université de BORDEAUX

Master Chimie QSE – S8

Cours de Statistiques Appliquées MSP

Objectifs : Inférence statistique et applications

Mettre en œuvre des méthodes statistiques pour permettre de **prendre une décision** à partir d'un ou plusieurs échantillon.s

Modalités :

Cours / TD / TP intégrés

Traitement de données avec le logiciel R

Evaluation sur machines (1,5 heure)

Florent ARNAL

florent.arnal@u-bordeaux.fr

http://flarnal.e-monsite.com

PLAN DU COURS

I. Rappels et compléments de statistiques descriptives

1. Paramètres fondamentaux de Statistiques à 1 variable

La moyenne d'une série statistique $(x_i)_{1\leq i\leq n}$ est un paramètre de position défini par :

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Sa variance est un paramètre de dispersion défini par :

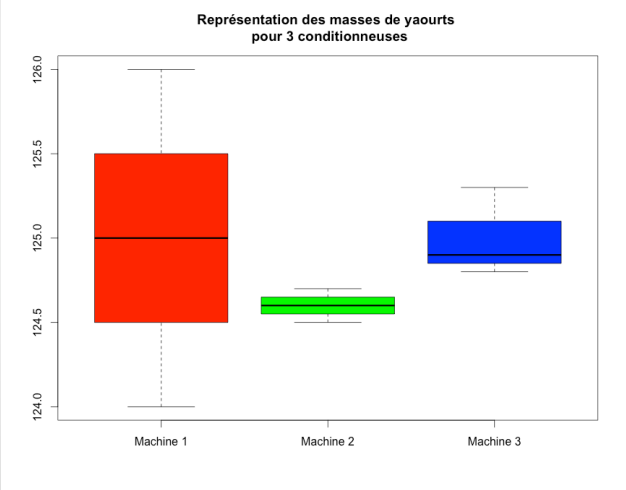
$$s^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2 = \frac{SCE}{n}$$

s correspond à l'écart-type de cette série statistique.
Plus s est voisin de 0, plus la série est homogène

3

2. Représentations graphiques

Données quantitatives : le boxplot



4

Données qualitatives (Répartition des défauts d'une production) :

Le diagramme de Pareto

Ce diagramme, basé sur des fréquences cumulées, permet d'évaluer l'importance de différents facteurs sur un processus.

Historique : V. Pareto, économiste italien, avait fait une étude sur la répartition des richesses mettant en évidence que 80 % des richesses étaient détenues par 20 % de la population (loi des 80/20).

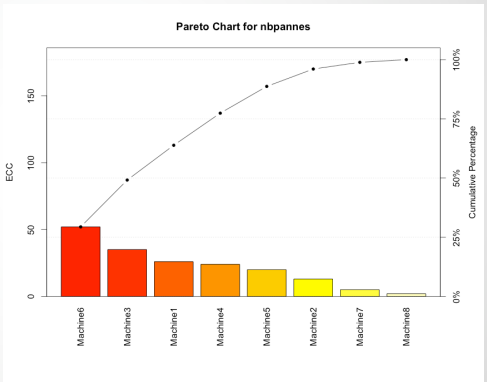
Juran en tire l'idée que, pour un phénomène, 20 % des causes produisent 80 % des effets (d'où l'intérêt de s'intéresser à la répartition des défauts d'une production).

5

Données qualitatives (Répartition des défauts d'une production) :

Le diagramme de Pareto

Numéro De machine	Nombre de pannes sur 1 an
1	26
2	13
3	35
4	24
5	20
6	52
7	5
8	2



Avec R : pareto.chart( ) via le package qcc

6

3. Statistiques à 2 variables : Introduction

Afin de procéder à l'étalonnage d'un nouvel appareil de mesures, on effectue 5 mesures (grandeurs obtenues) avec cet appareil associées à 5 valeurs de référence X (grandeurs théoriques).
Pour chaque mesure effectuée, on calcule l'écart, noté Y, entre la valeur obtenue et la valeur de référence.

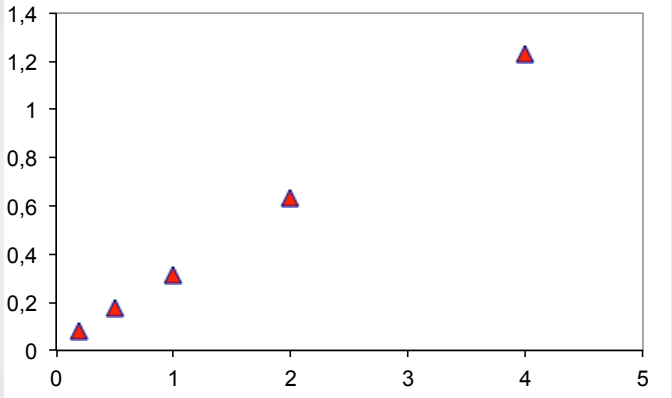
Valeurs de référence x_i	0,2	0,5	1	2	4
Écarts y_i	0,083	0,176	0,311	0,631	1,231

Pour une valeur de référence de 6, peut-on donner un ordre de grandeur de la mesure avec cet appareil ?

7

Statistiques à 2 variables : Introduction

Représentation du nuage de points associé à (X, Y)



Objectif : Trouver une relation affine entre Y et X de la forme

$$Y = aX + b$$

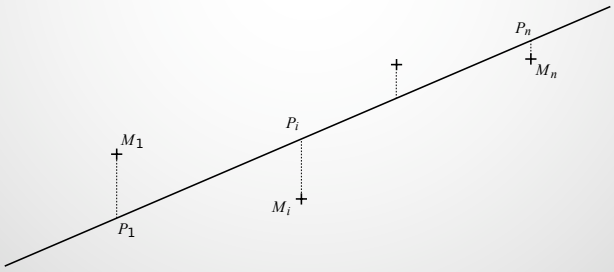
8

Statistiques à 2 variables : Méthode des moindres carrés

Soit une série statistique double représenté par un nuage de points $Mi\,(x_i;y_i)_{1\leq i\leq n}$.

Objectif : Déterminer, suivant la méthode des moindres carrés, une équation de la droite (D) passant "le plus proche" possible des points du nuage.

Pour tout entier naturel i tel que $1 \leq i \leq n$, on note P_i le projeté de M_i sur la droite (D) parallèlement à l'axe des ordonnées.
Ajuster ce nuage de points par la méthode des moindres carrés, c'est déterminer la droite (D) pour que la somme $\sum_{i=1}^{n} M_iP_i^2$ soit minimale.



Minimiser la somme $\sum_{i=1}^{n} M_iP_i{}^2$ revient à déterminer le minimum de la fonction φ définie sur $\mathbb{R}^2$ par

$$\varphi(a,b) = \sum_{i=1}^{n} [(ax_i + b) - y_i]^2$$

En considérant que les dérivées partielles doivent s'annuler en l'extrêmum, on obtient :

$$\sum_{i=1}^{n} x_iy_i = a\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i$$

$$\sum_{i=1}^{n} y_i = a\sum_{i=1}^{n} x_i + nb$$

Notons X et Y les variables prenant respectivement les valeurs (x_i) et (y_i)
La résolution du système associé conduit à :

$$a = \frac{\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y}}{\frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{cov(X,Y)}{\sigma_X^2}$$

$$b = \bar{y} - a\bar{x}$$

En fixant $b = \bar{y} - a\bar{x}$, on constate que cet extremum est un minimum.

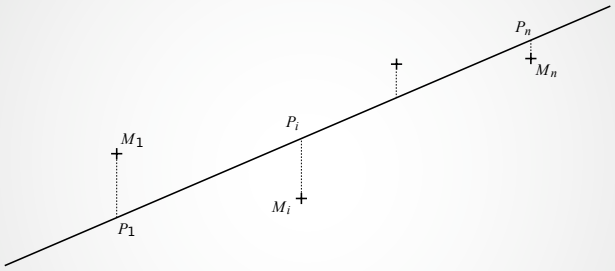
La droite d'équation $Y = aX + b$ est la droite d'ajustement du nuage de points par la méthode des moindres carrés.

11

Les points P_i appartenant à la droite d'ajustement (D) ont pour ordonnée

$$\widehat{y_i} = ax_i + b$$

La distance entre les points M_i et P_i correspond à $|y_i - \widehat{y_i}|$.



On définit les résidus, notés e_i, par :

$$e_i = y_i - \widehat{y_i}$$

Les résidus sont de somme et de moyenne nulle.
La méthode des moindres carrés conduit à avoir la somme $\sum e_i^2$ minimale.

12

Indicateur de qualité d'un ajustement : Coefficients de corrélation et détermination

Considérons les SCE suivantes :

$SCE_{totale} = \sum_{i=1}^{n} (y_i - \bar{y})^2$ $\quad SCE_{exp} = \sum_{i=1}^{n} (\widehat{y_i} - \bar{y})^2$ $\quad SCE_{res} = \sum_{i=1}^{n} {e_i}^2$

Relation fondamentale :

$$SCE_{totale} = SCE_{exp} + SCE_{res}$$

Plus la variabilité résiduelle est faible, plus la part expliquée est importante.
Le coefficient de détermination est un indicateur de la qualité de la régression défini par :

$$R^2 = \frac{SCE_{exp}}{SCE_{totale}}$$

Plus R^2 est voisin de 1, plus la relation affine entre Y et X est "significative" ...

13

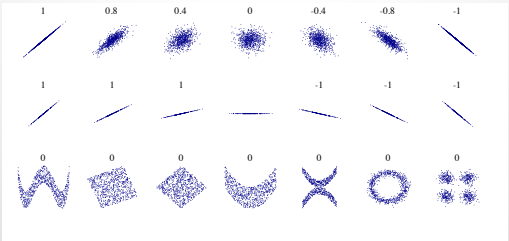
Indicateur de qualité d'un ajustement : Coefficients de corrélation et détermination

Le nombre R est appelé coefficient de corrélation linéaire entre Y et X.

On peut montrer que :

$$R = \frac{cov(X;Y)}{\sigma_X \sigma_Y}$$

Exemples de nuages avec les valeurs du coefficient de corrélation linéaire pour les deux premières lignes.
Les graphiques de la dernière ligne ne sont pas associés à des relations linéaires.



14

II. Généralités sur les Variables aléatoires

15

Généralités sur les variables aléatoires discrètes

Définition :
Soit X une variable aléatoire discrète sur Ω, muni d'une probabilité P telle que $X(\Omega) = \{x_1, x_2, \cdots, x_n\}$.
On appelle loi de probabilité de X, la suite de réels $(p_i)_{1 \leq i \leq n}$ définis par :

$$p_i = P(\{\omega \in \Omega \text{ tq } X(\omega) = x_i)\}) = P(X = x_i)$$

Remarque : Cette définition s'étend à des variables prenant des valeurs dans $\mathbb{N}$

Exemple : On lance 3 fois une pièce de monnaie non truquée.
Gain de 2 euros si le résultat est Pile et perte de 1 euro si le résultat est Face.
On note X la variable égale au gain d'un joueur.

$$X(\Omega) = \{-3; 0; 3; 6\}$$
$$P(X = -3) = P(X = 6) = \frac{1}{8}$$
$$P(X = 0) = P(X = 3) = \frac{3}{8}$$

16

Généralités sur les variables aléatoires discrètes

Propriété : $\sum_{k\in X(\Omega)} \mathbb{P}(X=k) = 1$

Définition : La fonction de répartition d'une variable X est la fonction F_X définie sur $\mathbb{R}$ par

$$F_X(t) = \mathbb{P}(X \leq t)$$

Définition : L'espérance de X est le réel noté $\mathbb{E}(X)$ défini par

$$\mathbb{E}(X) = \sum_{k\in X(\Omega)} k\mathbb{P}(X=k)$$

Propriété-définition : La variance de X est le réel (positif) défini par

$$\mathbb{V}(X) = \sum_{k\in X(\Omega)} [k-\mathbb{E}(X)]^2 \mathbb{P}(X=k)$$

$$\mathbb{V}(X) = \sum_{k\in X(\Omega)} k^2\mathbb{P}(X=k) - [\mathbb{E}(X)]^2 = \mathbb{E}\left(X^2\right) - [\mathbb{E}(X)]^2$$

17

Généralités sur les variables aléatoires discrètes
Propriétés de l'espérance et de la variance

Théorème de transfert : Soit $f : I \to \mathbb{R}$ une application avec $X(\Omega) \subset I$.

$$\mathbb{E}(f(X)) = \sum_{k\in X(\Omega)} f(k)\mathbb{P}(X=k)$$

Propriété : Soient a et b deux réels.

$$\mathbb{E}(aX+b) = a\mathbb{E}(X) + b \quad ; \quad \mathbb{V}(aX+b) = a^2\mathbb{V}(X)$$

Propriété : Soient X et Y deux variables aléatoires.

$$\mathbb{E}(X+Y) = \mathbb{E}(X) + \mathbb{E}(Y)$$

$\mathbb{V}(X+Y) = \mathbb{V}(X)+\mathbb{V}(Y)+2\,\mathrm{Cov}(X;Y)$ où $\mathrm{Cov}(X;Y) = \mathbb{E}(XY)-\mathbb{E}(X)\,\mathbb{E}(Y)$

Si X et Y sont indépendantes alors

$$\mathbb{V}(X+Y) = \mathbb{V}(X) + \mathbb{V}(Y)$$

18

III. Lois usuelles discrètes

19

Lois usuelles discrètes
Loi de Bernoulli

Définition : Une variable aléatoire X est distribuée suivant la loi de Bernoulli de paramètre p si :

- $X(\Omega) = \{0; 1\}$
- $\mathbb{P}(X = 0) = 1 - p$ et $\mathbb{P}(X = 1) = p$

Ces lois sont utilisées dans le cas d'une expérience avec deux issues possibles (épreuve de Bernoulli).
Le 1 est associé au succès et le 0 à l'échec.

Propriété : Soit X une variable aléatoire de Bernoulli de paramètre p.

- $\mathbb{E}(X) = p$
- $\mathbb{V}(X) = p(1 - p)$

20

Lois usuelles discrètes
Loi binomiale

Définition d'un schéma de Bernoulli : Epreuve de Bernoulli répétée n fois dans des conditions identiques et indépendantes.
Dans ces conditions, la variable aléatoire X égale au nombre de succès est distribuée suivant la loi binomiale $\mathcal{B}(n;p)$.

Propriété : La variable X peut s'écrire

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

où les variables X_i sont des variables de Bernoulli de paramètre p, deux à deux indépendantes.

21

Lois usuelles discrètes
Loi binomiale

Propriété :

- $X(\Omega) = \{0; 1; \cdots ; n\}$
- $\mathbb{P}(X = k) = \left(\begin{array}{c} n \\ k \end{array} \right) p^k (1-p)^{n-k}$

Propriété :

- $\mathbb{E}(X) = np$
- $\mathbb{V}(X) = np(1-p)$

22

IV. Lois usuelles continues

Lois usuelles continues
1. Généralités

Définition : Une fonction f définie sur $\mathbb{R}$ est une densité de probabilité si :

- f est continue presque partout
- f est positive
- $\int_{\mathbb{R}} f = 1$

Propriété :

$\mathbb{P}(a \leq X \leq b) = \int_a^b f(x)\,dx = F_X(b) - F_X(a)$

où F_X est la fonction de répartition associée à X

Remarque : Pour tout réel t, on a : $P(X = t) = 0$

On utilise indifféremment des inégalités strictes ou larges avec des lois continues.

Lois usuelles continues
Généralités

Définition : L'espérance d'une variable aléatoire continue X est définie par

$$\mathbb{E}(X) = \int_{\mathbb{R}} x f(x) dx$$

Définition : La variance d'une variable aléatoire continue X est définie par

$$\mathbb{V}(X) = \mathbb{E}\left([X - \mathbb{E}(X)]\right)^2 = \int_{\mathbb{R}} (x - \mathbb{E}(X))^2 f(x) dx$$

Propriété : $\mathbb{V}(X) = \int_{\mathbb{R}} x^2 f(x) dx - [\mathbb{E}(X)]^2 = \mathbb{E}\left(X^2\right) - [\mathbb{E}(X)]^2$

25

Lois usuelles continues
2. Lois uniformes

Définition :
La loi uniforme sur $[a; b]$ a pour densité de probabilité la fonction f définie par

$$f(x) = \begin{cases} \dfrac{1}{b-a} & \text{si} \quad x \in [a; b] \\ 0 & \text{sinon} \end{cases}$$

Ces lois sont utilisées pour modéliser des temps d'attente (à un arrêt de tramway par exemple).

Propriété : La loi uniforme sur $[a; b]$ a pour espérance

$$\frac{a+b}{2}$$

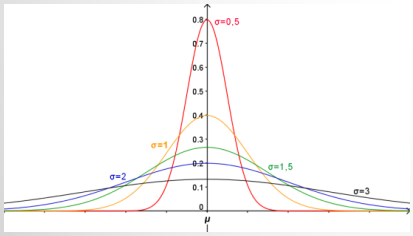
26

Lois usuelles continues
3. Lois normales

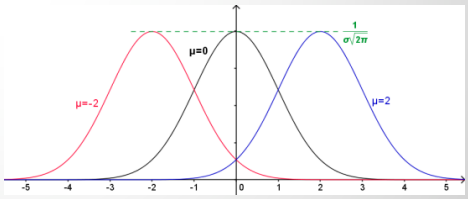
Définition :
La loi normale $\mathcal{N}(\mu;\sigma)$ a pour densité de probabilité la fonction $f_{\mu,\sigma}$ définie par

$$f_{\mu,\sigma}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Cas de différentes valeurs de σ



Cas de différentes valeurs de μ



27

Lois usuelles continues
Lois normales

Propriété : Si $X \hookrightarrow \mathcal{N}(\mu;\sigma)$ alors $\mathbb{E}(X) = \mu$ et $\mathbb{V}(X) = \sigma^2$

Propriété : Soit X est une variable aléatoire continue de densité f.
La fonction de répartition F_X de X caractérise la loi de X et F admet comme dérivée f presque partout.

Théorème : Si $X \hookrightarrow \mathcal{N}(\mu;\sigma)$ alors $X^\star = \dfrac{X-\mu}{\sigma} \hookrightarrow \mathcal{N}(0;1)$.

Définition : La loi $\mathcal{N}(0;1)$ est appelée loi normale centrée réduite.
Sa fonction de répartition se note généralement Φ.

Théorème : La somme de variables normales indépendantes suit une loi normale.

28

Lois usuelles continues
Loi normale centrée réduite

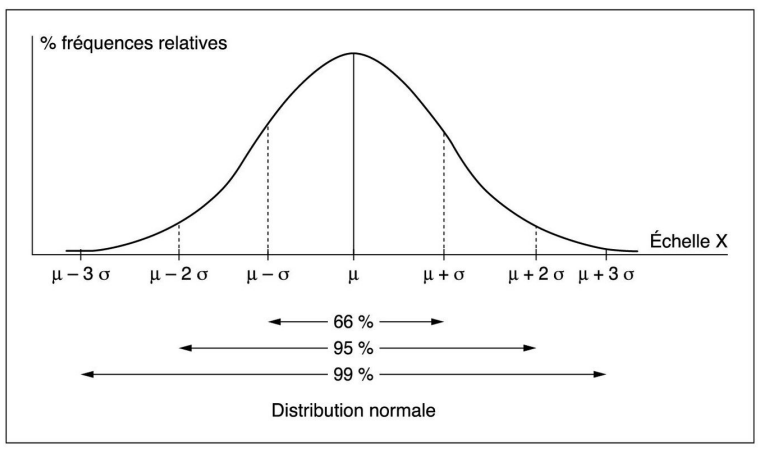
Rappels sur l'utilisation de la fonction de répartition :
$\mathbb{P}(X \leq b) = F_X(b)$
$\mathbb{P}(a \leq X \leq b) = F_X(b) - F_X(a)$

Les fonctions de répartition associées aux lois normales étant strictement croissantes, on peut utiliser :

$$F_X(b) = F_X(a) \Leftrightarrow a = b$$

29

Lois usuelles continues
Lois normales



30

Un critère pour décider de la normalité d'une distribution
Le graphique Quantile-Quantile de normalité (QQ-norm)

Notons x_i les différentes valeurs à considérer et f_i les fréquences cumulées définies par :

$$f_i = \frac{\text{nb val} \leq x_i}{N+1}$$

Soit le quantile u_i^* tel que

$$\mathbb{P}(X^* \leq u_i^*) = f_i \text{ ie } u_i^* = \phi^{-1}(f_i)$$

Si $X \sim \mathcal{N}(\mu, \sigma)$, on doit avoir

$$\mathbb{P}(X \leq x_i) \simeq f_i$$

ie

$$\mathbb{P}\left(X^* \leq \frac{x_i - \mu}{\sigma}\right) \simeq \mathbb{P}(X^* \leq u_i^*)$$

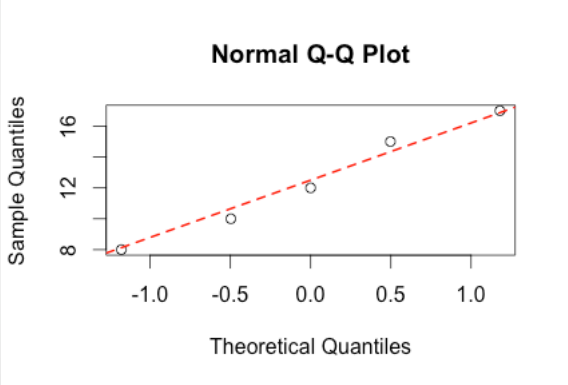
Il en résulte que

$$\frac{x_i - \mu}{\sigma} \simeq u_i^*$$

En conséquence, les points de coordonnées (x_i, u_i^*) doivent être sensiblement alignés.

31

Un critère pour décider de la normalité d'une distribution
Le graphique Quantile-Quantile de normalité (QQ-norm)



32

Applications de la normalité en production : Règle des 6 Sigma & Capabilité

Considérons une fabrication de pièce avec une longeur cible égale à L.
La pièce est utilisable si sa longueur appartient à $[L - \Delta L; L + \Delta L]$.

Si on considère que : $\Delta L = 3\sigma$ alors

$$P\left(X \in [L - \Delta L; L + \Delta L]\right) \simeq 0,9973$$

soit 2700 pièces défectueuses par million.

Si on considère que : $\Delta L = 6\sigma$ alors

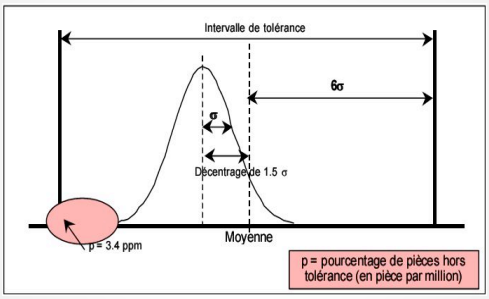
$$P\left(X \notin [L - \Delta L; L + \Delta L]\right) \simeq 2 \times 10^{-9}$$

soit 2 pièces défectueuses par milliard.

33

Règle des 6 Sigma & Capabilité

L'approche Six Sigma tient compte d'une éventuelle déviation de $1,5\sigma$



$$P\left(X \notin [L - 4,5\sigma; L + 7,5\sigma]\right) \simeq 3,4 \times 10^{-6}$$

soit 3,4 pièces défectueuses par million.

34

Règle des 6 Sigma & Capabilité

La capabilité d'un processus associé à un intervalle de tolérance $[T_I; T_S]$ est définie par :

$$C_p = \frac{T_S - T_I}{6\sigma}$$

Dans le cas d'un processus avec une dérive sur la valeur cible, on introduit :

$$C_{pk} = \min\left\{\frac{T_S - \mu}{3\sigma}; \frac{\mu - T_I}{3\sigma}\right\}$$

On considère une machine "capable" lorsque ces coefficients sont supérieurs à **1,33** ...

35

Capabilité

En phase de pré-industrialisation, ces indices font référence à la capabilité Machine. On les note alors :

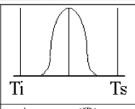
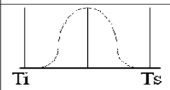
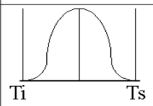
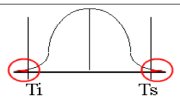
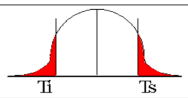
$$C_m \quad \text{et} \quad C_{mk}$$

En phase de production, les indices de capabilité ne reflètent la qualité du procédé que lorsque celui-ci est sous contrôle (stable en moyenne et dispersion).

Fonction sous R :
process.capability( ) via le package qcc

36

Capabilité

1	Cp > 1.67		Plus que suffisant	Non préoccupant, chercher à simplifier la gestion pour réduire les coûts.
2	1.67 > Cp > 1.33		suffisant	Situation idéale. A maintenir.
3	1.33 > Cp > 1.00		trop juste	Nécessite de l'attention, Cp proche de 1 signifie qu'une dérive peut créer des défauts.
4	1.00 > Cp > 0.67		insuffisant	Existence de Non Conformes. il faut contrôler à 100%, analyser le processus et si possible l'améliorer.
5	0.67 > Cp		très insuffisant	Analyse immédiate des causes, urgence de mise en place de contre-mesures, révision des tolérance.

37

V. Echantillonnage & Estimations

38

1. Échantillonnage
1.1 Variables d'échantillonnage

On suppose que les prélèvements sont effectués de manière aléatoire et avec remise ou peuvent être considérés comme tels.

On considère un caractère quantitatif X sur une population d'espérance μ et d'écart-type σ.
On prélève un échantillon aléatoire de taille n et on obtient une suite de valeurs $x_1, x_2, \ldots, x_n$.
On note X_i la variable aléatoire égale à la i-ème réalisation sur un échantillon de taille n.

Définition : Soit $(X_i)_{1\leq i\leq n}$ n variables aléatoires indépendantes de même loi, d'espérance μ et d'écart-type σ.
La variable aléatoire égale à la moyenne est définie par :

$$\overline{X} = \frac{X_1 + X_2 + ... + X_n}{n}$$

39

Échantillonnage
Variables d'échantillonnage

Définition : Soit $(X_i)_{1\leq i\leq n}$ n variables aléatoires indépendantes de même loi, d'espérance μ et d'écart-type σ.

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

Théorème : Soit une population qui contient une proportion p d'individus présentant un caractère qualitatif donné.

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

où les variables X_i sont mutuellement indépendantes et de loi de Bernoulli de paramètre p.

40

Échantillonnage
1.2 Estimation ponctuelle

Définition : Soit θ un paramètre d'un modèle.
Une variable Y est un estimateur sans biais de θ si

$$\mathbb{E}(Y) = \theta$$

Une estimation ponctuelle de ce paramètre θ, notée $\widehat{\theta}$, est alors y, réalisation de la variable Y sur l'échantillon considéré.
On note

$$\widehat{\theta} = y$$

41

Échantillonnage
Estimation ponctuelle

Estimation ponctuelle de la moyenne μ d'une population

Théorème :

$$\mathrm{E}\left(\bar{X}\right) = \mu$$

Conséquences :
- $\bar{X}$ est un estimateur de μ.
- Une estimation ponctuelle de la moyenne μ est donc

$$\hat{\mu} = \bar{x}$$

42

Échantillonnage
Estimation ponctuelle

Estimation ponctuelle de la variance σ^2 d'une population

On rappelle que $S^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$ où les n variables aléatoires $(X_i)_{1\leq i\leq n}$ sont indépendantes, d'espérance μ et d'écart-type σ

Théorème : Soient $(X_i)_{1\leq i\leq n}$ des variables aléatoires indépendantes et identiquement distribuées (i.i.d.) de loi $\mathcal{N}(\mu;\sigma)$. On a :

- $\frac{nS^2}{\sigma^2} \sim \chi^2(n-1)$
- $\mathbb{E}(S^2) = \frac{n-1}{n}\sigma^2$

Ainsi, en posant $\hat{S}^2 = \frac{nS^2}{n-1} = \frac{SCE}{n-1}$, on a :

$$\mathbb{E}\left(\hat{S}^2\right) = \sigma^2$$

43

Échantillonnage
Estimation ponctuelle

Estimation ponctuelle de la variance σ^2 d'une population

Conséquences :

- $\hat{S}^2$ est un estimateur de la variance σ^2.
- Une estimation ponctuelle de la variance σ^2 est donc

$$\hat{s}^2 = \frac{ns^2}{n-1} = \frac{SCE}{n-1}$$

Cette estimation de la variance s'obtient directement sous R avec la fonction *var*( ).

On considère fréquemment qu'une estimation de l'écart-type de la population est $\widehat{\sigma} = \sqrt{\frac{n}{n-1}}\, s$ mais il faut savoir qu'il s'agit d'une estimation biaisée ...

44

Échantillonnage
Estimation ponctuelle

Estimation ponctuelle de la proportion p d'un caractère dans une population

On rappelle que la variable aléatoire égale à la fréquence d'apparition de ce caractère sur un échantillon est définie par

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

où les variables X_i sont mutuellement indépendantes et de loi de Bernoulli de paramètre p.

Une estimation ponctuelle de la proportion dans une population est donnée par

$$\widehat{p} = f$$

où f représente la proportion observée sur l'échantillon.

45

Échantillonnage
1.3 Lois associées aux distributions des moyennes et proportions

Quelles sont les lois associées à ces variables aléatoires ?

Commençons par nous intéresser à la variable des moyennes $\bar{X}$

Attention, le choix du modèle dépend de la situation. Pensez à vous poser les questions suivantes :

- L'écart-type de la population est-il connu ou inconnu ?
- Est-on en présence d'un grand échantillon ?

46

Échantillonnage
1.4 Notions d'estimation

Théorème : Si le caractère X est distribué suivant la loi normale $\mathcal{N}(\mu;\sigma)$ alors

$$\overline{X} \hookrightarrow \mathcal{N}\left(\mu;\frac{\sigma}{\sqrt{n}}\right)$$

Théorème central limite (TCL) :

Si n est grand alors la loi de $\overline{X}$ se rapproche de la loi $\mathcal{N}\left(\mu;\frac{\sigma}{\sqrt{n}}\right)$.

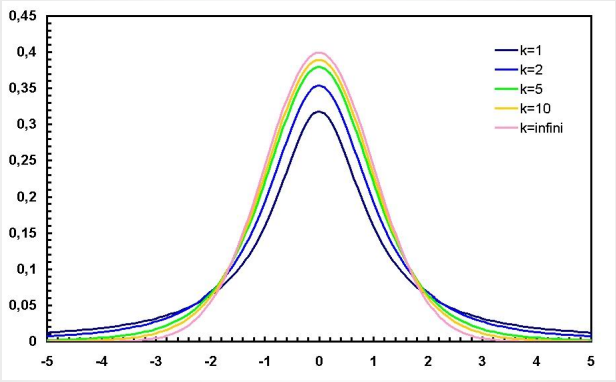
Ce théorème est à utiliser lorsqu'on connait l'écart-type de X et que l'on considère de grands échantillons ($n \geq 30$).

Théorème : (Distribution de Student)
Si la variable X est distribuée normalement alors

$$\frac{\bar{X}-\mu}{\frac{\widehat{S}}{\sqrt{n}}} \hookrightarrow T(n-1)$$

Ce théorème est à utiliser lorsque l'écart-type de la population est inconnu. 47

Échantillonnage
Notions d'estimation



Théorème : La loi de Student $\mathcal{T}(n)$ converge vers la loi normale $\mathcal{N}(0;1)$.

Cette convergence (en loi) n'a de sens que lorsque n est très grand ... 48

Échantillonnage
Notions d'estimation

Théorème : Soit une population qui contient une proportion p d'individus présentant un caractère qualitatif donné.

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

où les variables X_i sont mutuellement indépendantes et de loi de Bernoulli de paramètre p.

Quelle est la loi de F ?

Théorème : (Application du TCL)

Si n est grand, la loi de F se rapproche de la loi $\mathcal{N}\left(p; \sqrt{\frac{p(1-p)}{n}}\right)$

Ce théorème est à utiliser lorsque l'on considère de grands échantillons ($n \geq 30$).

49

Échantillonnage
1.5 Estimation par intervalle de confiance

Nous avons vu précédemment qu'il était possible d'avoir une estimation ponctuelle des paramètres de la population en considérant des observations sur un échantillon.

Cependant, différents échantillons nous donneraient différentes estimations d'un paramètre donné.

L'estimation ponctuelle étant trop liée à l'échantillon choisi, il est souvent intéressant de déterminer des intervalles appelés intervalles de confiance.

50

Échantillonnage
Estimation par intervalle de confiance

Estimer un paramètre θ par intervalle, au niveau de confiance $1-\alpha \in]0;1[$, à partir d'un estimateur Y revient à déterminer deux variables aléatoires θ_{min} et θ_{max} tels que

$$\mathbb{P}\left(\theta_{min} \leq \theta \leq \theta_{max}\right) = 1-\alpha$$

Dans le cas où Y est gaussienne, chercher un intervalle de confiance revient à déterminer un réel positif η tel que

$$\mathbb{P}\left(Y-\eta \leq \theta \leq Y+\eta\right) = 1-\alpha$$

L'intervalle $[y-\eta; y+\eta]$ est une estimation de θ par intervalle de confiance au niveau $1-\alpha$.

Déterminons une estimation par intervalle de confiance de la moyenne μ lorsque la distribution dans la population est normale, d'espérance μ et d'écart-type σ connu.

51

Échantillonnage
Estimation par intervalle de confiance

Théorème :
On considère un estimateur Y tel que

$$Y \sim \mathcal{N}(mean, sd)$$

Une estimation par intervalle de confiance au niveau $(1-\alpha)$ de $\mathbb{E}(Y) = mean$ est donnée par

$$I_{1-\alpha} = \left[y - u_{1-\frac{\alpha}{2}} \times sd; y + u_{1-\frac{\alpha}{2}} \times sd\right]$$

où y est la réalisation de Y sur l'échantillon.

52

Échantillonnage
Estimation par intervalle de confiance

Une estimation par intervalle de confiance au niveau $0,95$ est donnée par :

$$I_{0,95} = \left[\overline{x} - 1,96\frac{\sigma}{\sqrt{n}}; \overline{x} + 1,96\frac{\sigma}{\sqrt{n}}\right]$$

Il est très vraisemblable que la moyenne de la population μ appartienne à cet intervalle ...

Dans le cas où l'écart-type σ est inconnu, sous réserve de normalité de la distribution, on est amené à utiliser la même démarche que précédemment avec

$$\frac{\bar{X} - \mu}{\frac{\hat{S}}{\sqrt{n}}} \hookrightarrow \mathcal{T}(n-1)$$

Avec R, on peut utiliser : t.test( )$conf.int

53

VI. Cartes de contrôle

Principe

Graphique sur lequel on reporte, dans l'ordre chronologique, les valeurs d'une statistique calculée sur des échantillons, en général de même effectif, issus de la fabrication.

Une telle carte comporte :

- une ligne centrale (valeur cible)
- des limites de surveillance et de contrôle

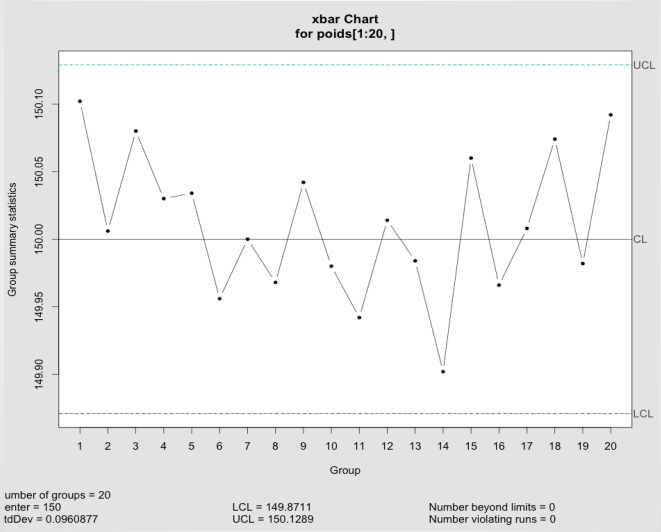
Les points placés ont pour :

- Abscisse : Numéro de l'échantillon
- Ordonnée : Valeur de la statistique calculée sur cet échantillon

54

Cartes de contrôle

Carte de contrôle de Shewart sur la moyenne



55

Cartes de contrôle

Principe pour la carte de contrôle de Shewart de la moyenne

La valeur cible correspond à la vraie valeur du paramètre (ou son estimation ponctuelle).

Les limites de contrôle inférieures (LIC ou LCL en anglais) et supérieures (LSC ou UCL en anglais) sont situées à 3 écarts-types de la cible.

$$\text{LCL} = \text{LIC} = \mu - 3\frac{\sigma}{\sqrt{n}} \; ; \; \text{UCL} = \text{LSC} = \mu + 3\frac{\sigma}{\sqrt{n}}.$$

Étant donné que $\overline{X}$ est distribuée suivant la loi normale $\mathcal{N}\left(\mu; \frac{\sigma}{\sqrt{n}}\right)$, on a

$$P\left(\mu - 3\frac{\sigma}{\sqrt{n}} \leq \overline{X} \leq \mu + 3\frac{\sigma}{\sqrt{n}}\right) \simeq 0,9973$$

Ainsi : la plupart des moyennes observées sont comprises entre ces 2 limites.

56

Cartes de contrôle

Principe pour la carte de contrôle de Shewart de la moyenne

On peut rajouter des limites de surveillance situées à 2 écarts-types de la cible.

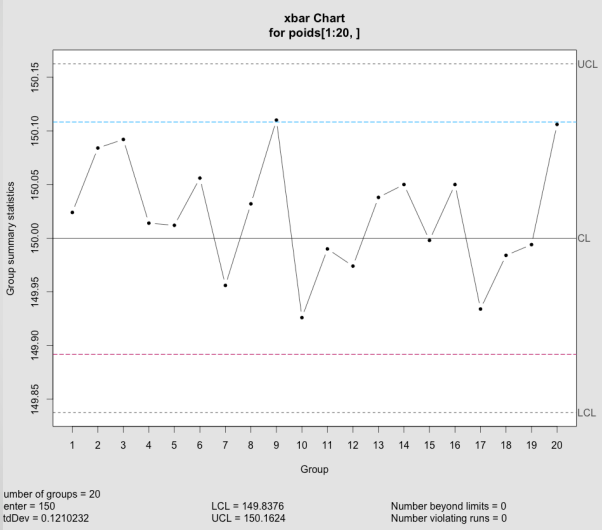
Étant donné que $\overline{X} \sim \mathcal{N}\left(\mu; \frac{\sigma}{\sqrt{n}}\right)$, on a

$$\mathbb{P}\left(\mu - 2\frac{\sigma}{\sqrt{n}} \leq \overline{X} \leq \mu + 2\frac{\sigma}{\sqrt{n}}\right) \simeq 0,954$$

Une observation comprise entre les limites de surveillance et de contrôle (inf ou sup) doit inciter à une attention particulière (prélèvement d'un nouvel échantillon par exemple).

57

Cartes de contrôle



Carte de contrôle de la moyenne avec limites de contrôle et surveillance

Avec R :

Package qcc

qcc(valeurs, type="xbar")

58

Cartes de contrôle

Étude et utilisation de la carte de contrôle de Shewart de la moyenne
Norme NF 06-031

1. Vérifier que la distribution des valeurs est normale.

2. Estimation de l'écart-type de la production en utilisant les variances ou étendues (voir diapo suivante).

3. Calculs des limites de contrôle (et surveillance).

4. Suivi de la production
   On peut considérer qu'il y a un déréglage lorsque :
   - Point à l'extérieur des limites de contrôle ;
   - 9 points consécutifs du même côté de la ligne centrale ;
   - 6 points consécutifs tous « ascendants » ou « descendants ».

59

Cartes de contrôle

Estimation de l'écart-type par les étendues

1. Prélever plusieurs échantillons de même taille n (issus d'une population gaussienne).

2. Calculer pour chaque échantillon l'étendue R_i ainsi que l'étendue moyenne $\bar{R}$.

3. Déterminer d_2 en fonction de n.
   On a alors
   $$\widehat{\sigma} = \frac{\bar{R}}{d_2}$$

Table 1: Control Chart Coefficients

Subgroup Size n	d_2	D_1	D_2	D_3	D_4	A_2
2	1.128	0	3.686	0	3.267	1.880
3	1.693	0	4.358	0	2.575	1.023
4	2.059	0	4.698	0	2.282	0.729
5	2.326	0	4.918	0	2.115	0.577
6	2.534	0	5.078	0	2.004	0.483
7	2.704	0.205	5.203	0.076	1.924	0.419
8	2.847	0.387	5.307	0.136	1.864	0.373
9	2.970	0.546	5.394	0.184	1.816	0.337
10	3.078	0.687	5.469	0.223	1.777	0.308
11	3.173	0.812	5.534	0.256	1.744	0.285
12	3.258	0.924	5.592	0.284	1.716	0.266
13	3.336	1.026	5.646	0.308	1.692	0.249
14	3.407	1.121	5.693	0.329	1.671	0.235
15	3.472	1.207	5.737	0.348	1.652	0.223
20	3.735	1.548	5.922	0.414	1.586	0.180
25	3.931	1.804	6.058	0.459	1.541	0.153

60

Cartes de contrôle

Pour compléter l'étude sur les moyennes : Étude de la variabilité

1. Carte de contrôle de l'étendue (qcc(valeurs, type= "R"))

$$\text{Carte } (\bar{x}, R)$$

1. Carte de contrôle de l'écart-type (qcc(valeurs, type= "S"))

$$\text{Carte } (\bar{x}, s)$$

Pour contrôler des proportions ou nombre de défectueux (qualitatif) : Carte de contrôle aux attributs

1. Carte de contrôle de la proportion (p) (qcc(valeurs, type= »"))
2. Carte de contrôle du nombre de défectueux (np) (qcc(valeurs, type= "np"))
3. Carte de contrôle du nombre de défauts (c) (qcc(valeurs, type= "c"))

61

VII. LES TESTS STATISTIQUES

OBJECTIFS :
A partir d'un ou plusieurs échantillon(s), on souhaite prendre des décisions concernant les populations dont sont extraits les échantillons.

Exemples : QN, Proportion de défectueux, ...

62

1. Généralités sur les tests d'hypothèse
Une aide à la prise de décision

Exemples

1. Afin de tester une solution toxique, on fait des injections à un groupe de 80 souris.
   On fait l'hypothèse que l'injection est mortelle dans 80% des cas.
   58 souris sont mortes.
   Cette étude remet-elle en cause l'hypothèse initiale ?

2. On lance 10 fois une pièce. On obtient 10 fois Pile.
   La probabilité d'obtenir un Pile est-elle supérieure à 50 % ?
   Quelle est la conclusion avec 9 Pile ? 8 Pile ? 7 Pile ?

63

Généralités sur les tests d'hypothèse
Hypothèse nulle et alternative

La première étape d'un test consiste à déterminer les hypothèses à tester.

Le contexte donne fréquemment $\mathcal{H}_1$.
Cette hypothèse est appelée l'hypothèse alternative et se présente souvent sous la forme :
soit " $\ldots \neq \ldots$ " ; on dit que le test est bilatéral.
soit " $\ldots > \ldots$ " ou " $\ldots < \ldots$ " ; on dit alors que le test est unilatéral.

La "première" hypothèse, notée $\mathcal{H}_0$, est appelée hypothèse nulle dans le sens "to be nullified" c'est-à-dire "à réfuter".
Il s'agit, en général, d'une égalité.

64

Généralités sur les tests d'hypothèse
Autour de la p-value

La p-valeur (p-value en anglais) est parfois nommée probabilité critique.

Elle correspond à la probabilité que la statistique de test soit égale à la valeur observée ou encore plus extrême, tout en supposant que l'hypothèse nulle $\mathcal{H}_0$ est vraie.
Elle correspond à la **probabilité d'avoir une observation égale ou "pire" que celle que l'on a obtenue.**
Dans le cas d'un test bilatéral, la p-value correspond au double de celle qui aurait été obtenue lors de la mise en place d'un test unilatéral.

La p-valeur est une probabilité calculée a posteriori, en fonction des données.

65

Généralités sur les tests d'hypothèse
Méthodologie d'un test (Fisher-Neyman-Pearson)

• Formuler les hypothèses $\mathcal{H}_0$ (hypothèse nulle) et $\mathcal{H}_1$ (hypothèse alternative).

• Préciser la loi de probabilité de la variable de décision (statistique de test) en justifiant précisément son utilisation.
(Normalité des distributions, indépendance, ...)

• Calcul de la p-valeur sous $\mathcal{H}_0$.
Cette valeur correspond à la probabilité d'avoir une observation égale ou pire que celle que l'on a obtenue.

• Conclusion en comparant la p-valeur au seuil de signification (usuellement 5 %).
On rejette $\mathcal{H}_0$ si la p-valeur est inférieure au seuil de signification.

66

Généralités sur les tests d'hypothèse
Risques associés

1er cas : Si on rejette une hypothèse vraie, on commet une erreur de première espèce. Le risque associé est noté α.
On l'appelle aussi risque du vendeur.

2nd cas : Si on "accepte" une hypothèse fausse, on commet une erreur de second espèce. Le risque associé est noté β.
On l'appelle aussi risque de l'acheteur.

REALITE / DECISION	H_0 VRAIE	H_0 FAUSSE
H_0 REJETEE (H_1 VRAIE)	α	$1-\beta$
H_0 ACEPTEE (H_1 FAUSSE)	$1-\alpha$	β

Lorsqu'on met en place un test, il faut se fixer un seuil de signification α.
On le prend généralement égal à 5%.

67

Généralités sur les tests d'hypothèse

Comment utiliser les tests ?

Se limiter à un simple calcul de p-value pour décider d'un éventuel effet est trop réducteur et source de beaucoup d'erreurs.
Par exemple, supposons que sur 1000 tests, seuls 10 % aient un effet réel. On fixe $\alpha = 5\%$ et $\beta = 20\%$ soit $1-\beta = 80\%$.
Ainsi, $80 + 45 = 125$ tests s'avèreront positifs (rejet de $\mathcal{H}_0$) et seulement 80 seront de "vrais" positifs. La probabilité, lorsqu'on a déclaré un effet positif (rejet de $\mathcal{H}_0$) qu'il n'y ait pas d'effet est

$$\frac{45}{125} \simeq 0,36$$

Près de $\frac{1}{3}$ des effets annoncés comme significatifs ne le sont pas ... Il est donc important d'accompagner un test de mesures ou analyses telles que :

- Calcul de la puissance d'un test (ajustement éventuel de la taille d'échantillon)
- Détermination d'intervalles de confiance
- Calculs de la taille de l'effet.

68

Généralités sur les tests d'hypothèse

Test et puissance

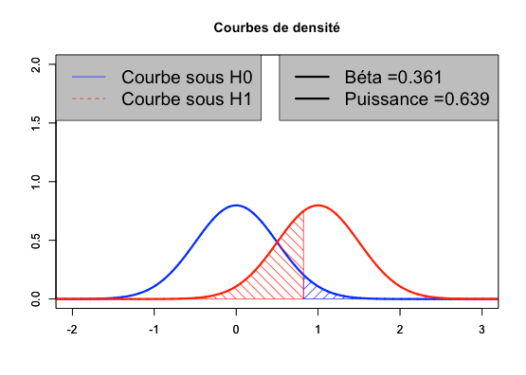
La puissance d'un test correspond à la probabilité de mettre en évidence un effet lorsque celui existe.
Il s'agit donc de la capacité d'un test à prendre la bonne décision lorsqu'il existe un effet.

1. Pour un même risque α et une même taille d'échantillon, on constate que, si *l'écart* Δ entre la valeur du paramètre posée en $\mathcal{H}_0$ et celle supposée dans l'hypothèse vraie $\mathcal{H}_1$ augmente, le risque β diminue ce qui induit une augmentation de $1 - \beta$.
2. Une diminution de la *variabilité* (cf. écart-type) peut également induire une augmentation de la puissance du test.
3. Enfin, une *augmentation de la taille du ou des échantillons* aura pour effet de donner une meilleure précision.
   Le test est alors plus puissant.

Il est recommandé de choisir une taille d'échantillon conduisant, a priori, à une **puissance de test au moins égale à 80 %**.
Avec R, on peut utiliser des fonctions implémentées de base ('power.t.test' ou 'power.anova.test') ou accessibles via le package 'pwr'.

69

Tests d'hypothèse



70

Tests d'hypothèse

Taille de l'effet

Pour un grand échantillon, un effet minime (et sans réelle signification) suffira à faire rejeter l'hypothèse $\mathcal{H}_0$.
On peut donc introduire la notion de *taille de l'effet* désignant à quel degré un phénomène donné est présent dans la population (*Revue des sciences de l'éducation, volume 31, 2009*).
Ainsi, **ne pas rejeter l'hypothèse nulle revient à considérer que la taille de l'effet est nulle**.

Soit Δ l'écart entre la moyenne de la population et une valeur cible, ou entre les moyennes de deux populations.
La taille de l'effet est déterminé par

$$\text{Taille de l'effet} = \frac{\Delta}{\sigma}$$

où σ correspond à l'écart-type des populations.
On peut utiliser des niveaux définis par Cohen pour qualifier un effet.

- $d = 0.2$: Effet faible
- $d = 0.5$: Effet moyen
- $d = 0.8$: Effet fort

71

2. Test du Khi-deux d'homogénéité & d'indépendance

Exemple :
Trois produits sont évalués par des consommateurs qui doivent répondre à la question suivante :
Achèteriez-vous ce produit ?
Les résultats obtenus sont les suivants :

	P1	P2	P3
OUI	40	50	60
NON	60	50	40

L'appréciation est-elle liée au produit ?

72

Déterminons les effectifs "théoriques" si l'appréciation est indépendante du produit.
On rappelle que si A et B sont indépendants alors $\mathbb{P}(A \cap B) = \mathbb{P}(A).\mathbb{P}(B)$.

73

$\mathcal{H}_0$: Les caractères étudiés sont indépendants
$\mathcal{H}_1$: Les caractères étudiés ne sont pas indépendants

On suppose que les deux caractères prennent respectivement p et q modalités.

Sous $\mathcal{H}_0$, la variable et la loi de décision sont les suivantes :

$$\sum_{i,j} \frac{\left(N_{ij} - N_{ij}^{th}\right)^2}{N_{ij}^{th}} \text{ suit la loi } \chi^2\left((p-1)\times(q-1)\right)$$

N_{ij} et N_{ij}^{th} correspondent respectivement aux effectifs observés et théoriques (cf. indépendance).
Ce test est par nature **unilatéral**.

Contrainte : les effectifs théoriques doivent être supérieurs à 5.
Dans le cas contraire, on doit effectuer des regroupements de classes ...

74

Mise en place du test

Avec R (Voir doc Begin'R)

```
Tableau = matrix(c(40, 50, 60, 60, 50, 40), nrow=2, ncol=3, byrow=TRUE)
colnames(Tableau)=c("P1","P2","P3")
rownames(Tableau) = c("OUI","NON")
Tableau
chisq.test(Tableau)
```

Pearson's Chi-squared test

data: Tableau
X-squared = 8, df = 2, p-value = 0.01832

75

3. Tests de conformité

On se propose de tester l'hypothèse selon laquelle dans une population, il y a une moyenne ou une proportion égale à une valeur donnée (par ex : QN).

On a donc $H_0 : \mu = \cdots$ ou $H_0 : p = \cdots$

Cas d'un test de conformité d'une moyenne :

Premier cas : Ecart-type de la population connu

Théorème : Si le caractère X est distribué suivant la loi normale $\mathcal{N}(\mu;\sigma)$ alors

$$\overline{X} \hookrightarrow \mathcal{N}\left(\mu; \frac{\sigma}{\sqrt{n}}\right)$$

Théorème central limite (TCL) :

Si n est grand alors la loi de $\overline{X}$ se rapproche de la loi $\mathcal{N}\left(\mu; \frac{\sigma}{\sqrt{n}}\right)$.

Ce théorème est à utiliser lorsqu'on connait l'écart-type de X et que l'on considère de grands échantillons ($n \geq 30$).

76

Second cas : Ecart-type de la population inconnu

Théorème : (Distribution de Student)
Si la variable X est distribuée normalement alors

$$\frac{\bar{X}-\mu}{\frac{\widehat{S}}{\sqrt{n}}} \hookrightarrow T(n-1)$$

Mise en place du test avec R :
t.test(x=… , mu= …. , alternative="less »)

Sous réserve de normalité de la distribution !!!

77

Cas d'un test de conformité d'une proportion

Théorème : (Application du TCL)
Si n est grand, la loi de F se rapproche de la loi $\mathcal{N}\left(p; \sqrt{\frac{p(1-p)}{n}}\right)$

Ce théorème est à utiliser lorsque l'on considère de grands échantillons ($n \geq 30$).

Mise en place du test avec R :
binom.test(x= … , n=… , p=… , alternative = …..)

Dans le cas d'un grand échantillon :
prop.test(x=…, n=….,p=….,alternative=…. , correct=FALSE)

78

4. Tests de comparaisons

On se propose de tester l'hypothèse selon laquelle dans deux populations P_1 et P_2 , il y a la même moyenne (μ inconnue) pour la variable étudiée à partir d'échantillons extraits de tailles respectives n_1 et n_2.

On a donc l'hypothèse nulle suivante :
Ho : « $\mu_1 = \mu_2$ »

Plusieurs cas peuvent se présenter
(il faut également s'intéresser à la taille de l'échantillon) :
- les **variances** sont **connues**
- les **variances** sont **inconnues**
- les **échantillons** sont **appariés** (pas indépendants)

79

Tests de COMPARAISON de 2 MOYENNES
Cas de distributions normales

• Cas où les échantillons sont **indépendants** et les **variances connues** :

$$\bar{X}_1 - \bar{X}_2 \sim \mathcal{N}\left(\mu_1 - \mu_2; \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$$

• Cas où les échantillons sont **indépendants** et les **variances inconnues mais** supposées **égales** *(hypothèse d'homoscédasticité)* :

$$\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{n_1 S_1^2 + n_2 S_2^2}{n_1 + n_2 - 2} \times \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim \mathcal{T}\,(n_1 + n_2 - 2)$$

$\frac{SCE_1 + SCE_2}{n_1 + n_2 - 2} = \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}$ est une **estimation de la variance commune**

Avec R :
t.test(x= , y = , var.equal = TRUE, …)

80

Tests de COMPARAISON de 2 MOYENNES

• Cas où les échantillons sont **indépendants** et les **variances inconnues mais différentes** :

Utilisation du test de Welch (adaptation d'un test de Student)
t.test(x= , y = , var.equal = FALSE, ...)

• Cas où les échantillons ne sont **pas indépendants** (échantillons appariés) avec $n_1=n_2=n$:
On raisonne sur la distribution des différences entre les mesures et on utilise que :

$$\frac{\bar{X}-\mu}{\frac{\hat{S}}{\sqrt{n}}} \hookrightarrow T(n-1)$$

Mise en place du test avec R :
t.test(x= , y = , paired = TRUE, …)
Sous réserve de *normalité de la distribution des différences* !

81

Tests de COMPARAISON de 2 PROPORTIONS

On se propose de tester l'hypothèse selon laquelle dans deux populations, il y a la même proportion (p inconnue) pour la variable mesurée.

On a donc l'hypothèse nulle suivante :
Ho : « $p_1 = p_2$ »
Nous ne considérerons que le cas de **grands échantillons**.

On a, sous H_0 :

$$F_1 - F_2 \sim \mathcal{N}\left(0; \sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}\right)$$

où $\hat{p}$est une estimation de la proportion commune.

Avec R : prop.test(x= , y= , …)

82

Tests de COMPARAISON de 2 VARIANCES

On se propose de tester l'hypothèse selon laquelle dans deux populations **gaussiennes indépendantes**, il y a la même variance pour la variable mesurée.

$$H_0 : {\sigma_1}^2 = {\sigma_2}^2$$

On utilise un test de Fisher avec, sous H_0 :

$$\frac{{\widehat{S}_1}^2}{{\widehat{S}_2}^2} \hookrightarrow F\left(n_1 - 1; n_2 - 1\right)$$

Mise en place du test avec R :
var.test(x= , y = , …)

83

VIII.ANOVA à 1 facteur

84

L'ANOVA (Analysis of variance) :
Modèle linéaire gaussien dans lequel toutes les variables explicatives sont qualitatives.
Elles sont appelées **facteurs** (cf. plans factoriels) et leurs modalités sont appelées **niveaux** (contrôlés ou provoqués).
La variable aléatoire réponse est toujours **quantitative** et supposée **gaussienne**.

Exemple : (Effet d'une molécule activateur d'enzyme)
- 5 groupes de 10 souris.
- Doses : 1ng, 10 ng, 50 ng et 100 ng. Le 5e groupe ne reçoit que le solvant utilisé.

85

Pourquoi faire une ANOVA ?

On considère un facteur prenant p niveaux de moyennes $\mu_1, \cdots, \mu_p$.
On teste l'hypothèse selon laquelle le facteur étudié n'a pas d'effet.
Ainsi, on a
$\mathcal{H}_0 : \mu_1 = \cdots = \mu_p$
$\mathcal{H}_1$: "deux moyennes au moins sont différentes"

→ Dans le cas de 4 modalités :
6 tests de Student a priori à mettre en place

Risque alpha au global trop élevé

86

Répétition de tests et Risque Alpha

Calcul du risque α_{global} lorsque l'on effectue 6 tests :
On considère que $\mathcal{H}_0$ est vraie c'est-à-dire qu'il n'y a pas d'effet significatif du facteur étudié (toutes les moyennes sont égales).

$1-\alpha_{global}$ correspond à la probabilité qu'aucun des 6 tests ne mette en évidence (à tort) un effet significatif.
Avec un seuil de signification de 5 %, pour 6 tests indépendants, on a :

$$1 - \alpha_{global} = 0,95^6$$

$$\alpha_{global} \simeq 26,5\%$$

On ne peut donc envisager de réaliser plusieurs tests de Student ...

Plus généralement, pour n tests, avec un risque α, on a

$$\alpha_{global} = 1 - (1-\alpha)^n$$

87

Exemple : L'acide citrique a-t-il une influence sur le brunissement de pâtes alimentaires ?

	Doses d'acide citrique		
	5 ppm	**10 ppm**	**20 ppm**
Indices de brun	**25,2**	**22,1**	**18,4**
	24,3	**23,8**	**19,5**
	26,8	**21,9**	**18,9**
	25,9	**22,6**	**19,9**

Facteur étudié : Quantité d'acide citrique
Nombre de niveaux : 3
Plan équilibré (même nombre de répétitions)

88

Hypothèses

- Indépendances des *p* populations dont sont extraits les échantillons (cf. niveaux)
- **Homoscédasticité** et **Distributions gaussiennes** vérifiées sur les **résidus** usuellement

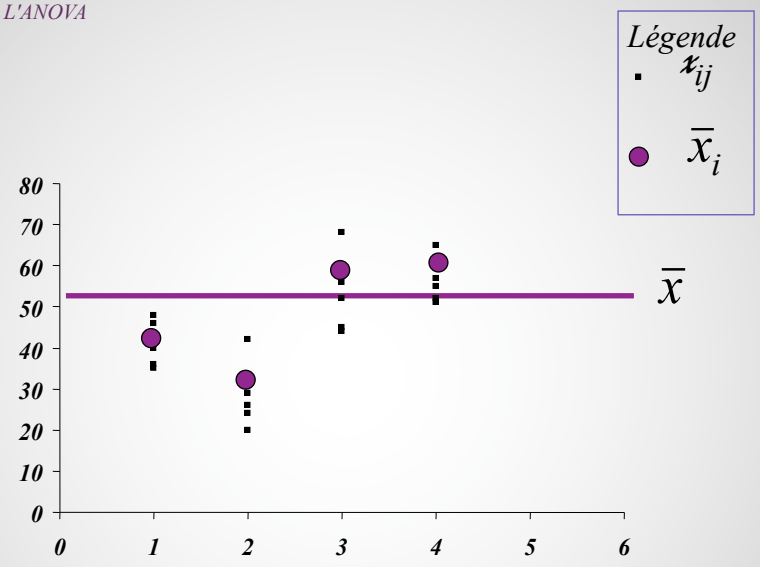
Formulation des hypothèses

H_0 : « $\mu_1 = \mu_2 = ... = \mu_p$ »

ie « il n'y a pas d'effet du facteur étudié »

H_1 : « Au moins deux moyennes sont différentes »

ie « il y a un effet du facteur étudié »



$$x_{ij} - \bar{x} = (x_{ij} - \bar{x}_i) + (\bar{x}_i - \bar{x}) \text{ avec } e_{ij} = x_{ij} - \bar{x}_i$$

donc $x_{ij} = \bar{x} + (\bar{x}_i - \bar{x}) + e_{ij}$

Observation

Moyenne globale

Part expliquée par l'effet du traitement

Résidu e_{ij}

Les résidus $e_{ij} = x_{ij} - \bar{x}_i$, par niveau (et donc au global), sont de somme et de moyenne nulle.

MODELE LINEAIRE

$$X_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

où

- μ est la moyenne (globale) associée aux p populations.
- X_{ij} est une variable aléatoire (réponse quantitative). Les X_{ij} sont indépendantes.
- ε_{ij} est une variable aléatoire d'erreur (non observées). On a : $\varepsilon_{ij} \sim \mathcal{N}(0; \sigma)$, σ étant un paramètre inconnu (à estimer).
- α_i correspond à l'effet associé à la modalité i avec $\sum_{i=1}^{p} \alpha_i = 0$.

  Une estimation de cet effet est donnée par

$$\hat{\alpha}_i = \bar{x}_i - \bar{x}$$

Le modèle peut également s'écrire

$$X_{ij} = \mu_i + \varepsilon_{ij}$$

où $\mu_i = \mu + \alpha_i$ dont une estimation est $\hat{\mu}_i = \bar{x}_i$.

En posant

- $SCE_{fact} = SCE_{inter} = \sum_{i,j} (\bar{x}_i - \bar{x})^2$
- $SCE_{res} = SCE_{intra} = \sum_{i,j} (x_{ij} - \bar{x}_i)^2$
- $SCE_{totale} = \sum_{i,j} (x_{ij} - \bar{x})^2$.

On a

$$SCE_{totale} = SCE_{fact} + SCE_{res}$$

Source	Df	SCE	CM	F_{obs}	p-valeur
Factorielle	$p-1$	SCE_{fact}	$CM_{fact} = \frac{SCE_{fact}}{p-1}$	f_{obs}	$\mathbb{P}(F_{obs} > f_{obs})$
Résiduelle	$N-p$	SCE_{res}	$CM_{res} = \frac{SCE_{res}}{N-p}$		
Totale	$N-1$	SCE_{totale}			

pour p modalités, N observations. Sous $\mathcal{H}_0$,

$$F_{obs} = \frac{CM_{fact}}{CM_{res}} \sim \mathcal{F}(p-1; N-p)$$

93

Conditions requises pour utiliser l'ANOVA

Similaires à un test de Student

Egalité des variances (estimation d'une variance commune par le **CMres**)

Indépendance

Normalité des distributions

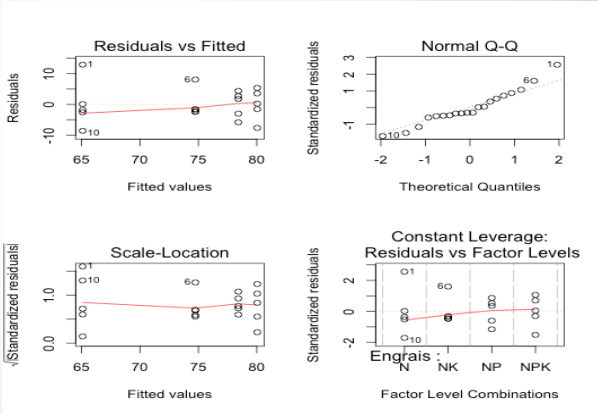
Vérifications sur les résidus (cf. Modèles)

94

Fonction plot sous R

Obtention de 4 graphiques permettant d'obtenir des informations sur :

- ✓ l'homoscédasticité (dispersion des résidus),
- ✓ la normalité (QQ-plot),
- ✓ l'indépendance des résidus.



Avec R :
anova = aov(var ~ facteur)
par(mfrow = c(2,2))
plot(anova)

Pour le modèle linéaire d'ANOVA, les **résidus** doivent être **indépendants** de loi **normale de moyenne nulle**

Egalité des variances

Test de Bartlett
Test de Levene
Méthodes graphiques
...

Indépendance

Méthodes graphiques

Normalité

Test sur coefficients
Test du Khi²
Test de Kolmogorov-Smirnoff
Test de Shapiro-Wilks
Méthodes graphiques

Hypothèse d'homoscédasticité
Comparaison des variances : Test de Bartlett

$$H_0 : \sigma_1^2 = \sigma_1^2 = \ldots = \sigma_p^2$$

$$\chi_0^2 = \frac{(N-p)\ln CM_R - \sum_{i=1}^{p}\left[(n_i - 1)\ln \hat{\sigma}_i^2\right]}{1 + \frac{1}{3(p-1)}\left[\sum_i \frac{1}{n_i - 1} - \frac{1}{N-p}\right]} \sim \chi^2(p-1)$$

Code R :
bartlett.test(Variable ~ Facteur)

97

Principe de l'Anova

Si f_0 est proche de 0 alors la variabilité factorielle est inférieure à la variabilité résiduelle.
Il n'y a donc pas d'effet factoriel significatif puisque la variabilité totale est due essentiellement à la variabilité résiduelle.

On rejette $\mathcal{H}_0$ si le bruit factoriel est bien plus grand que le bruit résiduel i.e.

$$f_0 >> 1$$

Ce test est donc, par nature, unilatéral.

Si on rejette $\mathcal{H}_0$ alors on peut dire que le facteur étudié a un effet significatif. Il est alors intéressant de comparer ces différentes moyennes à l'aide d'un test post hoc (par exemple, test de Bonferroni ou de Newman-Keuls).

98

Exemple d'ANOVA

ANALYSE DE VARIANCE

Source des variations	*Somme des carrés*	*Degré de liberté*	*Moyenne des carrés*	*F*	*Prob*
Entre Groupes	81,43				
A l'intérieur des groupes	6,8575				
Total					

99

ANALYSE DE VARIANCE

Source des variations	*Somme des carrés*	*Degré de liberté*	*Moyenne des carrés*	*F*	*Prob*
Entre Groupes	81,43	2	40,72	53,44	1E-05
A l'intérieur des groupes	6,8575	9	0,76		
Total	88,29	11			

Avec R :
anova = aov(var ~ facteur)
summary(anova)

100

Test post hoc : Comparaison de moyennes 2 à 2

Ces tests sont à réaliser lorsque l'on a mis en évidence un effet du facteur étudié.

Le test de Bonferroni est basé sur un test de Student de comparaison de deux moyennes.
Présentation dans le cas d'échantillons de même taille n :

Le carré moyen résiduel CMres est un bon estimateur de la variance commune.
Sous l'hypothèse d'égalité des moyennes :

$$T_0 = \frac{\bar{X}_i - \bar{X}_j}{\sqrt{CM_{res}}\sqrt{\frac{1}{n_i} + \frac{1}{n_j}}} = \frac{\bar{X}_i - \bar{X}_j}{\sqrt{\frac{2CM_{res}}{n}}} \hookrightarrow T(N-p)$$

101

Pour chaque comparaison, on effectue un test avec un seuil de signification égal à

$$\frac{\alpha}{\text{nombre de tests}} = \frac{\alpha}{\binom{p}{2}}$$

où $\alpha = 0,05$ (usuellement) et $\binom{p}{2} = \frac{p(p-1)}{2}$.

Avec R :
pairwise.t.test(variable, facteur, p.adj = "bonferroni")
LSD.test( ) via le package agricolae

Un test plus puissant (surtout si plan équilibré) basé sur les plus petites amplitudes significatives(ppas) :
Test de Student-Newman-Keuls
Avec R :
SNK.test( ) via le package agricolae

102

IX. Tests non paramétriques

À utiliser si les conditions d'applications des tests ne sont pas vérifiées ?

Pour comparer deux moyennes :
Tests des rangs de Wilcoxon
wilcox.test( )

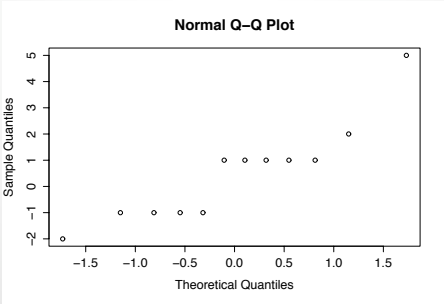
Pour comparer plus de 2 moyennes :
Test des rangs de Kruskal-Wallis
kruskal.test( )

103

Exemple avec le test des signes

Notes attribuées par 12 personnes d'un jury sur 2 produits

Produit 1	6	6	7	6	7	8	5	6	8	8	8	8
Produit 2	5	7	6	5	8	9	0	7	6	7	10	7



104

Exemple avec le test des signes

On va considérer les différences entres les notes des deux produits.
$\mathcal{H}_0$: "les notes sont similaires"

Sous $\mathcal{H}_0$, le nombre X de différences positives (ou négatives) est tel que

$$X \sim \mathcal{B}(n; 0,5)$$

Le test des signes se ramène à un test binomial.

Ici, on observe 7 différences positives pour tester l'hypothèse selon laquelle le produit 1 serait "meilleur".

- binom.test(x= …. , n= …. , p= …. , alternative= "greater")

Exact binomial test
data: 7 and 12
number of successes = 2, number of trials = 12,
p-value = 0.381

- **Avec le test de Wilcoxon**

wilcox.test(x, y, paired=…….. , alternative = "greater")
data: x and y
p-value = 0.255