StatL3S6
Statistique appliquée à la psychologie L3

Christian Lavergne

Université Paul Valéry - Montpellier 3
http://moodle-miap.univ-montp3.fr
http://www.univ-montp3.fr/miap/ens

Année universitaire 2013-2014

StatL3S6 Chapitre 1

Lois limites de la Statistique et Estimation :

1. Loi des grands nombres et estimation ponctuelle
2. Théorème central limite et estimation par intervalle

Loi des grands nombres et estimation ponctuelle

Soient $X_1, X_2, \ldots, X_n$ n variables aléatoires indépendantes associées aux répétitions d'une même expérience aléatoire X telle que $E(X) = \mu$) alors :

La moyenne des observations est aussi proche que possible
de la vraie valeur μ
à condition que n soit grand

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \quad \overset{n \text{ grand}}{\longrightarrow} \quad \mu$$

Exemple 1 : Un jeu de la Française des jeux.

Exemple 2 : 2 personnes A et B jouent au dé (à 6 faces) et on propose les gains suivants :

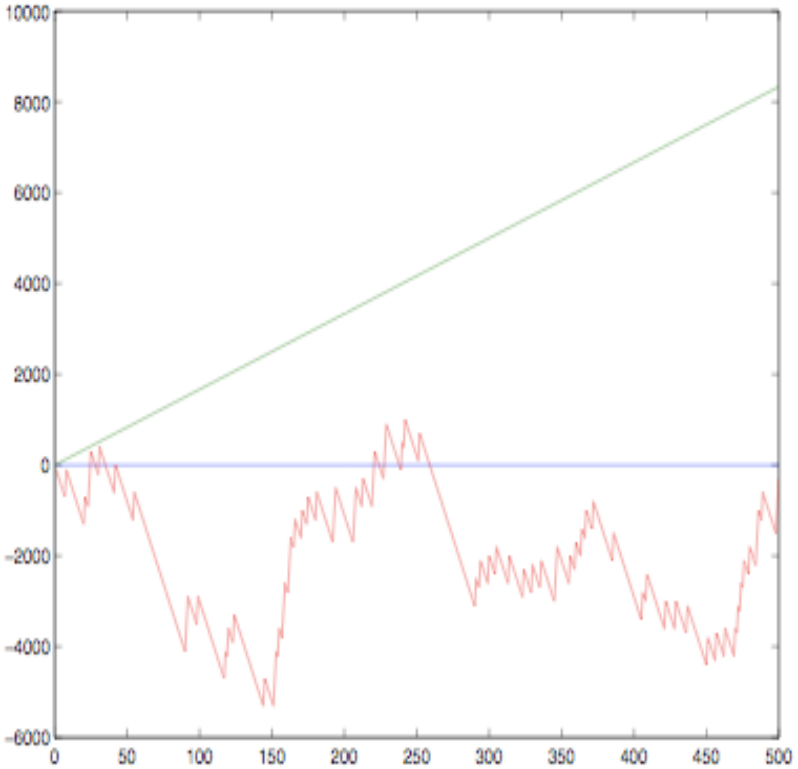
Si "le résultat est la face 6", B paie 600 euros à A
sinon A paie 100 euros à B.

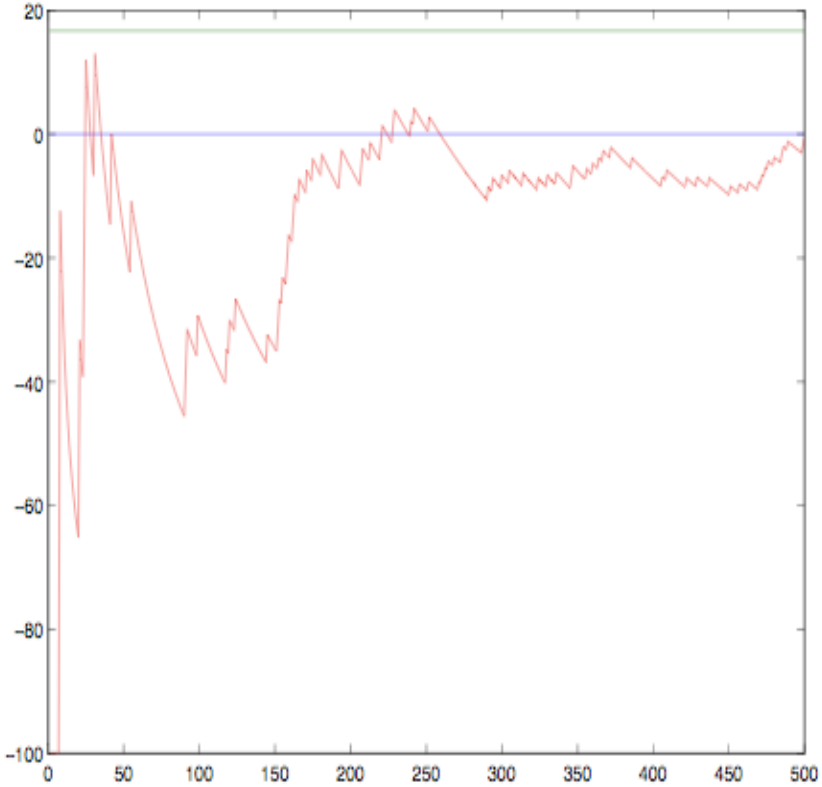
À chaque coup, quelle est l'espérance de gain de A ?
P(A gagne 600) $= \frac{1}{6}$ et P(A perd 100) $= \frac{5}{6}$
donc A a pour espérance de gain :

$$600 \times \frac{1}{6} - 100 \times \frac{5}{6} = \frac{100}{6} = 16.67 \ \textit{euros}$$

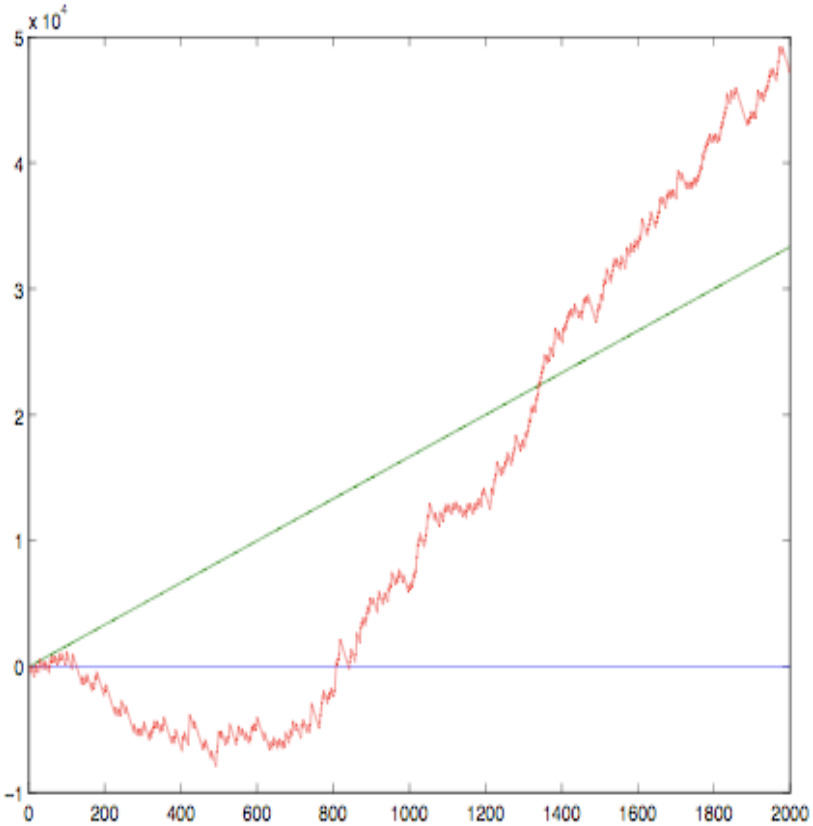
À chaque coup, A gagne 600 ou perd 100, et son espérance de gain est de 16.67.
Donc si A joue un très grand nombre de fois, A est sûr de gagner.

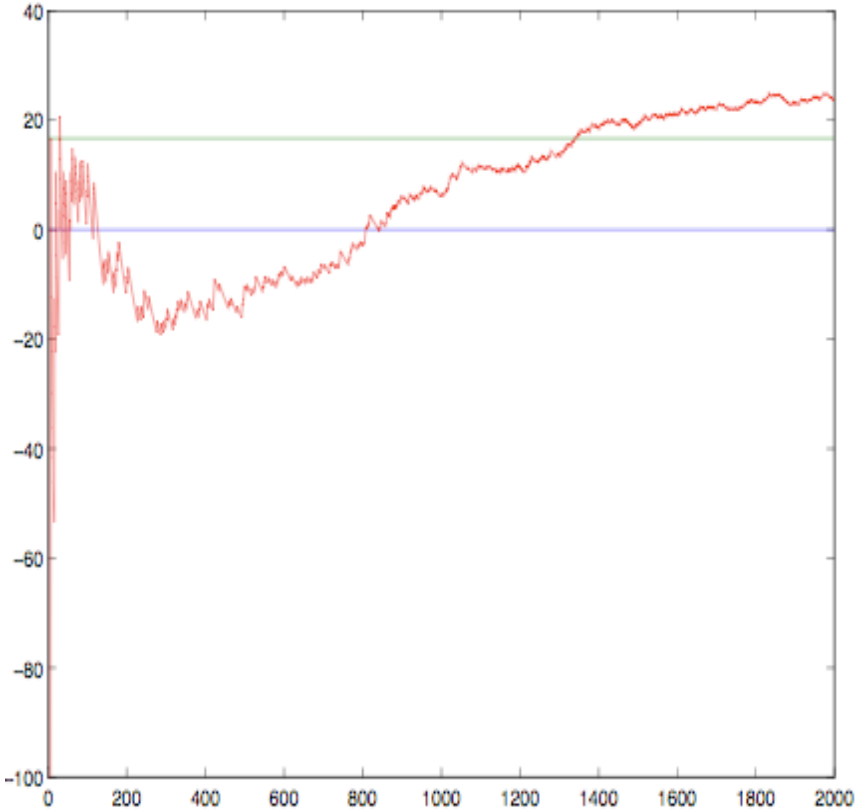
La loi des grands nombres : jeu du dé





La loi des grands nombres : jeu du dé





Propriété de la moyenne $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

de n répétitions indépendantes $X_1, X_2, \ldots, X_n$ d'une même expérience aléatoire X telle que $E(X) = \mu$ et $V(X) = \sigma^2$:

- son espérance mathématique

$$E(\overline{X}) = \mu$$

- sa variance

$$V(\overline{X}) = \frac{\sigma^2}{n}$$

- son écart-type

$$\sigma(\overline{X}) = \frac{\sigma}{\sqrt{n}}$$

La dispersion de la moyenne se réduit quand n grandit : c'est la loi des grands nombres

Estimation ponctuelle

La moyenne $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ [1] est donc un bon prétendant pour approcher le paramètre inconnu μ.

• La variable aléatoire $\overline{X}$ sera appelée l'Estimateur du paramètre μ.
notation : $\hat{\mu} = \overline{X}$
• Au vu d'observations, la valeur prise par $\overline{X}$ et notée $\overline{x}$ sera appelée une estimation du paramètre μ (notée aussi $\hat{\mu}$).
• En pratique il y a UN (voir deux ou trois) Estimateur naturel du paramètre inconnu ; mais il y a toujours une infinité d'estimations possible de ce paramètre.

1. $X_1, X_2, \ldots, X_n$ sont n répétitions indépendantes d'une même expérience aléatoire X telle que $E(X) = \mu$ et $V(X) = \sigma^2$

Forme générale et propriétés

Soient $X_1, X_2, \ldots, X_n$ n répétitions d'une expérience aléatoire et T_n une combinaison (ou fonction) de ces répétitions.
T_n sera un bon prétendant pour approcher un paramètre inconnu θ ; donc un Estimateur raisonnable de θ ; (T_n sera un $\hat{\theta}$) si :

- L'espérance mathématique de l'estimateur est aussi proche que possible du paramètre inconnu θ, idéalement on souhaite que

$$E(T_n) = \theta$$

  et on dira que T_n est un estimateur sans biais de θ ;
  mais on peut se contenter de $E(T_n) \overset{n \text{ grand}}{\to} \theta$
- La variance de l'estimateur diminue avec le nombre de répétitions :

$$V(T_n) \overset{n \text{ grand}}{\to} 0$$

Exemple : dans la cas de n répétitions indépendantes $X_1, X_2, \ldots, X_n$ d'une même expérience aléatoire X telle que $E(X) = \mu$ et $V(X) = \sigma^2$.

- Si μ est inconnu ; la moyenne
$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \text{ est un estimateur sans biais de } \mu \text{ ; } \hat{\mu}.$$
- Si μ est connu et σ^2 inconnu ; la moyenne des dispersions
$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 \text{ est un estimateur sans biais de } \sigma^2 \text{ ; noté } \hat{\sigma}^2_{\mu}.$$
- Si μ est inconnu et σ^2 inconnu ; la variance empirique
$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 \text{ est un estimateur biaisé de } \sigma^2 \text{ ; noté } \hat{\sigma}^2.$$
- Si μ est inconnu et σ^2 inconnu ;
$$\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 \text{ est un estimateur sans biais de } \sigma^2 \text{ ; noté } \hat{\sigma}^2_{SB} \ .$$

Exercice 1 : On suppose avoir observé n variables aléatoires $Y_1, Y_2, \ldots Y_n$ et on propose 3 estimateurs d'un paramètre θ : T_1, T_2, T_3 ayant les propriétés suivants (λ est un autre paramètre inconnu) :

$$E(T_1) = \theta + \lambda \quad \text{et} \quad V(T_1) = (\theta * \lambda)/n$$

$$E(T_2) = \theta + \lambda/n \quad \text{et} \quad V(T_2) = (\theta * \lambda)/n$$

$$E(T_3) = \theta \quad \text{et} \quad V(T_3) = \lambda$$

Donner les propriétés :

- de biais ("estimateur sans biais" ; "son biais diminue quand n, le nombre d'observations grandit")
- et de variance ("sa variance diminue quand n, le nombre d'observations grandit")

Lequel des 3 est il raisonnable de garder ?

La loi de bernoulli et le sondage

Soit X une expérience aléatoire à 2 états (codés 1/0) : $X \sim \text{Ber}(p)$

$$P(X = 1) = p \ ; \ \ P(X = 0) = 1 - p$$

- son espérance mathématique $E(X) = p$
- sa variance $V(X) = p(1-p)$ et son écart-type $\sqrt{p(1-p)}$

La moyenne de loi de Bernoulli (de n répétitions indépendantes) :

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

- son espérance mathématique est : $E(\overline{X}) = p$.
- sa variance $V(\overline{X}) = \dfrac{p(1-p)}{n}$ et son écart-type $\sigma(\overline{X}) = \sqrt{\dfrac{p(1-p)}{n}}$

La loi des grands nombres pour la loi de Bernoulli

Soient $X_1, X_2, \ldots, X_n$ n répétitions indépendantes d'une même expérience aléatoire de Bernoulli
($P(X = 1) = p$)) alors :

La proportion de 1 est aussi proche que possible de p
à condition que n soit grand

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \quad \overset{n \text{ grand}}{\longrightarrow} \quad p$$

Estimation ponctuelle

Soient $X_1, X_2, \ldots, X_n$ n répétitions indépendantes d'une même expérience aléatoire de Bernoulli
($P(X = 1) = p$) , p paramètre inconnu) alors :

• La variable aléatoire $\overline{X}$ (ici la proportion ou la fréquence de 1) sera donc un Estimateur sans biais du paramètre p ; $\hat{p} = \overline{X}$.

• Après avoir effectué le sondage, la valeur prise par $\overline{X}$ et notée $\overline{x}$ sera donc une estimation du paramètre p.

• En pratique on a toujours le même Estimateur du paramètre inconnu p ; mais chaque sondage pratiqué amène une estimation différente de ce paramètre.

Théorème central limite et estimation par intervalle

Variable centrée, réduite : *définition*

- variable centrée : $X - E(X)$
- variable centrée, réduite : $\frac{X-E(X)}{\sigma(X)}$

Une variable centrée réduite a pour espérance 0 et écart-type 1

Pour la moyenne de n répétitions indépendantes d'une même expérience aléatoire d'espérance μ et de variance σ^2 alors

$$\frac{\overline{X}-\mu}{\sqrt{\frac{\sigma^2}{n}}} \text{ est une variable centrée réduite}$$

Pour la moyenne de répétitions de même loi de Bernoulli indépendantes :

$$\frac{\overline{X}-p}{\sqrt{\frac{p(1-p)}{n}}} \text{ est une variable centrée réduite}$$

Théorème central limite

Soient $X_1, X_2, \ldots, X_n$ n répétitions indépendantes d'une même expérience aléatoire alors :

La moyenne se comporte comme une loi normale
à condition que n soit grand
donc centrée et réduite se comporte comme une $\mathcal{N}$ (0,1)

$$\frac{\overline{X} - \mu}{\sqrt{\frac{\sigma^2}{n}}} \quad \overset{n \text{ grand}}{\rightarrow} \quad \mathcal{N}(0,1)$$

Le sondage

Soient $X_1, X_2, \ldots, X_n$ n répétitions indépendantes d'une même expérience aléatoire de Bernoulli
$(P(X = 1) = p))$ alors :

La proportion de 1 se comporte comme une loi normale
à condition que n soit grand
donc centrée et réduite se comporte comme une $\mathcal{N}$ (0,1)

$$\frac{\overline{X} - p}{\sqrt{\frac{p(1-p)}{n}}} \overset{n \text{ grand}}{\rightarrow} \mathcal{N}(0,1)$$

Notion élémentaire d'intervalle de dispersion

On cherche à construire un intervalle de grande probabilité de la réalisation d'une expérience aléatoire dont la loi de probabilité est connue.
Pour cela on se fixe une faible probabilité α, (en pratique α vaut 1%, 5% parfois 10%, 0.1% et on construit un

Intervalle de dispersion de probabilité 1-α

(l'expérience aléatoire a une probabilité 1-α de se réaliser dans l'intervalle.

Exemple 1 : si on jette 500 fois une même pièce, on cherche l'intervalle de dispersion à 95% du nombre de faces.

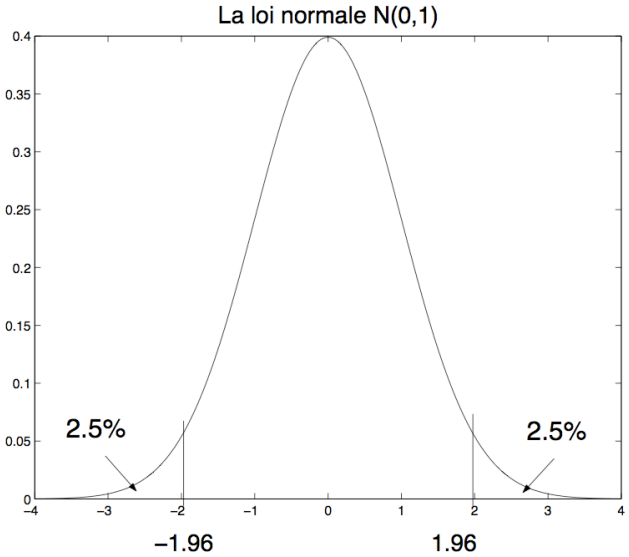
$$\mathrm{P}(\mathrm{Nfaces} \in [B_1, B_2]) = 95\%$$

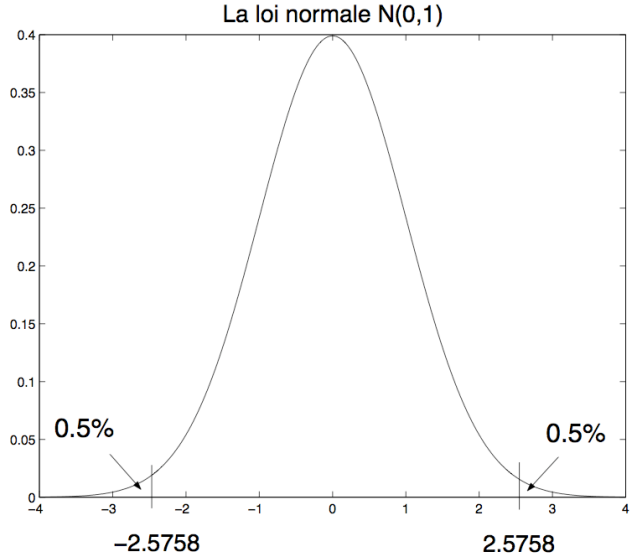
Est ce [220 , 280] ? ou alors [240 , 260], ou alors [180 , 320] ?

Exemple 2 : Courbes dans les carnets de santé !

Intervalle de dispersion de la loi normale

Soit Z une variable aléatoire de loi normale N(0,1) alors :





$$P(-1.96 < Z < 1.96) = .95 \qquad P(-2.576 < Z < 2.576) = .99$$

Donc $[-1.96\ ;\ 1.96]$ (resp. $[-2.576\ ;\ 2.576]$) est un intervalle de dispersion de Z à 95% (resp. 99%)

De façon générale on notera $ID_{1-\alpha}(Z) = [-\ell_{\frac{\alpha}{2}}\ ;\ \ell_{\frac{\alpha}{2}}]$ l'intervalle de dispersion de probabilité $1-\alpha$ d'une v.a. Z de loi normale $\mathcal{N}(0,1)$

$$P(-\ell_{\frac{\alpha}{2}} < Z < \ell_{\frac{\alpha}{2}}) = 1-\alpha$$

Intervalle de dispersion d'une somme de Bernoulli, notée S_n

Soient $X_1, X_2, \ldots, X_n$ n réalisations indépendantes d'une même expérience aléatoire de Bernoulli ($P(X = 1) = p_0$), p_0 connu) alors grâce au TCL :

$$\frac{\sum_{i=1}^{n} X_i - np_0}{\sqrt{np_0(1-p_0)}} = \frac{S_n - np_0}{\sqrt{np_0(1-p_0)}} \overset{n \text{ grand}}{\rightarrow} \mathcal{N}(0,1)$$

Et comme pour Z, v.a. N(0,1) on sait que $P(-\ell_{\frac{\alpha}{2}} < Z < \ell_{\frac{\alpha}{2}}) = 1 - \alpha$; on a

$$P(-\ell_{\frac{\alpha}{2}} < \frac{S_n - np_0}{\sqrt{np_0(1-p_0)}} < \ell_{\frac{\alpha}{2}}) \approx 1 - \alpha$$

Et, $P\left(np_0 - \ell_{\frac{\alpha}{2}}\sqrt{np_0(1-p_0)} < S_n < np_0 + \ell_{\frac{\alpha}{2}}\sqrt{np_0(1-p_0)}\right) \approx 1 - \alpha$

Un **I.D.** de $\sum_{i=1}^{n} X_i$ de probabilité approximative $(1-\alpha)$ est donc

$$ID_{1-\alpha}(\sum_{i=1}^{n} X_i) = \left[np_0 - \ell_{\frac{\alpha}{2}} \sqrt{np_0(1-p_0)} \ , \ np_0 + \ell_{\frac{\alpha}{2}} \sqrt{np_0(1-p_0)} \right]$$

Exemple : Sur 500 naissances, le nombre de garçons sera compris à 95% entre :

$$\left[500 \times \frac{1}{2} - 1.96 \sqrt{500 \times \frac{1}{2} \times (1 - \frac{1}{2})} \ , \ 500 \times \frac{1}{2} + 1.96 \sqrt{500 \times \frac{1}{2} \times \frac{1}{2}} \right]$$

$$= [250 - 21.91 \ , \ 250 + 21.91] \approx [228 \ , \ 272]$$

L'intervalle à 99% serait : [221 , 279].

Intervalle de dispersion d'une moyenne ; d'une somme de v.a.

Soient $X_1, X_2, \ldots, X_n$ n réalisations indépendantes d'une même expérience aléatoire d'espérance μ_0 connue, de variance σ_0^2 connue alors (TCL) :

$$\frac{\overline{X} - \mu_0}{\sqrt{\frac{\sigma_0^2}{n}}} \overset{n \text{ grand}}{\to} \mathcal{N}(0,1) \qquad \frac{S_n - n\mu_0}{\sqrt{n\sigma_0^2}} \overset{n \text{ grand}}{\to} \mathcal{N}(0,1)$$

$$P(-\ell_{\frac{\alpha}{2}} < \frac{\overline{X} - \mu_0}{\sqrt{\frac{\sigma_0^2}{n}}} < \ell_{\frac{\alpha}{2}}) \approx 1-\alpha \qquad P(-l_{\frac{\alpha}{2}} < \frac{S_n - n\mu_0}{\sqrt{n\sigma_0^2}} < \ell_{\frac{\alpha}{2}}) \approx 1-\alpha$$

$$P\left(\mu_0 - \ell_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_0^2}{n}} < \overline{X} < \mu_0 + \ell_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_0^2}{n}}\right) \approx 1-\alpha$$

$$ID_{1-\alpha}(\overline{X}) = \left[\mu_0 - \ell_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_0^2}{n}}\ , \mu_0 + \ell_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_0^2}{n}}\right]$$

$$P\left(n\mu_0 - \ell_{\frac{\alpha}{2}}\sqrt{n\sigma_0^2} < S_n < n\mu_0 + \ell_{\frac{\alpha}{2}}\sqrt{n\sigma_0^2}\right) \approx 1-\alpha$$

$$ID_{1-\alpha}(\sum_{i=1}^{n} X_i) = \left[n\mu_0 - \ell_{\frac{\alpha}{2}}\sqrt{n\sigma_0^2}\ ,\ n\mu_0 + \ell_{\frac{\alpha}{2}}\sqrt{n\sigma_0^2}\right]$$

Exercice 2 : Donner l'intervalle de dispersion pour la fréquence d'observations de Bernoulli de paramètre p_0 connue.

Exercice 3 : Une chaîne de supermarché décide de supprimer dans ses prix toutes références aux centimes d'euros par l'arrondi suivant :
$\{0,00;\ 0,01;\ 0,02\}$ donne 0,00 ; $\{0,08;\ 0,09\}$ donne 0,10 ;
$\{0,03;\ 0,04;\ 0,05;\ 0,06;\ 0,07\}$ donne 0,05

À quoi peut-elle s'attendre après la vente de 20 000 produits ?
Soit A la v.a. résultat pour le supermarché de l'arrondi, A prend pour valeur -2, -1, 0, 1, 2 en centime d'euros.

1. La loi de A est uniforme. Pourquoi ?
2. Quelle est son espérance, sa variance, son écart type (en euros)
3. Même question pour le résultat après la vente de 20 000 produits.
4. Donner alors l'ID pour le résultat de l'arrondi après la vente de 20 000 produits.

Notion élémentaire d'intervalle de confiance

On cherche à construire un intervalle de grande probabilité d'une grandeur particulière inconnue (en général le paramètre d'intérêt d'une loi de probabilité) .
Pour cela, on se fixe une faible probabilité α, appelée "risque" (en pratique α vaut 1%, 5% parfois 10%, 0.1%) et on construit

L'intervalle de confiance à 1-α

que l'on notera : $IC_{1-\alpha}$(paramètre) représentera un intervalle (dont les bornes seront aléatoires) qui aura une probabilité $1-\alpha$ de contenir le paramètre inconnu recherché.

Exemple 1 : à la veille d'une élection, on réalise un sondage pour connaître le plus précisément possible le paramètre associé au score d'un candidat.

$$\mathrm{P}(\text{score} \in [B_1, B_2]) = 95\%$$

Exemple 2 : on mesure les performances de n individus ($n = 200$) afin de proposer un intervalle pour l'ensemble de la population.

$$\mathrm{P}(\text{performance} \in [B_1, B_2]) = 99\%$$

les bornes B_1, B_2 de l'intervalle sont des v.a. et seront calculées en pratique à l'aide des observations.

Intervalle de confiance du paramètre p

d'une loi de Bernoulli

Soient $X_1, X_2, \ldots, X_n$ n répétitions indépendantes d'une même expérience aléatoire X de Bernoulli ($X \sim \text{Ber}(p)$) alors on sait que :

$$\frac{\overline{X} - p}{\sqrt{\frac{p(1-p)}{n}}} \overset{n \text{ grand}}{\rightarrow} \mathcal{N}(0,1)$$

Et comme pour Z v.a. N(0,1) on sait que $P(-\ell_{\frac{\alpha}{2}} < Z < \ell_{\frac{\alpha}{2}}) = 1 - \alpha$; on a

$$P(-\ell_{\frac{\alpha}{2}} < \frac{\overline{X} - p}{\sqrt{\frac{p(1-p)}{n}}} < \ell_{\frac{\alpha}{2}}) \approx 1 - \alpha$$

Et, $P\left(\overline{X} - \ell_{\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}} < p < \overline{X} + \ell_{\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}}\right) \approx 1 - \alpha$

$$P\left(\overline{X} - \ell_{\frac{\alpha}{2}}\sqrt{\frac{\overline{X}(1-\overline{X})}{n}} < p < \overline{X} + \ell_{\frac{\alpha}{2}}\sqrt{\frac{\overline{X}(1-\overline{X})}{n}}\right) \approx 1-\alpha$$

D'où

$$IC_{1-\alpha}(p) = \left[\overline{X} - \ell_{\frac{\alpha}{2}}\sqrt{\frac{\overline{X}(1-\overline{X})}{n}}\ ,\ \overline{X} + \ell_{\frac{\alpha}{2}}\sqrt{\frac{\overline{X}(1-\overline{X})}{n}}\right]$$

est donc un intervalle de confiance de p de probabilité approximativement $(1-\alpha)$

Exemple :

$n = 1000$

$\overline{x}$	$\sqrt{\frac{\overline{x}(1-\overline{x})}{n}}$	IC à 95%	IC à 99%
.2	.01265	[0.1752 , 0,2248]	[0.1674 , 0.2326]
.5	.01581	[0.4690 , 0.5310]	[0.4593 , 0.5407]
.9	.00949	[0.8814 , 0.9186]	[0.8756 , 0.9244]

Intervalle de confiance du paramètre μ

d'une loi gaussienne $\mathcal{N}(\mu, \sigma^2)$

Avec $X_1, X_2, \ldots, X_n$ n répétitions indépendantes d'une même expérience aléatoire de loi gaussienne $\mathcal{N}(\mu, \sigma^2)$, on sait que :

$$\frac{\overline{X} - \mu}{\sqrt{\frac{\sigma^2}{n}}} \sim \mathcal{N}(0,1)$$

Et comme pour Z v.a. N(0,1) on sait que $P(-\ell_{\frac{\alpha}{2}} < Z < \ell_{\frac{\alpha}{2}}) = 1 - \alpha$; on a

$$P(-\ell_{\frac{\alpha}{2}} < \frac{\overline{X} - \mu}{\sqrt{\frac{\sigma^2}{n}}} < \ell_{\frac{\alpha}{2}}) = 1 - \alpha$$

Et $P\left(\overline{X} - \ell_{\frac{\alpha}{2}}\sqrt{\frac{\sigma^2}{n}} < \mu < \overline{X} + \ell_{\frac{\alpha}{2}}\sqrt{\frac{\sigma^2}{n}}\right) = 1 - \alpha$

$$IC_{1-\alpha}(\mu) = \left[\overline{X} - \ell_{\frac{\alpha}{2}} \sqrt{\frac{\sigma^2}{n}} \ , \ \overline{X} + \ell_{\frac{\alpha}{2}} \sqrt{\frac{\sigma^2}{n}} \right]$$

est donc un intervalle de confiance de μ de probabilité $(1-\alpha)$ lorsque σ^2 est connu.

Dans le cas où σ est inconnu, on remplacera σ^2 par un estimateur : $\hat{\sigma}^2$ la variance empirique des observations $\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$.

$$IC_{1-\alpha}(\mu) = \left[\overline{X} - \ell_{\frac{\alpha}{2}} \sqrt{\frac{\hat{\sigma}^2}{n}} \ , \ \overline{X} + \ell_{\frac{\alpha}{2}} \sqrt{\frac{\hat{\sigma}^2}{n}} \right]$$

Ce qui nécessite un grand nombre d'observations n.

Exercice 4 :

On envisage de construire un intervalle de confiance de probabilité 94%

a) Quelle est la valeur du $\ell_{\alpha/2}$ correspondant : $\ell_{\alpha/2} =$

b) Si $X_1, X_2, ...X_n$ sont n variables aléatoires indépendantes d'une même loi Exponentielle $\mathcal{E}(\lambda)$, λ étant un paramètre inconnu.
Rappeler : si $X \sim \mathcal{E}(\lambda)$, $\mathrm{E}(X) =$ et $\mathrm{Var}(X) =$
ainsi que $\mathrm{E}(\bar{X}) =$ et $\mathrm{Var}(\bar{X}) =$

Donner l'expression d'un intervalle de confiance du paramètre λ qui tienne compte de la loi des X.

c) Si $X_1, X_2, ...X_n$ sont n variables aléatoires indépendantes d'une même loi de Poisson de paramètre inconnu λ.
Donner l'expression d'un intervalle de confiance du paramètre λ qui tienne compte de la loi des X.

StatL3S6 Chapitre 2

Le test statistique :

1. Introduction et Vocabulaire
2. Un test dans le cas continu
3. Un test dans le cas du sondage

Notion élémentaire du test statistique

Terminologie de base : on cherche à construire un test d'une hypothèse H_0 (appelée hypothèse nulle) contre une hypothèse H_1 (appelée hypothèse alternative) dans le but de rejeter l'hypothèse H_0.

Construire le test de H_0 contre H_1

C'est adopter une règle de décision qui amène :

à rejeter H_0 ou ne pas rejeter H_0.

Exemple 1 : avant son entrée en campagne, la popularité d'un candidat était stabilisée à 40%. Un récent sondage lui donne une popularité de 43%. Est-ce que la popularité du candidat a effectivement augmenté ou est-ce dû à une fluctuation du sondage ?

On construira le test de

- l'hypothèse H_0 *"status quo"* contre
- l'hypothèse alternative H_1 *"augmentation de la popularité"*.

Exemple 2 : un fabriquant de composant assure la qualité de son produit par la phrase suivante *"la fiabilité de ma production est supérieure ou égale à 99%"*. Lors d'un contrôle, on relève 1,09% de pièces défectueuses sur un échantillon de 1000 pièces. Doit-on supprimer l'accréditation au fabriquant ?
On construira le test de :
• l'hypothèse H_0 *"le taux de pièces défectueuses est de 1%"* contre
• l'hypothèse alternative H_1 *"le taux de pièces défectueuses est* $> 1\%$*"*.

Exemple 3 : on connaît les résultats d'un test (éducatif, psychologique, ...) sur l'ensemble de la population. On fait alors pratiquer ce test sur un groupe particulier et on observe sur ce groupe une augmentation des résultats. Est-ce que ce groupe a effectivement de meilleurs résultats ou est-ce dû à une fluctuation d'échantillonnage ?
On construira le test de :
• l'hypothèse H_0 *"le résultat au test est identique à celui de la population générale"*
• l'hypothèse alternative H_1 *"le résultat au test est supérieur"*.

Erreurs du test statistique

• On peut se tromper en déclarant H_1 vraie alors que H_0 est vraie : c'est l'erreur de 1^{re} espèce.

• On peut se tromper en déclarant H_0 vraie alors que H_1 est vraie : c'est l'erreur de 2^{e} espèce.

	Choix de H_0	Choix de H_1
H_0 vraie	Décision juste	Erreur de 1^{re} espèce
H_1 vraie	Erreur de 2^{e} espèce	Décision juste

En pratique, l'erreur de 1^{re} espèce ou niveau, notée α est fixée par l'utilisateur

et on construit donc

un test de H_0 contre H_1 de niveau α

Dans les exemples précédents l'erreur de 1^{re} espèce correspond à :

Exemple 1 : P(de croire à une "augmentation de la popularité" alors qu'il n'en est rien)

Exemple 2 : P(supprimer l'accréditation au fabriquant alors que le taux de pièces défectueuses est conforme)

Exemple 3 : P(penser à une amélioration dû au groupe alors qu'il n'en est rien)

Test du paramètre d'espérance d'une loi gaussienne $\mathcal{N}(\mu, \sigma^2)$ de l'hypothèse H_0 "$\mu = \mu_0$" contre H_1 "$\mu > \mu_0$"

Exemple 3 : on connaît les résultats d'un test (éducatif, psychologique, ...) sur l'ensemble de la population. On fait alors pratiquer ce test sur un groupe particulier et on observe sur ce groupe une augmentation des résultats. Est-ce que ce groupe a effectivement de meilleurs résultats ou est ce dû à une fluctuation du sondage ?
On construira le test de :

- l'hypothèse H_0 *"le résultat au test est identique à celui de la population générale"*
- l'hypothèse alternative H_1 *"le résultat au test est supérieur"*.

Exemple 3 : Les résultats d'un test (éducatif, psychologique, ...) sur l'ensemble de la population sont supposés $\mathcal{N}(\mu_0 = 100, \sigma^2)$. On fait alors pratiquer ce test sur un groupe particulier de 80 individus et on observe sur ce groupe une augmentation moyenne des résultats : sur ces 80 individus, la moyenne au test est de 110. Les résultats du groupe sont toujours supposés gaussiens $\mathcal{N}(\mu_{groupe}, \sigma^2)$; avec la même variance σ^2 .
Y-a-t-il une augmentation significative entre les résultats du groupe et la population totale ?

On construira le test de :

- l'hypothèse H_0 : $\mu_{groupe} = 100$
- l'hypothèse alternative H_1 : $\mu_{groupe} > 100$.

Test de l'hypothèse H_0 "$\mu = \mu_0$" contre H_1 "$\mu > \mu_0$"

Si les n observations sont gaussiennes $\mathcal{N}(\mu, \sigma^2)$ alors :

leur moyenne $\overline{X}$ est gaussienne $\mathcal{N}(\mu, \frac{\sigma^2}{n})$

et on rejettera l'hypothèse H_0 si $\overline{X}$ est grand par rapport à μ_0.

On se place alors sous H_0 vraie ("$\mu = \mu_0$") alors :

$$\frac{\overline{X} - \mu_0}{\sqrt{\frac{\sigma^2}{n}}} \sim \mathcal{N}(0, 1)$$

Donc,

$$P\left(\overline{X} < \mu_0 + \ell_\alpha \sqrt{\frac{\sigma^2}{n}}\right) = 1 - \alpha$$

Donc le test de H_0 "$\mu = \mu_0$" contre H_1 "$\mu > \mu_0$", d'erreur de première espèce (ou niveau) α : *Rejet de H_0 si*

$$\overline{x} > \mu_0 + \ell_\alpha \frac{\sigma}{\sqrt{n}}$$

C'est un test unilatéral.
Mais le fait de connaître la moyenne $\overline{x}$ ne permet aucune conclusion, la connaissance de la valeur de σ est fondamentale.

Exemple 3 : $\mu_0 = 100$; $n = 80$; pour $\alpha = 5\%$; $\ell_\alpha = 1.65$
Rejet de H_0 si $\overline{X} > 100 + .184 \times \sigma$.
Donc si $\sigma = 40$, on rejettera si $\overline{x} > 107.38$ et si $\sigma = 80$, on rejettera si $\overline{x} > 114.76$.
Dans l'énoncé, il est donné que $\overline{x}_{groupe} = 110$, on peut donc pas encore conclure puisque le paramètre σ n'est pas donné !

• Cas 1 : σ connu. Il n'y a donc pas de difficulté, mais ce n'est pas un cas réaliste en pratique.

• Cas 2 : σ inconnu. On remplacera σ^2 par la variance empirique des observations $\widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$.

En contre-partie, il est nécessaire que le nombre d'observations n soit grand.

En pratique, il est donc nécessaire d'avoir conservé les données et il est conseillé de prendre une erreur de 1^{re} espèce plus faible que lorsque l'on connaît σ.

• Cas 3 (fréquemment rencontré dans la littérature) : σ inconnu . On remplace σ^2 par

$$\widehat{\sigma^2} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

qui est un estimation (sans biais) du paramètre σ^2.
Sous l'hypothèse H_0 la statistique de test devient une loi de Student à $(n-1)$ degrés de liberté.

$$\frac{\overline{X} - \mu_0}{\sqrt{\frac{\widehat{\sigma^2}}{n}}} \sim \mathcal{S}_{n-1}$$

En pratique, il n'est pas nécessaire que le nombre d'observations n soit grand mais demande une *grande confiance* en l'hypothèse de normalité des observations.

Test du paramètre d'espérance d'une loi gaussienne

de l'hypothèse H_0 "$\mu = \mu_0$" contre H_1 "$\mu \neq \mu_0$"

Si H_0 est vraie alors :

$$\frac{\overline{X} - \mu_0}{\sqrt{\frac{\sigma^2}{n}}} \sim \mathcal{N}(0,1)$$

Donc,

$$P\left(\mu_0 - \ell_{\frac{\alpha}{2}}\sqrt{\frac{\sigma^2}{n}} < \overline{X} < \mu_0 + \ell_{\frac{\alpha}{2}}\sqrt{\frac{\sigma^2}{n}}\right)$$

$$= 1 - \alpha$$

Donc le test de H_0 "$\mu = \mu_0$" contre H_1 "$\mu \neq \mu_0$", d'erreur de première espèce (ou niveau) α.

Rejet de H_0 si

$$\overline{x} < \mu_0 - \ell_{\frac{\alpha}{2}} \sqrt{\frac{\sigma^2}{n}} \quad \text{ou} \quad \overline{x} > \mu_0 + \ell_{\frac{\alpha}{2}} \sqrt{\frac{\sigma^2}{n}}$$

C'est un test bilatéral, en remplaçant σ^2 par la variance empirique des observations $\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$.

Exercice 5 : Dans un magasin d'habillement, on a noté ces dernières années le prix moyen d'un article du rayon "enfant" qui est de 18 euros. La direction souhaite communiquer sur sa politique de promotion dans ce rayon par la baisse du prix moyen d'un article.
Pour cela, elle relève le prix de 121 articles du rayon (pris au hasard).

1. Construire le test statistique associé à cette démarche.
2. Calculer la zone de rejet de ce test, sachant que sur les 121 articles on a calculé que la dispersion $\sum_{i=1}^{121}(p_i - \bar{p})^2 = 5929$ où p_i est le prix de l'article i et $\bar{p}$ le prix moyen observé.
3. Sur ces 121 articles le prix moyen observé au mois de mai 2008 est de 16.8 euros. Quel est la décision à prendre concernant ce test statistique ? Que proposeriez-vous alors à la direction ?
4. Idem pour un prix moyen observé au mois de juin de 16.2 euros.
5. Construire alors un intervalle du prix moyen d'un article du rayon "enfant" (choisir un risque $\alpha = 3.6$ %). Quel est le nom de cet intervalle ?

Test du paramètre d'une loi de Bernoulli

de l'hypothèse H_0 "$p = p_0$" contre H_1 "$p \neq p_0$"

Si H_0 est vraie alors : $\dfrac{\overline{X} - p_0}{\sqrt{\dfrac{p_0(1-p_0)}{n}}} \overset{n \text{ grand}}{\rightarrow} \mathcal{N}(0,1)$

D'où, $P\left(p_0 - \ell_{\frac{\alpha}{2}} \sqrt{\dfrac{p_0(1-p_0)}{n}} < \overline{X} < p_0 + \ell_{\frac{\alpha}{2}} \sqrt{\dfrac{p_0(1-p_0)}{n}} \right) \approx 1 - \alpha$

Donc rejet de H_0 si

$$\overline{x} < p_0 - \ell_{\frac{\alpha}{2}} \sqrt{\frac{p_0(1-p_0)}{n}} \quad \text{ou} \quad \overline{x} > p_0 + \ell_{\frac{\alpha}{2}} \sqrt{\frac{p_0(1-p_0)}{n}}$$

C'est un test bilatéral.

Test du paramètre d'une loi de Bernoulli

de H_0 "$p = p_0$" contre H_1 "$p > p_0$"

Si H_0 est vraie alors :

$$P\left(\frac{\overline{X} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} < \ell_\alpha\right) = P\left(\overline{X} < p_0 + \ell_\alpha\sqrt{\frac{p_0(1-p_0)}{n}}\right) \approx 1 - \alpha$$

Donc le test de H_0 "$p = p_0$" contre H_1 "$p > p_0$", d'erreur de première espèce (ou niveau) α est :

Rejet de H_0 si

$$\overline{x} > p_0 + \ell_\alpha\sqrt{\frac{p_0(1-p_0)}{n}}$$

C'est un test unilatéral.

Exercice 6 :

1. Un fabricant de disques compacts affirme qu'au moins 99% de ses disques n'ont aucun défaut. Pour vérifier cette affirmation une association de défense des consommateurs teste 500 disques de ce fabriquant et en trouve 10 défectueux.
   Avec un seuil de 1%, l'association peut-elle contester l'affirmation du fabricant ?
2. Le même fabricant de disques compacts affirme toujours qu'au moins 99% de ses disques n'ont aucun défaut.
   Il cherche à conquérir un nouveau marché à l'aide de cette affirmation et doit donc rassurer son nouveau client.
   Il lui propose de venir contrôler sa chaîne de production en relevant 1000 ou 2000 exemplaires.
   Quel sera la démarche de ce nouveau client ?

StatL3S6 Chapitre 3

Construction de tests en présence de 2 distributions :

1. Test sur la corrélation
2. Test d'adéquation
3. Analyse de la variance

Deux variables quantitatives : le coefficient de corrélation linéaire

Le coefficient de corrélation linéaire cherche à mesurer la liaison linéaire qui peut exister entre deux variables X et Y observées sur les mêmes individus.

$$\rho_{XY} = E\left(\frac{X - E(X)}{\sigma_X} \times \frac{Y - E(Y)}{\sigma_Y}\right)$$

qui sera estimé par le coefficient de corrélation linéaire empirique :

$$\hat{\rho}_{XY} = \frac{1}{n}\sum_{i=1}^{n} \frac{(x_i - \bar{x})}{\hat{\sigma}_X} \times \frac{(y_i - \bar{y})}{\hat{\sigma}_Y}$$

et on a $-1 \leq \hat{\rho}_{XY} \leq 1$.
Plus ce coefficient se rapproche de 1, plus les variables sont corrélées positivement, c'est-à-dire qu'elles varient dans le même sens. Plus il se rapproche de -1, plus elles varient en sens opposé. S'il se rapproche de 0, leurs variations ne sont pas liées linéairement.

Exemple 1 : Deux séries de notes observées sur 12 individus

X :	14	13	17	15	14	15	16	12	14	12	13	13
Y :	13	11	16	15	12	13	15	10	14	12	13	12

Ex1 : $\hat{\rho} = 0.874$.

Exemple 2 : “fichier notes"

	Math	Phys	Chim	Angl	Fran	Hist
1	15	17	18	9	8	10
2	6	7	5	10	7	5
3	7	4	4	13	15	19
4	18	19	19	18	14	16
5	8	12	10	10	11	9
6	15	14	19	12	6	8
7	6	10	5	19	13	16
8	14	16	12	17	11	15
9	8	7	8	9	10	10
10	7	9	7	7	5	5
11	9	10	11	12	14	13
12	14	18	15	6	7	5
13	5	7	9	18	16	16
14	6	11	10	9	5	8
15	14	16	18	16	11	10
16	16	12	12	14	11	14
17	14	16	15	8	7	10
18	9	8	13	12	9	10
19	6	4	7	15	14	17
20	8	4	3	8	9	8
21	12	15	13	15	12	13
22	7	4	3	17	15	13
23	5	8	7	9	9	9
24	16	14	17	7	9	6
25	7	11	12	11	10	9

Variable	Moyenne	Écart-type	Minimum	Maximum
Mathém	10.08	4.17	5.00	18.00
Physique	10.92	4.69	4.00	19.00
Chimie	10.88	5.07	3.00	19.00
Anglais	12.04	3.94	6.00	19.00
Français	10.32	3.22	5.00	16.00
Histoire	10.96	4.03	5.00	19.00

	Math.	Phys.	Chim.	Angl.	Fran.	Hist.
Mathém	1.00	0.82	0.83	-0.00	-0.15	-0.05
Physique	0.82	1.00	0.87	-0.04	-0.29	-0.18
Chimie	0.83	0.87	1.00	-0.05	-0.25	-0.17
Anglais	-0.00	-0.04	-0.05	1.00	0.76	0.80
Français	-0.15	-0.29	-0.25	0.76	1.00	0.85
Histoire	-0.05	-0.18	-0.17	0.80	0.85	1.00

Ce tableau s'appelle le tableau ou matrice des corrélations.

!! Attention

Un coefficient de corrélation ne traduit pas nécessairement une relation de cause à effet :

"Une bonne note en math n'implique pas une bonne note en chimie."

Autre exemple : la corrélation entre le revenu et le débit de carte bancaire est fortement positif. Il existe ici une relation évidente : plus le revenu est élevé plus le débit de carte bancaire va augmenter et pas le contraire!!!

!! La relation n'est pas contenue dans les données.

Quand doit-on considérer le coefficient de corrélation ρ comme significativement non nul ?

C'est un problème de test statistique avec H_0 "$\rho = 0$".
En se plaçant sous l'hypothèse "absence de lien linéaire" entre 2 variables quantitatives et sous les hypothèses du théorème central limite :

$$\frac{\rho}{\sqrt{\frac{1}{n-1}}} \sim \mathcal{N}(0,1)$$

On décidera de la dépendance entre 2 variables quantitatives lorsque

$$\sqrt{n-1}\,\rho \text{ est grand en valeur absolue}$$

Rejet de l'hypothèse H_0 "$\rho = 0$" : "absence de lien linéaire" contre H_1 "$\rho \neq 0$" si

$$|\hat{\rho}| > \frac{\ell_{\frac{\alpha}{2}}}{\sqrt{n-1}}$$

Exemple : pour $\alpha = 5\%$, on rejettera l'hypothèse "absence de lien linéaire" si $|\hat{\rho}| > \frac{1.96}{\sqrt{n-1}}$. C'est à dire, rejet :

si $\lvert\hat{\rho}\rvert > 0.653$	$(n = 10)$	si $\lvert\hat{\rho}\rvert > 0.197$	$(n = 100)$
si $\lvert\hat{\rho}\rvert > 0.450$	$(n = 20)$	si $\lvert\hat{\rho}\rvert > 0.0877$	$(n = 500)$
si $\lvert\hat{\rho}\rvert > 0.364$	$(n = 30)$	si $\lvert\hat{\rho}\rvert > 0.0620$	$(n = 1000)$
si $\lvert\hat{\rho}\rvert > 0.314$	$(n = 40)$	si $\lvert\hat{\rho}\rvert > 0.0196$	$(n = 10\ 000)$
si $\lvert\hat{\rho}\rvert > 0.280$	$(n = 50)$	si $\lvert\hat{\rho}\rvert > 0.00196$	$(n = 1\ 000\ 000)$

Un test d'adéquation : le test du χ^2

Exemple 1 : "lancer de dés"

Une expérience consiste à lancer deux dés, et à relever la somme des chiffres lus. On fait l'expérience $n = 1000$ fois, et on obtient :

S	2	3	4	5	6	7	8	9	10	11	12
n_k	32	56	81	115	142	160	143	105	89	53	24

Exemple 2 : "Les familles de 8 enfants"

On a observé, en étudiant 53680 familles de 8 enfants, les résultats suivants (k désigne le nombre de garçons et n_k le nombre de familles ayant k garçons) :

k	0	1	2	3	4	5	6	7	8
n_k	215	1485	5331	10649	14959	11929	6678	2092	342

Exemple 3 : "10 000 premières décimales du nombre π"
La répartition de ces décimales est donnée dans le tableau suivant :

décimale	0	1	2	3	4	5	6	7	8	9
effectifs	968	1026	1021	974	1012	1046	1021	970	948	1014

Se répartissent-elles de manière uniforme ?

Exemple 4 : "Sondage sur le niveau d'acceptation d'un nouveau système"

Appréciation Les sondés	Très difficile	Assez difficile	Peu/pas difficile
Jeunes	81	138	132
Actifs	126	131	94
Retraités	203	78	69

Distance entre 2 tableaux : le tableau des observations $[n_k]$ et le tableau sous l'hypothèse d'un modèle [Eff théo$_k$]

$$\chi^2 = \sum_k \frac{(n_k - \text{Eff théo}_k)^2}{\text{Eff théo}_k}$$

Cette distance sera donc d'autant plus grande que le tableau des observations sera loin du tableau sous l'hypothèse du modèle

Pourquoi diviser par Eff théo$_k$?

Théo	1000	100	10
Obs	1010	110	20
écart	négligeable	faible	important
$\chi^2_{cellule}$	1/10	1	10

Quand doit-on considérer cette distance du χ^2 comme grande ?

C'est un problème décisionnel : donc un test statistique.
Pour chaque situation on a une hypothèse H_0 "adéquation à un modèle" et une alternative $H_1 =$ non H_0, et on aura :

rejet de H_0 si $\boxed{\sum_k \frac{(n_k - \text{Eff théo}_k)^2}{\text{Eff théo}_k} > \ell_{\nu,\alpha}}$

On décidera de rejeter l'adéquation au modèle lorsque la distance du χ^2 sera supérieure à une valeur limite, qui dépend d'un degré de liberté et de l'erreur de 1^{re} espèce choisie.

Pour répondre à cette question, on pourra utiliser la table du χ^2 qui donne cette valeur limite pour chaque degré de liberté et différentes erreurs de 1^{re} espèce.

Table du χ^2

ν : nombre de degrés de liberté

$P(\chi^2_\nu < \ell_{\nu,\alpha}) = p = 1 - \alpha$

Exemple : $P(\chi^2_4 < 11.1433) = 0.975$

p	0.8000	0.9000	**0.9500**	0.9750	0.9900	0.9950
ν						
1	1.6424	2.7055	3.8415	5.0239	6.6349	7.8794
2	3.2189	4.6052	5.9915	7.3778	9.2103	10.5966
3	4.6416	6.2514	7.8147	9.3484	11.3449	12.8382
4	5.9886	7.7794	9.4877	**11.1433**	13.2767	14.8603
5	7.2893	9.2364	11.0705	12.8325	15.0863	16.7496
6	8.5581	10.6446	12.5916	14.4494	16.8119	18.5476
7	9.8032	12.0170	14.0671	16.0128	18.4753	20.2777
8	11.0301	13.3616	15.5073	17.5345	20.0902	21.9550
9	12.2421	14.6837	16.9190	19.0228	21.6660	23.5894
10	13.4420	15.9872	18.3070	20.4832	23.2093	25.1882
11	14.6314	17.2750	19.6751	21.9200	24.7250	26.7568
12	15.8120	18.5493	21.0261	23.3367	26.2170	28.2995
13	16.9848	19.8119	22.3620	24.7356	27.6882	29.8195
14	18.1508	21.0641	23.6848	26.1189	29.1412	31.3193
90	101.0537	107.5650	113.1453	118.1359	124.1163	128.2989
100	111.6667	118.4980	124.3421	129.5612	135.8067	140.1695

Analyse de la variance ou Anova

1 - Introduction

Objectif : Étudier l'effet d'une ou plusieurs variables qualitatives sur une variable quantitative

Le cas d'une variable qualitatitive : on observe sur des individus à la fois une variable quantitative et une une variable qualitative. On cherhche alors à "savoir" si les différentes modalités de la variable qualitative influencent la variable quantitative.

Exemple : On considère 6 échantillons de patients correspondant à des localisations différentes. Pour chaque patient, on observe une donnée clinique :

1	2	3	4	5	6
1602	1472	1548	1435	1493	1585
1615	1477	1555	1438	1498	1592
1624	1485	1559	1448	1509	1598
1631	1493	1563	1449	1516	1604
	1496	1575	1454	1521	1609
	1504		1458	1523	1612
	1510		1467		
			1475		

Information disponible pour chaque patient :

- Y : donnée clinique
- X : code de la localisation

Question : peut-on considérer que la localisation a une influence sur la donnée clinique des patients ?

... ou encore : la variable X a-t-elle une influence sur la variable Y ?

... ou encore : la modélisation de l'espérance μ de Y doit-elle être différente selon les modalités de X ou non ?

Vocabulaire

On appelle facteur (ou facteur explicatif, cause contrôlée) la variable qualitative X qui sert à expliquer Y.
On parle de niveaux d'un facteur (ou traitements) pour désigner les différentes modalités de cette variable.
Lorsqu'on étudie l'effet de plusieurs facteurs sur Y, on peut regarder leurs effets cumulés mais aussi l'effet de leur interaction, i.e. le croisement de deux modalités a une influence particulière sur Y.
On appelle unité expérimentale le sujet que l'on soumet à un traitement et sur lequel on mesure Y.
De façon générale ici, on suppose que l'on ne soumet pas une unité expérimentale à plusieurs traitements, autrement dit on suppose qu'il n'y a pas de répétition de la mesure de Y. Il y a donc autant d'observations que d'unités expérimentales.

ANOVA : pourquoi ?

- Analyse de la variance :
  - variations inter-groupes : écart entre les moyennes des groupes, dispersion des moyennes autour de la moyenne globale.
  - variations intra-groupes : écart entre les données à l'intérieur des groupes, dispersion des données autour de leur moyenne de groupe.

- ANOVA
  - ↗ technique de comparaison de moyennes
  - ↘ décider de l'égalité des moyennes

Approche intuitive sur un exemple

Un étudiant a mesuré le temps de parcours pour ce rendre à la fac selon trois types de trajet.

T1	17.5	20.0	18.0	17.0	16.5
T2	15.1	16.0	13.0	12.0	14.5
T3	10.0	13.0	10.0	11.0	12.0

Structure générale des données :

On a le tableau des données suivant :

facteur	niveau 1	...	niveau i	...	niveau I
observations	y_{11}	...	y_{i1}	...	y_{I1}
indépendantes	y_{12}	...	y_{i2}	...	y_{I2}
de la variable	$\vdots$		$\vdots$		$\vdots$
quantitative	y_{1n_1}	...	y_{in_i}	...	y_{In_I}

- y_{ik} : k^e observation du niveau i
- n_i : nombre d'observations du niveau i

- $n = \sum_{i=1}^{I} n_i$: nombre total d'observations
- $\bar{y}_{i.} = \frac{1}{n_i} \sum_{k=1}^{n_i} y_{ik}$: moyenne des observations pour le niveau i
- $\bar{y}_{..} = \frac{1}{n} \sum_{i=1}^{I} \sum_{k=1}^{n_i} y_{ik}$: moyenne globale

! ! ! Attention
La moyenne des observations $\bar{y}_{..}$ n'est pas égale à la moyenne des moyennes par niveau du facteur $\bar{y}_{i.}$

Quelques hypothèses naturelles

Les Y_{ik} sont les résultats aléatoires de l'expérience étudiée et on suppose que leur "valeur espérée" est un paramètre noté μ_i qui ne dépend que du niveau du facteur contrôlé.

- μ_i est appelé l'effet fixe du niveau i du facteur contrôlé. C'est un paramètre inconnu : l'espérance de l'observation Y_{ik}
- les Y_{ik} sont aussi supposés être indépendants (donc associés à des sujets distincts.

Dans cette modélisation, il y a donc I paramètres inconnus liés à l'espérance : un paramètre pour chacun des niveaux du facteur.

Équation d'analyse de la variance :

$$\sum_{ik}(y_{ik} - \bar{y}_{..})^2 = \sum_{ik}(y_{ik} - \bar{y}_{i.})^2 + \sum_{i} n_i(\bar{y}_{i.} - \bar{y}_{..})^2$$

dispersion totale = dispersion **INTRA** + dispersion **INTER**

$$SS_T = SS_R + SS_F$$

$$\uparrow \quad \uparrow \quad \uparrow$$

Sum of Square Total = Sum of Square Résiduelle + Sum of Square due au facteur contrôlé

Degrés de liberté associés aux SS

$$\text{ddl}(SS_T) = \text{ddl}(SS_R) + \text{ddl}(SS_F)$$

$$n - 1 = n - I + I - 1$$

Données :

1	2	3	4	5	6	
1602	1472	1548	1435	1493	1585	
1615	1477	1555	1438	1498	1592	
1624	1485	1559	1448	1509	1598	
1631	1493	1563	1449	1516	1604	
	1496	1575	1454	1521	1609	
	1504		1458	1523	1612	
	1510		1467			
			1475			
4	7	5	8	6	6	n_i
1618	1491	1560	1453	1510	1600	$\bar{y}_{i.}$
470	1152	404	1296	760	534	$\sum_k (y_{ik} - \bar{y}_{i.})^2$

Analyse de la variance sur l'exemple des patients :

i	1	2	3	4	5	6
$\bar{y}_{i.}$	1618	1491	1560	1453	1510	1600

$$\bar{y}_{..} = \frac{1}{n}\sum_{ik} y_{ik} = \frac{1}{n}\sum_{i} n_i \bar{y}_{i.} = 1528 \qquad \text{ici } n = 36$$

Tableau d'analyse de la variance

Source de dispersion	Somme des carrés	ddl
INTER	125145 $\sum_i n_i(\bar{y}_{i.} - \bar{y}_{..})^2$	5 $I-1$
INTRA	4616 $\sum_{ik}(y_{ik} - \bar{y}_{i.})^2$	30 $n-I$
TOTALE	129761 $\sum_{ik}(y_{ik} - \bar{y}_{..})^2$	35 $n-1$

Test de l'égalité des espérances :

• l'hypothèse H_0 : absence d'effet du facteur contrôlé ; égalité des espérances $\mu_1 = \mu_2 \cdots = \mu_I$
• l'hypothèse alternative H_1 : effet significatif du facteur contrôlé ; différence des espérances $\mu_i,\ i = 1, \ldots, I$

$$\frac{\text{INTER}}{\text{INTRA}} * (n - I) = \frac{\sum_{ik}(y_{ik} - \bar{y}_{..})^2 - \sum_{ik}(y_{ik} - \bar{y}_{i.})^2}{(\sum_{ik}(y_{ik} - \bar{y}_{i.})^2)/(n - I)}$$

suit sous H_0 une loi de χ^2 à $(I - 1)$ degrés de liberté
dès que le nombre d'observations est grand

Re-écriture du paramètre μ_i et interprétation

$$\mu_i = \mu + \alpha_i \quad i = 1, \ldots, I \tag{2}$$

avec $I + 1$ paramètres pour l'espérance dont seulement I sont libres et identifiables $\Longrightarrow$ il y a sur-paramétrisation.
Différentes contraintes peuvent alors être envisagées :

- $\sum_i \alpha_i = 0$
  Lien avec la paramétrisation de l'équation (1) :

$$\mu = \frac{1}{I}\sum_{i=1}^{I} \mu_i = \bar{\mu}_. \qquad \text{et} \qquad \alpha_i = \mu_i - \bar{\mu}_.$$

  Le paramètre μ représente alors l'effet moyen général
  Le paramètre α_i représente alors l'effet différentiel du niveau i à la "moyenne"

- $\alpha_1 = 0$, par défaut dans de nombreux logiciels de statistiques.
  Lien avec la paramétrisation de l'équation (1) :

$$\mu = \mu_1 \qquad \text{et} \qquad \alpha_i = \mu_i - \mu_1$$

  Le paramètre μ représente alors l'effet du niveau 1 du facteur
  Le paramètre α_i représente alors l'effet différentiel du niveau i à l'effet du niveau 1
  Ici le traitement 1 sert de référence mais on peut prendre l'un quelconque des I traitements comme référence.

Estimation des paramètres :

Par moindres carrés : $\min_{\mu_i} \sum_{i=1}^{I} \sum_{k=1}^{n_i} (y_{ik} - \mu_i)^2$

- $\hat{\mu}_i = \bar{y}_{i.} = \frac{1}{n_i} \sum_{k=1}^{n_i} y_{ik}$
- contraste "sum" :

$$\hat{\mu} = \hat{\bar{\mu}}_{.} = \frac{1}{I} \sum_{i=1}^{I} \hat{\mu}_i = \frac{1}{I} \sum_{i=1}^{I} \bar{y}_{i.} \qquad \text{et} \qquad \hat{\alpha}_i = \hat{\mu}_i - \hat{\mu} = \bar{y}_{i.} - \frac{1}{I} \sum_{i=1}^{I} \bar{y}_{i.}$$

Remarque : si $\forall i \in \{1, \ldots, I\} \quad n_i = K$ alors $\hat{\mu} = \bar{y}_{..}$ et $\hat{\alpha}_i = \bar{y}_{i.} - \bar{y}_{..}$

- contraste "treatment" :

$$\hat{\mu} = \hat{\mu}_1 = \bar{y}_{1.} \qquad \text{et} \qquad \hat{\alpha}_i = \hat{\mu}_i - \hat{\mu}_1 = \bar{y}_{i.} - \bar{y}_{1.}$$