 personnes** sur leurs intentions de vote :

- 420 déclarent voter pour A
- 380 déclarent voter pour B

Estimer le résultat de l'élection, c'est estimer le pourcentage p de voix qu'obtiendra A le jour de l'élection, en inférant sur l'ensemble de la population. L'estimateur de p est la proportion $f_n = \frac{420}{800} = 52.5\%$. L'institut de sondage estime donc que le candidat A va gagner l'élection. Mais pour évaluer l'incertitude, on a besoin d'un intervalle de confiance de seuil disons 5% pour p. On obtient alors l'intervalle de confiance asymptotique suivant

$$\mathbf{[0.4904, 0.5596]}$$

Conclusion : on a une confiance de 95% dans le fait que le pourcentage de voix qu'obtiendra le candidat A sera compris entre 49% et 56%.

Crédits : O. Gaudoin

Estimation par intervalle de confiance d'une proportion

Obtenir une prédiction plus précise

À quelle condition l'intervalle de confiance pour p sera entièrement situé au dessus de 50% ?

$\Rightarrow$ Il s'agit donc de réduire l'intervalle de confiance, de largeur :

$$\ell = 2u_\alpha \sqrt{\frac{f_n(1-f_n)}{n}}$$

Pour diminuer cette largeur ℓ, on peut :

- Diminuer u_α, c'est-à-dire augmenter α, donc augmenter la probabilité de se tromper en affirmant que le candidat est élu ;
- Augmenter n, c'est-à-dire augmenter le nombre de personnes interrogées.

Crédits : O. Gaudoin

Estimation par intervalle de confiance d'une proportion

Taille de l'échantillon minimum

À un seuil de confiance $\alpha = 5\%$ fixé, combien de personnes n doit-on interroger pour que l'intervalle de confiance n'excède pas une largeur ℓ ?

On sait que $\forall p \in [0,1], p(1-p) \leq \frac{1}{4}$, donc

$$2u_\alpha \sqrt{\frac{f_n(1-f_n)}{n}} \leq \frac{u_\alpha}{\sqrt{n}}$$

Ainsi il suffit de déterminer n tel que

$$\frac{\mathbf{u}_\alpha}{\sqrt{\mathbf{n}}} < \ell \Longleftrightarrow \mathbf{n} > \frac{\mathbf{u}_\alpha^2}{\ell^2}$$

Pour $n = 800$, on a $\frac{u_\alpha}{\sqrt{n}} = \frac{1.96}{\sqrt{800}} \approx 7.5\%$ $\Rightarrow$ la précision sur l'estimation de p est donc avec une confiance de 95% de plus ou moins 3.5%, ce qu'on a constaté avec l'intervalle $[49\%, 56\%]$.

Crédits : O. Gaudoin

Estimation par intervalle de confiance d'une proportion

Taille de l'échantillon minimum

À un seuil de confiance $\alpha = 5\%$ fixé, combien de personnes n doit-on interroger pour que l'intervalle de confiance n'excède pas une largeur ℓ ?

On sait que $\forall p \in [0,1], p(1-p) \leq \frac{1}{4}$, donc

$$2u_\alpha \sqrt{\frac{f_n(1-f_n)}{n}} \leq \frac{u_\alpha}{\sqrt{n}}$$

Ainsi il suffit de déterminer n tel que

$$\frac{\mathbf{u}_\alpha}{\sqrt{\mathbf{n}}} < \ell \Longleftrightarrow \mathbf{n} > \frac{\mathbf{u}_\alpha^2}{\ell^2}$$

Si on veut, avec le même niveau de confiance, avoir une précision $<$ à 1%, il faudra interroger au moins :

$$n = \frac{u_\alpha^2}{\ell^2} = \frac{1.96^2}{0.01^2} = 38\,416 \text{ personnes}$$

Crédits : O. Gaudoin

Exercice

Afin d'étudier l'influence des rayons X sur la spermatogénèse de Bombyx Mori (vers à soie sous sa forme papillon), on a irradié des mâles au 2ème jour et au 4ème jour du stade larvaire. Ces mâles ont été accouplés avec des femelles non irradiées. On a compté le nombre d'œufs fertiles dans la ponte des femelles, et on a obtenu :

nombre d'œufs totals	nombre d'œufs fertiles
5646	4998

1. Donner l'estimation du pourcentage d'œufs fertiles
2. Calculer un intervalle de confiance approximatif du pourcentage d'œufs fertiles au niveau de confiance 0.9. On donne $u_{0.1} = 1.6448$.

Rappel théorème de la limite centrée

- Si $X_1, \ldots, X_n$ sont indépendantes et de même loi partageant d'espérance μ et de variance σ^2, l'estimateur sans biais de l'espérance μ est la moyenne empirique $\bar{X}_n$
- Le théorème central limite assure que pour n suffisamment grand :

$$\overline{\mathbf{X}}_{\mathbf{n}} \simeq \mathcal{N}\left(\mu, \frac{\sigma^{\mathbf{2}}}{\mathbf{n}}\right)$$

Rappel théorème de la limite centrée

Pour une loi normale

- Si $X_1, \ldots, X_n$ sont indépendantes et de même loi normale $\mathcal{N}(\mu, \sigma^2)$, l'estimateur sans biais de variance minimale (ESBVM) de μ est la moyenne empirique $\bar{X}_n$
- Les propriétés élémentaires de la loi normale permettent d'établir :

$$\sum_{i=1}^{n} X_i \sim \mathcal{N}(n\mu, n\sigma^2) \Rightarrow \overline{\mathbf{X}}_{\mathbf{n}} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{\mathbf{n}}\right)$$

Caractérisation de la loi de l'estimateur $\bar{X}_n$

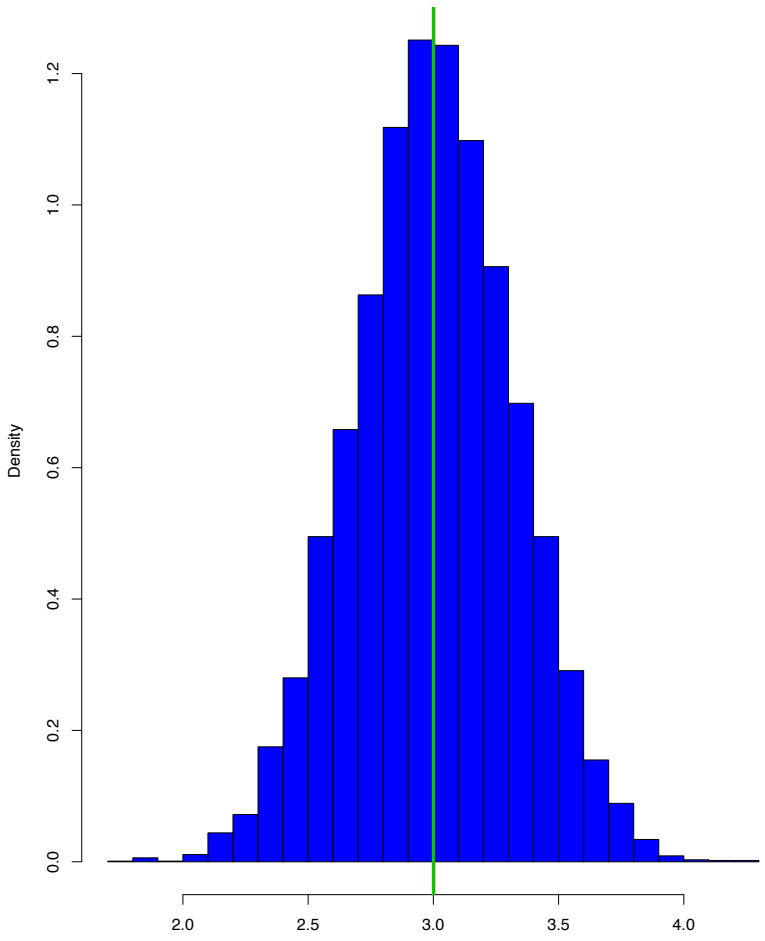
Pour une loi normale

```
K <- 10000
n <- 10
mu=3
sigma2=1
# On considère K échantillons composés de n valeurs issues de N(mu,sigma2)
# réparties sur chaque ligne de la matrice Xkn suivante
Xkn <- matrix(rnorm(n*K,mean=mu,sd=sqrt(sigma2)),ncol=n)
# On applique la moyenne selon les lignes, donnant K moyennes
Mn <- apply(Xkn,FUN=mean,MARGIN=1)
hist(Mn,nclass=30,probability=TRUE,main="Mn",col="blue")
abline(v=mean(Mn),col="red") ; abline(v=mu,col=3)
```

Caractérisation de la loi de l'estimateur $\bar{X}_n$

Pour une loi normale, effet de la taille de l'échantillon $n = 10$

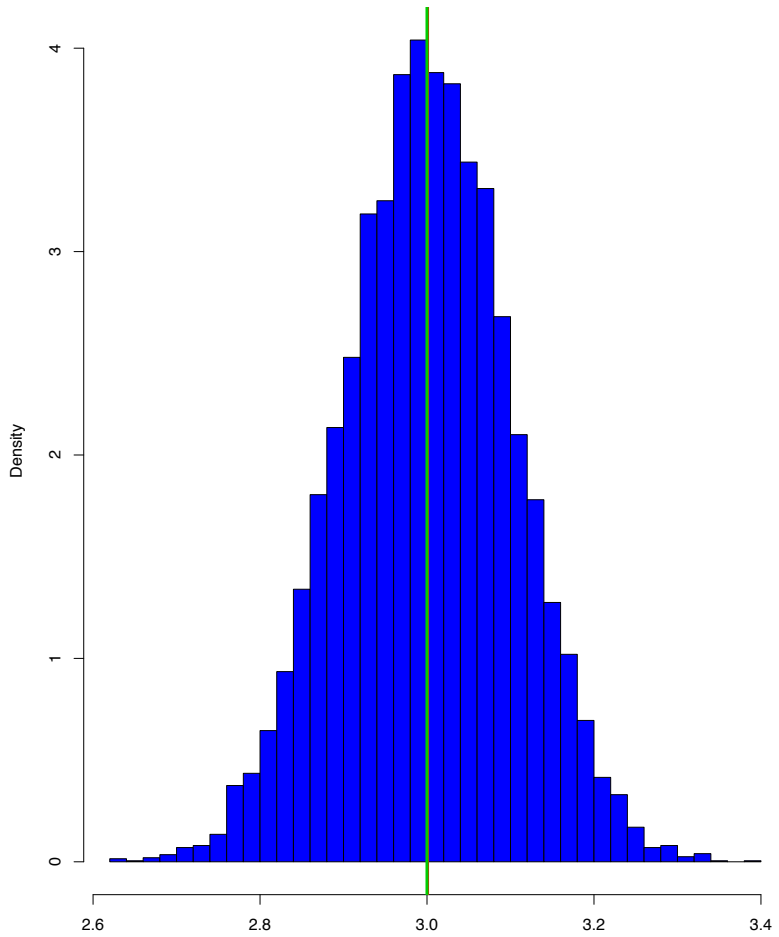
$$\sum_{i=1}^{n} X_i \sim \mathcal{N}(n\mu, n\sigma^2) \Rightarrow \overline{\mathbf{X}}_{\mathbf{n}} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{\mathbf{n}}\right)$$



Caractérisation de la loi de l'estimateur $\bar{X}_n$

Pour une loi normale, effet de la taille de l'échantillon $n = 100$

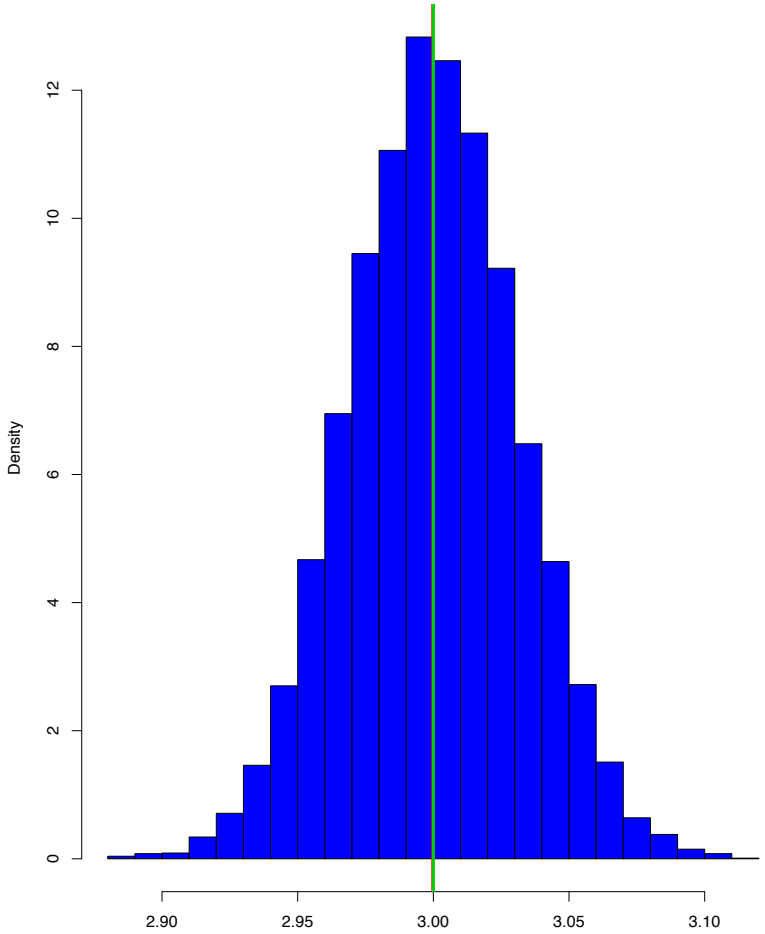
$$\sum_{i=1}^{n} X_i \sim \mathcal{N}(n\mu, n\sigma^2) \Rightarrow \overline{\mathbf{X}}_{\mathbf{n}} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{\mathbf{n}}\right)$$



Caractérisation de la loi de l'estimateur $\bar{X}_n$

Pour une loi normale, effet de la taille de l'échantillon $n = 1000$

$$\sum_{i=1}^{n} X_i \sim \mathcal{N}(n\mu, n\sigma^2) \Rightarrow \overline{\mathbf{X}}_{\mathbf{n}} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{\mathbf{n}}\right)$$



Estimation par intervalle de confiance de la moyenne

Pour une loi normale, lorsque la variance σ est connue

- Si $X_1, \ldots, X_n$ sont indépendantes et de même loi normale $\mathcal{N}(\mu, \sigma^2)$, l'estimateur sans biais de variance minimale (ESBVM) de μ est la moyenne empirique $\bar{X}_n$
- Les propriétés élémentaires de la loi normale permettent d'établir :

$$\sum_{i=1}^{n} X_i \sim \mathcal{N}(n\mu, n\sigma^2) \Rightarrow \overline{\mathbf{X}}_{\mathbf{n}} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{\mathbf{n}}\right)$$

$$\Rightarrow \mathbf{U} = \frac{\overline{\mathbf{X}}_{\mathbf{n}} - \mu}{\sqrt{\sigma^2/\mathbf{n}}} \sim \mathcal{N}(\mathbf{0}, \mathbf{1})$$

- On cherche un intervalle de confiance pour μ de la forme $[\bar{X}_n - \epsilon, \bar{X}_n + \epsilon]$, soit pour α fixé :

$$\mathbb{P}(|\bar{X}_n - \mu| \leq \epsilon) = 1 - \alpha$$

Crédits : O. Gaudoin

Estimation par intervalle de confiance de la moyenne

Pour une loi normale, lorsque la variance σ est connue

- Si $X_1, \ldots, X_n$ sont indépendantes et de même loi normale $\mathcal{N}(\mu, \sigma^2)$, l'estimateur sans biais de variance minimale (ESBVM) de μ est la moyenne empirique $\bar{X}_n$
- On cherche un intervalle de confiance pour μ de la forme $[\bar{X}_n - \epsilon, \bar{X}_n + \epsilon]$, soit pour α fixé :

$$1 - \alpha = \mathbb{P}(|\bar{X}_n - \mu| \leq \epsilon)$$

$$1 - \alpha = \mathbb{P}\left(|U| \leq \frac{\sqrt{n}\epsilon}{\sigma}\right)$$

$$1 - \mathbb{P}\left(|U| > u_\alpha\right) = 1 - \mathbb{P}\left(|U| > \frac{\sqrt{n}\epsilon}{\sigma}\right)$$

- On en déduit $u_\alpha = \frac{\sqrt{n}\epsilon}{\sigma}$ soit $\epsilon = \frac{\sigma}{\sqrt{n}} u_\alpha$, d'où l'intervalle de confiance :

$$\left[\overline{\mathbf{X}}_\mathbf{n} - \frac{\sigma}{\sqrt{\mathbf{n}}}\mathbf{u}_\alpha, \overline{\mathbf{X}}_\mathbf{n} + \frac{\sigma}{\sqrt{\mathbf{n}}}\mathbf{u}_\alpha\right]$$

Crédits : O. Gaudoin

Loi du χ^2

Définition :

Une v.a réelle suit une loi du χ^2_p (chi-deux) à p degrés de liberté si elle se réalise comme une somme

$$\chi^2_p = \sum_{i=1}^{p} U_i^2 \ ,$$

où $U_1, \ldots, U_p$ sont p variables normales $\mathcal{N}(0,1)$ indépendantes.

La densité de χ^2_p est donnée par

$$f_p(t) = \frac{1}{2^{p/2}\Gamma(p/2)} \exp(-t/2) t^{p/2-1} \ ,$$

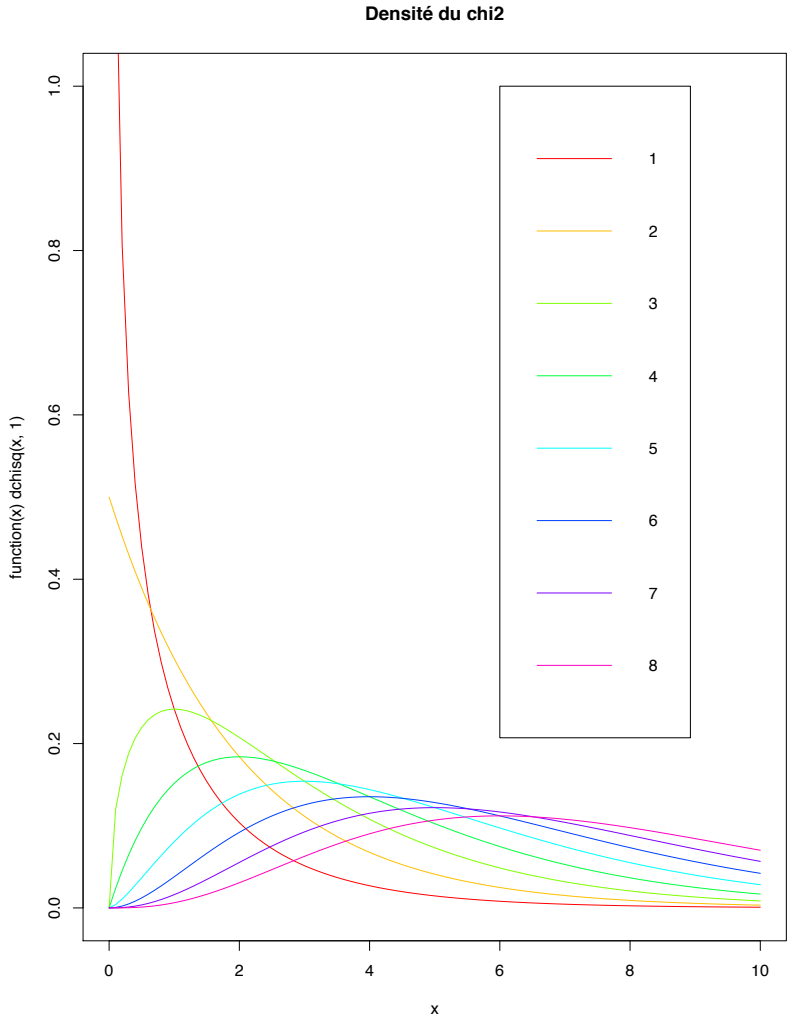
où

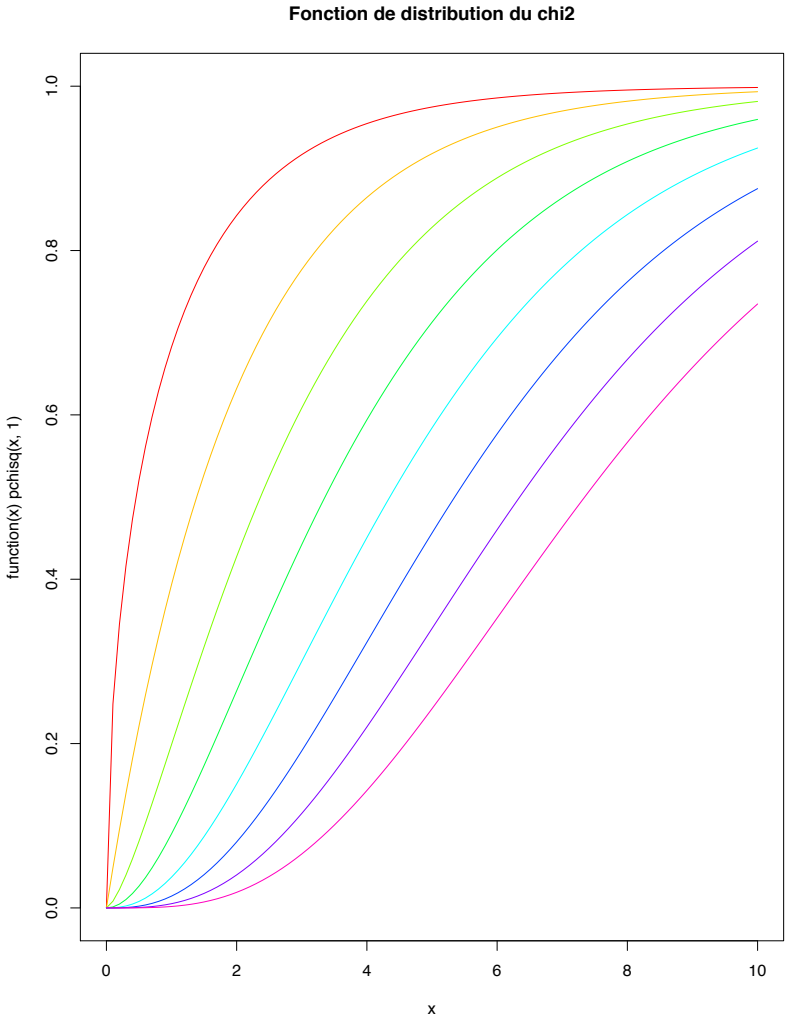
$$\Gamma : x \in \mathbb{R}^+ \mapsto \int_0^{+\infty} t^{x-1} \exp(-t) \mathrm{d}t \ ,$$

et son espérance vaut p et sa variance $2p$.

Loi du χ^2_p

Illustration de la densité et de la fonction de répartition selon p





Caractérisation de la loi de l'estimateur S_n

Théorème de Fischer

Théorème de Fischer

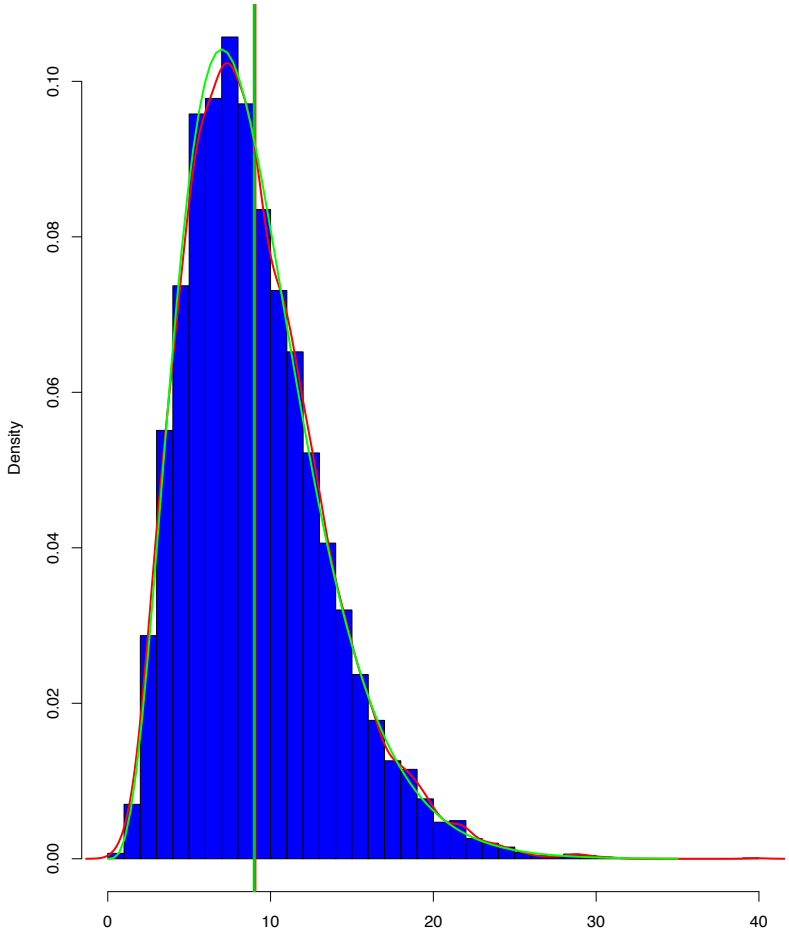
Si $X_1, \ldots, X_n$ sont indépendantes et de même loi normale $\mathcal{N}(\mu, \sigma^2)$, alors

$$\frac{\mathbf{n}\mathbf{S}_\mathbf{n}^2}{\sigma^2} = \frac{(\mathbf{n}-\mathbf{1})\mathbf{S}_\mathbf{n}'^2}{\sigma^2} \sim \chi^2_{\mathbf{n}-\mathbf{1}}$$

Caractérisation de la loi de l'estimateur S'_n

```
K <- 10000
n <- 10
mu=0
sigma2=1
# On considère K échantillons composés de n valeurs issues de N(mu,sigma2)
# réparties sur chaque ligne de la matrice Xkn suivante
Xkn <- matrix(rnorm(n*K,mean=mu,sd=sqrt(sigma2)),ncol=n)
# On applique la variance selon les lignes, donnant K variances
St <- apply(Xkn,FUN=var,MARGIN=1)
Ki <- (n-1)/sigma2*St
# On calcule l'histogramme qui doit approcher une loi chi2(n-1)
hist(Ki,nclass=30,probability=TRUE,main="Chi")
# On affiche les moyennes empiriques et vraies
abline(v=mean(Ki),col="red",lwd=2) ;
abline(v=n-1,col=3,lwd=2)
# On affiche les courbes de densités approchées et vraies
lines(density(Ki),col="red")
plot(function(x) dchisq(x,n-1),0,35,main="Chi square density
    ", col="green",ylim=c(0,1),xlim=c(0,10),add=TRUE)
```

Caractérisation de la loi de l'estimateur S'_n



Estimation par intervalle de confiance de la variance

Pour une loi normale

$$\mathbb{P}\left(a \leq \frac{nS_n^2}{\sigma^2} \leq b\right) = \mathbb{P}\left(\frac{nS_n^2}{b} \leq \sigma^2 \leq \frac{nS_n^2}{a}\right)$$
$$= F_{\chi^2_{n-1}}(b) - F_{\chi^2_{n-1}}(a)$$

$\Rightarrow$ **Il y a une infinité de façons possibles de choisir a et b de sorte à ce que cette probabilité soit égale à $1-\alpha$. Si on choisit :**

$$F_{\chi^2_{n-1}}(b) = 1 - \frac{\alpha}{2}, \quad F_{\chi^2_{n-1}}(a) = \frac{\alpha}{2}$$

alors, avec $z_{n-1,\alpha}$ dans la table de χ^2_{n-1}, on a

$$\mathbb{P}(Z > z_{n-1,\alpha}) = 1 - F_{\chi^2_n}(z_{n-1,\alpha}) = \alpha \ ,$$

et ainsi les valeurs suivantes conviennent :

$$b = z_{n-1,\alpha/2}, \quad a = z_{n-1,1-\alpha/2}$$

Crédits : O. Gaudoin

Estimation par intervalle de confiance de la variance

Pour une loi normale

Un intervalle de confiance de seuil α pour le paramètre σ^2 de la loi $\mathcal{N}(\mu, \sigma^2)$ est :

$$\left[\frac{nS_n^2}{z_{n-1,\alpha/2}}, \frac{nS_n^2}{z_{n-1,1-\alpha/2}}\right] = \left[\frac{(n-1)S_n'^2}{z_{n-1,\alpha/2}}, \frac{(n-1)S_n'^2}{z_{n-1,1-\alpha/2}}\right]$$

Loi de Student

Définition :

Une v.a réelle suit une loi de Student à p degrés de liberté si elle se réalise comme sous la forme

$$\frac{\mathbf{U}}{\sqrt{\frac{\chi^2_{\mathbf{p}}}{\mathbf{p}}}} \sim \mathbf{St(p)} \ ,$$

où U suit une loi $\mathcal{N}(0,1)$ et χ^2_p indépendantes.

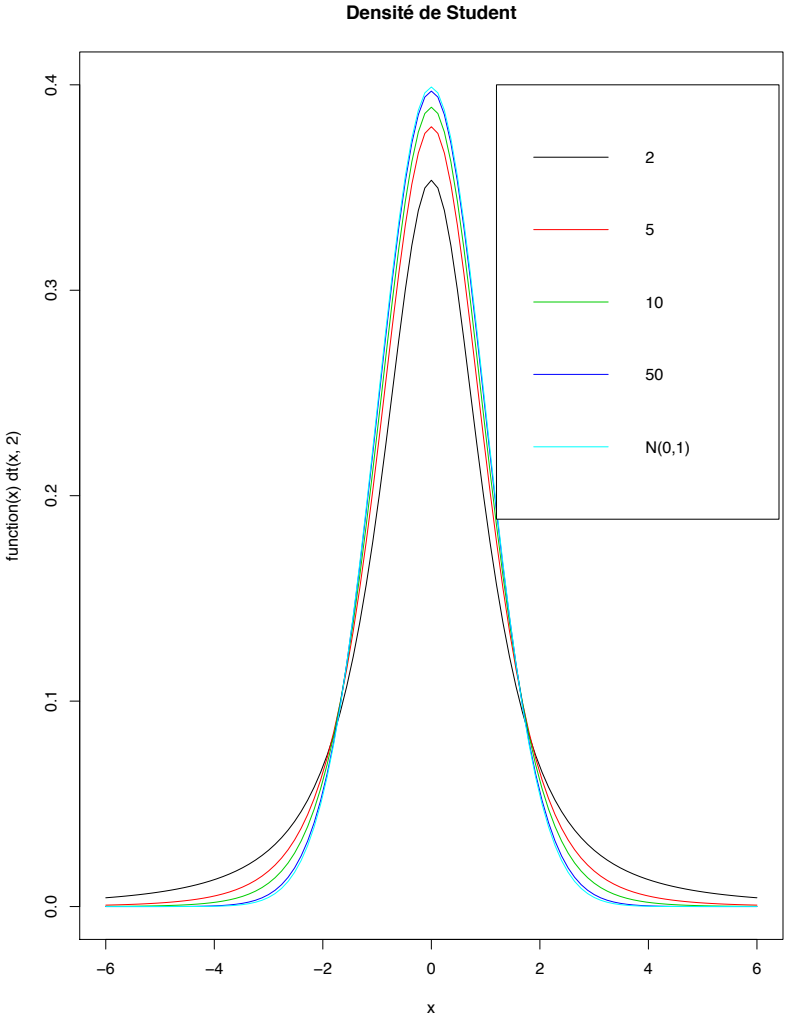
La densité de $\mathrm{St}(p)$ est donnée par

$$f_{\mathrm{St}(p)}(t) = \frac{1}{\sqrt{k\pi}} \frac{\Gamma(\frac{k+1}{2})}{\Gamma(\frac{k}{2})} \left(1 + \frac{t^2}{k}\right)^{-\frac{k+1}{2}} \ ,$$

son espérance n'est pas définie pour $p = 1$ et est nulle pour $p > 1$, sa variance est infinie pour $p \leq 2$ et vaut $\frac{p}{p-2}$ pour $p > 2$

Loi de Student

Illustration de la densité et normalité asymptotique



Estimation par intervalle de confiance de la moyenne

Pour une loi normale, lorsque la variance σ est connue

- Si $X_1, \ldots, X_n$ sont indépendantes et de même loi normale $\mathcal{N}(\mu, \sigma^2)$, l'ESBVM de μ est la moyenne empirique $\bar{X}_n$
- Un intervalle de confiance pour μ de la forme $[\bar{X}_n - \epsilon, \bar{X}_n + \epsilon]$, est pour un risque α fixé :

$$\left[\overline{\mathbf{X}}_\mathbf{n} - \frac{\sigma}{\sqrt{\mathbf{n}}}\mathbf{u}_\alpha, \overline{\mathbf{X}}_\mathbf{n} + \frac{\sigma}{\sqrt{\mathbf{n}}}\mathbf{u}_\alpha\right]$$

- **Une idée naturelle est de remplacer σ par son estimateur S'_n**
- On utilise non plus $\mathbf{U} = \frac{\overline{\mathbf{X}}_\mathbf{n} - \mu}{\sqrt{\sigma^2/\mathbf{n}}} \sim \mathcal{N}(\mathbf{0}, \mathbf{1})$, mais :

$$\frac{\overline{\mathbf{X}}_\mathbf{n} - \mu}{\sqrt{\mathbf{S}'^2_\mathbf{n}/\mathbf{n}}} = \sqrt{\mathbf{n}}\frac{\overline{\mathbf{X}}_\mathbf{n} - \mu}{\mathbf{S}'_\mathbf{n}} \not\sim \mathcal{N}(\mathbf{0}, \mathbf{1})$$

Crédits : O. Gaudoin

Caractérisation de la loi de l'estimateur T_n

Théorème de Fischer

Théorème de Fischer

Si $X_1, \ldots, X_n$ sont indépendantes et de même loi normale $\mathcal{N}(\mu, \sigma^2)$, alors

$$\sqrt{\mathbf{n}}\frac{\overline{\mathbf{X}}_\mathbf{n} - \mu}{\mathbf{S}'_\mathbf{n}} \sim \mathbf{St(n-1)}$$

Estimation par intervalle de confiance de la moyenne

Pour une loi normale, lorsque la variance σ est inconnue

- Si $X_1, \ldots, X_n$ sont indépendantes et de même loi normale $\mathcal{N}(\mu, \sigma^2)$, les estimateurs sans biais de variance minimale (ESBVM) de (μ, σ^2) sont $(\bar{X}_n, S_n'^2)$.
- On cherche un intervalle de confiance pour μ de la forme $[\bar{X}_n - \epsilon, \bar{X}_n + \epsilon]$, soit pour α fixé :

$$1 - \alpha = \mathbb{P}(|\bar{X}_n - \mu| \leq \epsilon)$$

$$1 - \alpha = \mathbb{P}\left(\left|\sqrt{n}\frac{\overline{X}_n - \mu}{S_n'}\right| \leq \frac{\sqrt{n}\epsilon}{S_n'}\right)$$

$$1 - \mathbb{P}\left(\left|\sqrt{n}\frac{\overline{X}_n - \mu}{S_n'}\right| > t_{n-1,\alpha}\right) = 1 - \mathbb{P}\left(\left|\sqrt{n}\frac{\overline{X}_n - \mu}{S_n'}\right| > \frac{\sqrt{n}\epsilon}{S_n'}\right)$$

- On en déduit $t_{n-1,\alpha} = \frac{\sqrt{n}\epsilon}{S_n'}$ soit $\epsilon = \frac{S_n'}{\sqrt{n}} t_{n-1,\alpha}$, d'où l'intervalle de confiance :

$$\left[\overline{\mathbf{X}}_\mathbf{n} - \frac{\mathbf{S}_\mathbf{n}'}{\sqrt{\mathbf{n}}}\mathbf{t}_{\mathbf{n-1},\alpha}, \overline{\mathbf{X}}_\mathbf{n} + \frac{\mathbf{S}_\mathbf{n}'}{\sqrt{\mathbf{n}}}\mathbf{t}_{\mathbf{n-1},\alpha}\right]$$

Crédits : O. Gaudoin

Liens avec les fonctions de distributions et quantiles

Pour une loi normale

Pour une variable aléatoire $U \sim \mathcal{N}(0,1)$ et un réel $\alpha \in [0,1]$ on a défini la valeur u_α vérifiant :

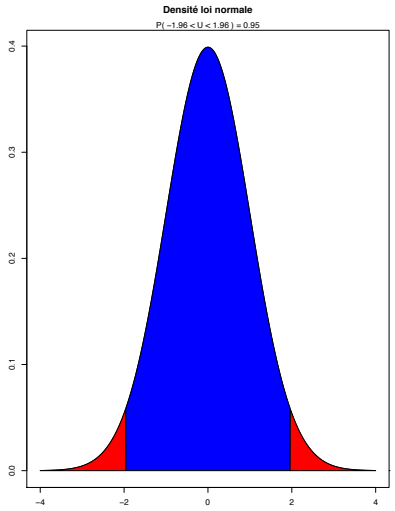
$$\mathbb{P}(|U| > u_\alpha) = \alpha$$

Par symétrie de la loi on a alors $\mathbb{P}(U > u_\alpha) = \frac{\alpha}{2}$ et $\mathbb{P}(U < -u_\alpha) = \frac{\alpha}{2}$.

$$F_U(u_\alpha) = \mathbb{P}(U \leq u_\alpha) = 1 - \mathbb{P}(U > u_\alpha) = 1 - \frac{\alpha}{2}\ ,$$

d'où $u_\alpha = F_U^{-1}(1 - \frac{\alpha}{2}) = \mathrm{qnorm}(1 - \frac{\alpha}{2})$. On a ainsi

$$\mathbb{P}(-u_\alpha \leq U \leq u_\alpha) = 1 - \alpha$$



Liens avec les fonctions de distributions et quantiles

Pour une loi du χ^2

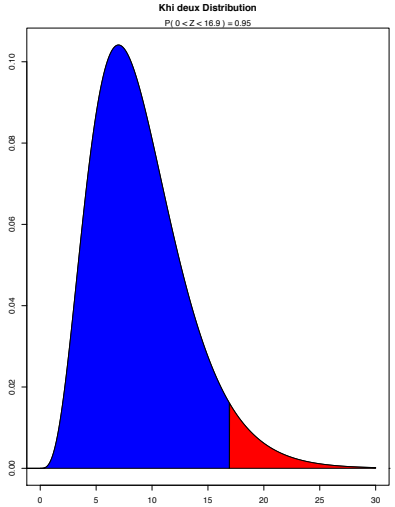
Pour une variable aléatoire $Z \sim \chi^2_n$ et un réel $\alpha \in [0,1]$ on a défini la valeur $z_{n,\alpha}$ vérifiant :

$$\mathbb{P}(Z > z_{n,\alpha}) = \alpha$$

$$F_Z(z_{n,\alpha}) = \mathbb{P}(Z \leq z_{n,\alpha}) = 1 - \mathbb{P}(Z > z_{n,\alpha}) = 1 - \alpha \ ,$$

d'où $z_{n,\alpha} = F_Z^{-1}(1-\alpha) = \mathrm{qchisq}(1-\alpha)$. On a ainsi

$$\mathbb{P}(0 \leq Z \leq z_{n,\alpha}) = 1 - \alpha$$



Liens avec les fonctions de distributions et quantiles

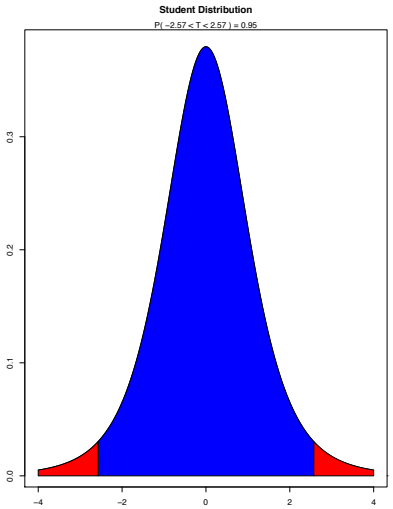
Pour une loi de Student

Pour une variable aléatoire $T \sim \mathrm{St}(n)$ et un réel $\alpha \in [0,1]$ on a défini la valeur $t_{n,\alpha}$ vérifiant :

$$\mathbb{P}(|T| > t_{n,\alpha}) = \alpha$$

Par symétrie de la loi on a alors $\mathbb{P}(T > t_{n,\alpha}) = \frac{\alpha}{2}$ et $\mathbb{P}(T < -t_{n,\alpha}) = \frac{\alpha}{2}$.

$$F_T(t_{n,\alpha}) = \mathbb{P}(T \leq t_{n,\alpha}) = 1 - \mathbb{P}(T > t_{n,\alpha}) = 1 - \frac{\alpha}{2}\ ,$$

d'où $t_{n,\alpha} = F_T^{-1}(1 - \frac{\alpha}{2}) = \mathrm{qt}(1 - \frac{\alpha}{2})$. On a ainsi

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