Eléments de statistique
Introduction - Analyse de données exploratoire

Louis Wehenkel

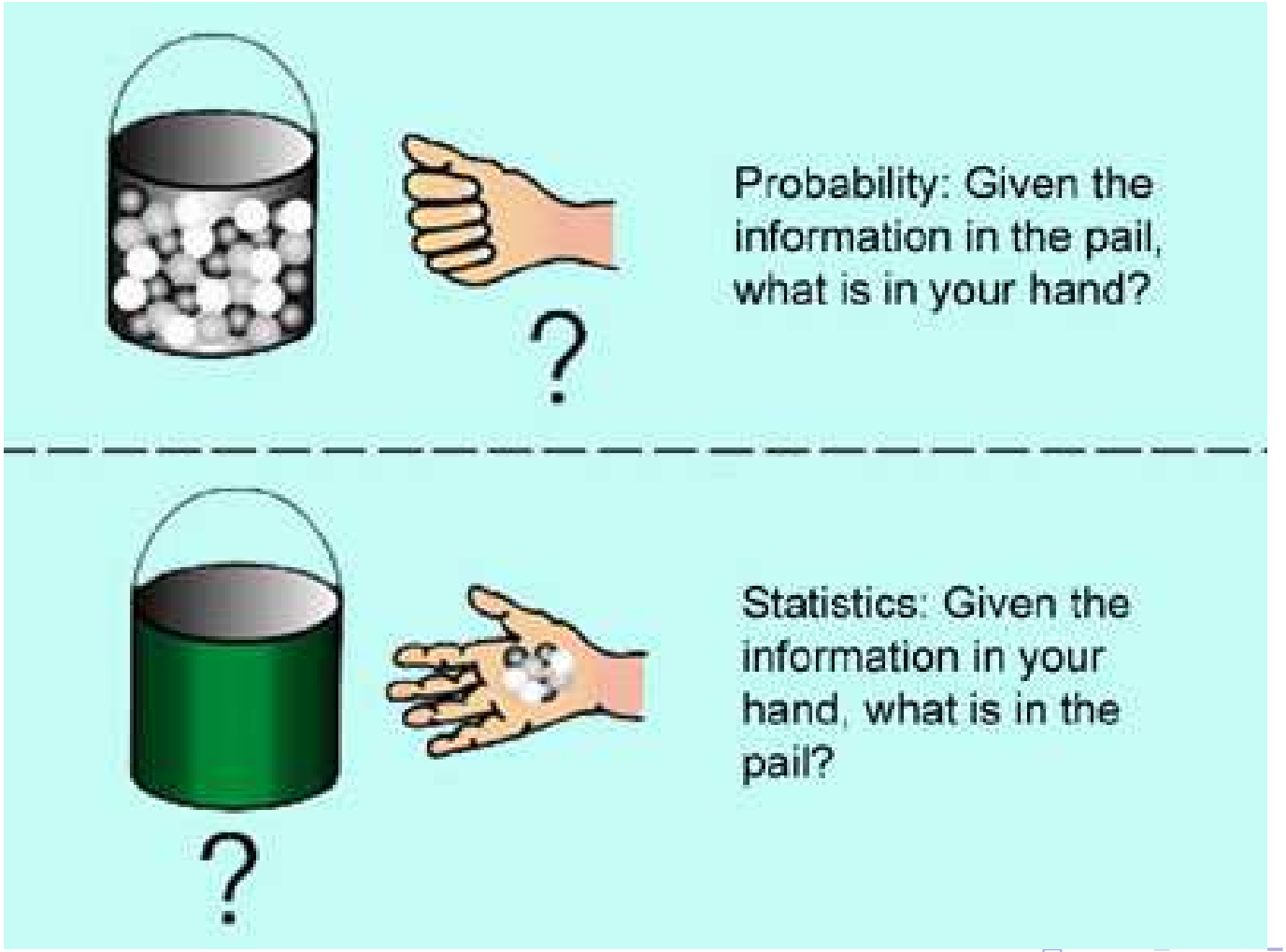
Département d'Electricité, Electronique et Informatique - Université de Liège
B24/II.93 - L.Wehenkel@ulg.ac.be

MATH0487-2 : 3BacIng, 3BacInf - 20/9/2016

Find slides: *http://montefiore.ulg.ac.be/∼lwh/Stats/*

Rappel: probabilités vs statistique

Rappel: probabilités vs statistique



Rappel: probabilités vs statistique

- Probabilités :
  - *le modèle $(\Omega, \mathcal{E}, P)$ est complètement spécifié*
  - le but essentiel est d'*exploiter le modèle* pour prendre des décisions
  - nécessite rigueur et cohérence
  - *Raisonnement déductif*
- Statistique :
  - *le modèle est inconnu*, mais on dispose d'*observations*
  - le but essentiel est de *compléter le modèle* à l'aide des observations
  - nécessite en plus intuition et sens physique
  - *Raisonnement inductif*

Rappel: probabilités vs statistique

- Probabilités :
  - *le modèle $(\Omega, \mathcal{E}, P)$ est complètement spécifié*
  - le but essentiel est d'*exploiter le modèle* pour prendre des décisions
  - nécessite rigueur et cohérence
  - *Raisonnement déductif*
- Statistique :
  - *le modèle est inconnu*, mais on dispose d'*observations*
  - le but essentiel est de *compléter le modèle* à l'aide des observations
  - nécessite en plus intuition et sens physique
  - *Raisonnement inductif*

NB.
- Dans de nombreuses applications, on peut très bien utiliser le calcul de probabilités sans faire appel à la statistique.
- Il est presque impossible de faire de la statistique sans faire appel au calcul des probabilités.

Exemple 1 de problème de statistique

Lors d'un sondage d'opinions, on interroge un échantillon de 500 personnes habitant des zones rurales ainsi que 500 personnes habitant des zones urbaines, sur leurs intentions de vote, au second tour des élections présidentielles en France.
La table suivante reprend le résultat

	Rural	Urbain	*Total*
Candidat 1	234	245	479
Candidat 2	266	255	521
Total	500	500	1000

Peut-on conclure que les intentions de vote des deux sous-populations dont sont issus les deux échantillons ont des préférences électorales différentes ?

Est-il probable que le candidat 2 va remporter les élections ?

Exemple 2 de problème de statistique

- On lance une pièce n fois, et on observe la suite des résultats.
- On souhaite à partir de ces résultats estimer la probabilité de tomber sur pile au prochain lancer de la même pièce
- On lance une seconde pièce n' fois, et on observe aussi les résultats.
- On souhaite choisir la pièce qui à la plus grande probabilité de tomber sur pile parmi ces deux pièces.

Exemple 3 de problème de statistique

- On dispose d'un tableau comprenant deux colonnes, la première reprenant les vraies valeurs d'une grandeur physique et la seconde les valeurs correspondantes mesurées par un instrument.
- On souhaite vérifier que l'erreur de mesure est bien indépendante des valeurs observées, et ensuite calibrer l'instrument en corrigeant la partie systématique de l'erreur, et caractériser la distribution de la partie aléatoire de l'erreur.

Exemple 4 de problème de statistique

- On dispose d'une base de données de patients, dont une moitié sont diagnostiqués comme souffrant d'une certaine maladie. De plus chaque patient est décrit par un certain nombre de valeurs numériques indiquant les résultats d'un examen sanguin.
- On souhaite identifier un sous-ensemble minimal des indicateurs sanguins qui sont en relation avec la maladie étudiée, et à partir de ces grandeurs formuler un modèle prédictif aussi précis que possible et permettant de décider si un patient est malade ou non.

Analyse(s) de données exploratoire(s)

- On dispose d'un tableau de données $\mathbf{D} \in \mathbb{R}^{n \times p}$, dont les n lignes sont des 'individus', et les p colonnes sont des variables utilisées pour mesurer les caractéristiques des individus.
  - NB: en pratique, on a souvent que n et/ou p sont grands (p.ex. plusieurs centaines, ou milliers).
  - On souhaite résumer l'information de $\mathbf{D}$, sous la forme d'une série de grandeurs (on utilisera le terme de 'statistique' pour désigner ces nombres), dont les valeurs peuvent être calculées à partir de $\mathbf{D}$ et ensuite interprétées par un 'expert' humain.
  - On souhaite créer des graphiques, qui résument les informations de $\mathbf{D}$ sous une forme 'visuellement parlante'.
  - Notations/terminologie: on appelle $\mathbf{D}$ l'échantillon.
    - élément (i, j) de $\mathbf{D}$: $\mathbf{D}_{i,j}$,
    - i-ème ligne de $\mathbf{D}$: $\mathbf{D}_{i,\cdot}$,
    - j-ème colonne de $\mathbf{D}$: $\mathbf{D}_{\cdot,j}$,

Exemple de tableau D

- Les résultats obtenus au cours de Probas en juin 2012.
- Individus: les students (ici $n = 182$)
- Variables: les cotes obtenues aux différents travaux et questions de l'examen (ici $p = 10$)
- Exemples d'analyses exploratoires: voir deuxième partie de cette leçon.
- NB: cet exemple servira tout au long du cours et des travaux pratiques, pour illustrer, et étudier les notions introduites dans ce cours.

Notion d'échantillon

- Le but de cette partie du cours est de formaliser mathématiquement la notion d'échantillon **D**.
- On se servira du modèle probabiliste $(\Omega, \mathcal{E}, P)$ pour définir la notion d'échantillon i.i.d. (ainsi que des variantes de ce modèle canonique).
- On mettra en oeuvre le calcul de probabilités pour caractériser les propriétés des statistiques calculées à partir d'un échantillon (aussi appelées distributions d'échantillonnage).
- On discutera les techniques dites de 'contrôle (ou design) d'expérience' visant à optimiser certaines de ces propriétés, quand le nombre d'individus n est limité (par le temps, ou le coût d'acquisition de données).

Estimation de paramètres de population

- On observe n fois les valeurs de p variables aléatoires issues d'un même modèle probabiliste $(\Omega, \mathcal{E}, P)$ inconnu (on utilise le terme de 'population' pour désigner ce modèle).
- On suppose que les n lignes de $\mathbf{D}$ sont obtenues de façon indépendante, en mesurant conjointement p variables $\{\mathcal{X}_j\}_{j=1}^{p}$ pour n résultats de l'expérience aléatoire modélisée.
- On suppose que les densités marginales et/ou conjointes des variables $\mathcal{X}_j$ sont connues à l'exception de la valeur de certains paramètres.
- On souhaite estimer la valeur de ces paramètres selon une procédure qui possède des propriétés souhaitées.
- On proposera des propriétés souhaitées et on en déduira des 'estimateurs' (ponctuels, et/ou par intervalle de confiance)

Test d'hypothèses et théorie de la décision

- Théorie de la décision:
  - En face de deux hypothèses (ou plus de deux), choisir celle qui est la plus appropriée. (Introduire le raisonnement bayesien).
- A partir de données, décider si une hypothèse est plausible ou non.
  - Est-ce que le producteur 1 fournit des équipements plus fiables que le producteur 2 ?
  - La consommation électrique, suit-elle une loi gaussienne ?
  - Est-ce que le taux d'échec a évolué dans le bon sens cette année ?
- Nous définirons un cadre 'cartésien' pour formuler et répondre à ce genre de questions à partir de données.

Régression, analyse de la variance, classification

- On sépare en deux groupes les variables $\mathcal{X}$ observées dans un échantillon :
  - les variables $\mathcal{Y}_k$ à expliquer (souvent une seule variable 'de sortie')
  - les variables $\mathcal{Z}_l$ 'explicatives' (souvent plusieurs, voir un très grand nombre de variables 'd'entrée')
- On souhaite construire un 'modèle' $\hat{\mathcal{Y}} = f(\mathcal{Z})$ à partir des données $\mathbf{D}$, qui explique (en un certain sens) les variations d'une variable cible en fonction des variables d'entrée.
  - Si $\mathcal{Y}$ est numérique on parle de régression et/ou d'analyse de la variance.
  - Si $\mathcal{Y}$ est discrète (un petit nombre modalités de type 'catégorie') on parle de classification.
- NB: dans ce cours nous nous limiterons aux méthodes les plus élémentaires (le traitement détaillé de ce volet est fait au cours d'apprentissage inductif appliqué).

Apperçu général

- Leçons de théorie
  - Louis Wehenkel, mardis matins 8h15, environ une fois sur deux
  - Voir http://www.montefiore.ulg.ac.be/~lwh/Stats/
- Séances de répétition (autres mardis matins 8h15)
  - Trois groupes (F. Van Lishout, H. Huaux, P. Lousberg)
  - Voir http://www.montefiore.ulg.ac.be/~vanlishout/stats.html
- Travail personnel (en deux parties)
  - Enoncés partie (a) postés le 2/10, explications le 4/10, à rendre pour le 31/10.
  - Enoncés partie (b) postés le 23/10, explications le 8/11, à rendre pour le 5/12.
  - Voir http://www.montefiore.ulg.ac.be/~vanlishout/stats.html

Agenda 2016 des leçons et répétitions

20/9	**Leçon 1:** Introduction - Analyse de données exploratoire	(8h15; Amphi 202/B7b)
27/9	**Congé: ni cours, ni répétition**	
4/10	**Leçon 2:** Notion d'échantillon (+ explications travail (a))	(8h15; Amphi 202/B7b)
11/10	**Répétition 1:** Notion d'échantillon	(8h15; S_{xx}, TP physique, selon groupes)
18/10	**Leçon 3:** Estimation I	(8h15; Amphi 202/B7b)
25/10	**Répétition 2:** Estimation I	(8h15; S_{xx}, TP physique, selon groupes)
1/11	**Congé: ni cours, ni répétition**	
8/11	**Leçon 4:** Estimation II (+ explications travail (b))	(8h15; Amphi 202/B7b)
15/11	**Répétition 3:** Estimation II	(8h15; S_{xx}, TP physique, selon groupes)
22/11	**Leçon 5:** Test d'hypothèses I	(8h15; Amphi 202/B7b)
29/11	**Répétition 4:** Test d'hypothèses I	(8h15; S_{xx}, TP physique, selon groupes)
6/12	**Répétition 5:** Test d'hypothèses II	(8h15; S_{xx}, TP physique, selon groupes)
13/12	**Leçon 6:** Test d'hypothèses II (+ QR travail)	(8h15; Amphi 202/B7b)
20/12	**Récapitulatif et Questions & Réponses Examen**	(8h15; Amphi 202/B7b)

Evaluation et supports écrits

- Evaluation
  - Cote pour le travail personnel
    - Evaluation des rapports reçus au premier quadrimestre
    - Pondération: 25% de la cote globale
  - Examen écrit en janvier (et/ou en septembre), compte pour 75% de la cote globale, à savoir
    - 25% théorie (deux questions)
    - 5% projet (une question)
    - 45% exercices (trois exercices)
- Supports écrits
  - Théorie :
    - Slides + lectures sur la page web du cours de Stats : http://www.montefiore.ulg.ac.be/~lwh/Stats/
  - Répétitions :
    - Enoncés + exercices suggérés + solutions finales : http://www.montefiore.ulg.ac.be/~vanlishout/stats.html

Analyse de données exploratoire, par l'exemple

- Le but de l'analyse de données exploratoire est de se familiariser avec la nature des données disponibles dans un tableau **D**, puis de commencer à poser des questions à traiter par les méthodes statistiques.
- Cela consiste d'abord à calculer un certain nombre de valeurs caractéristiques des données, à se poser des questions sur les distributions des valeurs observées, à identifier des valeurs qui paraissent anormales, et à faire un résumé de ces analyses sous la forme de graphiques.
- Nous allons illustrer cela au travers d'exemples 'parlants', en introduisant progressivement les grandeurs 'statistiques' calculées et divers types de représentations graphiques.

Extrait des 182 évaluations de juin 2012 en Probas

Projet 1	Projet 2	Projet 3	Q projet 3	Théorie 1	Théorie 2	Théorie 3	Exercice 1	Exercice 2	Exercice 3
0	0	0	1	2	3	1	1	0	0
7	8	19	0	6	6	2	3	8	3
14	15	19	16	12	17	14	6	5	17
11	2	0	0	6	11	5	6	14	11
18	19	19	5	16	10	10	2	18	15
18	9	17	7	18	18	5	12	4	6
19	20	17	12	16	11	16	8	16	15
16	12	20	6	18	12	17	16	15	13
19	10	14	11	15	7	6	5	19	1
14	18	19	0	4	2	5	8	3	2
0	0	0	0	5	8	1	11	0	0
0	12	20	3	20	20	16	5	18	0
17	18	14	7	20	15	14	12	12	14
11	14	16	0	18	14	14	4	9	5
20	19	19	14	18	13	17	7	12	13
20	16	20	20	16	13	15	17	12	7
20	14	20	13	13	12	4	0	14	15
13	16	16	1	12	12	10	4	3	1
16	16	19	7	10	9	2	18	0	7
20	14	0	0	20	5	0	6	15	0

. . . ($n = 20$ étudiants sur 182, $p = 10$ variables)

Statistiques descriptives uni-variées (1)

On entend par 'statistique uni-variée' une information numérique calculée à partir des valeurs d'une seule des colonnes du tableau **D**. Désignons par $x_i, i = 1, \ldots, n$ les n valeurs correspondantes.

- Moyenne et écart-type d'échantillon de la colonne x

$$m_x = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{et} \quad s_x = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - m_x)^2}.$$

- NB: on utilise les lettres latines m *et* s pour bien distinguer ces notions des notions d'espérance (μ) et de variance σ^2 de variables aléatoires (cf. cours de probas).

Sur nos 20 évaluations de juin 2012 en Probas

	Projet 1	Projet 2	Projet 3	Q projet 3	Théorie 1	Théorie 2	Théorie 3	Exercice 1	Exercice 2	Exercice 3
ind1	0	0	0	1	2	3	1	1	0	0
ind2	7	8	19	0	6	6	2	3	8	3
ind3	14	15	19	16	12	17	14	6	5	17
ind4	11	2	0	0	6	11	5	6	14	11
ind5	18	19	19	5	16	10	10	2	18	15
ind6	18	9	17	7	18	18	5	12	4	6
ind7	19	20	17	12	16	11	16	8	16	15
ind8	16	12	20	6	18	12	17	16	15	13
ind9	19	10	14	11	15	7	6	5	19	1
ind10	14	18	19	0	4	2	5	8	3	2
ind11	0	0	0	0	5	8	1	11	0	0
ind12	0	12	20	3	20	20	16	5	18	0
ind13	17	18	14	7	20	15	14	12	12	14
ind14	11	14	16	0	18	14	14	4	9	5
ind15	20	19	19	14	18	13	17	7	12	13
ind16	20	16	20	20	16	13	15	17	12	7
ind17	20	14	20	13	13	12	4	0	14	15
ind18	13	16	16	1	12	12	10	4	3	1
ind19	16	16	19	7	10	9	2	18	0	7
ind20	20	14	0	0	20	5	0	6	15	0
AVERAGE	13,65	12,60	14,40	6,15	13,25	10,90	8,70	7,55	9,85	7,25
STDEV.P	6,70	5,97	7,42	6,12	5,71	4,68	6,03	5,09	6,33	6,09

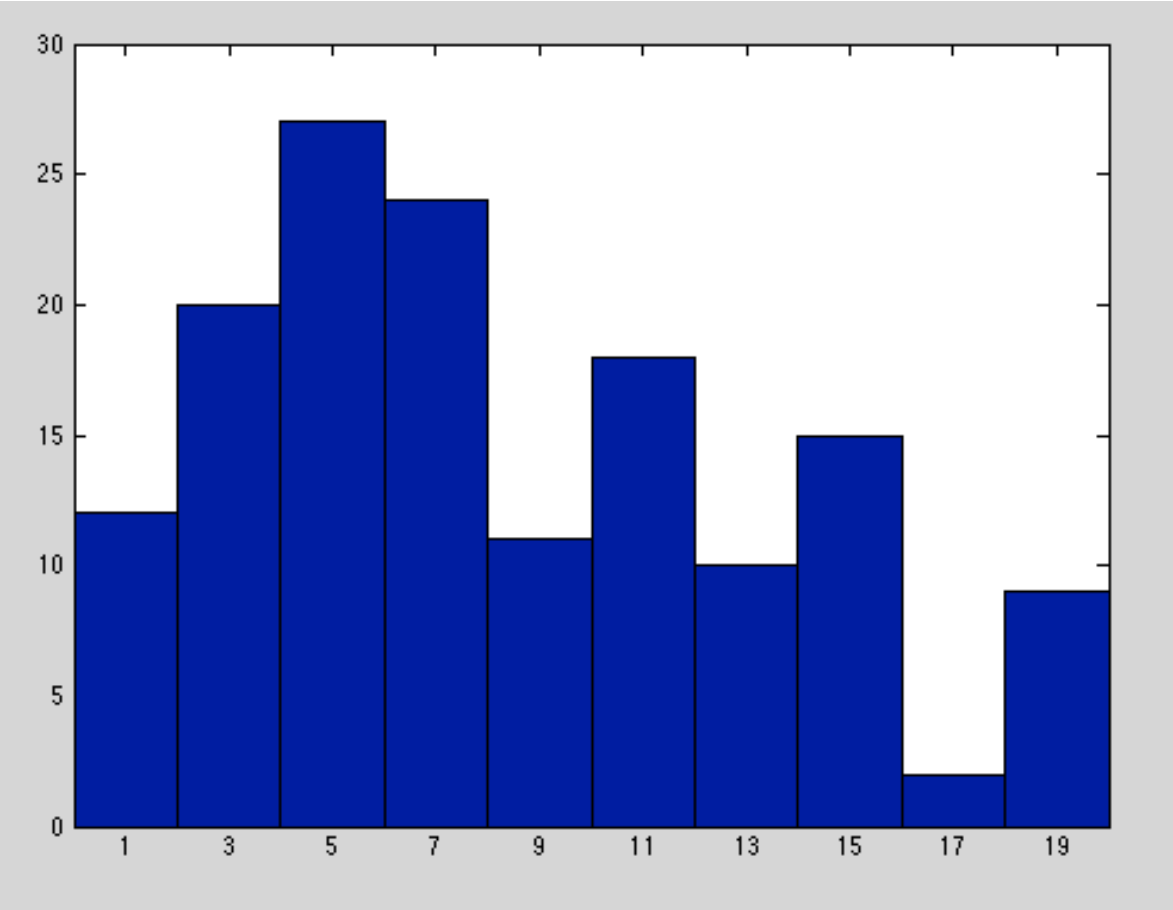
Statistiques descriptives uni-variées (2)

- Fréquence $n_x(y)$ absolue de la valeur y de x: nombre de fois que la valeur y est observée dans la colonne x
- Fréquence relative $\hat{f}_x(y)$ de la valeur y de x: $\hat{f}_x(y) = \frac{n_x(y)}{n}$, proportion des individus ayant la valeur y dans la colonne x.
- Mode de la colonne x: valeur y la plus souvent observée

$$\text{mode}_x = \arg\max_y \hat{f}_x(y) = \arg\max_y n_x(y).$$

- NB: lorsque les valeurs sont relevées avec grande précision, ces grandeurs sont généralement peu utiles; on applique alors au préalable un processus de 'regroupement des valeurs' (i.e. une forme d'arrondi), avant de les calculer.

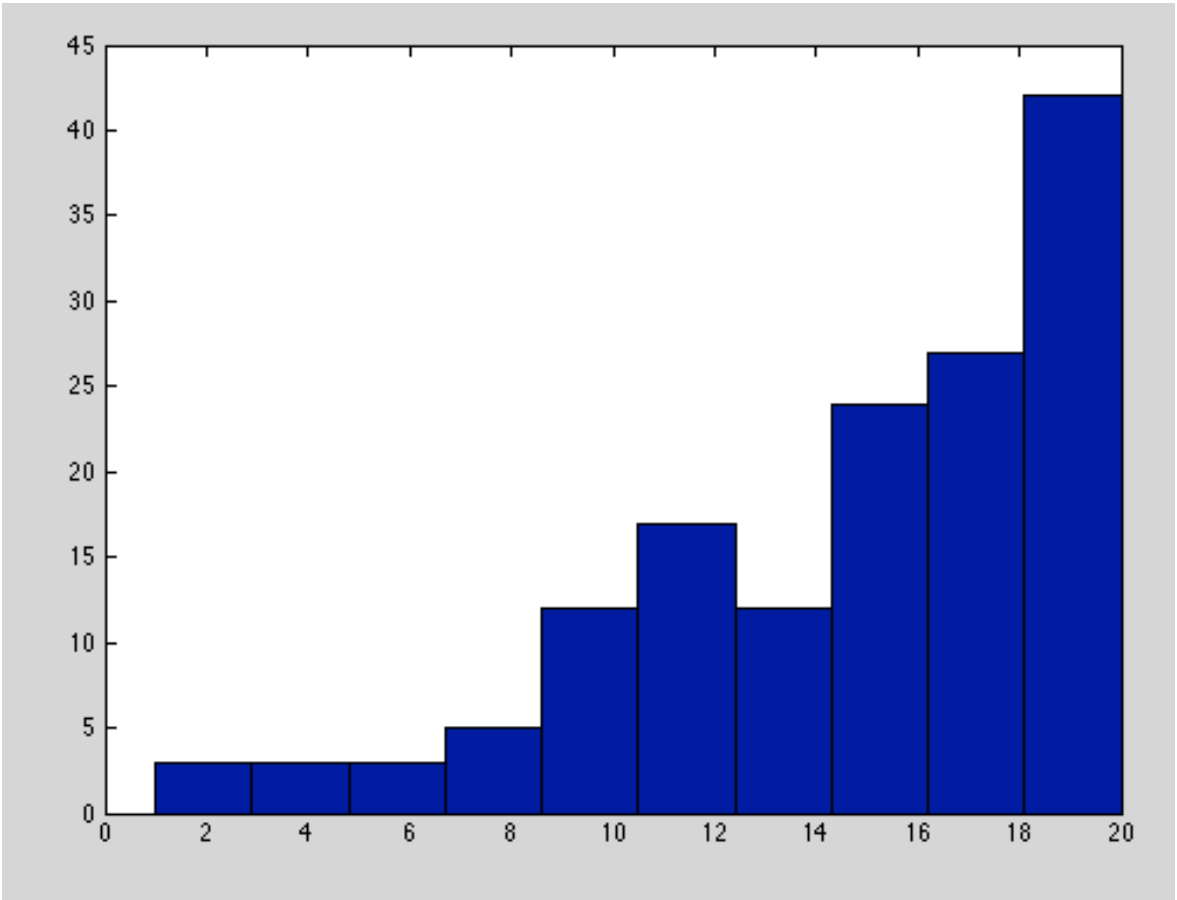
Illustration: diagramme de fréquences absolues (2.1)



Cette figure représente les fréquences absolues des cotes obtenues en juin 2013, pour l'exercice 1 de l'examen écrit, pour les 149 étudiants ayant présenté l'examen écrit.

Les données ont d'abord été regroupées par pas de 2 des valeurs de x.

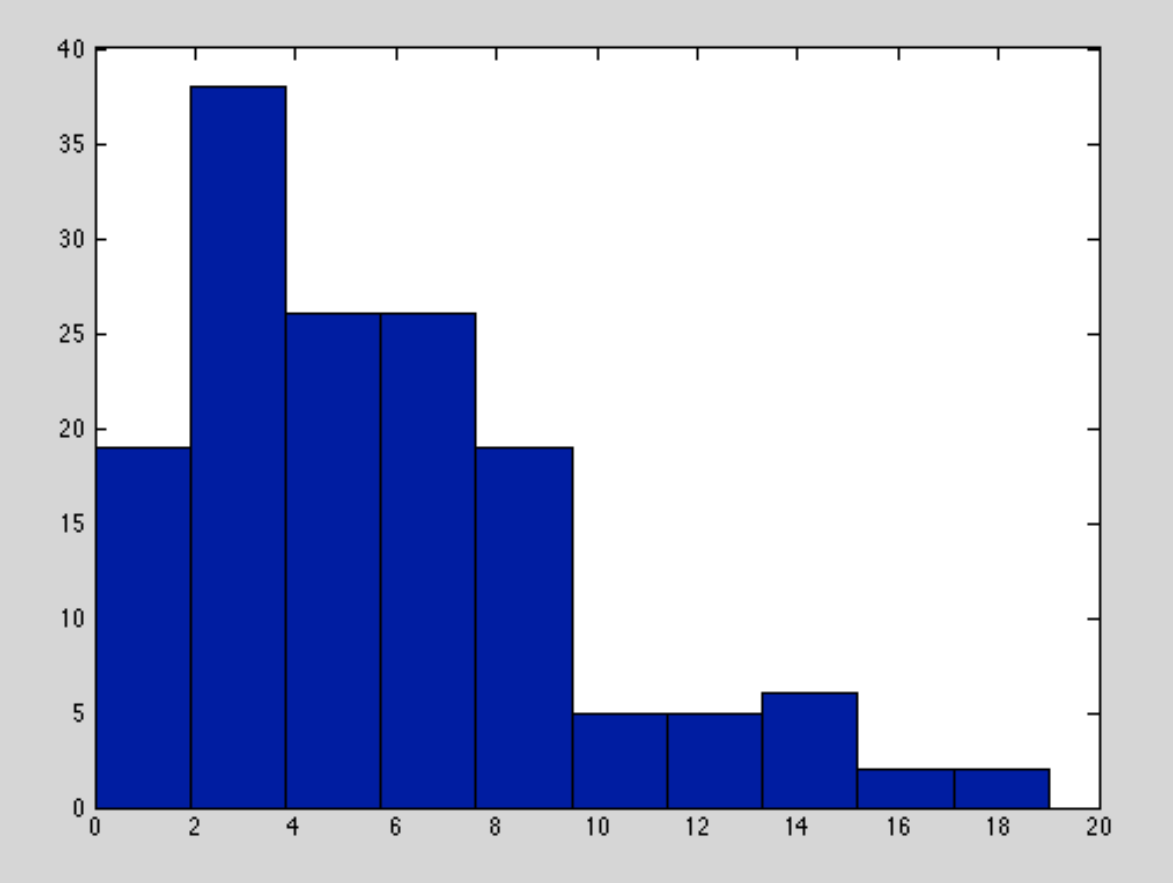
Illustration: diagramme de fréquences absolues (2.2)



Cette figure représente les fréquences absolues des cotes obtenues en juin 2013, pour l'exercice 2 de l'examen écrit, pour les 149 étudiants ayant présenté l'examen écrit.

Les données ont d'abord été regroupées par pas de 2 des valeurs de x.

Illustration: diagramme de fréquences absolues (2.3)



Cette figure représente les fréquences absolues des cotes obtenues en juin 2013, pour l'exercice 3 de l'examen écrit, pour les 149 étudiants ayant présenté l'examen écrit.

Les données ont d'abord été regroupées par pas de 2 des valeurs de x.

Statistiques descriptives uni-variées (3)

- Fréquences relatives cumulées (aussi appelé fonction de répartition empirique) des valeurs observées de x

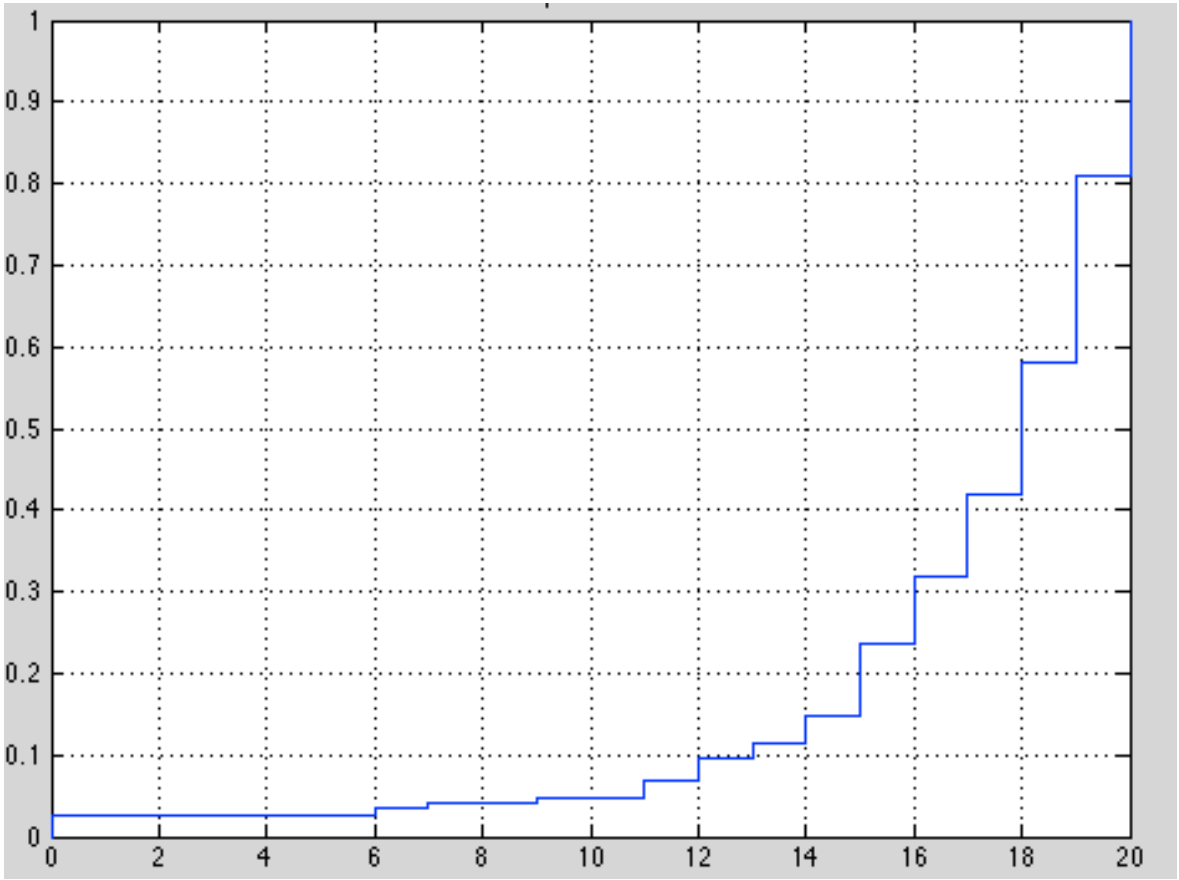
$$\hat{F}_x(y) = \frac{1}{n}\sum_{i=1}^{n} 1(x_i < y).$$

- NB: $\hat{F}_x(y)$ Calcule la proportion des individus présentant une valeur de $x < y$ (cf notre convention de continuité à gauche de la fonction de répartition introduite au cours de probas).
- Médiane

$$\text{Médiane}_x = \hat{F}_x^{-1}(0.5).$$

autrement dit, la valeur de x qui départage les observations en deux parts égales...

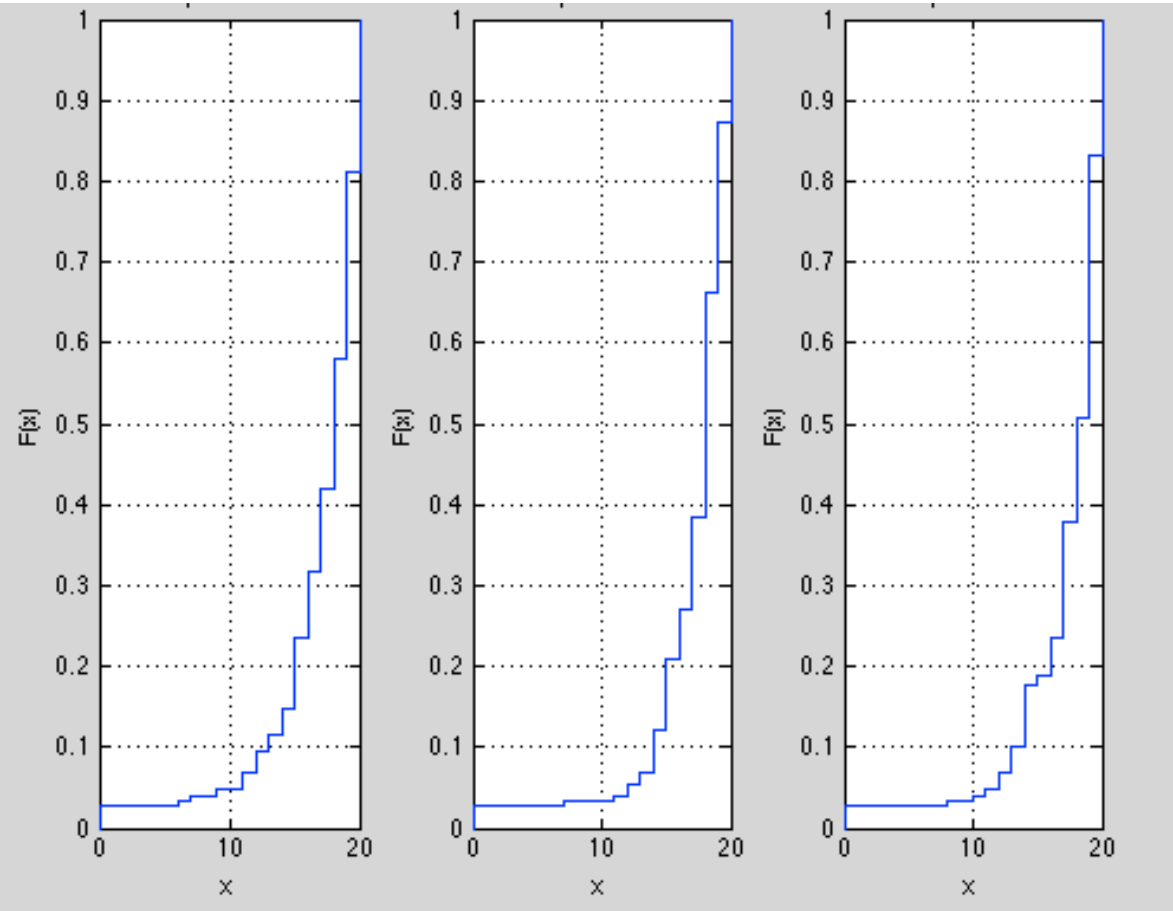
Illustration: fonction de répartition empirique (3.1)



Cette figure représente les fréquences relatives cumulées des cotes obtenues en juin 2013, pour le projet 1, pour les 149 étudiants ayant présenté l'examen écrit.

Les données n'ont pas été regroupées au préalable.

Illustration: fonction de répartition empirique (3.2)



Cette figure juxtapose les fréquences relatives cumulées des cotes obtenues en juin 2013, pour les projet 1, 2 et 3, pour les 149 étudiants ayant présenté l'examen écrit.

Les données n'ont pas été regroupées au préalable.

Statistiques descriptives uni-variées (4)

- Médiane :

$$\text{Médiane}_x = \hat{F}_x^{-1}(0.5).$$

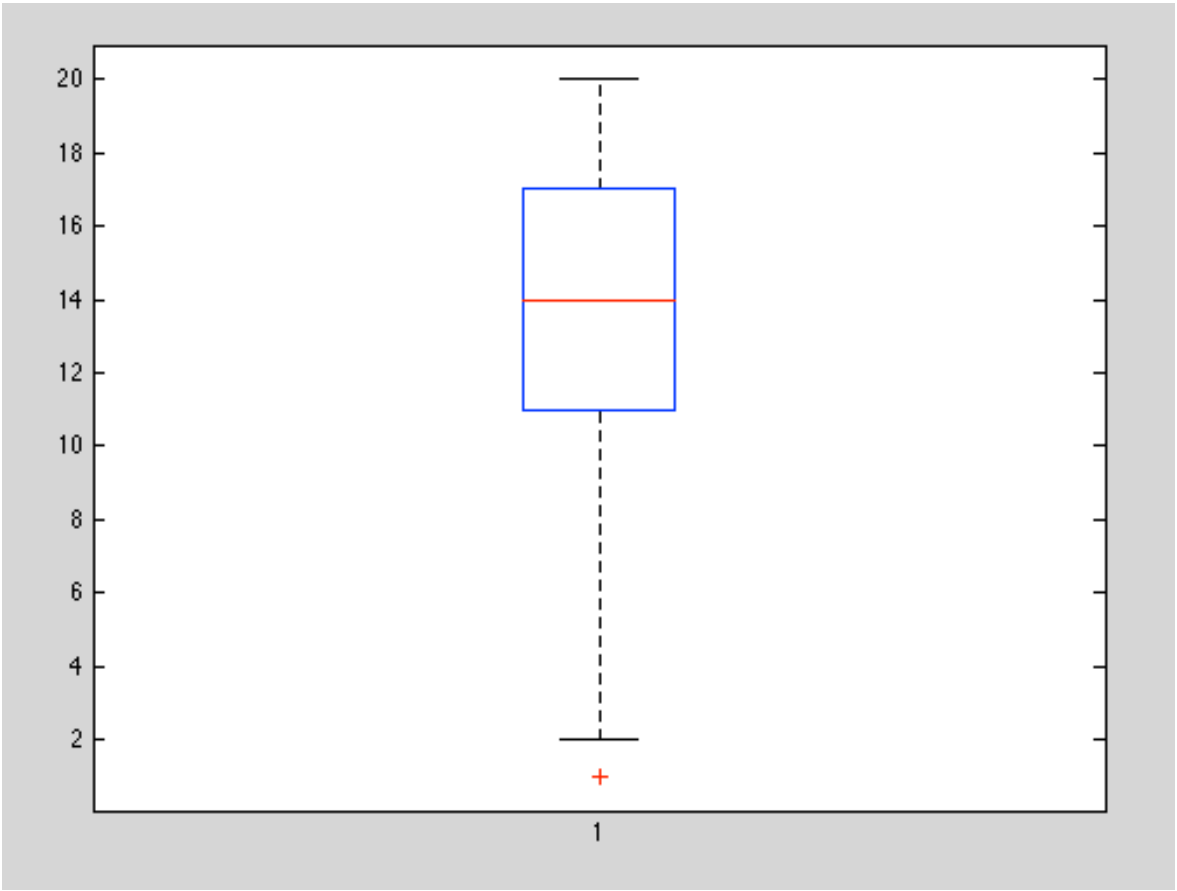
autrement dit, une valeur de x qui départage les observations triées selon x en deux parts égales...

- NB: cette définition est en réalité ambigüe. Il se pourrait en effet, que la valeur 0.5 sur l'axe vertical corresponde exactement à un point de discontinuité vu de l'axe horizontal; dans ce cas la médiane est définie comme étant la moyenne algébrique des deux valeurs définissant la 'marche horizontale' de l'escalier... Ce type de problème se produit seulement si le nombre d'individus est pair.

Statistiques descriptives uni-variées (5)

- Percentiles: la médiane est aussi appelée 'percentile 50' car elle correspond à un seuil tel que 50% des valeurs soient inférieures à ce seuil.
- On définit de façon analogue
  - Le percentile 25, ou premier quartile Q_x^1, par $Q_x^1 = \hat{F}_x^{-1}(0.25)$.
  - Le percentile 75, ou troisième quartile Q_x^3, par $Q_x^3 = \hat{F}_x^{-1}(0.75)$.
  - Et de façon générique, le percentile y par $\hat{F}_x^{-1}(\frac{y}{100})$.
- NB:
  - Mêmes remarques, en ce qui concerne les discontinuités.
  - La médiane, et dans une moindre mesure les quartiles 1 et 3, sont insensibles aux valeurs anormalement élevées des valeurs observées. Les valeurs qui sont inférieures à $Q^1 - 1.5(Q^3 - Q^1)$ ou supérieures à $Q^3 + 1.5(Q^3 - Q^1)$, sont pour cette raison souvent qualifiées d'aberrantes (outliers).

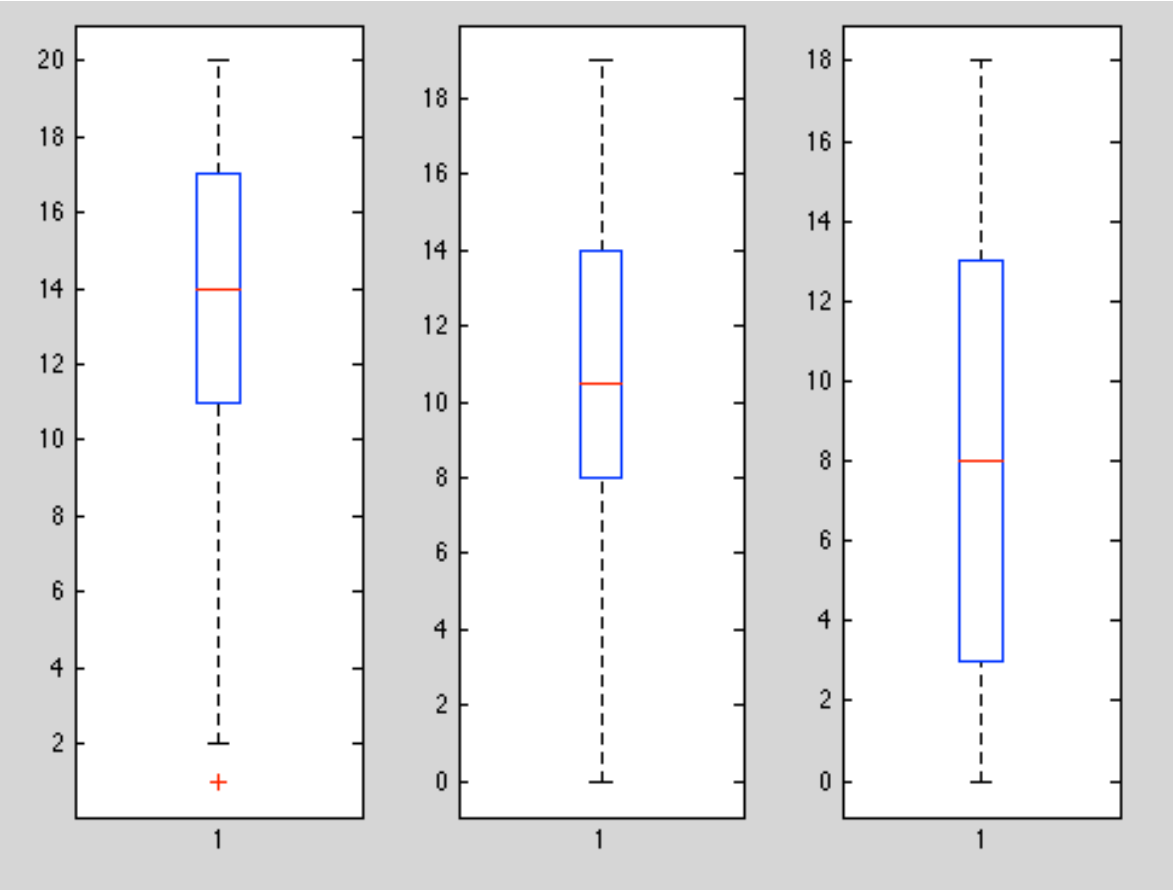
Illustration: boîtes à moustaches (5.1)



Boîte à 'moustaches' (ou box-plot) montrant la distribution des cotes obtenues en juin 2013, pour la question de théorie 1 (149 étudiants).

La ligne en rouge indique la médiane, la boîte en bleu indique l'intervalle inter-quartile délimité par Q^1 et Q^3, les 'moustaches' (i.e. barres horizontales) représentent en l'absence de données aberrantes les valeurs minimale $x_{\mathbf{min}}$ et maximale $x_{\mathbf{max}}$ observées; en cas de présence de données aberrantes (repérées par des $+$, comme dans cet exemple) la 'moustache' est positionnée en $Q^1 - 1.5(Q^3 - Q^1)$ (comme ici) et/ou en $Q^3 + 1.5(Q^3 - Q^1)$.

Illustration: boîtes à moustaches (5.2)



Cette figure juxtapose les box-plots des cotes obtenues par les 149 étudiants, pour les 3 questions de théorie (examen de juin 2013).

Attention: la plage couverte par l'axe vertical est différente pour la figure de gauche et les deux autres.

Statistiques descriptives bi- et multi-variées (1)

- Coefficient de corrélation linéaire entre les valeurs observées dans la colonne x et la colonne y:

$$r_{x,y} = \frac{\sum_{i=1}^{n}(x_i - m_x)(y_i - m_y)}{\sqrt{\sum_{i=1}^{n}(x_i - m_x)^2 \sum_{i=1}^{n}(y_i - m_y)^2}}.$$

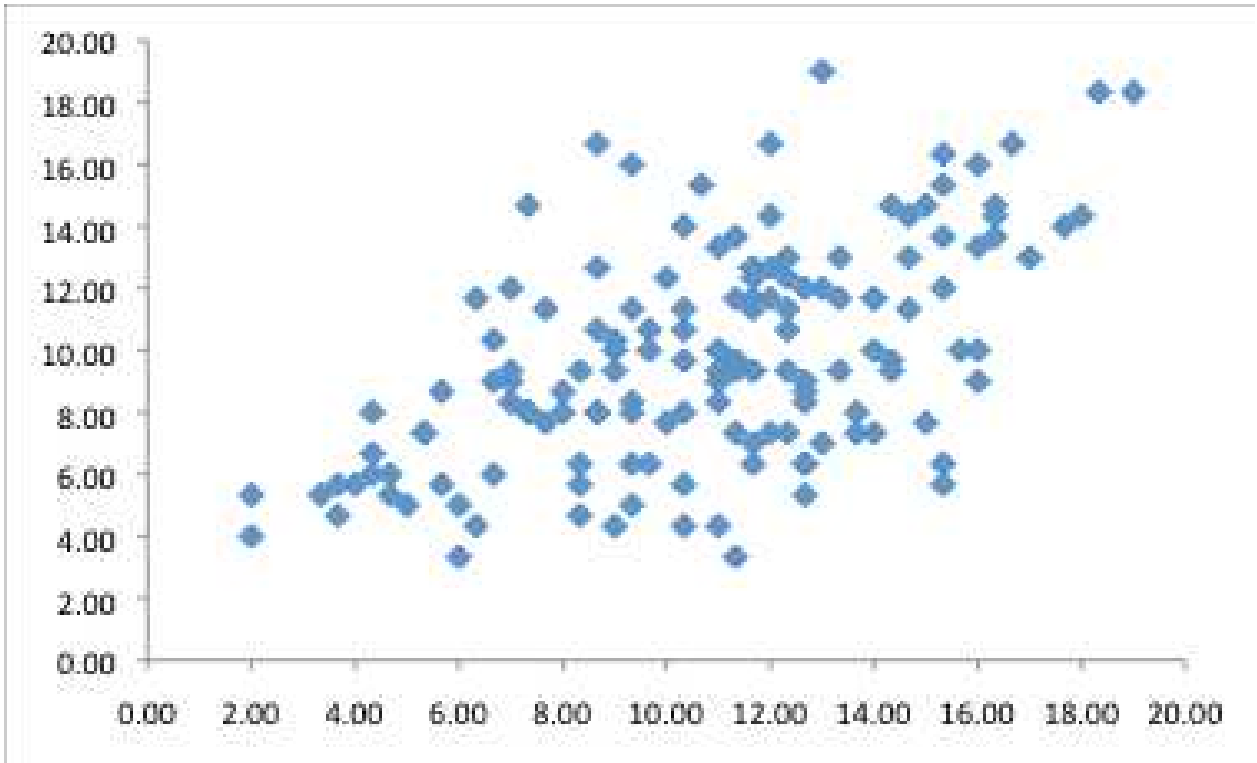
ou bien de façon équivalente, $$r_{x,y} = \frac{1}{n}\sum_{i=1}^{n}\frac{(x_i - m_x)}{s_x}\frac{(y_i - m_y)}{s_y}.$$

- On définit la covariance empirique des deux colonnes par

$$\widehat{cov}(x,y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - m_x)(y_i - m_y) = r_{x,y}s_x s_y.$$

- Remarque: $r_{x,x} = 1$ et $\widehat{cov}(x,x) = s_x^2$.

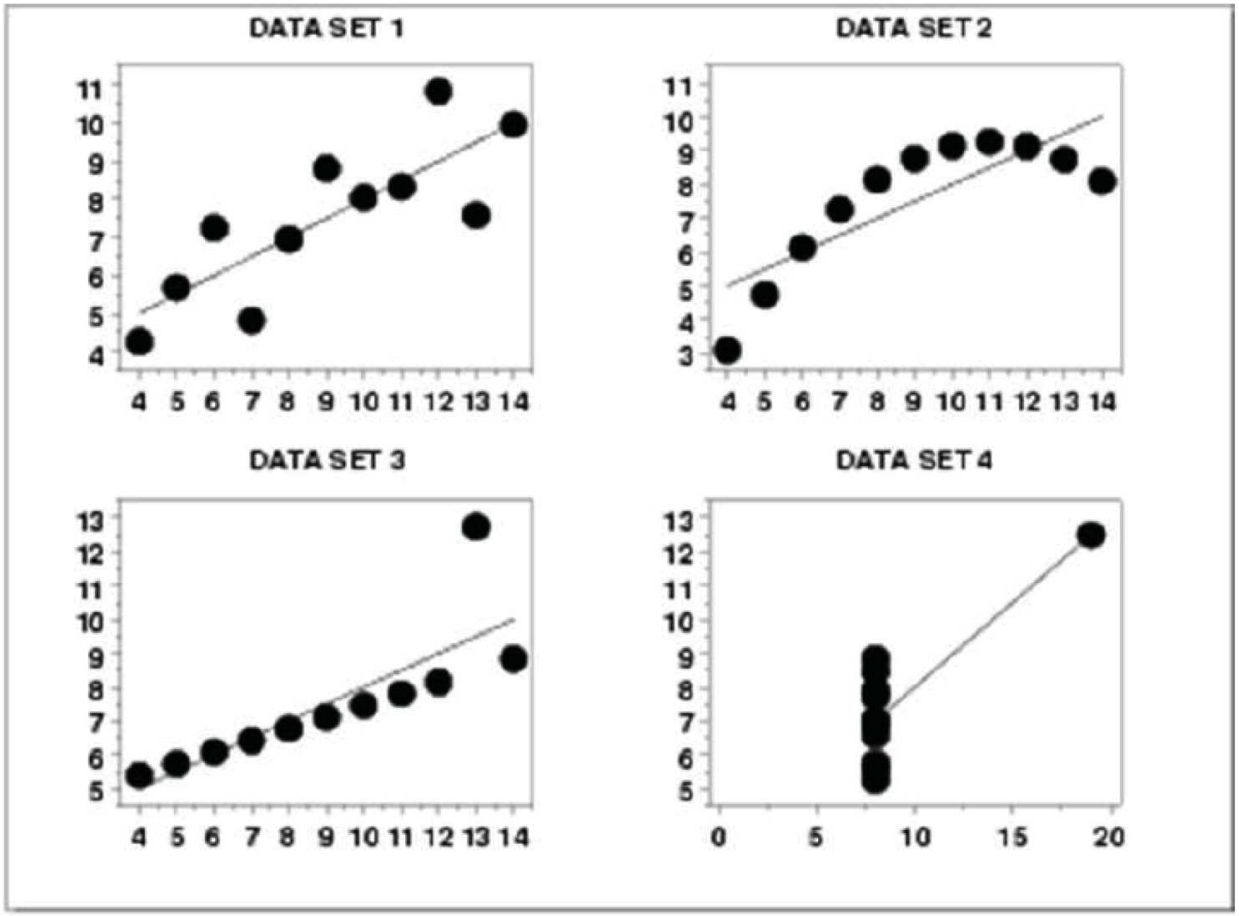
Illustration: un nuage de points (scatter-plot)



Cette figure représente pour les 149 étudiants, en axe horizontal leur moyenne algébrique des questions de théorie, et en axe vertical leur moyenne algébrique des questions d'exercice.

Le coefficient de corrélation entre ces deux moyennes vaut $r_{t,e} = 0.56$.

Illustration: d'autres nuages de points



Cette figure représente 4 nuages de points fort différents.

Cependant, il sont tous les quatre tels que

$$n = 11$$
$$m_x = 9.0 \text{ et } s_x = \sqrt{10}$$
$$m_y = 7.5 \text{ et } s_y = \sqrt{3.75}$$
$$r_{x,y} = 0.82$$

Statistiques descriptives bi- et multi-variées (2)

- Pour étudier plus de deux variables, une première approche consiste à les étudier deux à deux.
- On réalise donc $p(p-1)$ analyses bi-variées.
- On peut ainsi construire un vecteur des moyennes (de dimension p), un vecteur des écart-types (de dimension p), et une matrice (symétrique s.d.p.) des corrélations et/ou des covariances deux-à-deux (de dimension $p \times p$).
- Des techniques d'analyses de données plus sophistiquées existent cependant pour étudier conjointement plusieurs variables (voir suite du cours).

Autres représentations graphiques

- Il y a encore de nombreuses autres façons de représenter de l'information sous forme graphique; nous en illustrerons certaines dans la suite du cours.
- Ces techniques de visualisation sont d'autant plus utiles que les données sont nombreuses (n ou p très élevés).
- Elles aident le statisticien à se faire un premier avis sur la nature des données "à l'oeil".
- Cependant, des conclusions rigoureuses et reproductibles nécessitent une approche quantitative de l'analyse de données.
- Développer les outils de base pour faire cela est l'objectif principal des autres leçons de ce cours.

Propriétés théoriques - 1er pont avec les probas

- On peut définir une expérience aléatoire (discrète) $(\Omega, \mathcal{E}, P)$ à partir d'un tableau **D** de données:
  - les lignes de **D** correspondent aux n résultats possibles de Ω;
  - chaque ligne a une probabilité $\frac{1}{n}$ d'être choisie (P);
  - les variables aléatoires correspondent aux colonnes de **D**.
- Vu sous cet angle, les statistiques (telles que moyennes, écart-types, corrélations, fonctions de répartition ...) que nous avons définies, deviennent alors des caractéristiques du modèle probabiliste (telles qu'espérances, variances, densités, fonctions de répartition ....)
- Nous reviendrons sur ce pont entre statistiques et probabilités, dans la suite de ce cours. Il permet en particulier d'étendre les propriétés que nous avons vues au cours de probas aux grandeurs statistiques que nous venons d'introduire.

Exemples des propriétés qui résultent du pont

Sous la forme de homework pour la leçon suivante :

- Trouver la manière dont les inégalités de Markov et de Bienaymé-Tchebyshev (voir cours de Probas) peuvent s'appliquer à l'analyse d'un tableau de données.
- Montrer que s_x est invariant par translation.
- Montrer que $r_{x,y} \in [-1; 1]$ et que $r_{x,x} = 1$.
- Lire une première fois l'article suivant (voir page web du cours)

Nature Reviews Neuroscience | AOP, published online 10 April 2013; doi:10.1038/nrn3475

ANALYSIS

Power failure: why small sample size undermines the reliability of neuroscience

Katherine S. Button[1,2], John P. A. Ioannidis[3], Claire Mokrysz[1], Brian A. Nosek[4], Jonathan Flint[5], Emma S. J. Robinson[6] and Marcus R. Munafò[1]