Évaluation de Performances

Mesure et Analyse

Florence Perronnin

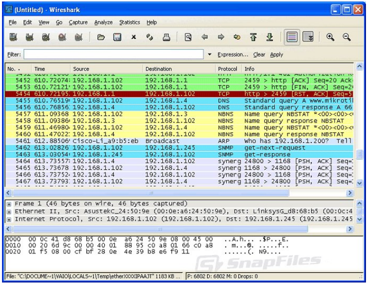
Université Joseph Fourier, LIG Laboratory

March 11, 2010

Outline

Mesure

- Que mesurer?
- Génération de charge
- Instrumentation: logicielle et/ou matérielle



Outils

Question

Quels outils de mesure logiciels connaissez-vous en réseaux?

Réseaux: ping, traceroute, netperf, perfping, netstat, pathchar, iperf, tcptrace, tstat...

Programmes, Système: gprof, iostat, vmstat...

Outils

Question

Quels outils de mesure logiciels connaissez-vous en réseaux?

Réseaux: ping, traceroute, netperf, perfping, netstat, pathchar, iperf, tcptrace, tstat...

Programmes, Système: gprof, iostat, vmstat...

Choix d'outils

1. Granularité des mesures (bits/trames/paquets/messages/fichiers...)
2. Précision de l'outil?
3. Technique de mesure de l'outil?
   - Passif/actif
   - échantillonnage
   - sonde
4. Sensibilité à d'autres facteurs?
5. Format de sortie de l'outil?
6. Automatisation possible? (API)
7. Coût d'utilisation (licence, formation, déploiement)
8. portabilité et support
9. pérennité?

Petite piqûre de rappel

- Une campagne d'expériences a un but (cf cours 1).
- Sans résultats chiffrés, elle est réputée non faite.
- Sans conclusion, elle est sans objet (cf DM...)
- Sans quantification de l'erreur, elle est inutilisable (cf DM...)
- Sans planification des conditions expérimentales, elle n'a aucun sens (cf DM!!)

Petite piqûre de rappel

- Une campagne d'expériences a un but (cf cours 1).
- Sans résultats chiffrés, elle est réputée non faite.
- Sans conclusion, elle est sans objet (cf DM...)
- Sans quantification de l'erreur, elle est inutilisable (cf DM...)
- Sans planification des conditions expérimentales, elle n'a aucun sens (cf DM!!)

Petite piqûre de rappel

- Une campagne d'expériences a un but (cf cours 1).
- Sans résultats chiffrés, elle est réputée non faite.
- Sans conclusion, elle est sans objet (cf DM...)
- Sans quantification de l'erreur, elle est inutilisable (cf DM...)
- Sans planification des conditions expérimentales, elle n'a aucun sens (cf DM!!)

Petite piqûre de rappel

- Une campagne d'expériences a un but (cf cours 1).
- Sans résultats chiffrés, elle est réputée non faite.
- Sans conclusion, elle est sans objet (cf DM...)
- Sans quantification de l'erreur, elle est inutilisable (cf DM...)
- Sans planification des conditions expérimentales, elle n'a aucun sens (cf DM!!)

Petite piqûre de rappel

- Une campagne d'expériences a un but (cf cours 1).
- Sans résultats chiffrés, elle est réputée non faite.
- Sans conclusion, elle est sans objet (cf DM...)
- Sans quantification de l'erreur, elle est inutilisable (cf DM...)
- Sans planification des conditions expérimentales, elle n'a aucun sens (cf DM!!)

Outline

Plan d'expériences

Remarque

Les expérimentations coûtent cher:

- mobilisation de matériel
- incidence sur les personnes (expériences sociales, médicales)
- durée des expériences

Principe

Méthodologie d'expérimentation permettant d'obtenir un maximum d'information en réalisant un minimum d'expériences.

Plan d'expériences

Remarque

Les expérimentations coûtent cher:

- mobilisation de matériel
- incidence sur les personnes (expériences sociales, médicales)
- durée des expériences

Principe

Méthodologie d'expérimentation permettant d'obtenir un maximum d'information en réalisant un minimum d'expériences.

Définitions

1. Variable dépendante / réponse
2. Facteurs
3. Niveaux
4. facteurs primaires/secondaires
5. Réplication
6. Interactions entre facteurs
7. Baseline

Plan d'expériences

Supposons k facteurs à N_i niveaux, $1 \leq i \leq k$

Plan simple

Config standard puis variation d'un paramètre à la fois.

$$\text{Nb Exp} = 1 + \sum_{i=1}^{k} N_i$$

- masque interactions

Plan factoriel complet

k facteurs à N_i niveaux, $1 \leq i \leq k$: toutes les combinaisons sont testées:

$$\text{Nb Exp} = \prod_{i=1}^{k} N_i$$

- coûteux

Plan d'expériences

Supposons k facteurs à N_i niveaux, $1 \leq i \leq k$

Plan simple

Config standard puis variation d'un paramètre à la fois.

$$\text{Nb Exp} = 1 + \sum_{i=1}^{k} N_i$$

- masque interactions

Plan factoriel complet

k facteurs à N_i niveaux, $1 \leq i \leq k$: toutes les combinaisons sont testées:

$$\text{Nb Exp} = \prod_{i=1}^{k} N_i$$

- coûteux

Plan d'expériences

Plan factoriel 2^k

k facteurs, 2 niveaux par facteur (min et max)

$$\text{Nb Exp} = 2^k$$

- facteurs monotones
- isolation des facteurs influents

Plan factoriel 2^{k-p}

k facteurs, 2 niveaux par facteur (min et max) mais moins de combinaisons testées

$$\text{Nb Exp} = 2^{k-p}$$

- perte d'information
- interactions potentiellement masquées

Plan d'expériences

Plan factoriel 2^k

k facteurs, 2 niveaux par facteur (min et max)

$$\text{Nb Exp} = 2^k$$

- facteurs monotones
- isolation des facteurs influents

Plan factoriel 2^{k-p}

k facteurs, 2 niveaux par facteur (min et max) mais moins de combinaisons testées

$$\text{Nb Exp} = 2^{k-p}$$

- perte d'information
- interactions potentiellement masquées

Comparaison de systèmes

La mise en oeuvre de plusieurs systèmes concurrents pour en déterminer le "meilleur" est périlleuse :

- Que signifie "meilleur"? (résultat optimal, plus rapide, plus économe...)
- Comparaison à armes égales : ne signifie pas forcément égalité des paramètres!
- Comparaison de choux et de carottes
- Interactions possibles

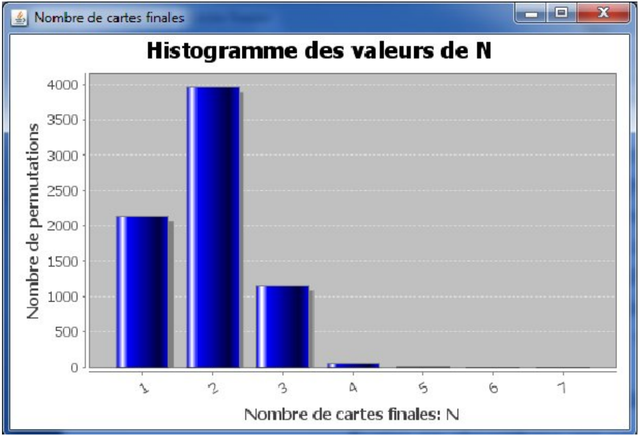
Example

Étude de différents schémas de décroissance de température dans l'algorithme de recuit simulé.

Outline

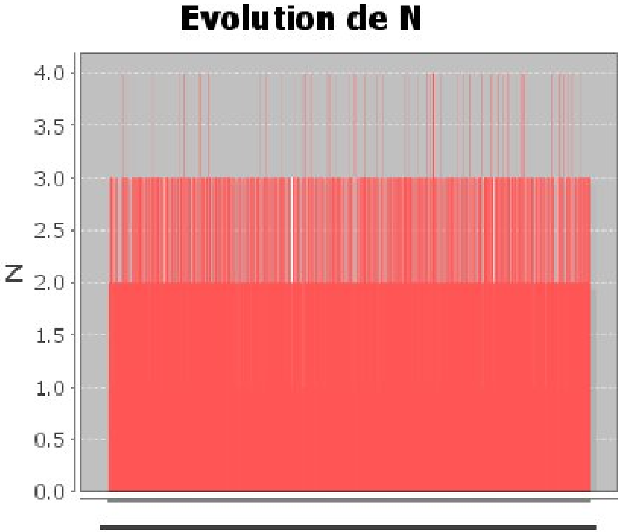
Histogrammes

Un histogramme représente sous forme de barres verticales de hauteur variable la fréquence, ou le nombre, d'occurrences d'échantillons dans chaque plage de valeurs.



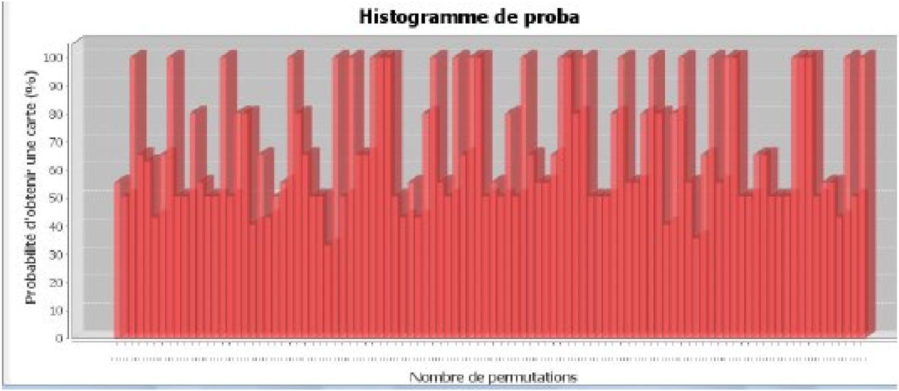
Histogrammes

Un histogramme n'est pas une représentation verticale de toutes les mesures d'un échantillon!

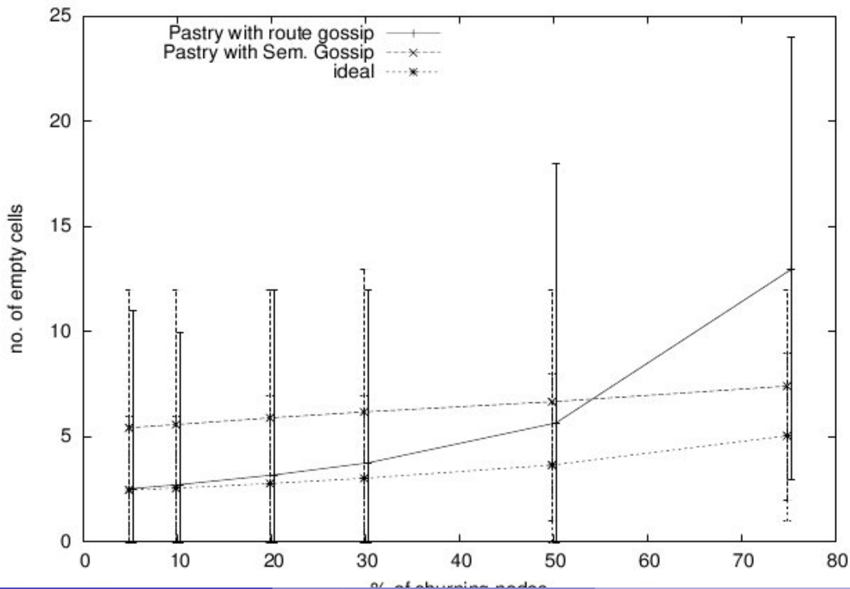


Histogrammes

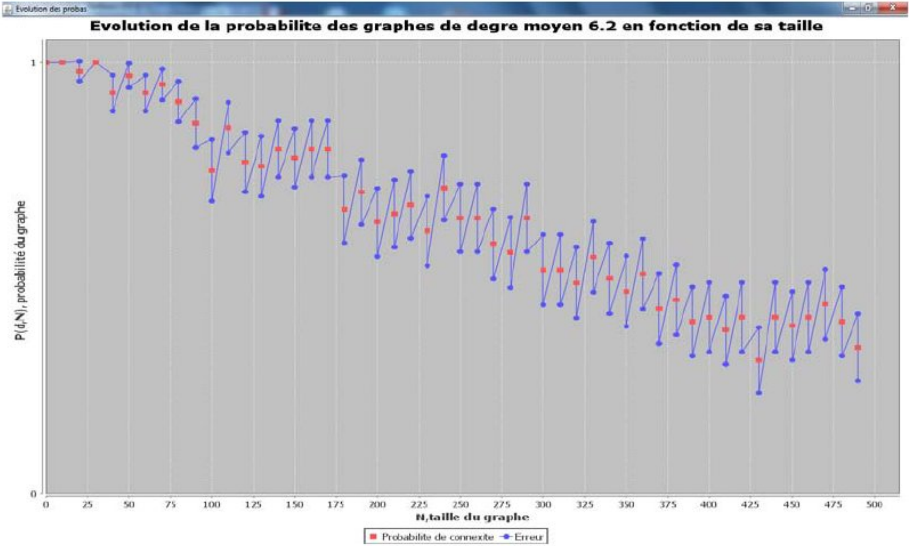
Voilà ce qui aurait dû être un nuage de points :



Intervalles de confiance



Intervalles de confiance



Outline

Visualisation

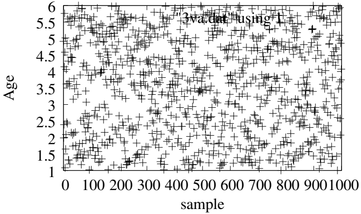
Nuage de points :

- détection de tendances
- identification de mesures aberrantes
- corrélations

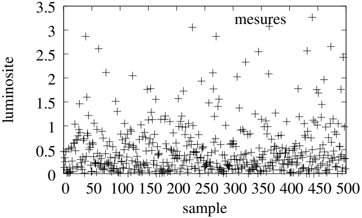
Distribution de probabilité :

- box-plots
- histogrammes
- qq-plots

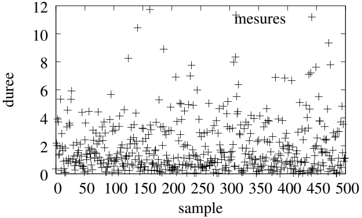
Nuage de points : échantillons séparés



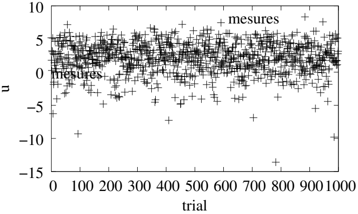
(a) Âge



(b) Luminosité

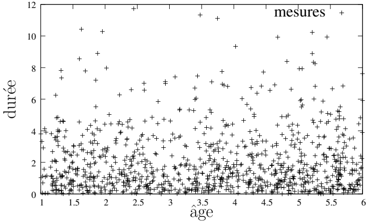


(c) Durée

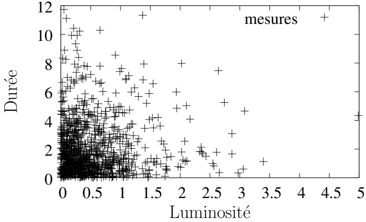


(d) Variable u

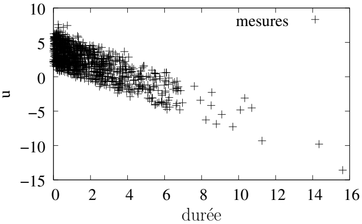
Etude de la corrélation



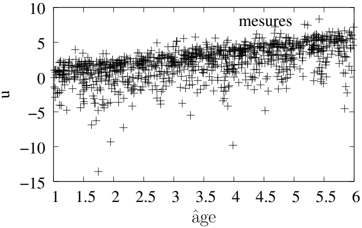
(a) Pas de corrélation



(b) Pas de corrélation



(c) Corrélation linéaire



(d) Corrélation linéaire

Coefficient de Corrélation linéaire

Example

Pour l'algorithme de recuit simulé, on ne peut pas considérer séparément le qualité du résultat et le temps de convergence!

Pour des variables aléatoires non indépendantes, on peut calculer un coefficient de corrélation linéaire:

$$r_{X,Y} = \frac{\sigma_{X,Y}}{\sigma_X \sigma_Y} \tag{1}$$

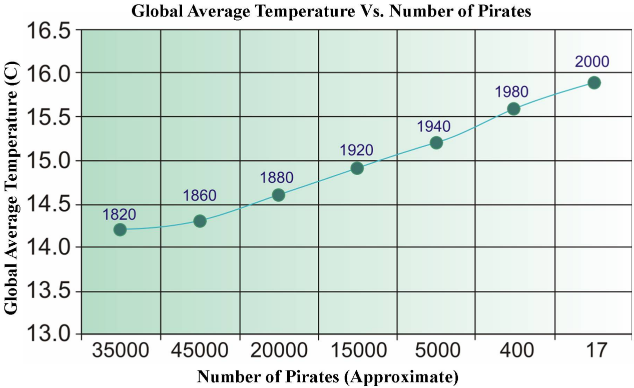
où:

- $\sigma_{X,Y} = \frac{1}{N} \sum_{i=1} N(X_i - \bar{X})(Y_i - \bar{Y})$ covariance de X et Y
- σ_X écart-type empirique de X
- σ_y écart-type empirique de Y

Ce coefficient $r_{X,Y}$ sans unité, compris dans $[-1, 1]$, est sensible aux valeurs aberrantes.

- si $r_{X,Y} \rightarrow 1$ (ou -1), les variables sont corrélées linéairement
- si $r_{X,Y} = 0$ les variables sont indépendantes

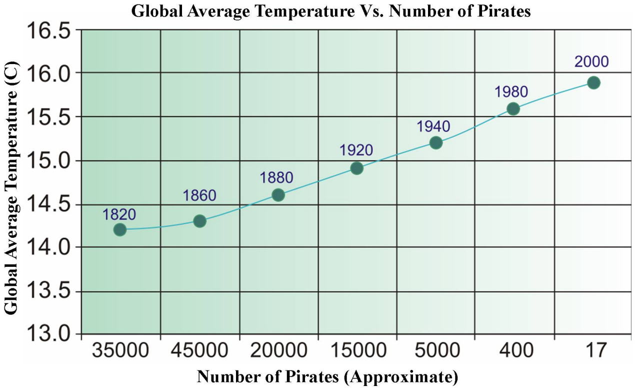
Corrélation et causalité



Conclusion

La disparition des pirates provoque donc le réchauffement climatique. (Ou est-ce le contraire?)

Corrélation et causalité



Conclusion

La disparition des pirates provoque donc le réchauffement climatique. (Ou est-ce le contraire?)

Corrélation et causalité (suite)

Example

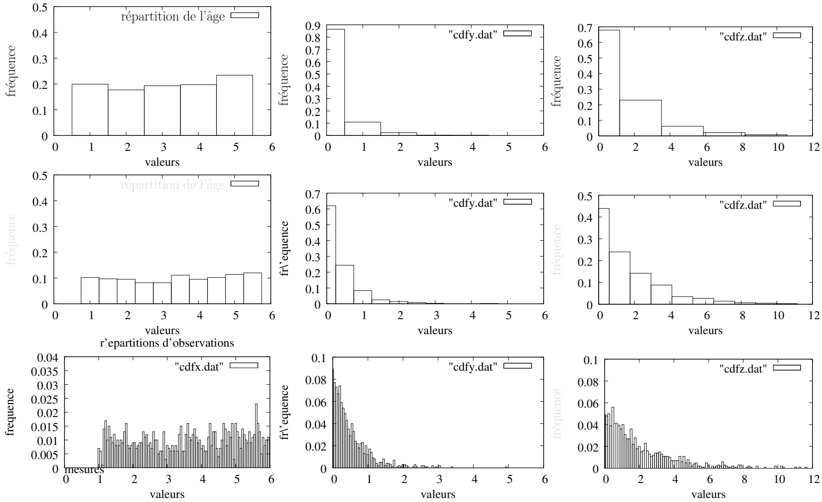
L'augmentation des ventes de crèmes glacées est corrélée linéairement à l'augmentation des morts par noyade.

Example

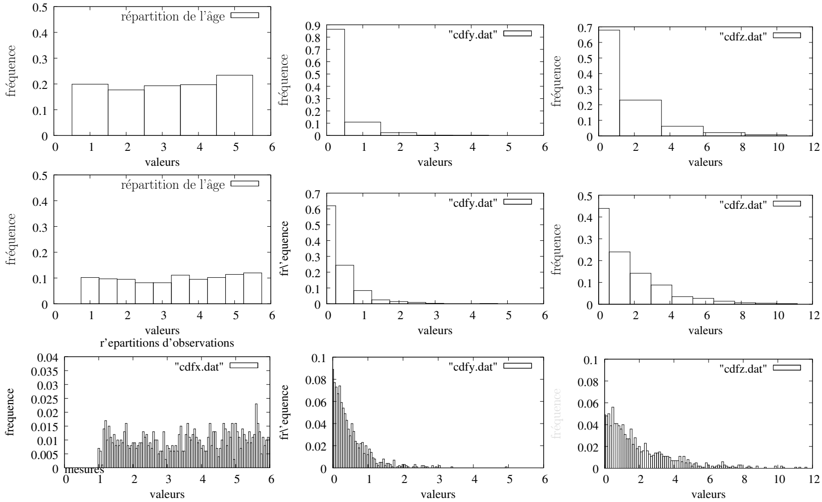
Dormir tout habillé est corrélé à la survenue de céphalées.

.

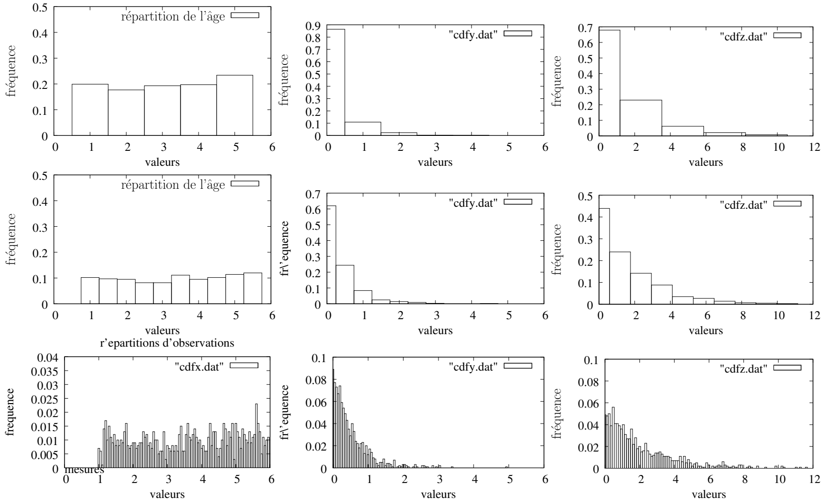
Représentation de la distribution: histogrammes



Représentation de la distribution: histogrammes

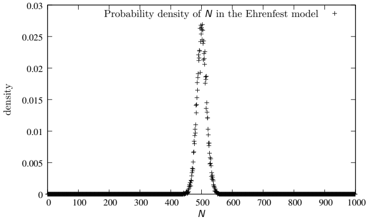


Représentation de la distribution: histogrammes

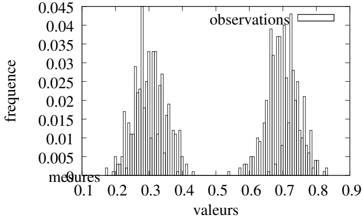


Représentation de la distribution: histogrammes

Autres exemples



(a) Ehrenfest



(b) Distribution bimodale

Figure: Autre visualisations d'histogrammes

- observation des modes
- symétrie (ou non) de la distribution
- écrasement de la courbe
- forme générale de la distribution (exponentielle, uniforme, gaussienne,...)

Indices de tendance centrale

Mode : valeur la plus probable

$$x_i = \mathrm{argmax}_{0 \leq i \leq n}(\mathbb{P}[x_i])$$

Moyenne empirique :

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

Médiane : m telle que $P(X \leq m) = \frac{1}{2}$

$$\text{soit } \sum_{i=1}^{n} \mathbb{1}_{[X_i \leq m]} = \frac{n}{2} \quad \text{(50\% des observations au-dessous)}$$

α-quantiles ou 100α-percentile : valeur x_α telle que $\mathbb{P}[X \leq x_\alpha] = \alpha$

$$\text{soit } \sum_{i=1}^{n} \mathbb{1}_{[X_i \leq x_\alpha]} = \alpha n \quad \text{(}100\alpha\text{\% des observations au-dessous)}$$

Moyenne: attention danger!

Signification hasardeuse

En moyenne une femme française a $1,98$ enfant(s).

En moyenne, une voiture islandaise a 3 roues motrices.

Variabilité

Le gain moyen par joueur à l'Euromillions est d'environ 39 euros en 2007.
Autres valeurs possibles[a] de gain moyen :

26,4 € /gagnant	15,15 € hors rang 1	1,10 € /grille	39,08 € /joueur/an	-44 € $\mathbb{E}[gain]$/joueur/an

$\mathbb{P}[\text{grille gagnante}] \approx 4\%$ et $\mathbb{P}[\text{jackpot}] \leq 10^{-6}$

[a] Sources: données FDJ 2007

Moyenne: attention danger!

Signification hasardeuse

En moyenne une femme française a $1,98$ enfant(s).
En moyenne, une voiture islandaise a 3 roues motrices.

Variabilité

Le gain moyen par joueur à l'Euromillions est d'environ 39 euros en 2007.
Autres valeurs possibles[a] de gain moyen :

26,4 € /gagnant	15,15 € hors rang 1	1,10 € /grille	39,08 € /joueur/an	-44 € $\mathbb{E}[gain]$/joueur/an

$\mathbb{P}[\text{grille gagnante}] \approx 4\%$ et $\mathbb{P}[\text{jackpot}] \leq 10^{-6}$

[a]Sources: données FDJ 2007

Moyenne: attention danger!

Signification hasardeuse

En moyenne une femme française a $1,98$ enfant(s).
En moyenne, une voiture islandaise a 3 roues motrices.

Variabilité

Le gain moyen par joueur à l'Euromillions est d'environ 39 euros en 2007.
Autres valeurs possibles[a] de gain moyen :

26,4 € /gagnant	15,15 € hors rang 1	1,10 € /grille	39,08 € /joueur/an	-44 € $\mathbb{E}[gain]$/joueur/an

$\mathbb{P}[\text{grille gagnante}] \approx 4\%$ et $\mathbb{P}[\text{jackpot}] \leq 10^{-6}$

[a] Sources: données FDJ 2007

Moyenne: attention danger!

Signification hasardeuse

En moyenne une femme française a $1,98$ enfant(s).
En moyenne, une voiture islandaise a 3 roues motrices.

Variabilité

Le gain moyen par joueur à l'Euromillions est d'environ 39 euros en 2007.
Autres valeurs possibles[a] de gain moyen :

26,4 € /gagnant	15,15 € hors rang 1	1,10 € /grille	39,08 € /joueur/an	-44 € $\mathbb{E}[gain]$/joueur/an

$\mathbb{P}[\text{grille gagnante}] \approx 4\%$ et $\mathbb{P}[\text{jackpot}] \leq 10^{-6}$

[a] Sources: données FDJ 2007

Moyenne: attention danger!

Signification hasardeuse

En moyenne une femme française a $1,98$ enfant(s).
En moyenne, une voiture islandaise a 3 roues motrices.

Variabilité

Le gain moyen par joueur à l'Euromillions est d'environ 39 euros en 2007.
Autres valeurs possibles[a] de gain moyen :

26,4 € /gagnant	15,15 € hors rang 1	1,10 € /grille	39,08 € /joueur/an	-44 € $\mathbb{E}$ [*gain*]/joueur/an

$\mathbb{P}[\text{grille gagnante}] \approx 4\%$ et $\mathbb{P}[\text{jackpot}] \leq 10^{-6}$

[a] Sources: données FDJ 2007

Moyenne: attention danger!

Mesures aberrantes

Classe A	Classe B
13	11
13	11
14	12
15	13
13	11
12	12
14	11
13	11
12	10
0	20
11,9	12,2

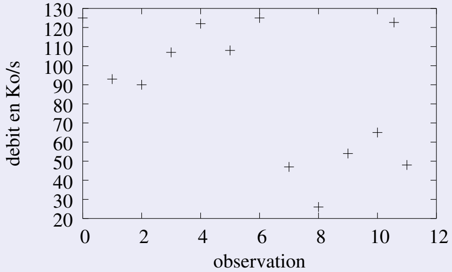
Classe A	Classe B
13	11
13	11
14	12
15	13
13	11
12	12
14	11
13	11
12	10
~~0~~	~~20~~
13,2	11,3

Moyenne: attention danger!

Multimodalité

FAI	download (Ko/s)
Alice	125,1
AOL	93,5
CI	90,3
Free	107,0
Noos	122,7
Orange	108,9
Tele2	125,7
AOL	47,9
Cegetel	26,1
Free	54,2
NC	65,0
Neuf	48,7

Débit moyen en download: 84,3 Ko/s Ne correspond à aucune réalité!



Données[a] du 03/02/2009 en Ko/s

[a] Source: www.grenouille.com

Moyenne: attention danger!

Ratios

Durée de mesure	Occupation CPU (%)
1	45
1	45
1	45
1	45
100	20

Moyenne 1: (45+45+45+45+20)/5=40%

Moyenne 2: (45*4+20*100)/104=21%

Moyenne: attention danger!

Ratios

Durée de mesure	Occupation CPU (%)
1	45
1	45
1	45
1	45
100	20

Moyenne 1: (45+45+45+45+20)/5=40%

Moyenne 2: (45*4+20*100)/104=21%

Moyenne : quantités particulières

Rappel

En général,

$$\mathbb{E}[X, Y] \neq \mathbb{E}[X]\,\mathbb{E}[Y]$$

L'égalité a lieu si les variables X et Y sont *indépendantes*.

Autres moyennes possibles :

Moyenne géométrique :

$$\sqrt[n]{\prod_{i=1}^{n} X_i}$$

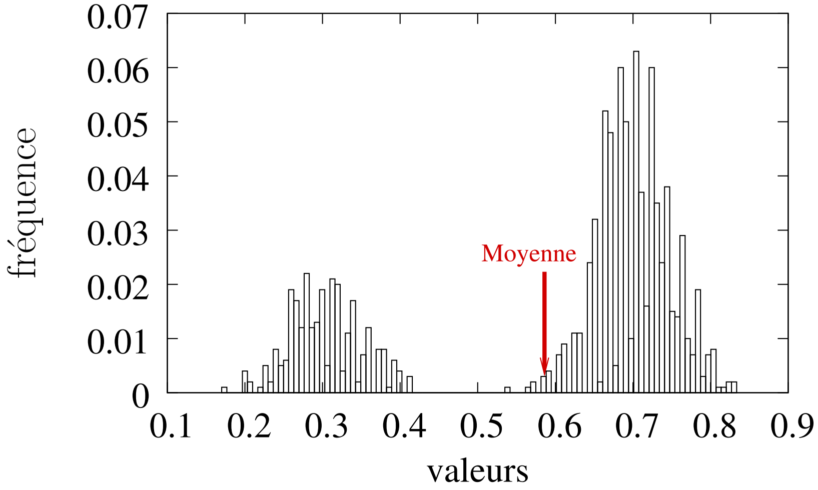
lorsque le cumul des X_i est multiplicatif (ex: pourcentage d'amélioration)

Moyenne harmonique :

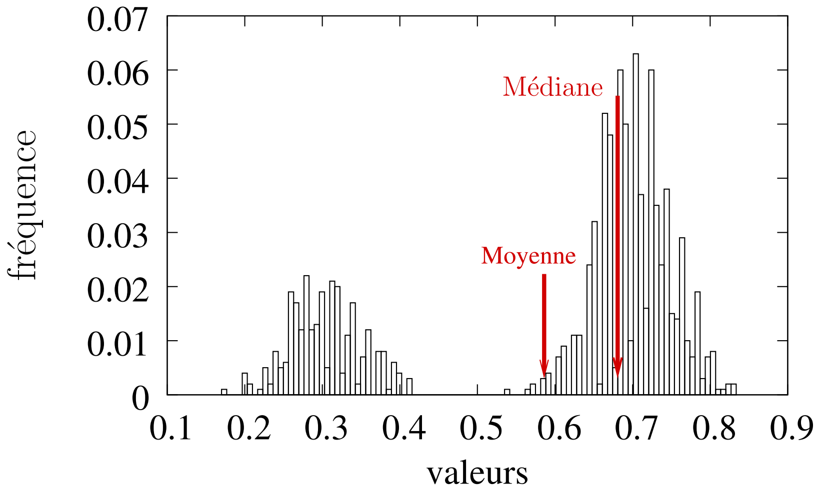
$$\frac{n}{\sum_{i=1}^{n} \frac{1}{X_i}}$$

lorsque la moyenne de $\frac{1}{x_i}$ a un sens (ex: débits pour 1 même fichier).

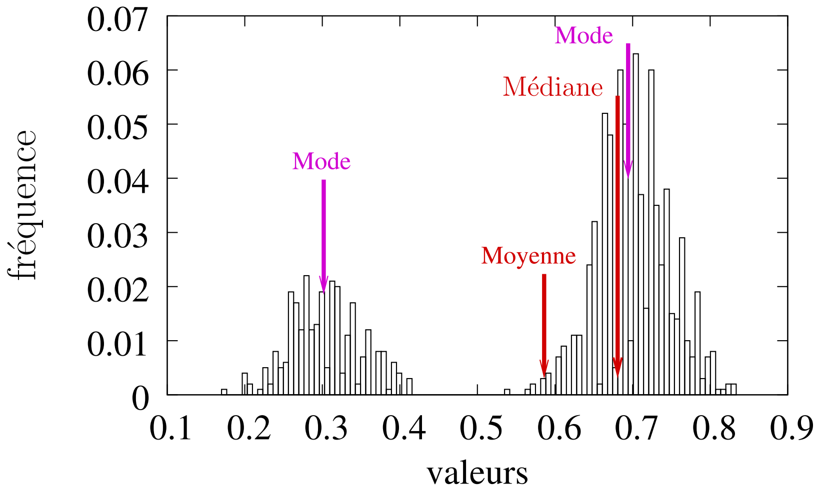
Représentation graphique



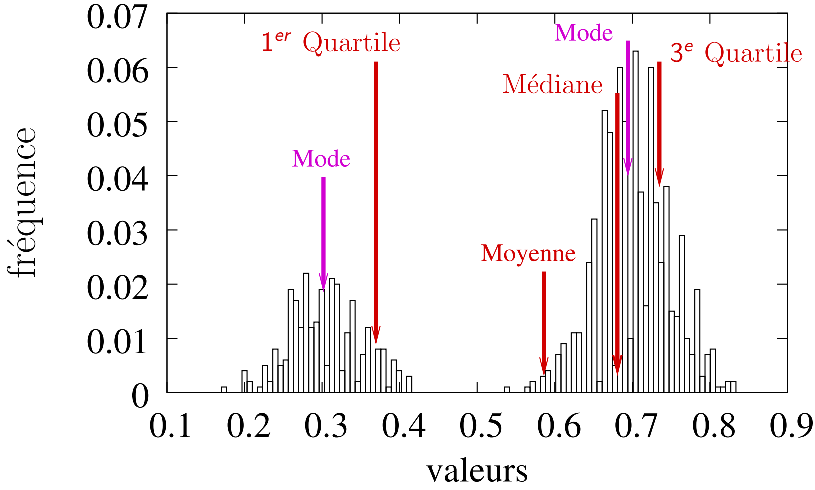
Représentation graphique



Représentation graphique



Représentation graphique



Indices de dispersion

Étendue :

$$[x_{min}, x_{max}]$$

écart absolu moyen :

$$\sum_{i=1}^{n} |x_i - \bar{X}|$$

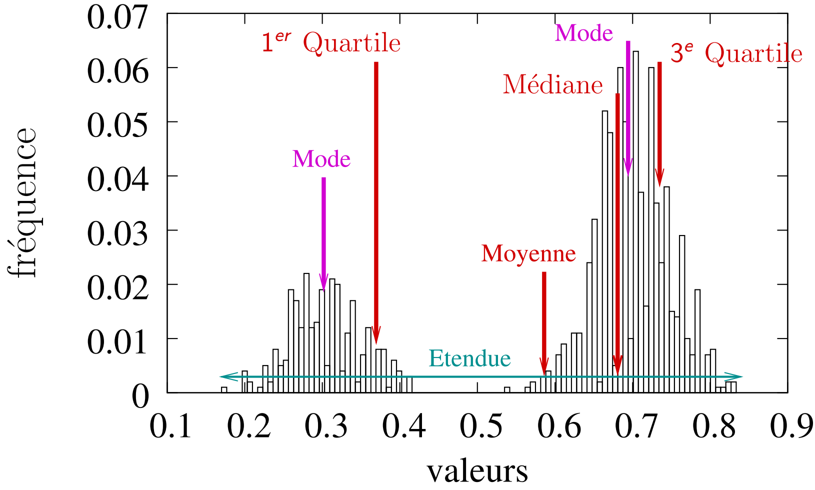
Écart inter-quartile :

$$q_{0.75} - q_{0.25}$$

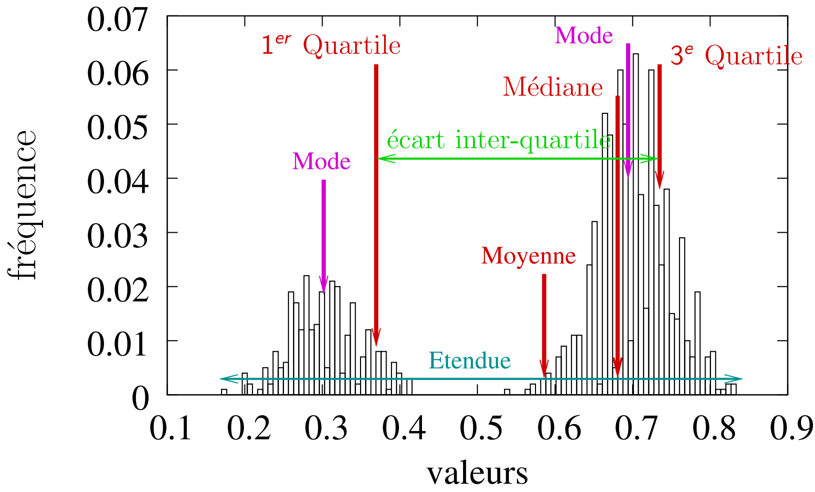
Variance empirique :

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

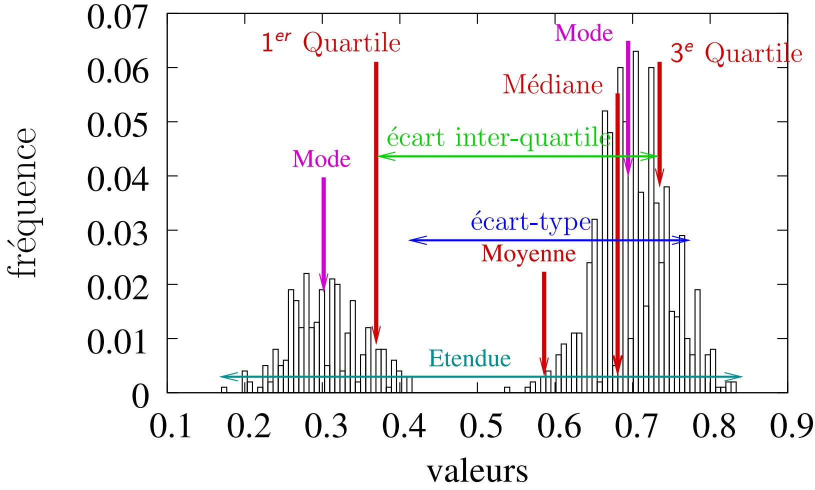
Représentation graphique



Représentation graphique



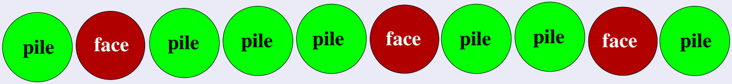
Représentation graphique



Intervalles de confiance (Rappels)

Problème

On veut tester si une pièce est truquée. On veut donc calculer $p = \mathbb{P}[\text{pile}]$. On effectue 10 lancers et l'on observe:



Soit 7 *pile* et 3 *face*. Peut-on dire que $p = 0.7$?

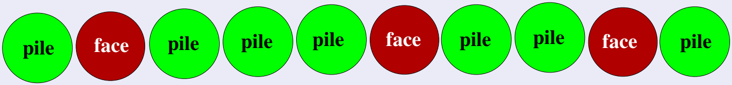
Réponse:

- $\mathbb{E}[X] = p$ se trouve dans l'intervalle $[0.41; 0.99]$ avec une probabilité de 95%.
- Si la pièce n'était pas truquée, la probabilité d'observer 7 pile en 10 lancers serait 0.12 (loi binômiale).

Intervalles de confiance (Rappels)

Problème

On veut tester si une pièce est truquée. On veut donc calculer $p = \mathbb{P}[\text{pile}]$. On effectue 10 lancers et l'on observe:



Soit 7 *pile* et 3 *face*. Peut-on dire que $p = 0.7$?

Réponse:

- $\mathbb{E}[X] = p$ se trouve dans l'intervalle $[0.41; 0.99]$ avec une probabilité de 95%.
- Si la pièce n'était pas truquée, la probabilité d'observer 7 pile en 10 lancers serait 0.12 (loi binômiale).

Problème général

Loi faible des grands nombres

Soit $X_1, X_2, \ldots$ une suite infinie de v.a. indépendantes et de même loi, et possédant un second moment. Alors, pour tout nombre $\epsilon > 0$ fixé on a

$$\lim_{n \to \infty} P\left(\left| \frac{X_1 + \cdots + X_n}{n} - \mathbf{E}(X_1) \right| \geq \epsilon \right) = 0.$$

Taille de l'échantillon

La fiabilité d'une mesure dépend fortement de la taille de l'échantillon.

Intervalles de confiance

Le but des intervalles de confiance est de quantifier la certitude d'une statistique.

Problème général

Loi faible des grands nombres

Soit $X_1, X_2, \ldots$ une suite infinie de v.a. indépendantes et de même loi, et possédant un second moment. Alors, pour tout nombre $\epsilon > 0$ fixé on a

$$\lim_{n\to\infty} P\left(\left|\frac{X_1 + \cdots + X_n}{n} - \mathbf{E}(X_1)\right| \geq \epsilon\right) = 0.$$

Taille de l'échantillon

La fiabilité d'une mesure dépend fortement de la taille de l'échantillon.

Intervalles de confiance

Le but des intervalles de confiance est de quantifier la certitude d'une statistique.

Problème général

Loi faible des grands nombres

Soit $X_1, X_2, \ldots$ une suite infinie de v.a. indépendantes et de même loi, et possédant un second moment. Alors, pour tout nombre $\epsilon > 0$ fixé on a

$$\lim_{n \to \infty} P\left(|\frac{X_1 + \cdots + X_n}{n} - \mathbf{E}(X_1)| \geq \epsilon\right) = 0.$$

Taille de l'échantillon

La fiabilité d'une mesure dépend fortement de la taille de l'échantillon.

Intervalles de confiance

Le but des intervalles de confiance est de quantifier la certitude d'une statistique.

Cas d'une variable gaussienne

Soit $X_1, \ldots, X_n$ une suite de v.a. indépendantes normales de moyenne μ et de variance σ^2. Alors :

- La variable $\bar{X} = \frac{1}{n}(X_1 + \ldots + X_n)$ suit la loi $\mathcal{N}(\mu, n\sigma^2)$.
- La variable $Z = \sqrt{n}\,\dfrac{\bar{X} - \mu}{\sigma}$ suit $\mathcal{N}(0, 1)$.

Méthode lorsque la variance σ^2 est connue

1. Choisir un niveau de confiance γ (typiquement 95% ou 99%)
2. Déterminer le $\frac{1+\gamma}{2}$ quantile c_γ correspondant dans la table $\mathcal{N}(0, 1)$
3. L'intervalle de confiance à γ est donné par:

$$\mu \in [\bar{X} - \frac{c_\gamma \sigma}{\sqrt{n}}; \bar{X} + \frac{c_\gamma \sigma}{\sqrt{n}}]$$

Preuve

Soit $F_{\mathcal{N}}$ la fonction de répartition de $\mathcal{N}(0,1)$: $F_{\mathcal{N}}(c) = \mathbb{P}[Z \leq c]$.
On cherche k_γ tel que:

$$\begin{aligned}
\gamma &= \mathbb{P}\left[|\bar{X} - \mu| \leq k_\gamma\right] \\
&= \mathbb{P}\left[|Z| \leq \frac{k_\gamma \sqrt{n}}{\sigma}\right] \qquad \text{car } Z = \frac{\bar{X} - \mu}{\sigma}\sqrt{n} \\
&= \mathbb{P}[Z \leq c_\gamma] - \mathbb{P}[Z \leq -c_\gamma] \qquad \text{en posant } c_\gamma = \frac{k_\gamma \sigma}{\sqrt{n}} \\
&= F_{\mathcal{N}}(c_\gamma) - F_{\mathcal{N}}(-c_\gamma) \\
&= 2F_{\mathcal{N}}(c_\gamma) - 1 \qquad \text{car par symétrie } F_{\mathcal{N}}(c_\gamma) = 1 - F_{\mathcal{N}}(-c_\gamma)
\end{aligned}$$

D'où

$$F_{\mathcal{N}}(c_\gamma) = \frac{1+\gamma}{2}$$

Ainsi, $k_\gamma = \frac{c_\gamma \sigma}{\sqrt{n}}$ où c_γ est le $\frac{(1+\gamma)}{2}$-quantile de $\mathcal{N}(0,1)$.

Preuve

Soit $F_{\mathcal{N}}$ la fonction de répartition de $\mathcal{N}(0,1)$: $F_{\mathcal{N}}(c) = \mathbb{P}[Z \leq c]$.
On cherche k_γ tel que:

$$\begin{aligned}
\gamma &= \mathbb{P}\left[|\bar{X} - \mu| \leq k_\gamma\right] \\
&= \mathbb{P}\left[|Z| \leq \frac{k_\gamma \sqrt{n}}{\sigma}\right] \qquad \text{car } Z = \frac{\bar{X} - \mu}{\sigma}\sqrt{n} \\
&= \mathbb{P}[Z \leq c_\gamma] - \mathbb{P}[Z \leq -c_\gamma] \qquad \text{en posant } c_\gamma = \frac{k_\gamma \sigma}{\sqrt{n}} \\
&= F_{\mathcal{N}}(c_\gamma) - F_{\mathcal{N}}(-c_\gamma) \\
&= 2F_{\mathcal{N}}(c_\gamma) - 1 \qquad \text{car par symétrie } F_{\mathcal{N}}(c_\gamma) = 1 - F_{\mathcal{N}}(-c_\gamma)
\end{aligned}$$

D'où

$$F_{\mathcal{N}}(c_\gamma) = \frac{1+\gamma}{2}$$

Ainsi, $k_\gamma = \frac{c_\gamma \sigma}{\sqrt{n}}$ où c_γ est le $\frac{(1+\gamma)}{2}$-quantile de $\mathcal{N}(0,1)$.

Preuve

Soit $F_{\mathcal{N}}$ la fonction de répartition de $\mathcal{N}(0,1)$: $F_{\mathcal{N}}(c) = \mathbb{P}[Z \leq c]$.
On cherche k_γ tel que:

$$\begin{aligned}
\gamma &= \mathbb{P}\left[|\bar{X} - \mu| \leq k_\gamma\right] \\
&= \mathbb{P}\left[|Z| \leq \frac{k_\gamma \sqrt{n}}{\sigma}\right] \qquad \text{car } Z = \frac{\bar{X} - \mu}{\sigma}\sqrt{n} \\
&= \mathbb{P}[Z \leq c_\gamma] - \mathbb{P}[Z \leq -c_\gamma] \qquad \text{en posant } c_\gamma = \frac{k_\gamma \sigma}{\sqrt{n}} \\
&= F_{\mathcal{N}}(c_\gamma) - F_{\mathcal{N}}(-c_\gamma) \\
&= 2F_{\mathcal{N}}(c_\gamma) - 1 \qquad \text{car par symétrie } F_{\mathcal{N}}(c_\gamma) = 1 - F_{\mathcal{N}}(-c_\gamma)
\end{aligned}$$

D'où

$$F_{\mathcal{N}}(c_\gamma) = \frac{1+\gamma}{2}$$

Ainsi, $k_\gamma = \frac{c_\gamma \sigma}{\sqrt{n}}$ où c_γ est le $\frac{(1+\gamma)}{2}$-quantile de $\mathcal{N}(0,1)$.

Preuve

Soit $F_{\mathcal{N}}$ la fonction de répartition de $\mathcal{N}(0,1)$: $F_{\mathcal{N}}(c) = \mathbb{P}[Z \leq c]$.
On cherche k_γ tel que:

$$\begin{aligned}
\gamma &= \mathbb{P}\left[|\bar{X} - \mu| \leq k_\gamma\right] \\
&= \mathbb{P}\left[|Z| \leq \frac{k_\gamma\sqrt{n}}{\sigma}\right] \qquad \text{car } Z = \frac{\bar{X}-\mu}{\sigma}\sqrt{n} \\
&= \mathbb{P}[Z \leq c_\gamma] - \mathbb{P}[Z \leq -c_\gamma] \qquad \text{en posant } c_\gamma = \frac{k_\gamma \sigma}{\sqrt{n}} \\
&= F_{\mathcal{N}}(c_\gamma) - F_{\mathcal{N}}(-c_\gamma) \\
&= 2F_{\mathcal{N}}(c_\gamma) - 1 \qquad \text{car par symétrie } F_{\mathcal{N}}(c_\gamma) = 1 - F_{\mathcal{N}}(-c_\gamma)
\end{aligned}$$

D'où

$$F_{\mathcal{N}}(c_\gamma) = \frac{1+\gamma}{2}$$

Ainsi, $k_\gamma = \frac{c_\gamma \sigma}{\sqrt{n}}$ où c_γ est le $\frac{(1+\gamma)}{2}$-quantile de $\mathcal{N}(0,1)$.

Preuve

Soit $F_{\mathcal{N}}$ la fonction de répartition de $\mathcal{N}(0,1)$: $F_{\mathcal{N}}(c) = \mathbb{P}[Z \leq c]$.
On cherche k_γ tel que:

$$\begin{aligned}
\gamma &= \mathbb{P}\left[|\bar{X} - \mu| \leq k_\gamma\right] \\
&= \mathbb{P}\left[|Z| \leq \frac{k_\gamma \sqrt{n}}{\sigma}\right] \qquad \text{car } Z = \frac{\bar{X} - \mu}{\sigma}\sqrt{n} \\
&= \mathbb{P}[Z \leq c_\gamma] - \mathbb{P}[Z \leq -c_\gamma] \qquad \text{en posant } c_\gamma = \frac{k_\gamma \sigma}{\sqrt{n}} \\
&= F_{\mathcal{N}}(c_\gamma) - F_{\mathcal{N}}(-c_\gamma) \\
&= 2F_{\mathcal{N}}(c_\gamma) - 1 \qquad \text{car par symétrie } F_{\mathcal{N}}(c_\gamma) = 1 - F_{\mathcal{N}}(-c_\gamma)
\end{aligned}$$

D'où

$$F_{\mathcal{N}}(c_\gamma) = \frac{1+\gamma}{2}$$

Ainsi, $k_\gamma = \frac{c_\gamma \sigma}{\sqrt{n}}$ où c_γ est le $\frac{(1+\gamma)}{2}$-quantile de $\mathcal{N}(0,1)$.

Preuve

Soit $F_{\mathcal{N}}$ la fonction de répartition de $\mathcal{N}(0,1)$: $F_{\mathcal{N}}(c) = \mathbb{P}[Z \leq c]$.
On cherche k_γ tel que:

$$
\begin{aligned}
\gamma &= \mathbb{P}\left[|\bar{X} - \mu| \leq k_\gamma\right] \\
&= \mathbb{P}\left[|Z| \leq \frac{k_\gamma \sqrt{n}}{\sigma}\right] \qquad \text{car } Z = \frac{\bar{X} - \mu}{\sigma}\sqrt{n} \\
&= \mathbb{P}[Z \leq c_\gamma] - \mathbb{P}[Z \leq -c_\gamma] \qquad \text{en posant } c_\gamma = \frac{k_\gamma \sigma}{\sqrt{n}} \\
&= F_{\mathcal{N}}(c_\gamma) - F_{\mathcal{N}}(-c_\gamma) \\
&= 2F_{\mathcal{N}}(c_\gamma) - 1 \qquad \text{car par symétrie } F_{\mathcal{N}}(c_\gamma) = 1 - F_{\mathcal{N}}(-c_\gamma)
\end{aligned}
$$

D'où

$$F_{\mathcal{N}}(c_\gamma) = \frac{1+\gamma}{2}$$

Ainsi, $k_\gamma = \frac{c_\gamma \sigma}{\sqrt{n}}$ où c_γ est le $\frac{(1+\gamma)}{2}$-quantile de $\mathcal{N}(0,1)$.

Cas d'une variable gaussienne: variance inconnue

Lorsqu'on ne connait pas a priori la variance σ^2 on peut utiliser la variance empirique $\hat{\sigma}^2$ sous certaines conditions:

- la variable $\frac{\bar{X}-\mu}{\hat{\sigma}}\sqrt{n-1}$ suit une loi de Student à $n-1$ degrés de libertés
- intervalle de confiance plus large pour des échantillons petits (< 100).

Méthode lorsque la variance σ^2 est inconnue

On dispose de la moyenne empirique $\bar{X}$ et de la variance empirique $\hat{\sigma}^2$.

1. Choisir un niveau de confiance γ (typiquement 95% ou 99%)
2. Déterminer le $\frac{1+\gamma}{2}$ quantile c_γ correspondant dans la table $T(n-1)$.
3. L'intervalle de confiance à γ est donné par:

$$\mu \in [\bar{X} - \frac{c_\gamma \hat{\sigma}}{\sqrt{n}}; \bar{X} + \frac{c_\gamma \hat{\sigma}}{\sqrt{n}}]$$

Cas général

Problème

Comment calculer l'intervalle de confiance si la variable X n'est pas gaussienne mais inconnue?

Théorème de la Limite Centrale (TCL)

Soit $\{X_n\}$ une suite de v.a. i.i.d telle que $\mathbb{E}[X_n] = \mu$ et $Var[X_n] = \sigma^2$. Alors

$$\sqrt{n}\,\frac{\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) - \mu}{\sigma} \xrightarrow[n\to\infty]{} \mathcal{N}(0,1) \tag{2}$$

Moralité

Pour un échantillon n suffisamment grand, la somme des X_n se comporte comme une gaussienne et l'on peut appliquer la méthode précédente.

Cas général

Problème

Comment calculer l'intervalle de confiance si la variable X n'est pas gaussienne mais inconnue?

Théorème de la Limite Centrale (TCL)

Soit $\{X_n\}$ une suite de v.a. i.i.d telle que $\mathbb{E}[X_n] = \mu$ et $Var[X_n] = \sigma^2$. Alors

$$\sqrt{n}\,\frac{\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) - \mu}{\sigma} \xrightarrow[n\to\infty]{} \mathcal{N}(0,1) \tag{2}$$

Moralité

Pour un échantillon n suffisamment grand, la somme des X_n se comporte comme une gaussienne et l'on peut appliquer la méthode précédente.

Cas général

Problème

Comment calculer l'intervalle de confiance si la variable X n'est pas gaussienne mais inconnue?

Théorème de la Limite Centrale (TCL)

Soit $\{X_n\}$ une suite de v.a. i.i.d telle que $\mathbb{E}[X_n] = \mu$ et $Var[X_n] = \sigma^2$. Alors

$$\sqrt{n}\,\frac{\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) - \mu}{\sigma} \xrightarrow[n\to\infty]{} \mathcal{N}(0,1) \qquad (2)$$

Moralité

Pour un échantillon n suffisamment grand, la somme des X_n se comporte comme une gaussienne et l'on peut appliquer la méthode précédente.