Fiche TD 1

Statistique descriptive

Série statistique simple

Dans le domaine des sciences de la vie, la statistique est omniprésente en raison d'une spécificité fondamentale des données biologiques : **la variabilité individuelle** (inter et intra-individuelle). Une des conséquences de la variabilité individuelle est la **fluctuation d'échantillonnage**.

L'analyse statistique permet de dépasser le désordre apparent des jeux de données créé par la variabilité individuelle. Elle donne des résultats moyens correspondant à la normalité et évalue le risque d'erreur lié à la fluctuation d'échantillonnage en rappelant explicitement son existence dans l'énoncé des conclusions

Objectifs pédagogiques :

1) Maîtriser le langage de base en statistique
2) Organiser les données sous forme de tableaux, de graphes
3) Calculer les paramètres de statistique descriptive
4) Réaliser l'analyse statistique descriptive avec le *logiciel R*

1. Le langage en statistique

Quelques fondamentaux à connaître (cf le glossaire pour plus de détail).
La variable est la caractéristique spécifique observées/mesurées. Elle constitue la série statistique à analyser. Il est important de déterminer la nature de la variable (qualitative ou continue) car elle détermine dans certain cas le choix de l'outil statistique.

Variable qualitative ou quantitative ?

- Une variable quantitative est mesurable.
- Une **variable qualitative** (synonyme « caractère ») peut prendre plusieurs valeurs appelées **modalités**. La variable qualitative est soit **nominale**, soit **ordinale** lorsque ses modalités présentent une hiérarchie. Par définition une variable qualitative est discrète.

Exemple : la couleur du pelage d'un animal (nominale), plus ou moins foncé (ordinale).

Variable quantitative discrète ou continue ?

- Une variable quantitative continue est analogique
  Exemple : la taille des embryons
- Une variable quantitative discrète est numérique
  Exemple : le nombre d'embryons par portée

Variable dépendante ou indépendante ?

Une variable indépendante est une variable explicative. Dans certain cas elle est parfaitement contrôlée

Exemple : la température d'incubation

Une variable dépendante est une variable expliquée.

Exemple : la croissance bactérienne

L'individu (synonyme « observation ») désigne l'unité statistique, sur laquelle les variables seront observées/mesurées.

La population est l'ensemble des individus étudiés. Elle peut être finie (effectif noté N) ou infinie. En général, il est impossible d'étudier la population dans sa totalité (infinie, trop long, trop coûteux, mesures destructives).

L'échantillon est un sous-ensemble de la population. Il représente correctement la population s'il est constitué de façon aléatoire et si son effectif, noté n, est suffisamment grand (n>30) pour éviter la fluctuation d'échantillonnage.

L'analyse statistique est réalisée dans le cadre d'une **hypothèse théorique** qui mentionne la variable indépendante, la variable dépendante et éventuellement le sens de la relation (hypothèse forte ou faible débouchant sur un test uni ou bilatéral).

La statistique descriptive (ou statistique exploratoire) permet d'organiser, de visualiser et de résumer l'information contenue dans de volumineux jeux de données.

2. Les tableaux : mettre de l'ordre dans les données

Tableau d'effectifs

Soit un échantillon d'effectif n. Chaque valeur x_i de la variable, on note n_i le nombre d'individus qui présente cette valeur dans l'échantillon. Si la variable est continue, il est nécessaire au préalable de définir des classes, c'est à dire un intervalle $[e_{i-1} ; e_i[$, défini par ses bornes inférieure et supérieure, son amplitude ($h_i=e_i - e_{i-1}$) et son centre de classe ($C_{i=} e_i + (e_i - e_{i-1})/2$).

Tableau de fréquences

Pour chaque valeur x_i de la variable, on note sa fréquence ($f_i = n_i / n$) dans l'échantillon.

Tableau des effectifs et des fréquences

Valeur de variable (x_i)	Effectif (n_i)	Fréquence (f_i)
x_1	n_1	f_1
x_2	n_2	f_2
…	…	…
x_n	n_n	f_n

Remarque : On peut s'intéresser aux **fréquences cumulées croissantes** jusqu'à une valeur x_i de la variable, la somme des fréquence de la classe x_i et de toutes les classes inférieures à x_i.

(Le tableau de contingence sera présenté en TD3)

3. Représentations graphiques : visualiser les données

Histogramme pour les variables continues

L'histogramme permet de visualiser la distribution des données. Chaque classe est représentée par un rectangle dont l'aire est proportionnelle à son effectif. A partir de l'histogramme, en joignant les centres des différentes classes par des segments, on obtient une ligne brisée appelée **le polygone de fréquences**. La courbe limite s'appelle **densité de probabilités**. On peut également tracer le **graphe de répartition** correspondant au graphe des fréquences cumulées croissantes.

Diagramme en boîte ou box-plot (J.W. Tukey, 1977)

On utilise un diagramme en boîte (box-plot) en représentation verticale ou horizontale. Les côtés de chaque boîte correspondent aux quartiles Q1 et Q3 et la ligne centrale indique la médiane. Les extensions (1.5 fois l'écart inter-quartile) montrent la dispersion des données. Les observations dont les valeurs vont au delà de ces limites sont considérées comme aberrantes (« outliers » en anglais). Cette représentation permet de comparer plusieurs séries statistiques.

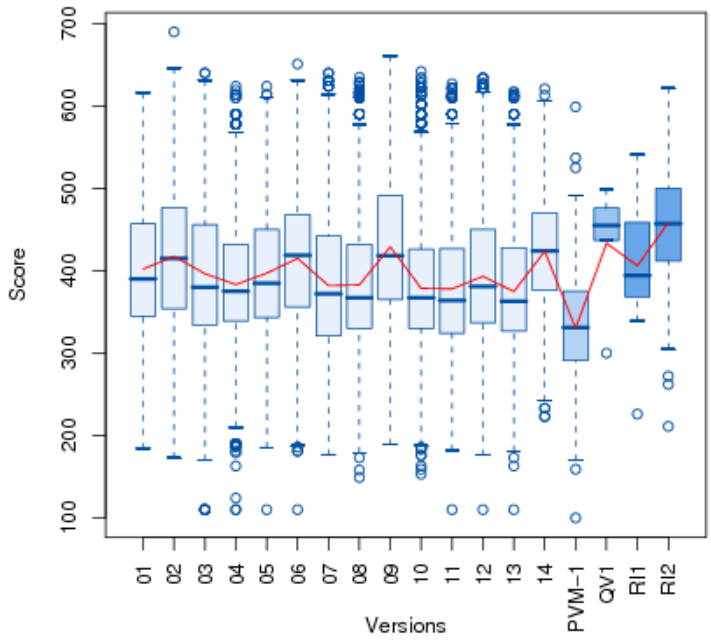
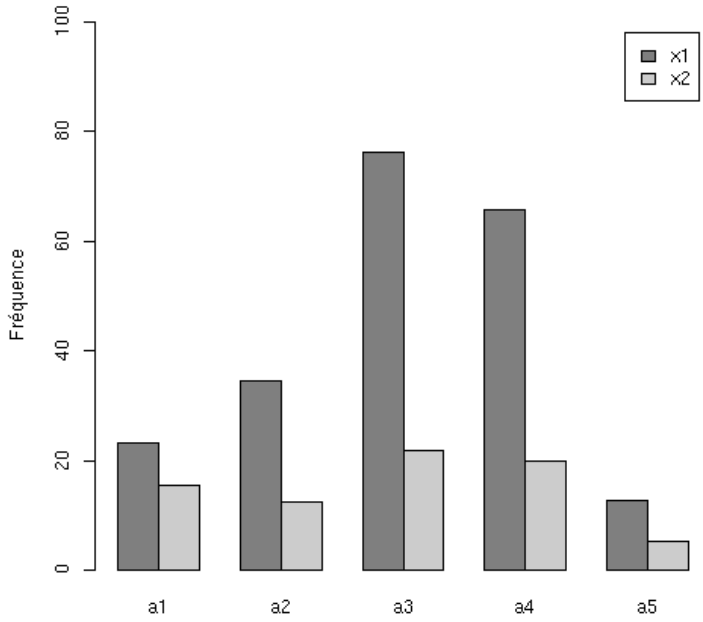


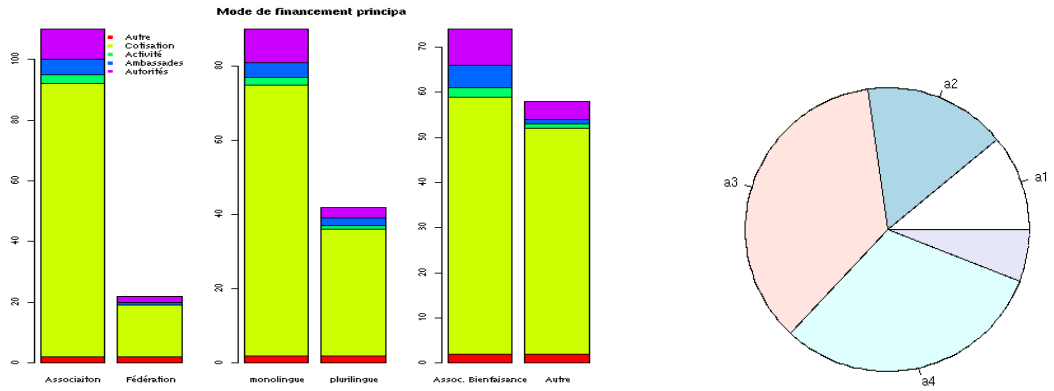
Diagramme en bâtons pour les variables quantitatives discrètes

Le diagramme en bâtons illustre la discontinuité de la variable. Il est réalisé en portant en abscisse les valeurs observées des modalités et en ordonnée l'effectif (ou la fréquence/le pourcentage). Les hauteurs des bâtons sont proportionnelles aux effectifs ou aux fréquences.



Diagrammes pour les variables qualitatives

Les diagrammes circulaires (camembert), à bandes ou en barres. Chaque modalité est représentée par un secteur ou un rectangle dont l'aire est proportionnelle à son effectif.



(source des figures - http://www.aliquote.org/articles/tech/r-graphics/index.html.old)

4. Paramètres de distribution : résumer une série statistique

L'information d'un jeu de données sera résumée par des indicateurs appelés paramètres de distribution. On distingue les paramètres de position et des paramètres de dispersion.

Les paramètres de position sont des indicateurs de la tendance centrale de la population : moyenne arithmétique, médiane, mode.

La moyenne arithmétique se calcule en effectuant la somme des valeurs des variables de chaque individu divisées par l'effectif de l'échantillon. Sa valeur est très sensibles aux valeurs extrêmes.

La médiane d'une série statistique est une valeur de la variable, telle qu'il y ait autant d'observations ayant une valeur supérieure à la médiane que d'observations ayant une valeur inférieure à la médiane. Elle partage donc l'échantillon en deux sous-ensembles de même effectif.

Le mode est la valeur que la variable prend le plus fréquemment, c'est à dire celle qui correspond à l'effectif maximum.

Les quartiles partagent la série statistique ordonnée en quatre sous-ensembles de taille égale. Le premier quartile noté Q1 est la valeur telle qu'au moins 25% des données soient inférieures ou égales à cette valeur. Le 2ème quartile correspond à la médiane (Q2 = Me). Le 3ème quartile noté Q3 est la valeur telle qu'au moins 75% des données soient inférieures ou égales à cette valeur.

Remarques : Les déciles partagent la série statistique ordonnée en 10 sous-ensembles de taille égale. Les centiles partagent la série statistique ordonnée en 100 sous-ensembles de taille égale.

Les paramètres de dispersion permettent de mesurer la dispersion des valeurs prises par la variable autour de la valeur centrale.

La variance est la moyenne arithmétique des carrés des écarts à la moyenne.

L'écart-type (« standard deviation » en anglais) est la racine carrée de la variance.

Le coefficient de variation est le rapport entre l'écart-type et la moyenne d'une série ; il est souvent exprimé en pourcentage.

L'écart interquartile (Q3 – Q1) permet d'apprécier la dispersion de la moitié des valeurs qui entourent la médiane.

Remarque importante :
Les indicateurs (médiane - intervalle interquartile) sont robustes par rapport aux valeurs extrêmes contrairement aux indicateurs (moyenne – écart-type).

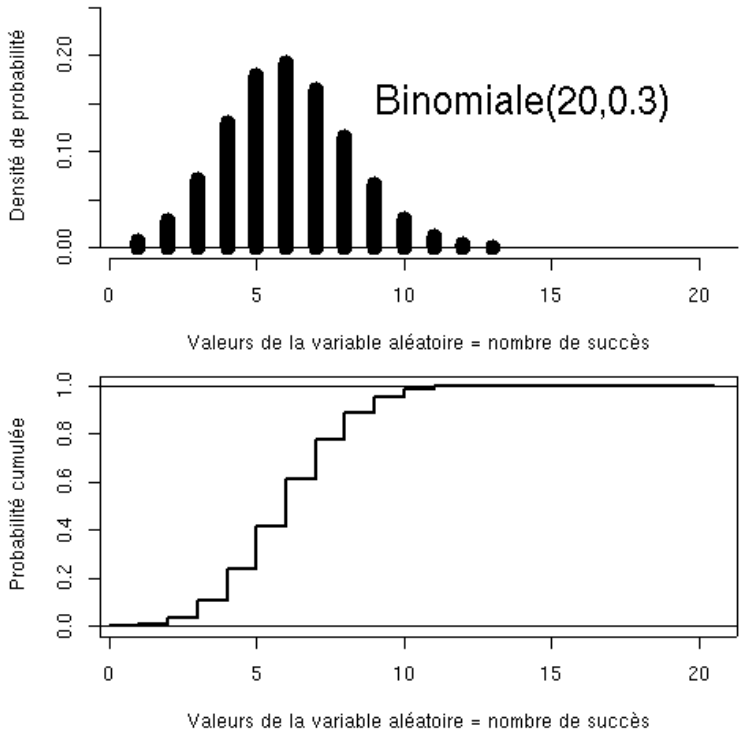
5. Distribution de probabilités

Loi binomiale (variable aléatoire discrète)

Une épreuve de Bernoulli est une expérience aléatoire qui ne donne que deux résultats (succès, échec) avec : **p+q = 1** (p la probabilité de succès et q la probabilité d'échec)
Les expériences de Bernoulli sont indépendantes : répétées à l'infini, le résultat de l'une n'influe pas sur celui des autres. Une **chaîne de Bernoulli** est une épreuve de Bernoulli répétée n fois. Le **coefficient binomial** correspond aux nombre de chaînes de Bernoulli (n). La variable aléatoire X comptabilisant le nombre de succès sur les n épreuves suit une loi binomiale de paramètre n et p. Cette loi est notée B(n,p). La probabilité d'avoir x succès parmi n est :

$$f(x) = \frac{n!}{x!(n-x!)} p^x (1-p)^{n-x}$$

Moments d'une loi binomiale *B(n,p)*
E(*X*) = n . p
V.X) = n . p . q



(source - http://w3.jouy.inra.fr/)

Remarque : On peut représenter une distribution soit par sa **fonction de densité,** soit par sa **fonction de répartition**. La densité renvoie la probabilité d'appartenance à un segment (=surface sous la courbe pour ce segment). La somme sous toute la courbe de densité est 1.
La fonction de répartition renvoie la probabilité que la variable soit inférieure à l'abscisse ; elle tend asymptotiquement vers 1.

Loi de Poisson (variable aléatoire discrète)

La loi de Poisson est le modèle probabiliste des situations qui voient un flux d'événements se produire à la suite des autres de façon aléatoire. C'est la loi des « événements rares ». Le résultat ne dépend que de l'état à l'instant t et pas de ce qui s'est passé avant.

Une variable aléatoire x suit une loi de Poisson de paramètre λ (λ>0) si :

$$f(x;\lambda)=\begin{cases}\lambda e^{-\lambda x} & ,x\geq 0\\ 0 & ,x<0\end{cases}$$

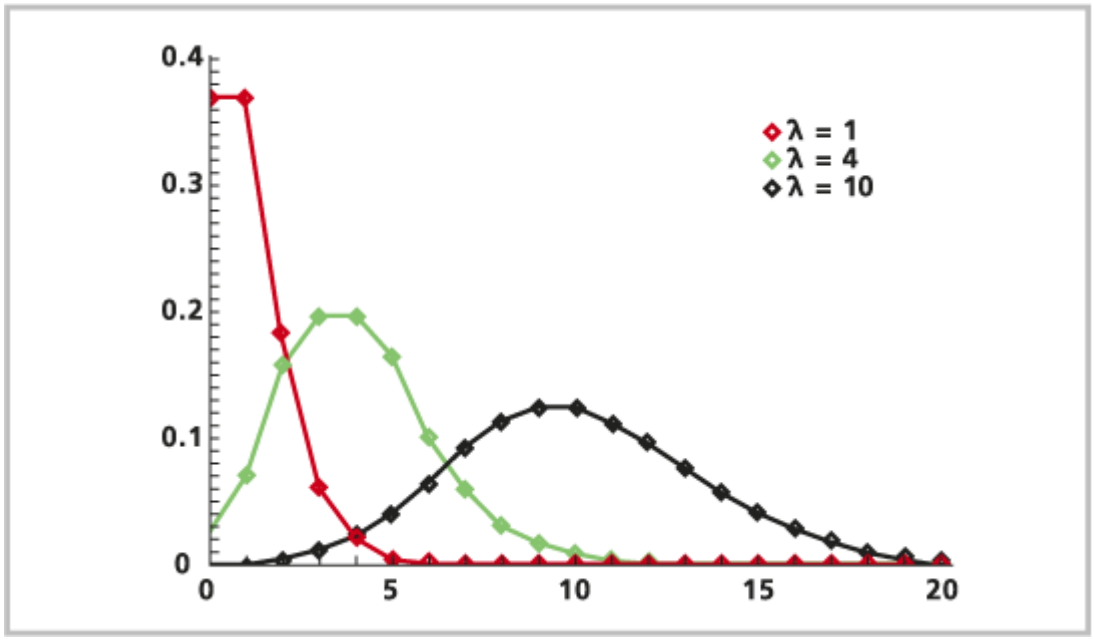
Moments d'une loi de Poisson

E(X) = V(X) = λ

Remarque :

Si λ<1, la fonction de densité est dissymétrique.

Plus λ est grand, plus la fonction de densité est symétrique.



(source - http://resources.arcgis.com/)

Loi de Gauss ou normale (variable aléatoire continue)

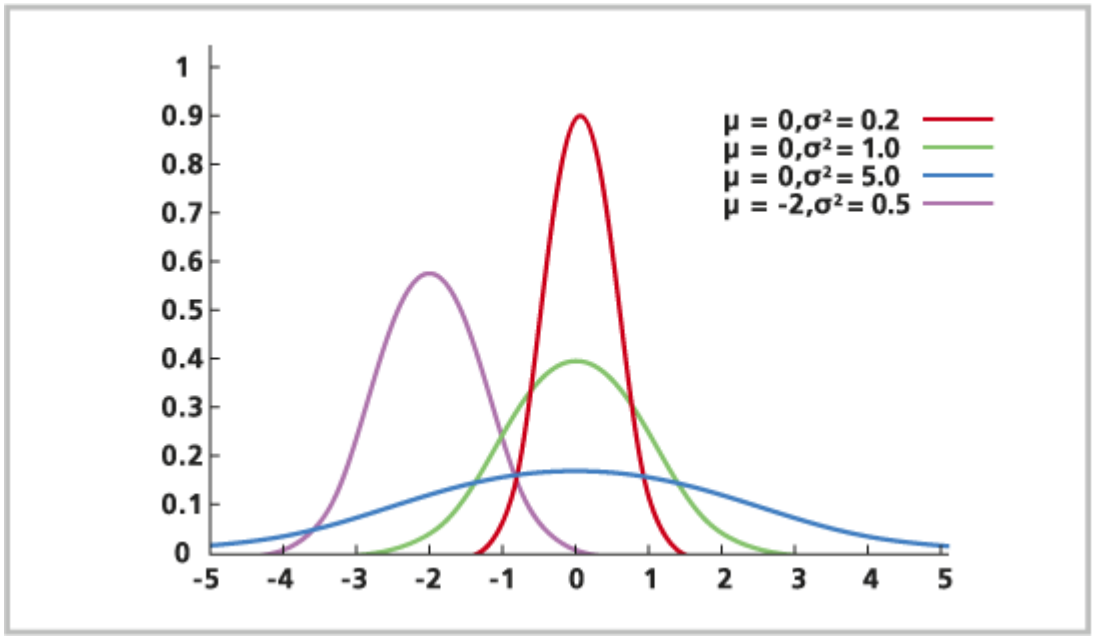
La distribution de probabilité continue peut être utilisée pour de nombreux phénomènes que l'on rencontre dans la nature. La question fondamentale est : les valeurs mesurées dans une expérience sont-elles celles d'une variable aléatoire qui suit une loi normale ?

Une variable aléatoire X de moyenne μ et d'écart type σ suit une loi normale si sa densité de probabilité est : $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

Moments de la loi normale standard

E(X) = 0

V(X) = 1

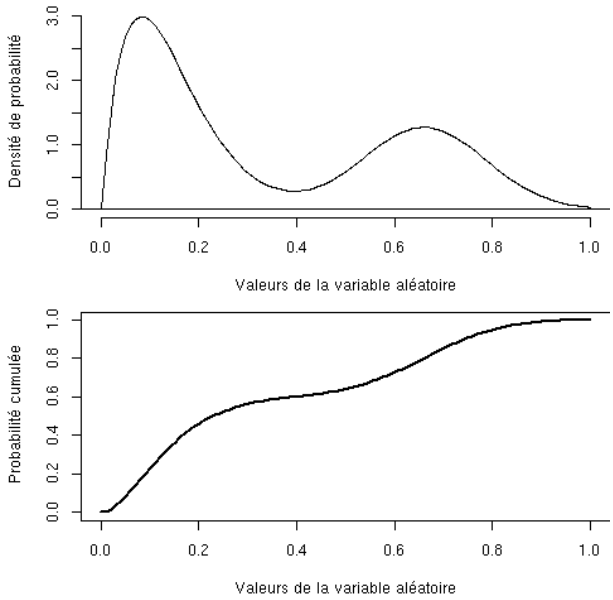


(source - http://resources.arcgis.com/)

Pour valider l'hypothèse d'une distribution normale, on peut utiliser le test de Kolmogorov-Smirnov ou Shapiro & Wilk selon la taille de l'échantillon.

Distribution multimodale (variable continue)

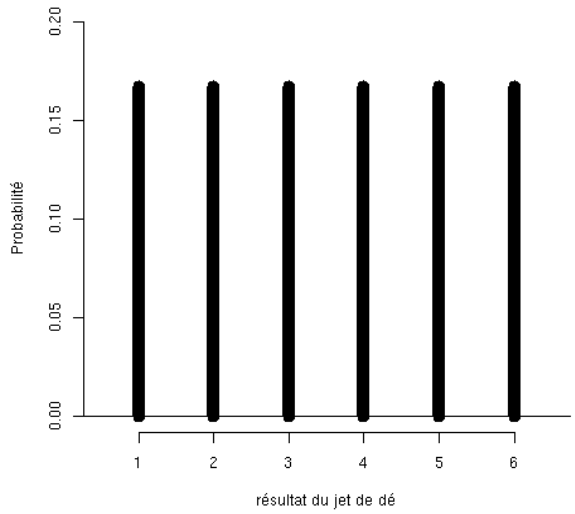
Une variable aléatoire continue suit une loi multimodale lorsque la densité de probabilités montre un ou plusieurs maximum locaux.



(source - http://w3.jouy.inra.fr/)

Distribution uniforme (variable discrète ou continue)

Une distribution est uniforme lorsque la variable aléatoire X peut prendre n'importe quelle valeur entre une borne minimale et une borne maximale avec la même probabilité. Par exemple, le jet d'un dé non truqué présente une distribution uniforme, car la probabilité d'apparition de l'une des six faces est équiprobable.



(source - http://w3.jouy.inra.fr/)

Références

A. Linderberg et I. Wagner (2007) *Les stats en bulles*, Collection Pearson Education