Cours Statistiques et Analyse de Données

Cours 1

FILIÈRE ISA - UV1

R. Billot (le bon, la brute) et G. Coppin (le truand)

2018 - 2019


Selon les derniers sondages, 47% des statistiques sont fausses.
G&W

Une petite blague

Un statisticien et un biologiste sont condamnés à mort. On leur accorde une dernière faveur.

> *- Je voudrais donner une grande conférence sur la statistique devant tout le monde, dit le statisticien.*
>
> *- Accordé, dit le juge*
>
> *Le biologiste n'exprime aucune hésitation :*
>
> *- Je souhaiterais être exécuté en premier.*

extrait de "Comprendre et réaliser les tests statistiques à l'aide de R", Gael Millot

Une question d'argent

Vous êtes sur le point d'être embauché(e) dans la société Macheprot. Marcel-Benoit Schblurb, de la promo précédente, vous indique que le salaire moyen des embauchés est de 35 keuros *. Après petite enquête, vous obtenez les données suivantes:

Emb1	Emb2	Emb3	Emb4	Emb5
34.5	36	35.2	33	34.3

Ce satané Marcel-Benoit vous roule-t-il une fois de plus dans la farine ?

* net !

Une question d'âge (mental)?

Les 50 élèves de la filière 3 ont une moyenne d'âge de 24,6 et une variance de 0,4 ans alors que les 22 élèves de la filière 1 ont une moyenne d'âge de 25,1 et une variance de 0,3 ans.

La filière ISA fait-elle du jeunisme ?

Exemple développé : effet de la lune sur les naissances

On souhaite étudier les effets de la lune sur les naissances (plus précisément l'effet supposé de la pleine lune sur l'augmentation des naissances). On relève dans une maternité les données suivantes:

Phase	Nouvelle lune	Premier quartier	Pleine lune	Dernier quartier	Total
Effectif	76	88	100	96	360
Fréquence	0,211	0,244	0,278	0,267	1

Peut-on valider l'hypothèse à partir de ces données ?

Effet de la lune (II)

Le cas est plus difficile que si l'on avait:

Phase	Nouvelle lune	Premier quartier	Pleine lune	Dernier quartier	Total
Effectif	89	88	92	91	360

Phase	Nouvelle lune	Premier quartier	Pleine lune	Dernier quartier	Total
Effectif	10	20	300	30	360

Effet de la lune (III)

- On pose l'hypothèse nulle H_0: les naissances sont équiprobables par rapport aux phases de la lune.
- Ceci peut se traduite par:

Phase	Nouvelle lune	Premier quartier	Pleine lune	Dernier quartier	Total
Effectif observé	89	88	92	91	360
Effectif théorique	90	90	90	90	360

Effet de la lune (IV)

La démarche statistique dévoilée :

- On décide d'une hypothèse de référence (dite *nulle*) H_0 portant sur une population donnée,
- On récupère un échantillon de la population en question,
- On "compare" la distribution de la population à celle qui serait obtenue en partant de l'hypothèse nulle,
- Si l'écart est trop grand, il ne peut pas être uniquement *lié au hasard* et on réfute l'hypothèse nulle.

Effet de la lune (V)

- Dans notre cas, on peut "comparer" les distributions à l'aide d'une mesure globale $M = \sum (Obs. - Theo)^2 / Theo$,
- Cette mesure est ensuite comparée à une valeur de référence définie en fonction du nombre de degrés de libertés des données (ν égal au (nombre de classes - 1)) et d'une marge d'erreur (classiquement 5%),
- Si la mesure est inférieure au seuil, on ne rejette pas l'hypothèse nulle.

Ici, $\nu = 3$, on peut montrer que le point critique vaut 7,82 et notre mesure vaut 3,83 : *l'écart est assez petit pour être justifié par le hasard, on conservera ici l'hypothèse nulle*.

Objectifs généraux du cours

Objectifs:

1. Maîtriser les principes de la démarche d'analyse statistique,
2. Maîtriser les principaux tests statistiques,
3. Maîtriser les principaux modèles statistiques linéaires (régression, ANOVA),
4. Savoir gérer les étapes principales d'une analyse de données (ACP).

Structure du module

- Cours magistraux (5)
- TP / TD (4)
  - avec "papier / crayon" - pour compréhension des principes (1 seul)
  - sur R, logiciel statistique opensource
- Projet statistique
  - par groupe (4 à 6 personnes - 6 c'est mieux ...), début 15 octobre matin
  - sur un sujet libre ou associé à un autre projet ou module
  - un sujet validé par les gentils enseignants et complété par leur soin (pour justifier l'effort de 6 personnes)
  - objectifs : dérouler tout ou partie de la démarche statistique présentée en cours et présenter les résultats principaux
  - livrables : un rapport de synthèse, une présentation devant les enseignants
  - une séance de synthèse faite par l'enseignant, commentant les points positifs et les points négatifs des projets

Exemples de projets

- Peut-on modéliser statistiquement le résultat d'un match de football à partir du temps de jeu au premier but marqué ?
- "La femme idéale" : étude statistique
- Analyse statistique des raisons de non-fréquentation du RAK
- Comparaison statistique des comportements des étudiants ingénieurs / école de commerce vis-à-vis de la technologie
- L'usage des technologies 3D est-il lié à l'âge ou au genre ?
- Le chiffre 7 est il bien le préféré des français ?
- Dites moi comment vous vous douchez, je vous dirai qui vous êtes ...
- etc.

Notation

- Contrôle de connaissances : 50%
- Projet : 50%
- Possibilité de quizz en début de cours : 10% (jamais pratiqué pour l'instant)

Tout ça pour 110% ...

Contenu du module

1. Cours 1 : Introduction et rappels - **4 Octobre matin**
2. Cours 2 : Echantillonnage et théorie de l'estimation. Sondages : méthodes d'échantillonnage aléatoires et non aléatoires. Intervalles de confiance - **9 Octobre matin**
3. TP 1 : échantillonnage et estimation- **9 Octobre après-midi**
   - "sur R"
4. Cours 3 : Tests statistiques - **10 Octobre matin**
5. TP 2 : découverte de R et stats descriptives - **10 Octobre après-midi**
   - sur papier
6. Cours 4 : tests non paramétriques, corrélation - **16 octobre matin**
7. Cours 5 : régression linéaire - **19 Octobre matin**
8. TP 3 : Tests/Régression/Corrélation - **22 Octobre matin**
   - sur R
9. TP 4 : Analyse en composantes principales - **23 Octobre matin**
   - sur R
10. Remise du projet : **17 Novembre**
11. Soutenance des projets : **journée du 19 novembre**
12. Contrôle de connaissances - **29 Novembre**
    - individuel
    - avec supports de cours
    - avec outils informatiques

Important

Vous devez effectuer l'apprentissage de R principalement par vous mêmes.

- une introduction à R et à ses fonctions statistiques est disponible sous Moodle
- il est possible (suggéré) d'utiliser l'outil RStudio (qui fournit un interface graphique plus facile à appréhender)
- un premier TP vous permet de jouer avec les primitives de base

N'oubliez pas que le cours ne vise pas la maîtrise de ces environnements informatiques, mais celle des concepts et principes statistiques.

Objectifs du cours 1

Objectifs:

1. Historique des statistiques
2. Se sensibiliser au raisonnement statistique,
3. Revoir les principales distributions de probabilité utiles pour les statistiques,
4. Comprendre le théorème central limite (et surtout à quoi il sert).

Généralités

Definition

Le mot "statistique" désigne à la fois un ensemble de données d'observations et l'activité qui consiste dans leur recueil, leur traitement et leur interprétation" (Encyclop. Univ.)

- Etude d'une population (ensemble d'individus ou unités statistiques), définie par des variables
- Recensement (population) vs. sondage (échantillon)
- Extraction de propriétés globales

Historique 1

- latin *status* (Etat), italien *stato statistica* relatif à l'état (1633) puis *statisticus* (1672) en latin moderne
- *statistik* (allemand)
- *statistics* (anglais)
- *statistique* (français)

Historique 2

- Paléolithique supérieur : liste d'objets
- -3000 : tablettes d'argile sumériennes pour des listes d'hommes et de biens

- -2238 : empereur chinois Yao organise le recensement de productions agricoles
- -2500 : on gère l'impôt et le cadastre en Egypte

Historique 3

- Empire romain : données de recensement

- ressources en hommes et en biens
- classer les citoyens pour répartir charges et avantages
- bilan de richesse de l'empire (soldats, navires, ressources)

Historique 4

- XIIIeme siècle : données sur le commerce extérieur
- XVIeme siècle : registres de naissance (François 1er)
- XVIIeme siècle : fort pouvoir central et stratégies de dénombrement

Historique 5

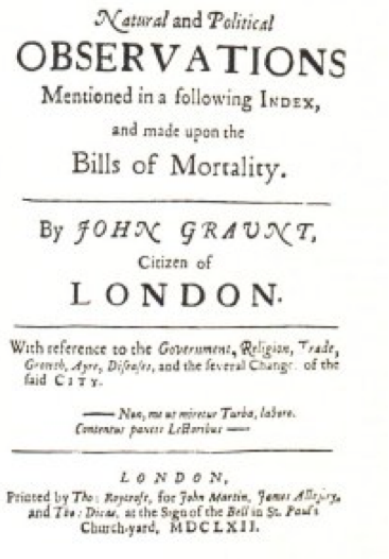
XVIIème et XVIIIème siècle

- coordonnées cartésiennes en 1637, représentation graphique et organisation des données en tableau
- Domination de l'école allemande : description globale des états, classification des savoirs, nomenclature

Historique 6

Influence anglaise

- 1662 : Graunt, point de départ de la démographie moderne. Méthode statistiques pour l'analyse des bulletins de mortalité
- 1690 : Petty promeut les approches quantitatives pour l'analyse des phénomènes socio-économiques. Arithmétique politique anglaise.



Natural and *Political*

OBSERVATIONS

Mentioned in a following INDEX,

and made upon the

Bills of Mortality.

By *JOHN GRAUNT*,

Citizen of

LONDON.

With reference to the *Government*, *Religion*, *Trade*, *Growth*, *Ayre*, *Diseases*, and the several Changes of the said CITY.

—— *Nec, me ut miretur Turba, laboro,*
Contentus paucis Lectoribus ——

LONDON,

Printed by *Tho: Roycroft*, for *John Martin*, *James Allestry*, and *Tho: Dicas*, at the Sign of the *Bell* in St. *Paul's* Church-yard, MDCLXII.

Historique 7

XIXème siècle

- Quetelet (1796-1874) : homme moyen
- Angleterre : biologie, hérédité, eugénisme et régression linéaire
- Galton (1822-1911) : taille parents / enfants

Historique 8

XXème siècle

- Formulations probabilistes : Fischer (1890-1962), Gosset (1876-1937)
- Fisher : *statistique inférentielle* - agronomie - nombre insuffisant données
- Gosset : statistiques industrielles - brasserie Guinness !

Démarche statistique classique

- Recueil de données
  - Sondages
  - Plans d'expériences
- Statistique exploratoire : synthèse de l'information contenue dans les données
  - Statistique descriptive
  - Analyse de données: classification, analyse en composantes principales, analyse des correspondances
- Statistique inférentielle
  - Construction d'estimateurs
  - Test d'hypothèses
  - Modélisation et prévision statistique

En pratique

- Identifier la population ciblée
- Déterminer le codage adapté (si pas naturel)
- Identifier et caractériser l'échantillon (statistiques descriptives)
- Modélisation statistique (identification de la loi ou modèle)
- Travailler les paramètres d'intérêt (estimation, intervalles de confiance)

Recueil de données

La population globale est généralement inaccessible.

- vote et élections
- test de produits manufacturés
- caractéristiques de produits et objets naturels (poids et taille des cornichons cueillis en Alabama du Sud, une question tellement essentielle ...)

Qualité d'un échantillon / d'une procédure de recueil

Voir cours 2 sur les méthodes d'échantillonnage.

.

Les échantillons doivent être constitués d'observations **indépendantes** et **identiquement distribuées** dans la population.
Attention aux biais de recueil :

- Biais liés aux sondages téléphoniques
- Biais liés aux heures et jours du recueil
- Biais de motivation
- Biais de manipulation

On peut se raccrocher aux **méthodes de quotas** ou autre.

Quelles propriétés globales ?

A quelles grandeurs peut-on se référer pour décrire l'échantillon - et par suite la population complète ?

- moyenne
- variance / écart-type
- droite de régression
- médiane
- mais aussi ...
- valeur extrême
- moments d'ordre supérieur (kurtosis, *skewness* par exemple)
- plus généralement **identification paramétrique** (modèle statistique)

Ce qui vous sert (ou sert le client final de l'analyse statistique).

Concepts de base de statistique descriptive

Vous devez maîtriser les concepts de :

- codage
- histogramme
- moyenne, médiane, classe modale
- variance
- quartiles
- boite à moustache
- quantiles

Ces concepts sont passés en revue dans les transparents fournis en annexe.

LE FIGARO.fr SOCIAL

Derrière le salaire moyen, de fortes disparités

http://www.lefigaro.fr/social/2014/03/13/09010-20140313ARTFIG00167-derriere-le-salaire-moyen-de-fortes-disparites.php

| Mis à jour le 13/03/2014 à 14:27 |

Les 1% des Français les mieux payés touchent plus de 7.800 euros bruts par mois. Crédits photo: François Bouchon/Le Figaro. *Crédits photo : François BOUCHON/Le Figaro*

Le salaire médian, à 1.675 euros bruts mensuels, sépare les 50% des Français qui gagnent moins que cette somme de l'autre moitié qui gagne plus.

Hier, l'Acoss[1], l'organisme qui fédère les organismes de Sécurité sociale[2], révélait que le salaire[3] moyen par tête avait augmenté de 0,3% au 4e trimestre 2013 dans le secteur privé, atteignant 2.449 euros mensuel. Un chiffre qui a étonné nombre d'internautes du *Figaro*. Certains le trouvant élevé, d'autres trop faible! Le fait est que ce chiffre n'est qu'un moyenne. Il ne signifie pas que la majorité des Français touchent tous les mois cette somme. Il représente seulement la masse salariale par rapport au cumul des rémunérations brutes des salariés, avec de très hauts et de très bas salaires.

7.800 euros par mois

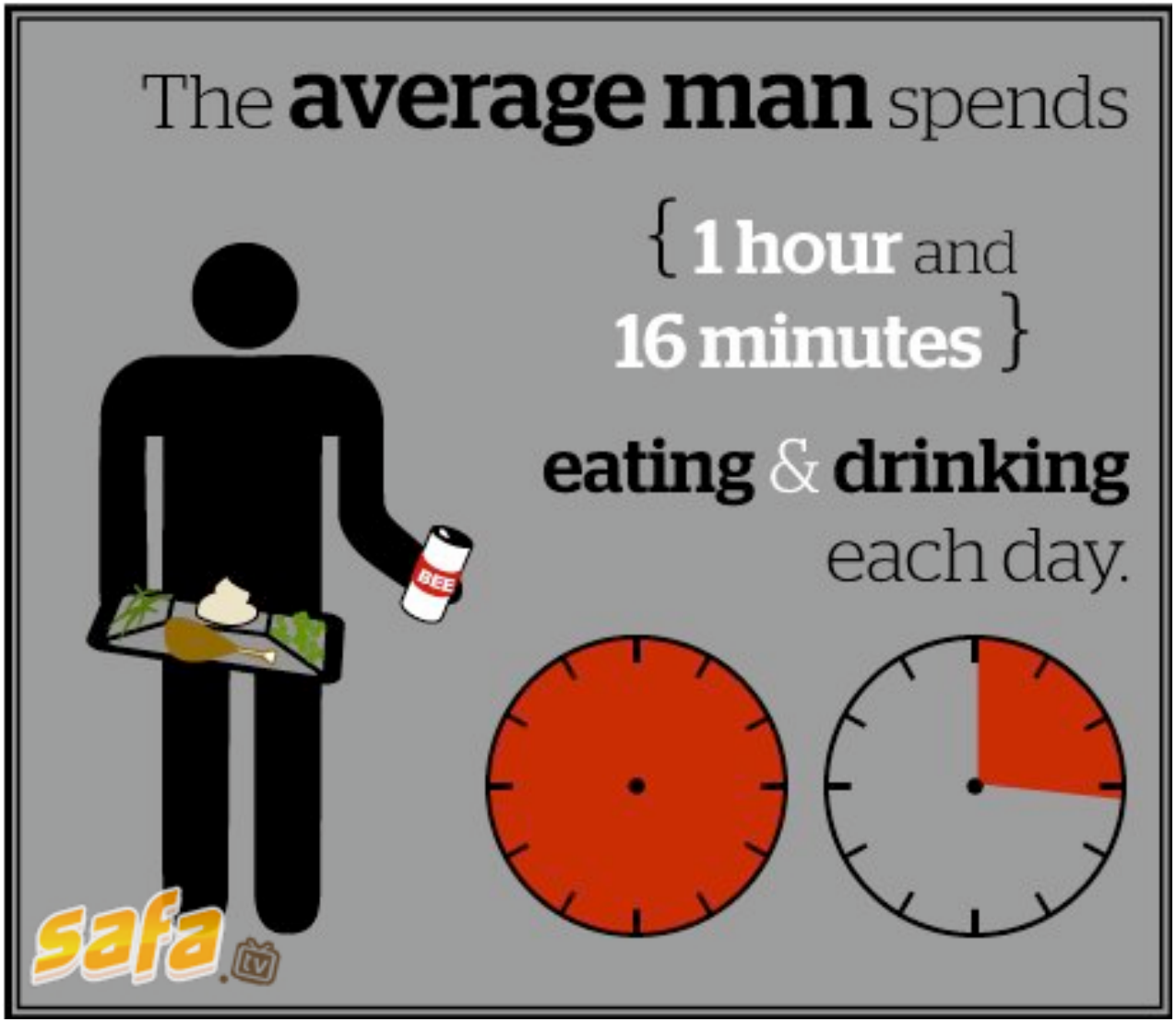
Plus significatif est le salaire médian. S'élevant, selon les dernières données disponibles, à 1.675 euros bruts mensuels, ce dernier sépare les 50% des Français qui gagnent moins que cette somme de l'autre moitié qui gagne plus.

Derrière ces chiffres, les disparités restent importantes. Selon l'Insee[4], les 10% de salariés les moins bien payés touchent en moyenne un salaire net mensuel de 1.170 euros. A l'inverse, les 10 % de salariés les mieux rémunérés disposent, eux, de plus de 3.400 euros. Pour faire partie du «top» - les 1 % des Français les mieux payés -, il faut afficher une feuille de paie supérieure à 7.800 euros par mois.

La rédaction vous conseille :

- Un bulletin de salaire coûte plus de 390 euros par salarié[5]
- Salaires impayés: les entreprises font de plus en plus appel aux crédits[6]
- Les salaires ne flamberont pas cette année[7]

Marie Visot


The average man spends
{ 1 hour and
16 minutes }
eating & drinking
each day.
BEE
safa

A quoi servent VRAIMENT les lois statistiques ?

On associe à un phénomène observé une *variable aléatoire* qui représente ce phénomène ou la mesure du phénomène, et dont on est supposé maîtriser la loi. Maîtriser la loi signifie, par exemple, en connaître la moyenne et la variance (ou tout autre moment).
On peut ensuite jouer avec :

- les probabilités (je connais la loi et je calcule la probabilité d'obtenir une valeur donnée)
- les statistiques (je connais une valeur et j'estime s'il est possible de la considérer comme "normale")

Loi binomiale (I)

- 20 pièces, X "faces" obtenues
- 36 dés, Y "1" obtenus
- 10 naissances, U filles
- 45% de la population en faveur d'un projet de loi ; dans un échantillon de 100 personnes, W se déclarent en faveur du projet

Loi binomiale $B(n,p)$ où n est le nombre d'essais et p est la probabilité (à chaque essai) d'obtenir un succès.

$$p(x) = C_n^x p^x (1-p)^{n-x}$$

Loi binomiale (II) - Exemple

Dans une manufacture, on inspecte les lots d'articles produits en série en utilisant des méthodes d'échantillonnage. Dans chaque lot, 10 articles sont choisis au hasard et le lot est rejeté si 2 articles (ou plus) sont défectueux. Si le lot contient 5% d'articles défectueux, quelle est la probabilité que le lot soit accepté ou refusé ?

X est $B(10; 0,05)$ et le lot est accepté si $X = 0$ ou $X = 1$.

$$P(lot\ accepte) = p(0) + p(1) = C_{10}^{0}(0,05)^{0}(0,95)^{10} + C_{10}^{1}(0,05)^{1}(0,95)^{9}$$

$$P(lot\ accepte) = 0,91386$$

Loi binomiale (III) - Espérance et Variance

Pour une loi binomiale $B(n,p)$,

$$E(X) = \mu = np$$

et

$$Var(X) = \sigma^2 = npq$$

Loi binomiale (IV) - exemple

Dans la présipauté de Groland 30% des gens sont partisans du vice-président Marcel. Lors d'un sondage auprès de 1000 personnes, X personnes se déclarent en faveur de Marcel. Au vu de la taille relative de l'échantillon et de la ville, on peut considérer X comme binomiale $B(1000\ ;\ 0,3)$.

$$E(X) = np = 1000.(0,3) = 300$$

$$Var(X) = npq = 1000.(0,3).(0,7) = 210$$

$$\sigma = \sqrt{(Var(X))} = 14,49$$

Si vous obtenez 700 supporters, c'est louche ...

Pourquoi c'est louche ?

Parce que Bienaymé-Chebychev !

$$P(|X - \mu| \geq \alpha) \leq \frac{\sigma^2}{\alpha^2}$$

Loi de Poisson (I)

La loi de Poisson exprime l'arrivée d'événements "au sens temporel". Plus précisément, X est le nombre d'événements qui se produisent durant un certain intervalle de temps. Si X suit une loi de Poisson de moyenne λ, on a:

$$p(x) = \frac{\exp^{(-\lambda)} \lambda^x}{x!}$$

Si X est une loi de Poisson

$$E(X) = \lambda$$

$$Var(X) = \lambda$$

L'égalité de l'espérance et de la variance constitue un indicateur empirique de présence d'une loi de Poisson.

Loi de Poisson (II)

Un enseignant-chercheur fais en moyainne 17 fautes de frappe part transparant de cour. Ile vient de taper 50 transparans, quelle est la probabillitté d'avoir moins de 2 fautes de frap ? Le nombre moyen de fautes attendu est $50.17 = 850$ et donc

$$P(X < 2) = p(0) + p(1) = \exp^{-850}(\frac{850^0}{0!} + \frac{850^1}{1!}) = ...$$

Loi de Poisson (III)

La loi de Poisson peut être approximée par une loi binomiale:

- n tend vers l'infini
- p tend vers 0

Ex: nombre d'appels à un standard entre 10h et 11h. Pour définir X binomiale, on découpe l'intervalle en 3600 secondes, donnant donc 3600 essais. La probabilité d'avoir un succès par essai est très faible et on peut raccourcir l'intervalle et diminuer la probabilité à loisir ... Convergence vers loi de Poisson (si conditions indépendance et résultat binaire validés).

Lois continues : loi exponentielle (I)

Une variable aléatoire X est de loi exponentielle (avec moyenne $\theta > 0$) si sa fonction de densité est :

$$f(x) = \frac{1}{\theta} \exp^{-\frac{x}{\theta}}$$

Ces lois modélisent les temps d'attente avant l'arrivée d'un événement. Le lien avec la loi de Poisson est immédiat : si le nombre d'événements survenant pendant un intervalle de temps t est régi par une loi de Poisson de moyenne $\lambda = ct$, le temps d'attente entre deux arrivées suivra une loi exponentielle avec $\theta = \frac{1}{c}$

Lois continues : loi exponentielle (II)

- Application au MTBF : si un enseignant-chercheur pète les plombs en moyenne toutes les 40 heures de cours, quelle est la probabilité qu'il reste normal sur un semestre avec 80 heures de cours ?
- On suppose le temps de bon fonctionnement en loi exponentielle de moyenne 40
- $P(X > 80) = \exp^{\frac{-80}{40}} = \exp -2 = 0,1353$

Lois continues : loi normale (ou loi de Gauss) (I)

La variable X est de loi normale si sa fonction de densité est:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

où μ et σ^2 sont l'espérance et la variance de la loi. On note la loi $N(\mu, \sigma^2)$. Loi réduite : pour une une loi $N(\mu, \sigma^2)$, on définit classiquement la loi réduite $Z = \frac{X-\mu}{\sigma}$ qui est de loi $N(0,1)$.

$$P(a < N(\mu, \sigma^2) < b) = P(\frac{a-\mu}{\sigma} < N(0,1) < \frac{b-\mu}{\sigma})$$

Lois continues : loi normale (ou loi de Gauss) (II)

- Si X est une loi normale $N(\mu, \sigma^2)$, $a + bX$ est une loi normale $N(a + b\mu, b^2\sigma^2)$
- si $X_1, X_2, .., X_n$ sont des lois normales indépendantes de lois respectives $N(\mu_i, \sigma_i^2)$, alors leur somme est normale de loi $N(\mu_1 + \mu_2 + .. + \mu_n, \sigma_1^2 + \sigma_2^2 + .. + \sigma_n^2)$

Lois continues : loi normale (ou loi de Gauss) (III)

- On suppose que le poids en kilogrammes d'un enseignant-chercheur se distribue selon une loi $N(70, 25)$. Quelle est la probabilité que le véhicule administratif de TB (qui admet une charge de 1500 kg) ne puisse transporter les 20 E/C du département LUSSI ?
- $E(Lussi) = E(X_1 + .. + X_{20}) = 1400$
- $Var(X) = Var(X_1 + .. + X_{20}) = 500$
- $P(X > 1500) \Leftrightarrow P(N(1400, 500) > 1500)$
- soit $P(N(0, 1)) > \frac{1500-1400}{\sqrt{500}}$
- soit $P(N(0, 1)) > 4,47$
- on est donc "normalement" sauvés...

Loi du χ^2

Soient $U_1, U_2, .., U_p$ p variables normales $N(0,1)$ indépendantes. On appelle loi du χ^2 à p degrés de liberté χ^2_p la loi de la variable $\sum_{i=1}^{p} U_i^2$.

La loi du χ^2 peut être approximée par une loi normale. Lorsque $p > 30$, on peut effectivement considérer que $\sqrt{(2\chi^2)} - \sqrt{(p-1)}$ est une loi de type $N(0,1)$.

Loi de Student

Soit une variable aléatoire U suivant une loi normale $N(0,1)$ et X indépendante de U suivant une loi χ^2_n. On définit la variable de Student T à n degrés de liberté par:

$$T_n = \frac{U}{\sqrt{\frac{X}{n}}}$$

Loi de Fisher-Snedecor

Si X et Y sont respectivement des lois de type χ^2_n et χ^2_p, alors on peut définir la loi suivante:

$$F(n;p) : \frac{X/n}{Y/p}$$

Cette loi sert de référence pour les analyses de variances.

Statistiques et probabilités

Une question clé (que nous retrouverons avec le théorème central limite) : pourquoi se ramène-t-on le plus souvent à une loi normale? 3 raisons bien différentes :

- La répartition statistique d'une variable est voisine d'un modèle théorique
- Les observations sont des données imprécises, donc entachées d'erreurs
- Le mode de constitution d'un échantillon permet d'aligner la distribution statistique sur des modèles

Théorème central limite (I)

- Théorème central limite (variables de même loi): Soit un grand nombre n de variables indépendantes $X_1, X_2, .., X_n$ de même loi. Alors leur somme $X = X_1 + X_2 + .. + X_n$ suit approximativement une loi normale, même si ces variables ne sont pas normales
- Théorème central limite généralisé: Soit un grand nombre n de variables indépendantes $X_1, X_2, .., X_n$. Alors sous certaines conditions, leur somme $X = X_1 + X_2 + .. + X_n$ suit approximativement une loi normale, même si ces variables ne sont pas normales.

Pour certains auteurs, n grand pour $n > 30$, pour d'autres $n > 100$.

Théorème central limite généralisé (II)

Conditions : essentiellement celle de Lindeberg, qui indique que les variables réduites $\frac{X_i - \mu_i}{S_n}$ soient "uniformément petites" avec une grande probabilité.

$$\lim_{n\to\infty} \frac{1}{S_n^2} \sum_{i=1}^{n} \int_{|x|>S_n} x^2 dF_i(x) = 0$$

avec F_i fonction de répartition de $X_i - \mu_i$ et

$$S_n^2 = \sum_{i=1}^{n} \sigma_i^2$$

Eléments de démonstration (variables de même loi)

La fonction caractéristique $\phi_x(t) = E[\exp^{itX}]$ d'une variable Y d'espérance 0 et de variance 1 peut être approximée par :

$$\phi_Y(t) = 1 - \frac{t^2}{2} + o(t^2)$$

Alors si la moyenne centrée réduite d'observations est :

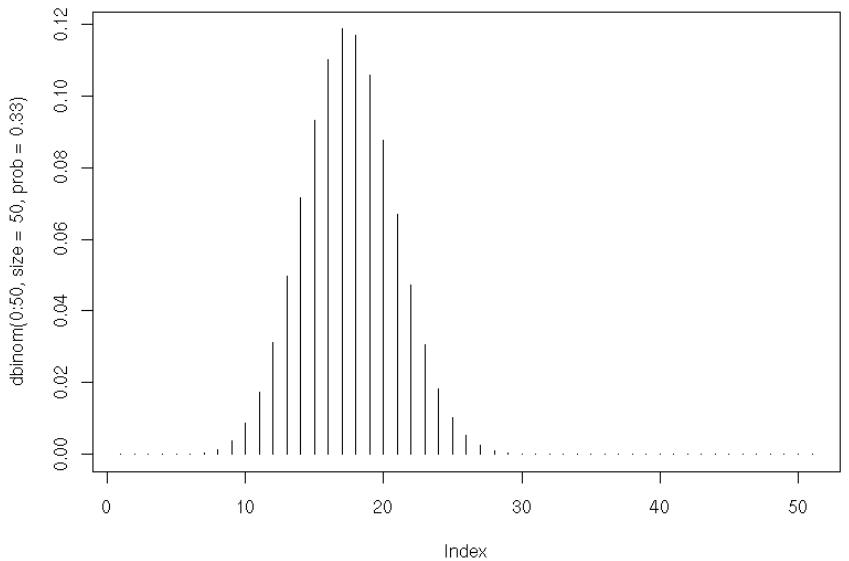
$$Z_n = \frac{\overline{X_n} - \mu}{\frac{\sigma}{\sqrt{n}}} = \Sigma_{i=1}^{n} \frac{Y_i}{\sqrt{n}}$$

la fonction caractéristique de Z_n est $(\phi_Y(\frac{t}{\sqrt{n}}))^n = [1 - \frac{t^{2n}}{2} + o(\frac{t^2}{n})]^n$

qui converge vers $\exp^{-\frac{t^2}{n}}$... qui est la fonction caractéristique de la loi normale.

Application à l'approximation de loi binomiale

En répétant une expérience de type binomial, plusieurs composantes indépendantes - toutes de même loi - sont ajoutées, de sorte que leur somme est proche d'une loi normale. Une loi binomiale $B(n,p)$ sera ainsi approximée par une loi normale $N(\mu = np, \sigma = \sqrt{(npq)})$.



Approximation loi binomiale (II)

On lance 16 pièces. On veut calculer la probabilité d'obtenir un nombre de "faces" entre 5 et 10 compris. Avec $B(16, 1/2)$, on obtient en prenant les valeurs exactes binomiales :

$$P(5 \leq X \leq 10) = p(5)+p(6)+p(7)+p(8)+p(9)+p(10) = 0,85654$$

Pour utiliser une approximation normale, on doit passer à une loi continue, et on ajuste les bornes de l'intervalle, ici à 4,5 et 10,5. On obtient:

$$\begin{aligned} P(5 \leq X \leq 10) &\simeq P(4,5 < N(8,4) < 10,5) \\ &= P(-1,75 < N(0,1) < 1,25) \\ &= 1 - (P(N(0,1) > 1,75) - P(N(0,1) > 1,25) \\ &= 0,8543 \end{aligned}$$

Approximation loi binomiale (III)

L'approximation sera d'autant meilleure que n sera grand, on estime l'approximation valable pour $npq > 5$. Sans la correction pour la continuité, on aurait obtenu pour l'exemple précédent une réponse de 0,7745, donc bien moins précise.

Approximation loi binomiale (IV) - Eléments de démonstration

On part de la fonction caractéristique de la variable aléatoire (transformée de Fourier de sa densité) soit

$$\varphi_X(t) = \int_{\mathbf{R}} \exp^{itx} f(x)dx = (p \exp^{it} + 1 - p)^n$$

alors la fonction caractéristique de $\frac{X_n - np}{\sqrt{npq}}$ peut être approximée par:

$$\varphi(t) = (p \exp^{\frac{it}{\sqrt{npq}}} + 1 - p)^n \exp^{-\frac{itnp}{\sqrt{npq}}}$$

Approximation loi binomiale (V) - Eléments de démonstration - suite

En passant au log, on obtient:

$$\ln \varphi = n \ln(p(\exp^{\frac{it}{\sqrt{npq}}}) - 1)) - \frac{itnp}{\sqrt{npq}}$$

En développant successivement au deuxième ordre l'exponentielle $(1 + px + px^2)$ puis le logarithme $(x - \frac{x^2}{2})$, on obtient:

$$\ln \varphi \simeq \frac{-t^2}{2q} + \frac{pt^2}{2q} = \frac{t^2}{2q}(p-1) = \frac{-t^2}{2}$$

Ce qui correspond à la fonction caractéristique de la loi normale centrée réduite $N(0,1)$. Ouf.

Retour sur le test du χ^2

Si on part d'un partitionnement de l'espace des événements en n classes $A_1, A_2, ..A_n$, et si on mesure leur taux de réalisation $N_1, N_2, .., N_n$, le vecteur suit une loi multinomiale (extension de la loi binomiale de 2 à n classes) d'effectif total n et de paramètres $p_1, p_2, ..., p_n$. On peut définir la loi conditionnelle de N_i connaissant $N_j = n_j$ comme $B(n - n_j, \frac{p_i}{1-p_j})$. Le théorème central limite (appliqué à chacune des composantes du vecteur) indique que chacune de ces composantes $N_i - np_i$ tend vers une loi normale $N(0, 1)$. Donc:

$$\lim_{n\to\infty} \sum_{i=1}^{k} \frac{(N_i - np_i)^2}{np_i} \to \chi^2_{k-1}$$

On retrouve ici le résultat de l'exemple précédent.

Distribution d'une moyenne

Grâce au théorème central limite, on a le résultat suivant :

- Si $\overline{X}$ est la moyenne de n observations indépendantes $X_1, X_2, .., X_n$ où $E(X_i) = \mu$ et $Var(X_i) = \sigma^2$, alors, pour n grand, $\overline{X}$ est approximativement de loi $N(\mu, \frac{\sigma^2}{n})$.
- Plus n est grand, plus l'estimation $\overline{X}$ sera près de la véritable valeur μ et plus la variance de la variable "moyenne estimée" sera petite.

Application à une loi exponentielle

On suppose la loi exponentielle d'espérance $\frac{1}{\lambda}$ et de variance $\frac{1}{\lambda^2}$.
Alors si on prend un échantillon de n tirages ce cette loi $X_1, X_2, .., X_n$,
la moyenne de ces échantillons satisfait :

$$\frac{\sqrt{n}}{\sigma}(X_n - \mu) \to N(0,1)$$

soit

$$\lambda\sqrt{n}(X_n - \frac{1}{\lambda}) \to N(0,1)$$

$$\sqrt{n}(X_n - \frac{1}{\lambda}) \to N(0, \frac{1}{\lambda^2})$$

Annexes

L'exemple fil-rouge...

Une population de 26 étudiants passant un contrôle...

Pour chaque candidat, on note :

- le temps mis à effectuer l'épreuve (variable x),
- le nombre d'erreurs commises (variable y).

Voici les résultats :

Candidat n°	1	2	3	4	5	6	7	8	9	10	11	12	13
x	15	15	20	10	15	30	10	10	5	5	5	10	10
y	4	5	10	0	4	10	2	5	0	1	0	3	3
Candidat n°	14	15	16	17	18	19	20	21	22	23	24	25	26
x	20	15	10	5	20	30	30	30	40	10	5	10	10
y	6	3	2	0	6	8	5	10	12	3	0	2	3

Distribution statistique

Definition

On appellera distribution statistique ou fonction de répartition de X, la donnée des couples $\{(c_1, n_1), \ldots, (c_i, n_i), \ldots, (c_k, n_k)\}$ tel que :

- les c_i forment une partition en k intervalles (appelés aussi classes) de l'ensemble des valeurs prises par la variable ($c_1 = [a_0, a_1]$, $c_i =]a_{i-1}, a_i]$, $c_k =]a_{k-1}, a_k]$),
- chaque n_i est le nombre de valeurs observées dans l'intervalle c_i.

Par convention le centre des intervalles est également noté c_i.

Remarque

Pour une variable discrète, la distribution statistique associée est telle que :

- les c_i représentent toutes les valeurs prises par la variable,
- chaque n_i est le nombre de fois que la valeur c_i a été prise.

Choix des intervalles

Il n'existe pas de choix pertinent du nombre et de l'amplitude des classes ! Cependant :

- il est plus aisé de prendre des classes de même amplitude,
- on peut utiliser la règle de Sturges comme choix de k :

$$k = 1 + \frac{10 \ln(n)}{3 \ln(10)}$$

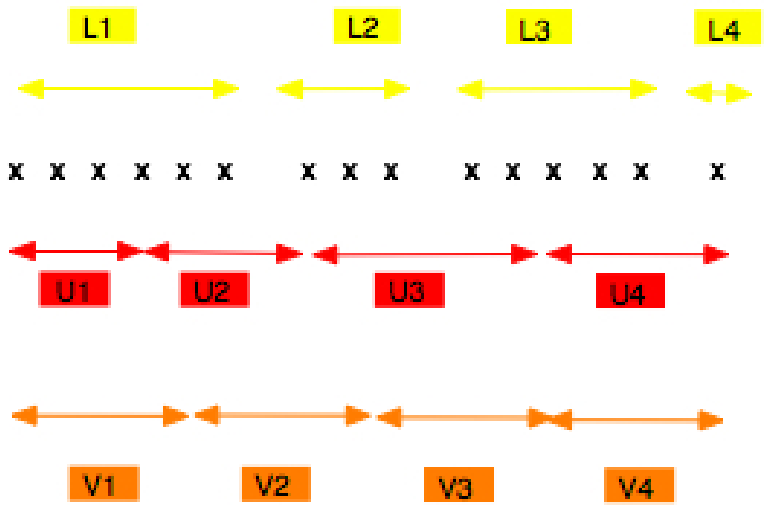
- ou celle de Huntsberger $k = 1 + 3.332 * log(n)$
- ou encore celle de Brooks-Carruthers $k = 5 * log(n)$
- ou celles de Scott ou Freedman qui tiennent compte de la variance des données

Parfois, la découpe en intervalles ira de soi, par exemple lorsque x ne prend que des valeurs entières (se ramène au cas d'une variable discrète).

Choix des intervalles (suite)

On peut forcer les intervalles :

- à être contigus et égaux.
- à contenir le même nombre d'individus
- à minimiser la variance



Histogramme (1/2)

Definition

Pour une distribution statistique donnée, on appellera fréquence de i le rapport :

$$f_i = \frac{n_i}{n}$$

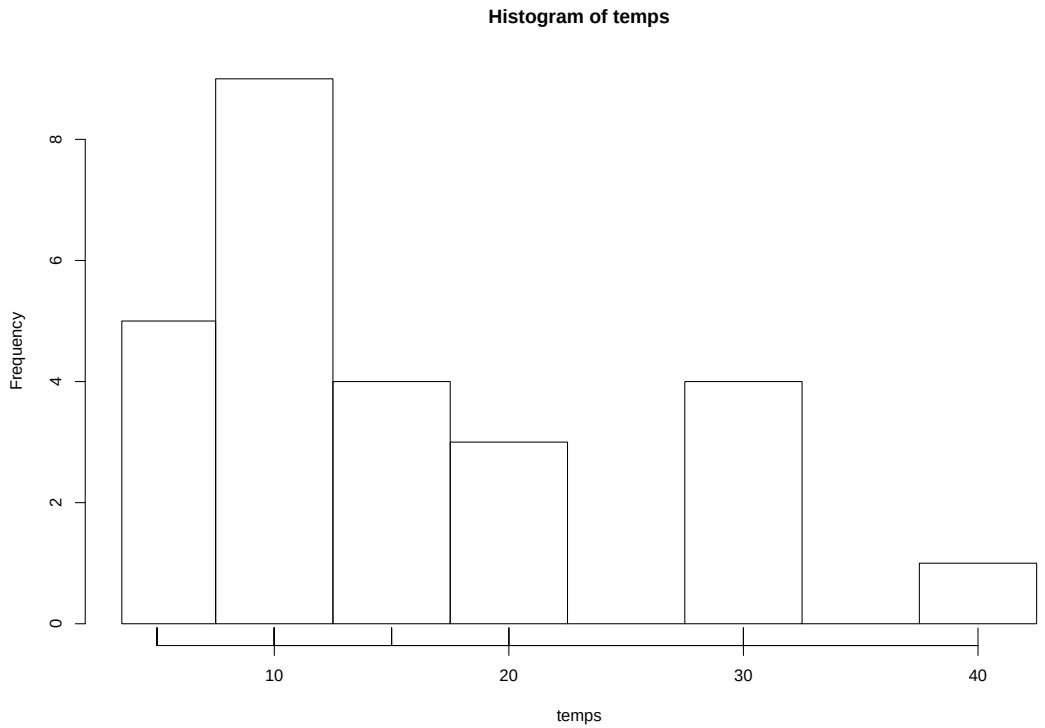
et fréquence cumulée la somme : $F_i = f_1 + f_2 + \cdots + f_i = \sum_{1 \leq j \leq i} f_j$

Definition

On appelle histogramme des fréquences d'une distribution statistique $(]a_{j-i}, a_j], n_j)$ pour $(1 \leq j \leq k)$, le graphique tel que les classes sont reportées en abscisse et au-dessus de chacune d'elle un rectangle d'*aire* égale ou proportionnelle à f_i est tracé.

Histogramme (2/2)

Histogramme de la variable x temps mis pour l'épreuve :

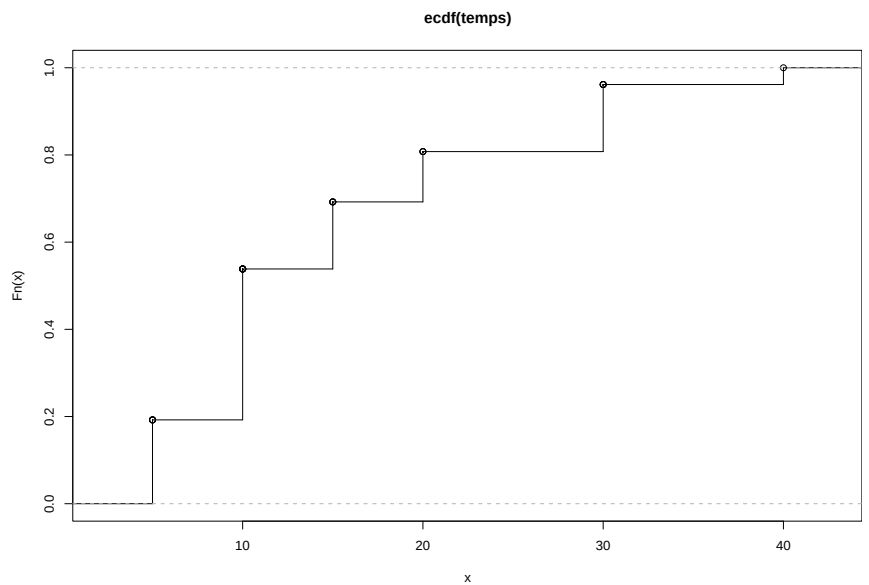


Dans le cas de l'exemple, nous n'avons pas utilisé la règle de Sturges puisqu'un découpage en intervalles centrés autour des notes possibles est plus naturel.

Graphique des fréquences cumulées

Definition

On appelle graphique des fréquences cumulées d'une distribution statistique $(]a_{j-i}, a_j], n_j)$ (pour $1 \leq j \leq k$), le graphique tel que les classes sont reportées en abscisse et au-dessus de chacune d'elle un rectangle de hauteur égal à F_i est tracé.



Moyenne

Definition

La moyenne $\bar{x}$ d'une variable x est définie par l'expression :

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

Par exemple, la moyenne de la variable x est égale à 15.19, et sa médiane vaut 10.0.

Médiane

On note $x_{(1)}, x_{(2)}, \ldots x_{(n)}$ les éléments $x_1, x_2, \ldots, x_n$ rangés par ordre croissant.

Definition

Si on note m et d la partie entière et décimale de $\frac{n+1}{2}$, la médiane $\text{me}(x)$ de la variable x est définie par :

$$\text{me}(x) = x_{(m)} + d(x_{(m+1)} - x_{(m)})$$

Par exemple, la médiane de la variable x est égale à 10.0.
Quel est donc l'intérêt de la médiane par rapport à la moyenne ?

Classe modale

Definition

On appelle classe modale $\text{mo}(x)$ d'une distribution statistique $(]a_{j-i}, a_j], n_j)$ (pour $1 \leq j \leq k$) l'intervalle $]a_{i-1}a_i]$ tel que $n_i = \max_{1 \leq j \leq n}\{n_j\}$

Pour la distribution statistique de la variable x, la classe modale est $]7.5, 12.5]$

Variance et écart-type

Definition

La variance d'une variable est le nombre $s^2(x)$ défini par l'expression :

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

La racine carrée de $s^2(x)$, notée $s(x)$ est appelé écart-type de la variable.

Attention : à ne pas confondre variance et variance corrigée

Quartiles

Definition

Soient m et d les parties entières et décimales de $\frac{n+1}{4}$, et m' et d' les parties entières et décimales de $\frac{3(n+1)}{4}$. On note $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ les valeurs de x rangées par ordre croissant.
Alors :

- le premier quartile noté $q_{0,25}(x)$ est défini par l'expression : $q_{0,25}(x) = x_{(m)} + d(x_{(m+1)} - x_{(m)})$,
- le deuxième quartile noté $q_{0,5}(x)$ est égal à la médiane de x,
- le troisième quartile noté $q_{0,75}(x)$ est défini par l'expression : $q_{0,75}(x) = x_{(m')} + d'(x_{(m'+1)} - x_{(m')})$.

L'étendue inter-quartile $\mathrm{IQR}(x)$ étant défini par
$\mathrm{IQR}(x) = q_{0,75} - q_{0,25}$.

Boîte à moustaches (1/3)

Definition

Une boîte à moustache est un graphique constitué de deux axes :

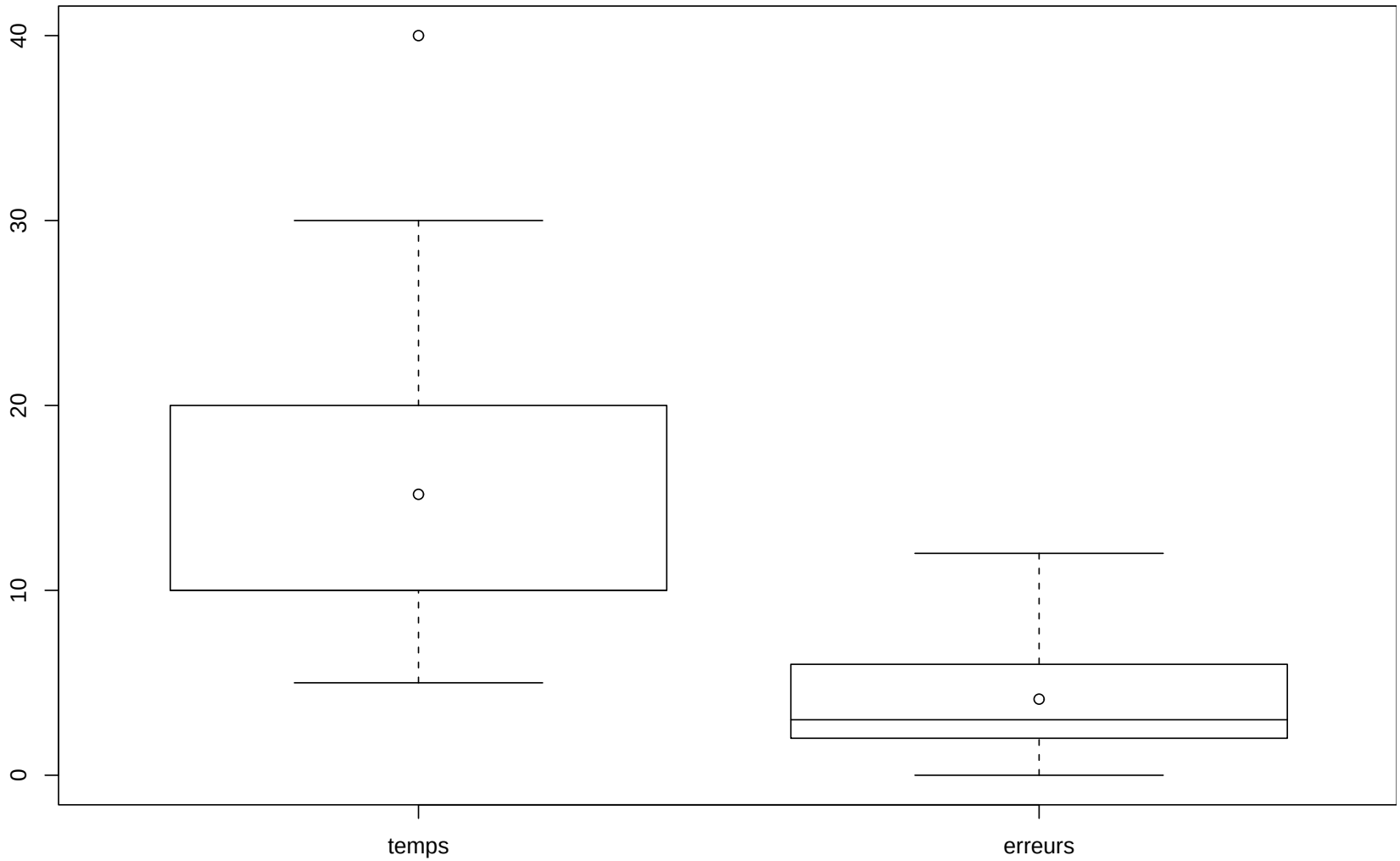
- l'axe vertical, muni d'une échelle numérique qui correspond aux valeurs de la variable observée
- et l'axe horizontal est sans échelle.

sur lesquels :

1. un segment horizontal (de longueur arbitraire) est tracé en regard de la médiane,
2. une boîte est reportée avec les côtés supérieur et inférieur en regard de $q_{0,75}$ et $q_{0,25}$ respectivement,
3. deux segments verticaux sont tracé vers l'extérieur de la boîte (les *moustaches*) joignant le milieu du côté supérieur (*resp.* inférieur) à la plus grande (*resp.* la plus petite) valeur inférieure ou égale (*resp.* supérieure ou égale) à $q_{0,75} + \frac{3}{2}\mathrm{IQR}(x)$ (*resp.* $q_{0,25} - \frac{3}{2}\mathrm{IQR}(x)$).

Boîte à moustaches (2/3)

Boîtes à moustaches des variables x et y :



Boîte à moustaches (3/3)

Boîtes à moustaches des variables "salaires" de l'enquête INSSEE 2004 (source G. Stolz, HEC) :

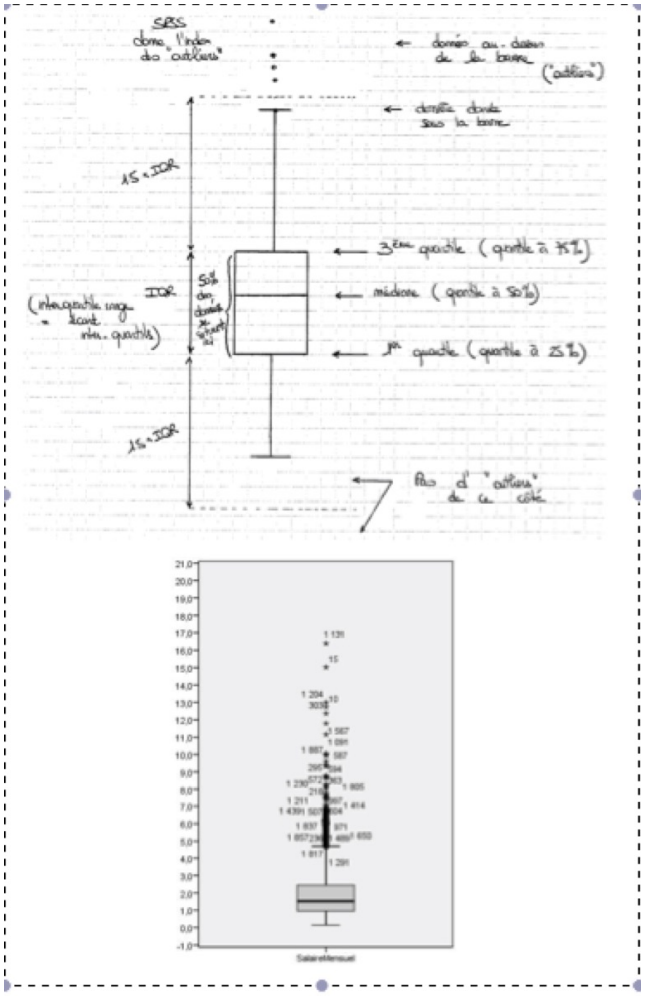
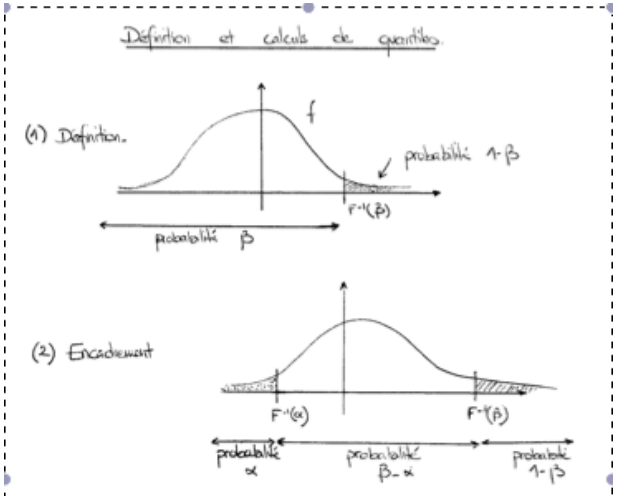


FIGURE 17. Boîtes à moustaches : principe et application.

Passage aux lois continues : quantiles



Definition

On appelle fonction **quantile** la fonction inverse F^{-1} de la fonction de répartition F. On a en particulier :

$$P\{X \leq F^{-1}\} = \beta$$

et

$$P\{X \geq F^{-1}\} = 1 - \beta$$