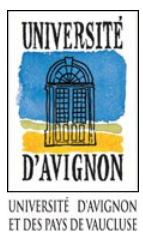
Pratiques des statistiques paramétriques

Denis Allard (I), Olivier Martin (II), Etienne Klein (III)

Biostatistique et Processus Spatiaux (BioSP), INRA, Avignon

11-13 avril 2012



Plan général

I Statistique inférentielle : estimation,test
Denis Allard

II Régression simple et multiple ; modèle linéaire ; test pour les modèles linéaires
Olivier Martin

III Analyse de la variance à un facteur ; à deux facteurs équilibrés
Etienne Klein

Statistique inférentielle

Définition

- Étendre les propriétés constatées sur l'échantillon à la population toute entière
- Valider ou infirmer des hypothèses a priori ou formulées après une *phase exploratoire*
- Utilise la théorie des probabilités

Objectifs de la séquence

- Savoir estimer une moyenne, une variance : ponctuellement, par intervalle de confiance
- Savoir tester une moyenne, une variance, un coefficient de corrélation
- Savoir tester une différence de moyenne, une différence de variance

Quelques définitions

Population

Une population est l'ensemble des objets ou des personnes (les individus) auxquels une étude statistique s'intéresse. Une population doit être définie par des critères ne laissant aucune équivoque : il faut être capable de dire sans erreurs possibles si tel objet ou telle personne appartient ou non à la population.

- Les étudiants inscrits en 2012 au module "Statistiques paramétriques" de l'ED 536
- Les étudiants régulièrement inscrits à l'université d'Avignon
- La population municipale (?)

Ex : population municipale

Décret n˚2003-485 publié au Journal officiel du 8 juin 2003

La population municipale comprend les personnes ayant leur résidence habituelle (au sens du décret) sur le territoire de la commune, dans un logement ou une communauté, les personnes détenues dans les établissements pénitentiaires de la commune, les personnes sans-abri recensées sur le territoire de la commune et les personnes résidant habituellement dans une habitation mobile recensée sur le territoire de la commune. La population municipale d'un ensemble de communes est égale à la somme des populations municipales des communes qui le composent.

- *Les étudiants majeurs vivant en internat*
- *Pas les mineurs (comptés chez leur parents)*
- *Les militaires logés dans un établissement militaire (caserne, quartier, base, camp militaire...)*
- *Les personnes détenues dans un établissement pénitentiaire de la commune*

Ne comporte pas de doubles comptes

Echantillons

Echantillon

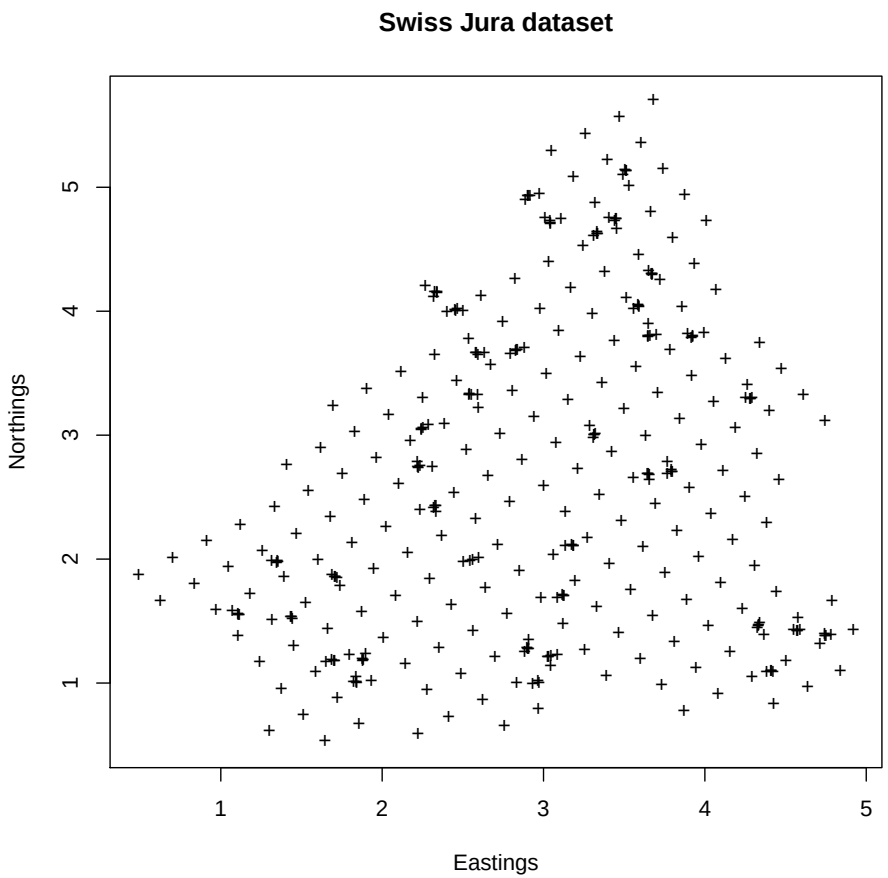
Un échantillon $X_1, \ldots, X_n$ est un sous-ensemble d'une population.

Echantillon aléatoire

Un échantillon est aléatoire si chaque individu qui le compose est tiré aléatoirement dans la population

- de façon équiprobable
- indépendamment les uns des autres

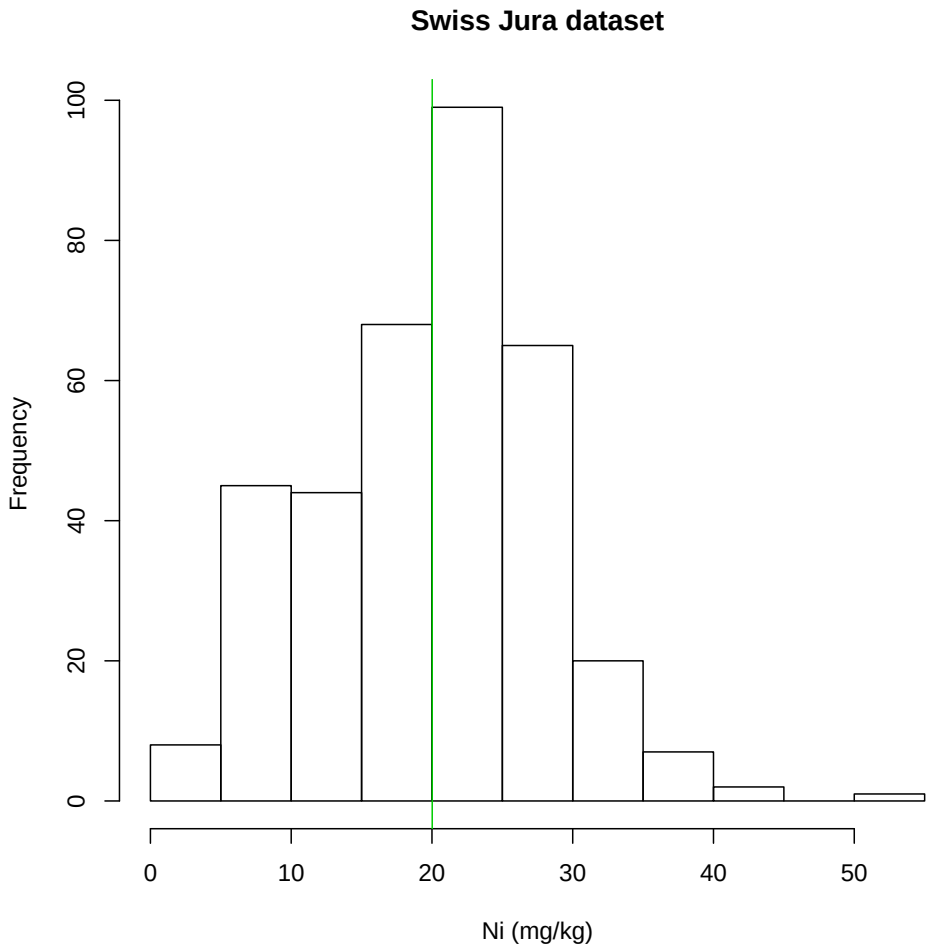
Jeu de données



359 échantillons dans le Jura Suisse, 5×5 km

- Grille régulière + densification aléatoire
- 7 teneurs de métaux lourds mesurés : Cd, Co, Cr, Cu, Ni, Pb, Zn
- 5 types de roches : Argovien, Kimmeridgien, Sequenien, Portlandien, Quaternaire
- 4 types d'usage : Forêt, pâtures, prairie, labours
- (x, y) non utilisé

Jeu de données



Ni : Moyenne = 20.02

Estimation de la moyenne

Echantillons (sans remise) de 30 valeurs

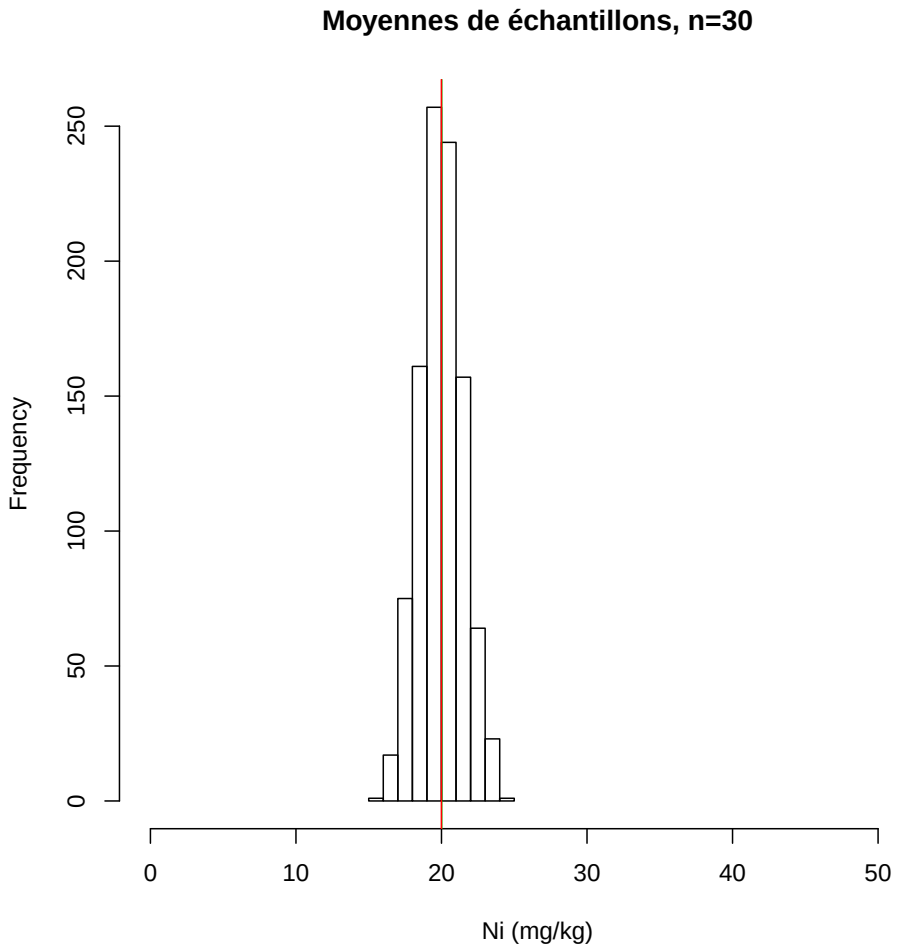
1. Moyenne # 1 : 20.95
2. Moyenne # 2 : 20.35
3. Moyenne # 3 : 20.47

   ⋮

4. Moyenne #100 : 18.12

Moyenne des moyennes : 19.91

Estimation de la moyenne



Comment quantifier cette variabilité ?

Rappels de probabilité

Variable aléatoire

Issue d'une expérience aléatoire décrite par une variable numérique, à valeur dans $\mathbf{R}$ ou dans une partie de $\mathbf{R}$: $\mathbf{R}^+$, intervalle,...

Espérance mathématique

Paramètre de position, qui indique où se trouve la variable aléatoire :

$$E[x] = \sum_k k.p_k \quad \text{ou} \quad E[X] = \int_{\mathbf{R}} x.f(x)dx$$

Variance

Paramètre qui indique comment la variable se répartit autour de l'espérance mathématique

$$\text{Var}(x) = \sum_k (k - E[X])^2.p_k \quad \text{ou} \quad E[X] = \int_{\mathbf{R}} (x - E[X])^2.f(x)dx$$

Rappels de probabilité

Variable aléatoire uniforme

$X \sim \mathcal{U}(a, b)$

$$f(x) = \frac{1}{b-a}, \quad x \in [a, b]$$

Alors

$$E[X] = (b+a)/2; \quad \mathrm{Var}(X) = (b-a)^2/12.$$

On peut aussi écrire

$$X \sim a + (b-a)\mathcal{U}(0, 1)$$

Variable aléatoire exponentielle

$X \sim \mathcal{E}(\lambda)$

$$f(x) = \lambda e^{-\lambda x}, \quad x \in \mathbf{R}^+$$

Alors

$$E[X] = 1/\lambda; \quad \mathrm{Var}(X) = 1/\lambda^2.$$

Rappels de probabilité

Variable aléatoire Gaussienne

$X \sim \mathcal{N}(\mu, \sigma^2)$

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{\sigma^2}}, \quad x \in \mathbf{R}$$

Alors

$$E[X] = \mu; \quad \text{Var}(X) = \sigma^2.$$

On peut aussi écrire

$$X \sim \mu + \sigma\mathcal{N}(0,1)$$

Quantiles

On notera $u_{1-\alpha}$ la valeur telle que

$$P(\mathcal{N}(0,1) \leq u_{1-\alpha}) = 1 - \alpha$$

Exemple : $1 - \alpha = 0.975 \Leftrightarrow \alpha = 0.025 \Leftrightarrow u_{1-\alpha} = 1.96$

Estimation de la moyenne

Soit $X_1, \ldots, X_n$ un échantillon aléatoire de taille n provenant d'une population d'espérance μ et de variance σ^2. La moyenne arithmétique

$$\hat{\mu} = \bar{X} = \frac{1}{n}(X_1 + \cdots + X_n) = \frac{1}{n}\sum_{i=1}^{n} X_i$$

est l'estimateur naturel de μ.
C'est une variable aléatoire.

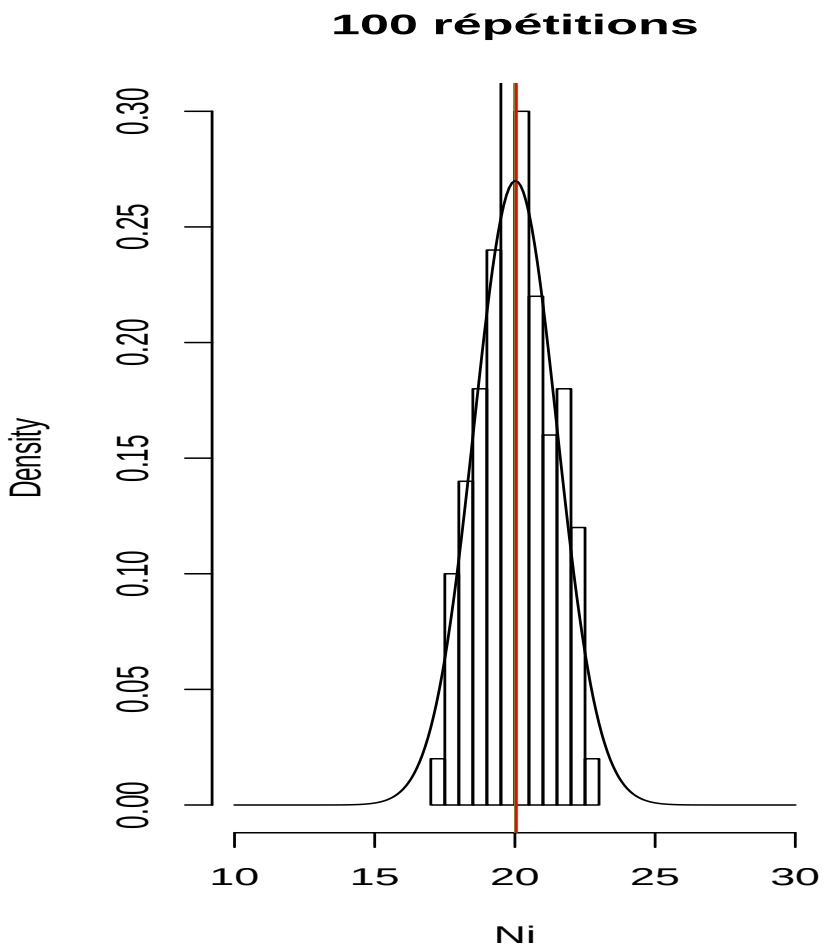
$$\begin{aligned} E[\bar{X}] &= \mu \\ \text{Var}(\bar{X}) &= \frac{\sigma^2}{n} \end{aligned}$$

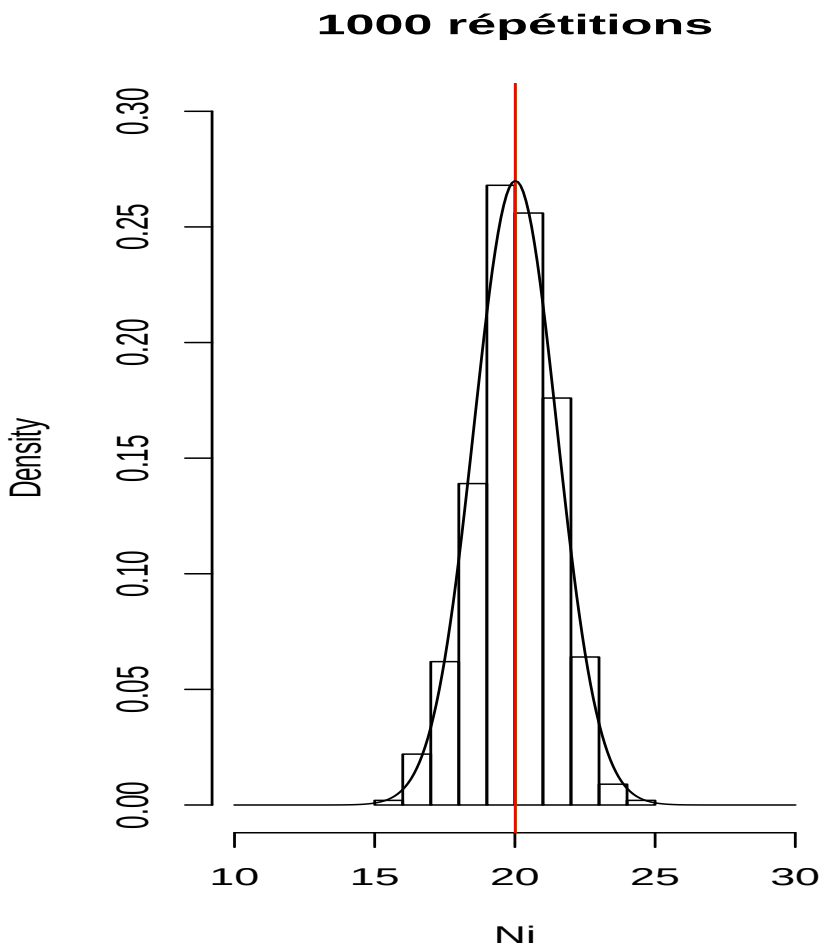
Théorème Central Limite

Lorsque $n \to \infty$

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \to \mathcal{N}(0, 1)$$

Retour sur l'estimation de la moyenne





Estimation par intervalle de confiance

1. On cherche un intervalle $[\hat{\mu}_{inf}, \hat{\mu}_{sup}]$ qui contient la vraie valeur $\mu = E[X]$ avec la probabilité $1 - \alpha$: c'est le niveau.
2. On va également poser

$$P(\mu < \hat{\mu}_{inf}) = P(\mu \geq \hat{\mu}_{sup}) = \alpha/2.$$

Intervalle de confiance : σ^2 connue

Faisons l'hypothèse (irréaliste !) que σ^2 est connue

$$[\hat{\mu}_{inf}, \hat{\mu}_{sup}] = [\bar{X} - u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}]$$

L'amplitude de l'intervalle

- Augmente avec le niveau
- Augmente avec la variance
- Diminue avec n

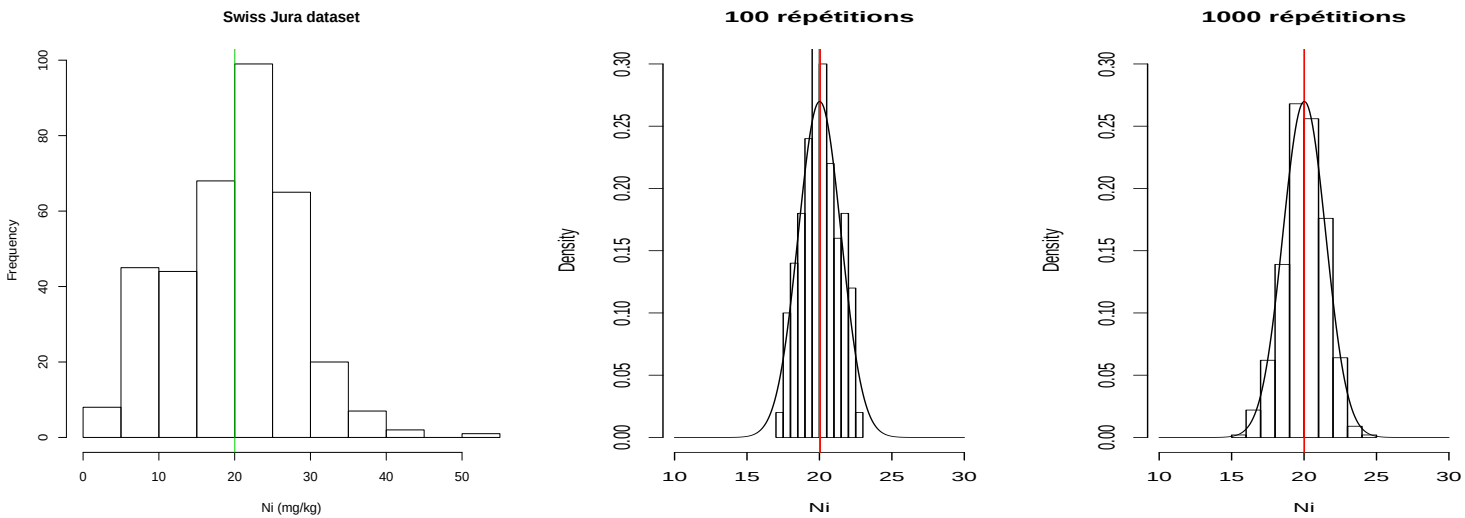
Preuve

Par le TCL,

$$\begin{aligned} 1-\alpha &= P\left(-u_{1-\alpha/2} \leq \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \leq u_{1-\alpha/2}\right) \\ &= P\left(-u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} - \bar{X} \leq -\mu \leq u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} - \bar{X}\right) \\ &= P\left(\bar{X} - u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right) \end{aligned}$$

CQFD.

Retour sur l'estimation de la moyenne



On pose $1 - \alpha = 0.95 \Leftrightarrow u_{1-\alpha/2} = 1.96$.
1000 échantillons de taille 30. On trouve

$$\#(\mu < \hat{\mu}_{inf}) = 24; \quad \#(\mu > \hat{\mu}_{sup}) = 19.$$

Valeur théorique attendue : 25.

Estimation de la variance

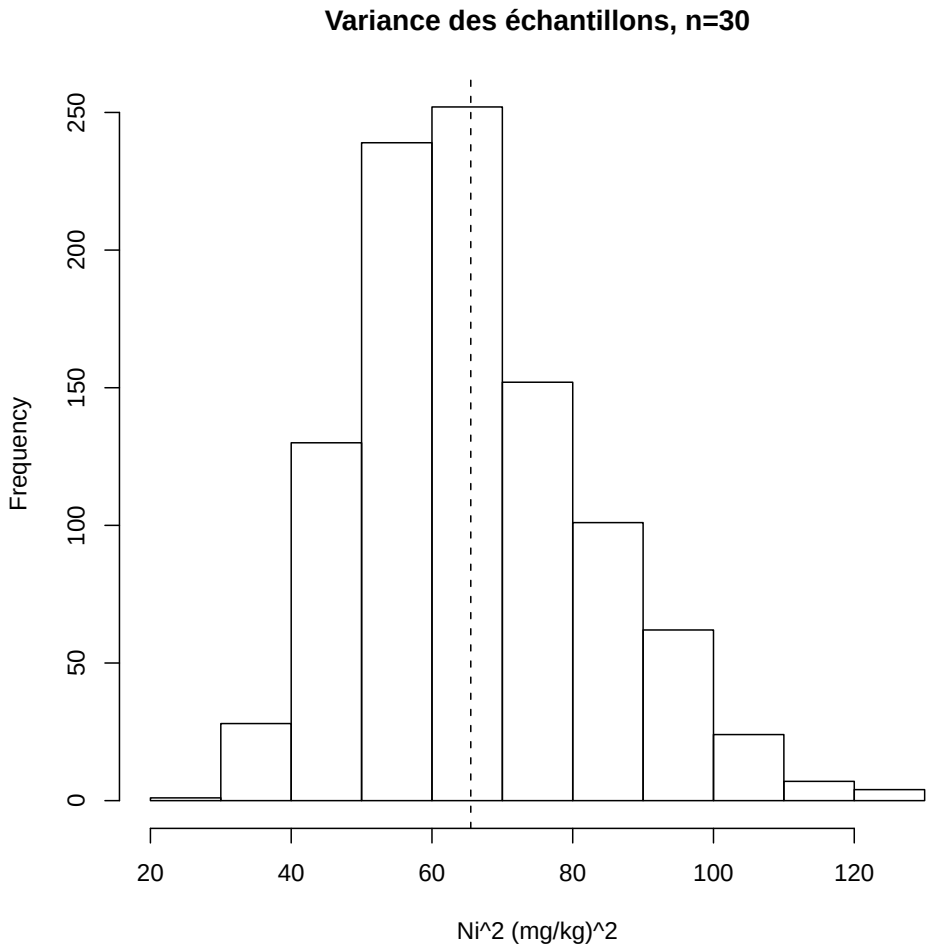
Echantillons (sans remise) de 30 valeurs

1. Variance # 1 : 72.2
2. Variance # 2 : 55.1
3. Variance # 3 : 107.8

   ⋮
4. Variance #100 : 95.3

Variances des 359 échantillons : 65.5

Estimation de la variance



Il faut également quantifier cette variabilité.

Encore des probabilités

σ^2 pas connue, mais estimée ! ! Besoin de nouveaux outils.

Estimation de la variance

L'estimateur

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

est sans biais, i.e. $E[\hat{\sigma}^2] = \sigma^2$.

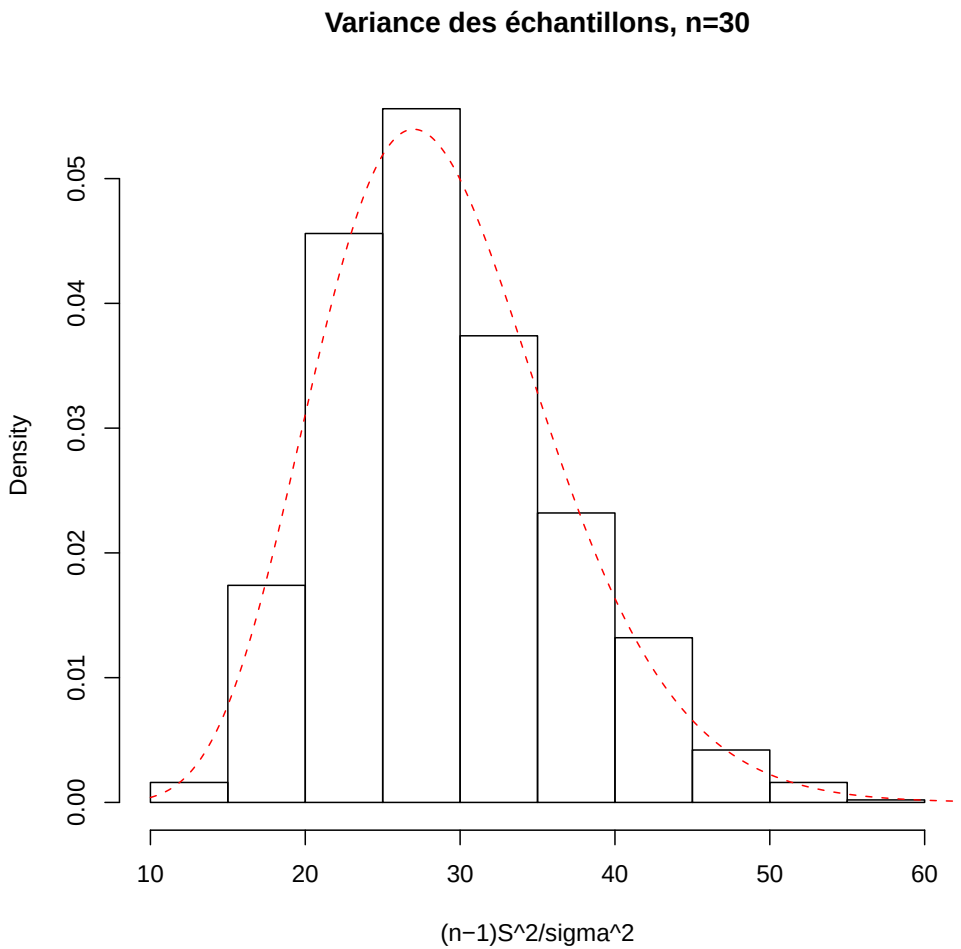
Loi khi-deux

Si $X_1, \ldots, X_n$ est un échantillon (indépendant) Gaussien,

$$(n-1)S^2/\sigma^2 \sim \chi^2_{(n-1)}.$$

Il y a $(n-1)$ d.d.l. car $(n-1)$ VA indépendantes dans S^2.

La variance comme une distribution khi-2



Encore des probabilités

Loi de Student

Si $X \sim \mathcal{N}(0,1)$ et $Y \sim \chi^2_{(n)}$ indépendantes, alors

$$\frac{X}{Y/\sqrt{n}} \sim t_n$$

loi de Student à n d.d.l.

Loi de Fisher

Si $X \sim \chi^2_{(n_X)}$ et $Y \sim \chi^2_{(n_Y)}$ indépendantes, alors

$$\frac{X/\sqrt{n_X}}{Y/\sqrt{n_Y}} \sim F^{n_X}_{n_Y}.$$

Lois tabulées dans R

Estimation par intervalle de confiance

Intervalle de confiance : σ^2 est inconnue

En utilisant la définition d'une loi de Student

$$[\hat{\mu}_{inf}, \hat{\mu}_{sup}] = [\bar{X} - t_{1-\alpha/2}\frac{S}{\sqrt{n}}, \bar{X} + t_{1-\alpha/2}\frac{S}{\sqrt{n}}]$$

où

$$P(t_{(n-1)} \leq t_{1-\alpha/2}) = 1 - \alpha/2$$

L'amplitude de l'intervalle est plus grande que lorsque σ est connu.
Mêmes 1000 échantillons de taille 30. On trouve

$$\#(\mu < \hat{\mu}_{inf}) = 29; \quad \#(\mu > \hat{\mu}_{sup}) = 24.$$

Valeur théorique attendue : 25. Hypothèse Gaussienne ?

Tests statistiques

Des études antérieurs et/ou mon expertise en science du sol me conduisent à penser que la moyenne de Ni dans la population est $\mu = 20$. J'observe une moyenne $\bar{X} = 22.2$ sur 30 échantillons, avec $S^2 = 52$.
$\hookrightarrow$ Différence due au hasard, ou différence significative ?

Définition

Tests statistiques = Outil d'aide à la décision pour vérifier une hypothèse. Il permet de trancher entre

- une hypothèse neutre, appelée hypothèse "nulle", notée H_0
- une hypothèse alternative, notée H_1 ou H_a.

Tests statistiques

Test

$$H_0 \text{ vs. } H_1$$

Il faut trancher entre deux hypothèses ; l'une seulement est vraie. La décision peut mener à deux risques d'erreurs :

	Décision	
	Retenir H_0	Retenir H_1
H_0 vraie	Choix correct	Erreur I
H_1 vraie	Erreur II	Choix correct

- Niveau $1-\alpha$ = Proba de ne pas commettre d'erreur de type I (se calcule sous l'hypothèse H_0 vraie)
- Puissance $1-\beta$ = Proba de ne pas commettre d'erreur de type II (se calcule sous l'hypothèse H_1 vraie)

Conduite d'un test : un exemple

Ex : je pense que la moyenne de Ni dans la population est $\mu = 20$.
J'observe $\bar{X} = 22.2$ *sur 30 échantillons avec* $S^2 = 52$.

$\bar{X}$ compatible avec mon hypothèse ?

1. Choisir le niveau $1 - \alpha = 0.95$
2. On définit les hypothèses $H_0 : \ \mu = 20$; $H_1 : \ \mu > 20$
3. On utilise $(\bar{X} - \mu)/(S/\sqrt{n-1}) \sim t_{(n-1)}$
4. Si $(\bar{X} - \mu)/(S/\sqrt{n-1})$ est "trop grand" on doit rejeter H_0
5. On trouve dans les tables $P(t_{(n-1)} \leq 1.70) = 0.95$
   Donc $t_{1-\alpha} = 1.7$ est la valeur critique
6. Or $(\bar{X} - \mu)/(S/\sqrt{n-1}) = (22.2 - 20)/\sqrt{52/29} = 1.64 < 1.7$
7. On ne rejette pas H_0

Conduite d'un test : cas général

Le présupposé est que l'hypothèse nulle est vraie. On doit prouver le contraire. Les calculs ne font sous H_0.

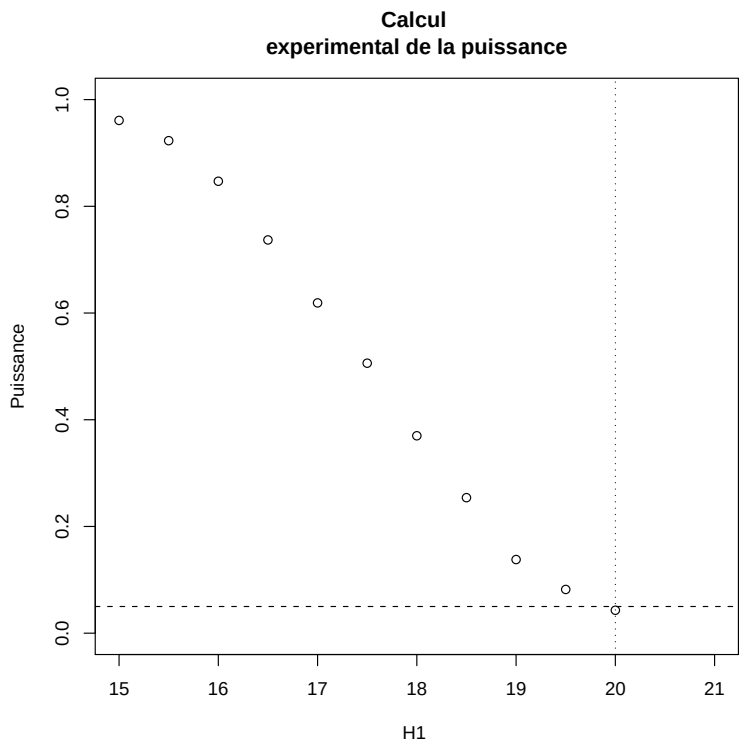
1. Choisir le niveau $1 - \alpha$
2. Choisir les hypothèses nulles et alternatives
3. Déterminer une statistique de test, notée T, (laquelle ? "la meilleure" ... on verra plus tard)
4. Déterminer la valeur critique de T, en fonction de α, n,...
5. Calcul de la valeur expérimentale de la statistique
6. Rejet ou non de H_0

Retour sur le test de la moyenne

- Le niveau est fixé par l'utilisateur

$$1 - \alpha = P(H_0 \mid H_0) = \text{Ne pas rejeter } H_0 \text{ à tort}$$

- Puissance : nécessite de spécifier H_1 ; calcul parfois difficile ; accessible par la simulation



Significativité

Définition

- La significativité, p, aussi appelée p-valeur est la probabilité de rejeter H_0 pour la statistique observée T
- Cette probabilité est calculée sous l'hypothèse H_0
- L'utilisateur compare p avec sa propre valeur α

Ex : je pense que la moyenne de Ni dans la population est $\mu = 20$.

J'observe $\bar{X} = 22.2$ sur 30 échantillons avec $S^2 = 52$.

$$\begin{aligned} p &= 1 - P\left(t_{(n-1)\leq}(\bar{X} - \mu)/(S/\sqrt{n-1})\right) \\ &= 1 - P\left(t_{(29)} \leq (22.2 - 20)/\sqrt{52/29}\right) \\ &= 1 - P(t_{(29)} \leq 1.643) \\ &= 0.0556 \end{aligned}$$

Sous H_0, $p \sim \mathcal{U}(0, 1)$. 1000 échantillons de 30 valeurs : 40 p-valeurs < 0.05.

Test d'une variance

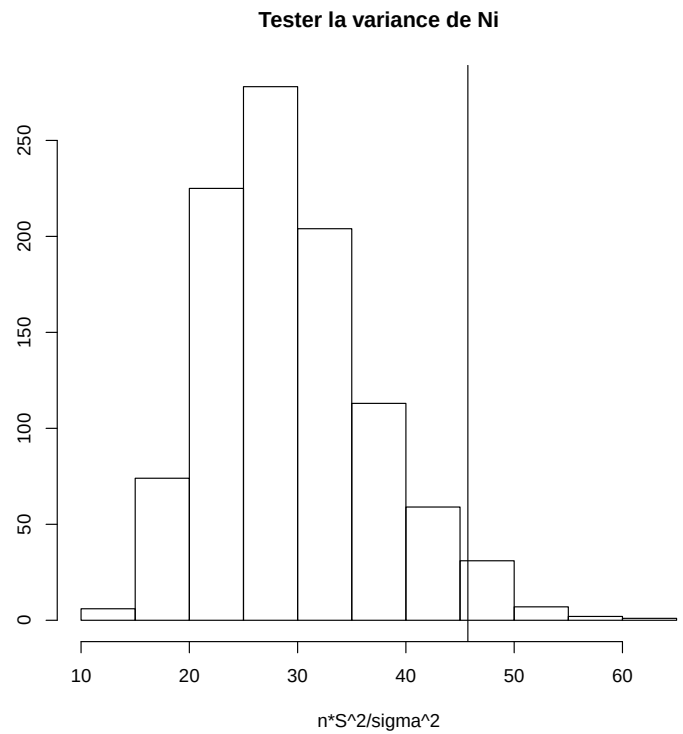
Ex : je pense que la variance de Ni dans la population est $\sigma^2 = 65.5$.
J'observe $S^2 = 72.2$ *sur 30 échantillons.*

S^2 *compatible avec mon hypothèse ?*

1. Choisir le niveau $1 - \alpha = 0.95$
2. On définit les hypothèses $H_0 : \sigma^2 = 65.5$; $H_1 : \sigma^2 \neq 65.5$
3. On utilise $S^2/(\sigma^2/n) = nS^2/\sigma^2 \sim \chi^2_{(n-1)}$
4. Si nS^2/σ^2 est "trop grand" on doit rejeter H_0
5. On trouve dans les tables $P(\chi^2_{(n-1)} \leq 45.72) = 0.975$
   Donc $\chi^2_{(n-1)} = 45.72$ est la valeur critique
6. Or $nS^2/\sigma^2 = 30 * 72.2/65.5 = 33.06 < 45.72$
7. On ne rejette pas H_0

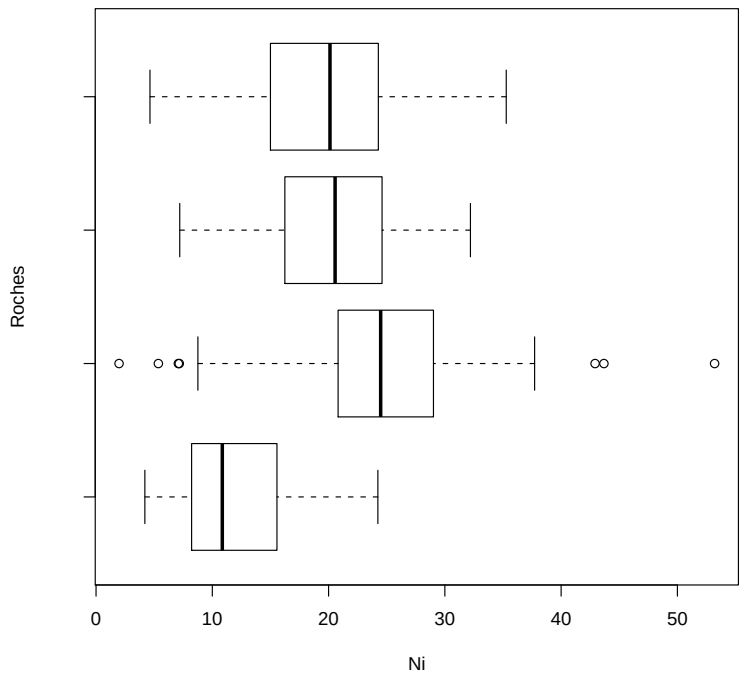
Nécessite hypothèse Gaussienne.

Test d'une variance



1000 répétitions d'un échantillon de taille 30 : taux de rejet = 0.031.

Retour sur les données



Un type de roche semble très différent des autres. Différence significative ?

Tester deux moyennes lorsque les variances sont identiques

Test

$$H_0 : \mu_1 = \mu_2; \quad H_1 : \mu_1 \neq \mu_2$$

ou encore

$$H_0 : \mu_1 - \mu_2 = 0; \quad H_1 : \mu_1 - \mu_2 \neq 0$$

avec $\sigma_1^2 = \sigma_2^2 = \sigma^2$.

Sous hypothèse Gaussienne, on a

$$\frac{n_1 S_1^2}{\sigma^2} \sim \chi^2_{n_1-1} \text{ et } \frac{n_2 S_2^2}{\sigma^2} \sim \chi^2_{n_2-1}$$

et

$$\bar{X}_1 \sim \mathcal{N}(\mu_1, \sigma^2/\sqrt{n_1}) \text{ et } \bar{X}_2 \sim \mathcal{N}(\mu_2, \sigma^2/\sqrt{n_2})$$

Tester deux moyennes lorsque les variances sont identiques

Et donc,

$$\frac{n_1 S_1^2 + n_2 S_2^2}{\sigma^2} \sim \chi^2_{n_1+n_2-2}$$

et

$$\bar{X}_1 - \bar{X}_2 \sim \mathcal{N}\left(\mu_1 - \mu_2 = 0, \sigma^2(1/n_1 + 1/n_2)\right)$$

On utilise finalement la statistique de test

$$T = \frac{\bar{X}_1 - \bar{X}_2}{(n_1 S_1^2 + n_2 S_2^2)(1/n_1 + 1/n_2)} \sqrt{n_1 + n_2 - 2} \sim t_{(n_1+n_2-2)}$$

Tester si deux moyennes sont identiques, lorsque les variances son identiques

Ex : Pour les roches de type 1 (Argovien), on observe $\bar{X}_1 = 12.38$, $n_1 = 76$ *et* $S_1^2 = 30.98$.
Pour les roches de type 3 (Séquenien), on observe $\bar{X}_3 = 20.42$, $n_3 = 89$ *et* $S_3^2 = 32.04$.

Différence significative ou non ?

1. Choisir le niveau $1 - \alpha = 0.95$
2. On définit les hypothèses $H_0 : \mu_1 - \mu_2 = 0$ et $H_1 : \mu_1 - \mu_2 \neq 0$
3. On considère que les variances sont semblables, et que $\sigma_1^2 = \sigma_2^2 = \sigma^2$
4. On utilise la statistique de Student, qui ne doit pas être "trop différente de 0"
5. On trouve dans les tables $P(t_{(n_1+n_2-2)} \leq 1.97) = 0.975$
6. Or $T = 9.28$
7. On rejette H_0
8. p-valeur de $T = 9.28$ est 0 !!

Rappel : Nécessite hypothèse Gaussienne

Tester si deux variances sont identiques

Test

$$H_0 : \sigma_1^2 = \sigma_2^2; \quad H_1 : \sigma_1^2 \neq \sigma_2^2$$

ou encore

$$H_0 : \sigma_1^2/\sigma_2^2 = 1; \quad H_1 : \sigma_1^2/\sigma_2^2 \neq 1$$

Sous hypothèse Gaussienne, on a

$$\frac{n_1 S_1^2}{\sigma^2} \sim \chi^2_{n_1-1} \text{ et } \frac{n_2 S_2^2}{\sigma^2} \sim \chi^2_{n_2-1}$$

et donc

$$\frac{n_1 S_1^2}{n_1 - 1} \Big/ \frac{n_2 S_2^2}{n_2 - 1} \sim F_{n_1-1} n_2 - 1.$$

Tester si deux variances sont identiques

Ex : Pour les roches de type 1 (Argovien), on observe $n_1 = 76$ *et* $S_1^2 = 30.98$.
Pour les roches de type 2 (Kimmeridgien), on observe $n_2 = 124$ *et* $S_2^2 = 54.59$.

Différence significative ou non ?

1. Choisir le niveau $1 - \alpha = 0.95$
2. On définit les hypothèses $H_0 : \sigma_1^2/\sigma_2^2 = 1$ et $H_1 : \sigma_1^2/\sigma_2^2 \neq 1$
3. On utilise la statistique de Fisher, qui ne doit pas être "trop différente de 1"
4. On trouve dans les tables $P(F_{(n_1-1,n_2-1)} \leq 0.66) = 0.025$ et $P(F_{(n_1-1,n_2-1)} \geq 1.49) = 0.975$
5. Or $F = 0.57$
6. On rejette H_0
7. p-valeur de $F = 0.57$ est 0.0042

Rappel : Nécessite hypothèse Gaussienne

Test de comparaison(s)

- Lorsque n modéré ($n \leq 30$), attention à l'ordre : tester les variances d'abord ; si variances identiques non rejeté, tester les moyennes
- Nécessite une hypothèse Gaussienne dans la définition des lois de Student et de Fisher ;
  mais, si n est élevé, alors
  - par le TCL $\bar{X} \to \mathcal{N}$ lorsque $n \to \infty$
  - par le TCL, $nS^2/\sigma^2 \to \chi^2_{n-1} \to \mathcal{N}$ lorsque $n \to \infty$
  - par le TCL $t_{n-1} \to \mathcal{N}$ lorsque $n \to \infty$

$\hookrightarrow$ on peut quand même tester les moyennes en utilisant la formule de Student lorsque n grand (quelques dizaines)

En dehors de ces cas, utiliser d'autres tests, p.ex. tests non paramétriques

Un cas particulier : estimer/tester une fréquence

Dans une population, une proportion p des individus partagent une caractéristique.
$X_1, \ldots, X_n$: échantillon qui vaut 1 si caractéristique est observée, 0 sinon.

1. Estimer p à l'aide de $Y = X_1 + \cdots + X_n$
2. Tester $H_0 : p = p_0$ vs. $H_1 : p \neq p_0$.

Rappel de probabilité

X est une variable aléatoire de Bernouilli(p)

$$E[X] = p \text{ et } Var(x) = p(1-p)$$

Y est une variable aléatoire Binomiale(n, p)

$$E[X] = np \text{ et } Var(Y) = np(1-p)$$

Un cas particulier : estimer/tester une fréquence

Estimateur

$$\hat{p} = \frac{Y}{n}$$

avec

$$E[\hat{p}] = p \text{ et } Var(\hat{p}) = \frac{p(1-p)}{n}$$

TCL : lorsque $n \to \infty$,

$$\frac{\hat{p} - p}{\sqrt{p(1-p)/n}} \to \mathcal{N}(0, 1)$$

Fréquences : estimation par intervalle de confiance

On choisit un niveau $1-\alpha$; on en déduit $u_{1-\alpha/2}$. Alors,

$$\begin{aligned} 1-\alpha &= P\left(-u_{1-\alpha/2} \leq (\hat{p}-p)/\sqrt{p(1-p)/n} \leq u_{1-\alpha/2}\right) \\ &= P\left(-u_{1-\alpha/2}\sqrt{p(1-p)/n} - \hat{p} \leq -p \leq \hat{p} + u_{1-\alpha/2}\sqrt{p(1-p)/n}\right) \\ &= P\left(\hat{p} - u_{1-\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n} \leq p \leq \hat{p} + u_{1-\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}\right) \end{aligned}$$

Intervalle de confiance

IC au niveau $1-\alpha$ est

$$\hat{p} \pm u_{1-\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}$$

Amplitude augmente avec le niveau ; diminue avec n ; dépend de p

Ex : $p \simeq 0.5$, $n = 900$, $1-\alpha = 0.95 \Rightarrow u_{1-\alpha/2} \simeq 2$. Alors IC $= [\hat{p} \pm 0.033]$

Fréquences : test

Deux échantillons, taille n_1 et n_2. On observe $\hat{p}_1$ et $\hat{p}_2$.

Test

$$H_0 : \ p_1 = p_2 = p \ \ vs. \ \ H_1 : p_1 \neq p_2$$

Sous H_0,

$$\hat{p}_1 \sim \mathcal{N}(p, p(1-p)/n_1); \quad \hat{p}_2 \sim \mathcal{N}(p, p(1-p)/n_2)$$

Donc,

$$\hat{p}_1 - \hat{p}_2 \sim \mathcal{N}(0, p(1-p)(1/n_1 + 1/n_2))$$

On utilise comme statistique de test

$$D = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{p(1-p)(1/n_1 + 1/n_2)}}$$

qui est comparé à $u_{1-\alpha/2}$.

Fréquences : test

On utilise comme statistique de test

$$D = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{p(1-p)(1/n_1 + 1/n_2)}}$$

qui est comparé à $u_{1-\alpha/2}$.

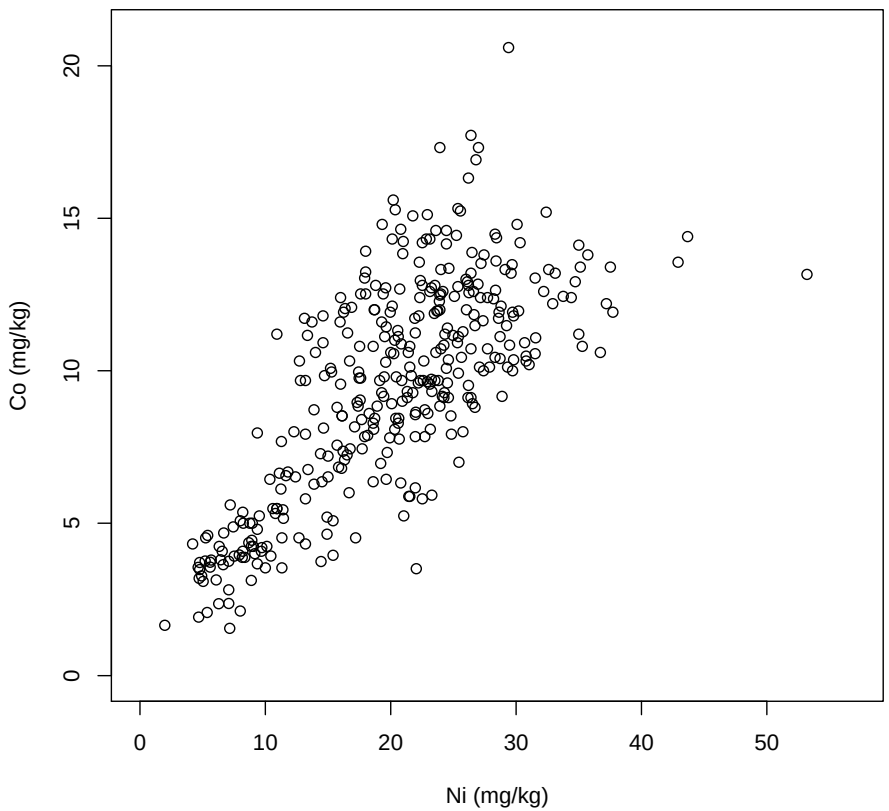
Exemple : sondages

Sondage 1 pour un candidat $n_1 = 987$ et $\hat{p}_1 = 27\%$; sondage 2 donne $n_2 = 1625$ et $\hat{p}_2 = 33\%$.
Différence significative ou non ?

- On trouve
  $D = (0.27 - 0.33)/\sqrt{1/987 + 1/1625} = -1.49 \in [-1.96, 1.96]$.
- On ne rejette pas H_0 au niveau 0.95
- p-valeur du test unilatéral : 0.068

Corrélation entre deux variables

Une série bivariée : $(X_1, Y_1), \ldots, (X_n, Y_n)$.



Corrélation entre deux variables

Une série bivariée : $(X_1, Y_1), \ldots, (X_n, Y_n)$.

Définition

$$R = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{S_X S_Y}$$

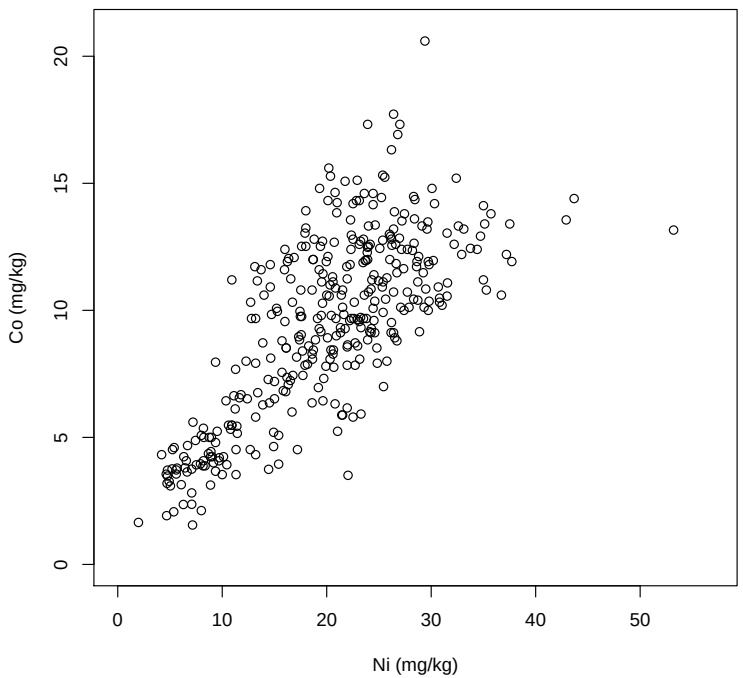
avec

$$S_X^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2; \quad S_Y^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

$R \in [-1; 1]$.

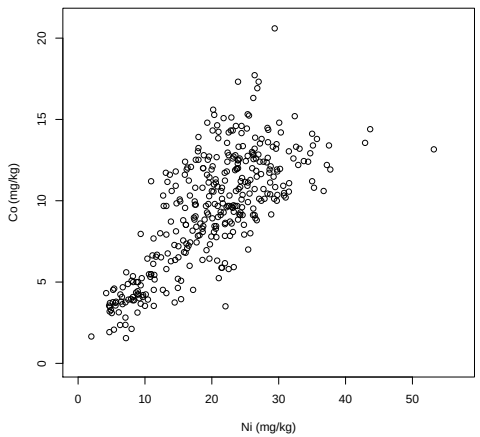
Corrélation entre deux variables

Une série bivariée : $(X_1, Y_1), \ldots, (X_n, Y_n)$.

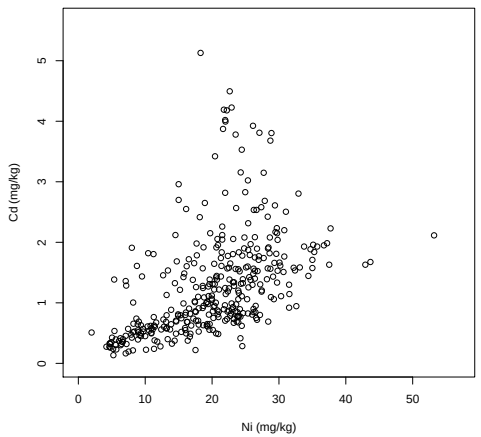


$R = 0.74$

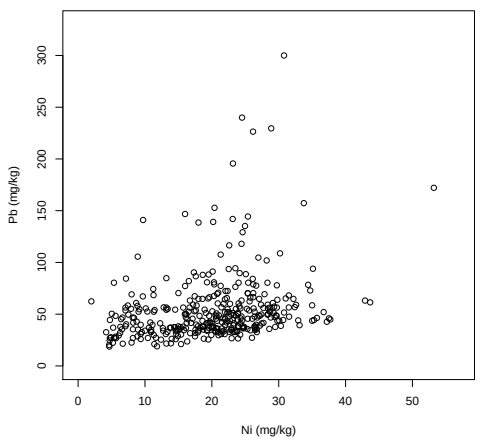
Corrélation entre deux variables



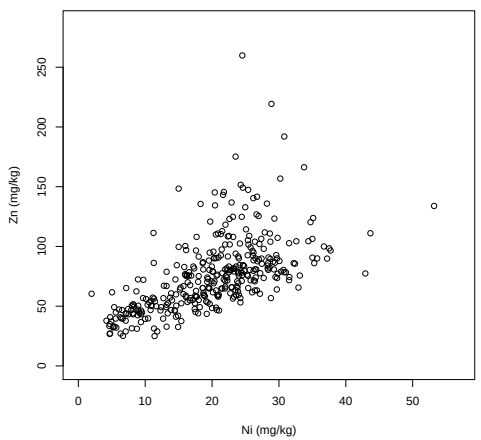
$R = 0.74$



$R = 0.49$



$R = 0.27$



$R = 0.60$

Corrélation entre deux variables

Attention !

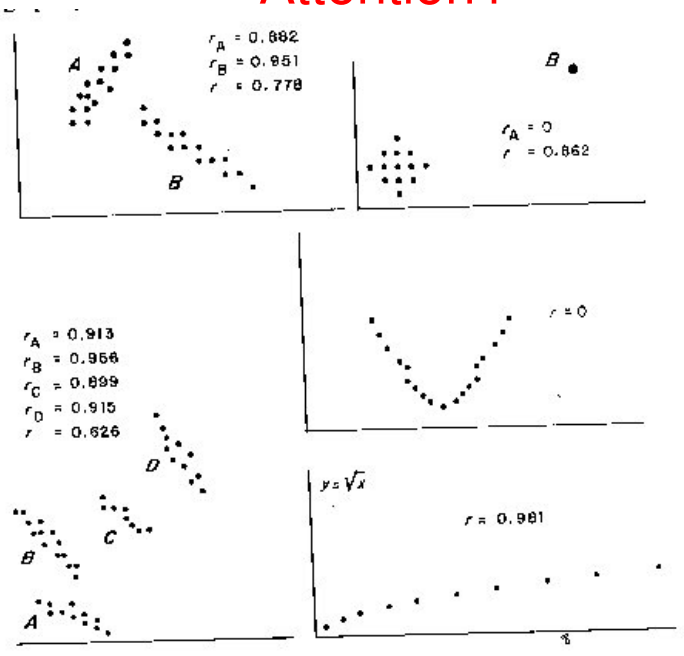
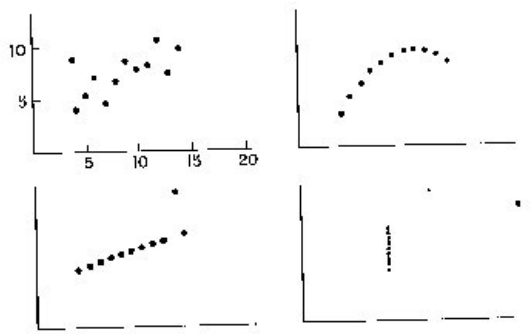


Fig. 7.2



Coefficient de corrélation linéaire

Entre Ni et Pb, on observe $\rho = 0.27$. Intervalle de confiance ?

Résultat

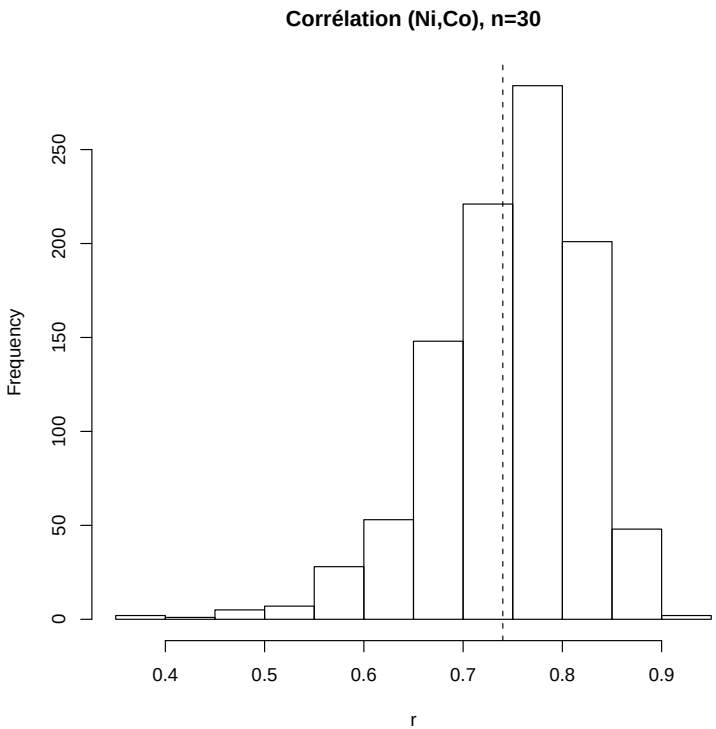
Sous hypothèse Gaussienne,

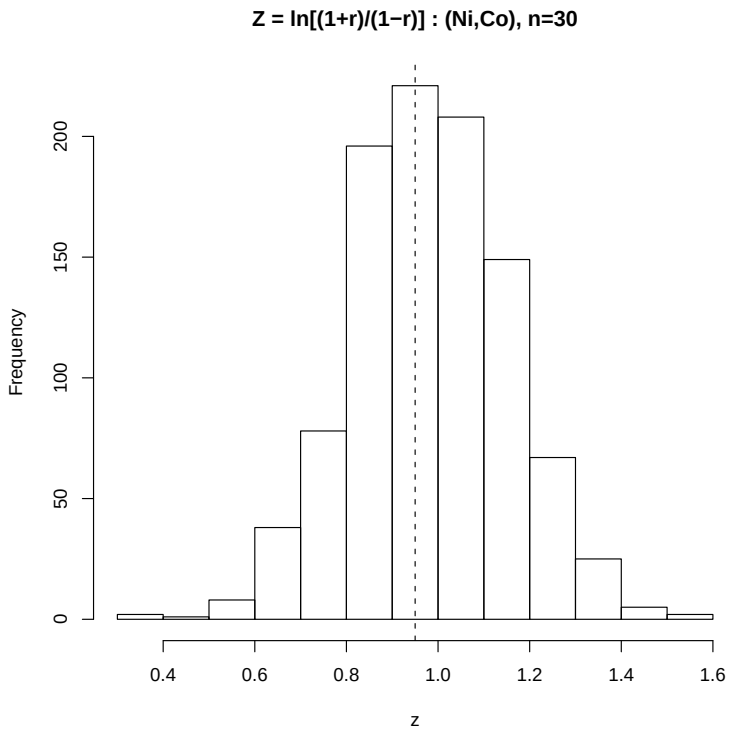
$$Z = \frac{1}{2}\ln\left(\frac{1+R}{1-R}\right) \to \mathcal{N}\left(\frac{1}{2}\ln(\frac{1+\rho}{1-\rho}); \frac{1}{n-3}\right)$$

lorsque $n \to \infty$.

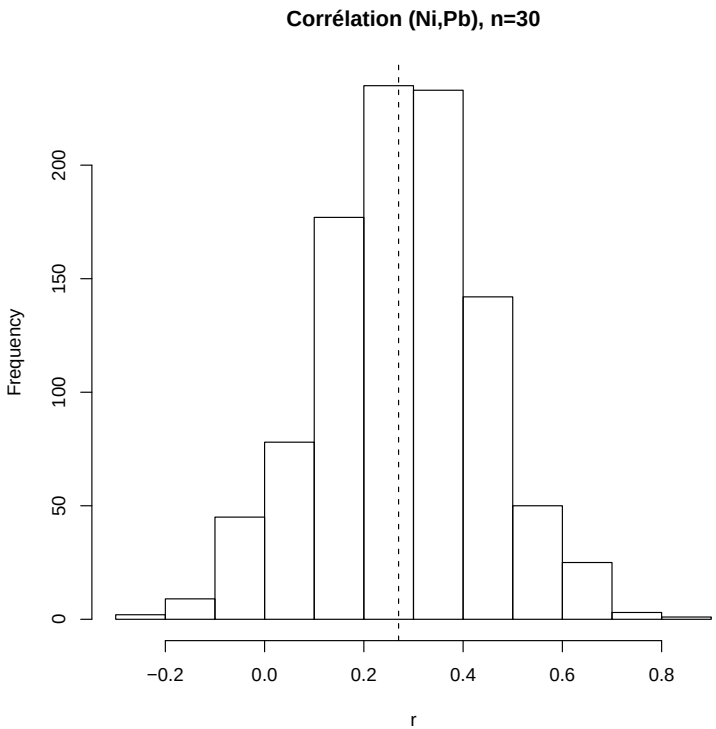
Si le couple (X, Y) n'est pas Gaussien, il faut n grand (> 30) pour utiliser cette convergence.

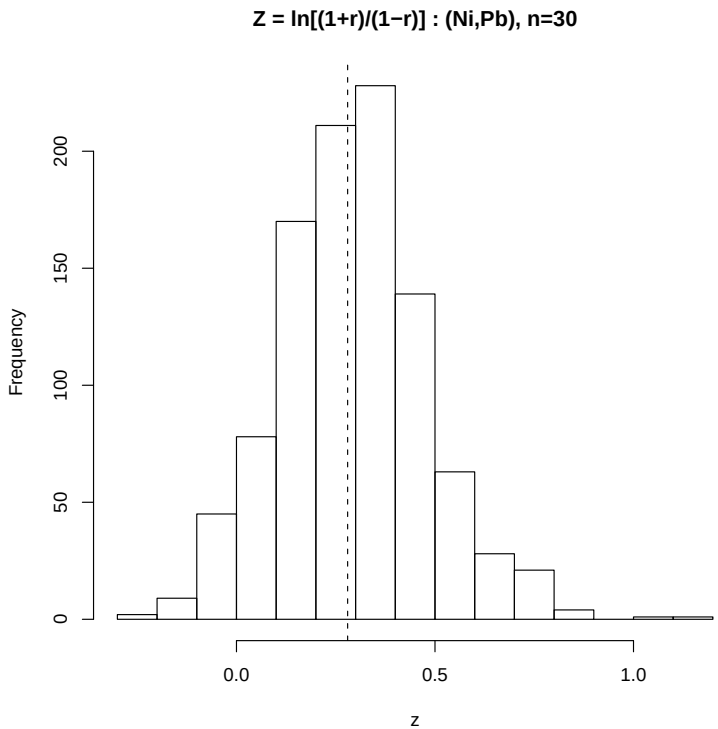
Coefficient de corrélation linéaire





Coefficient de corrélation linéaire





Coefficient de corrélation linéaire

$$Z = \frac{1}{2}\ln\left(\frac{1+R}{1-R}\right) \to \mathcal{N}\left(\frac{1}{2}\ln(\frac{1+\rho}{1-\rho}); \frac{1}{n-3}\right)$$

Utilisation

Notons $z = 0.5\ln((1+\rho)/(1-\rho))$. IC pour z est

$$z \in [z_1 = Z - u_{1-\alpha/2}/\sqrt{n-3}; z_2 = Z + u_{1-\alpha/2}/\sqrt{n-3}].$$

et donc

$$\rho \in [(e^{2z_1}-1)/(e^{2z_1}+1); (e^{2z_2}-1)/(e^{2z_2}+1)]$$

ρ(Ni,Pb) $\in [0.23, 0.31]$ au niveau 0.95

Merci