Analyse des données et argumentation numérique

Initiation à la Recherche
Master 1 ISiDIS

Sébastien Verel
verel@lisic.univ-littoral.fr
http://www-lisic.univ-littoral.fr/~verel

Université du Littoral Côte d'Opale
Laboratoire LISIC
Equipe OSMOSE

27 septembre 2017

Plan

Observer

- Les sens et la mémoire humaine ne sont pas toujours fiables :
  - Pas de référence, ni d'échelle absolue (mesure de la température)
  - Observation rapide (< 100 ms) ou trop longue (mouvement bâtiment)
  - Fiabilité de l'observation (illusion d'optique, sensorielle, etc.), impression biaisée, etc.
  - Impossibilité d'observer (ultra-son, rayon-X, etc.)
  - Beaucoup d'observations
  - Parfois quand même, analyse d'image, reconnaissance de "forme", odorat, capacité de synthèse des sens (observation critique), etc.

Observer

- Les sens et la mémoire humaine ne sont pas toujours fiables :
  - Pas de référence, ni d'échelle absolue (mesure de la température)
  - Observation rapide (< 100 ms) ou trop longue (mouvement bâtiment)
  - Fiabilité de l'observation (illusion d'optique, sensorielle, etc.), impression biaisée, etc.
  - Impossibilité d'observer (ultra-son, rayon-X, etc.)
  - Beaucoup d'observations
  - Parfois quand même, analyse d'image, reconnaissance de "forme", odorat, capacité de synthèse des sens (observation critique), etc.

- Utilisation d'une mémoire papier ou électronique pour récolter les observations
- Utilisation d'instruments de mesure

Observations de la vie courante

Observer, mesurer

- Un instrument de mesure parfaitement fiable n'existe pas.
- Le phénomène observé peut être soumis à des variabilités (conditions expérimentales)

Observations de la vie courante

$23^o C$ ou $24^o C$ ou ... ?

Observer, mesurer

- Un instrument de mesure parfaitement fiable n'existe pas (cause exogène)
- Le phénomène observé peut être soumis à des variabilités aléatoires (cause endogène)

$\Rightarrow$ Mesures "entachées" d'erreurs ou de variabilités aléatoires

Observations de la vie courante

$23^o C$ ou $24^o C$ ou ... ?

Observer, mesurer

- Un instrument de mesure parfaitement fiable n'existe pas (cause exogène)
- Le phénomène observé peut être soumis à des variabilités aléatoires (cause endogène)

$\Rightarrow$ Mesures "entachées" d'erreurs ou de variabilités aléatoires

Problème

Comment faire pour connaitre la température dans la pièce ?

Moyenne empirique

Moyenne empirique

Echantillon de n observations $\{x_1, \ldots, x_n\}$

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

Modélisation par une loi de probabilité

Variable aléatoire

Une **variable aléatoire** X est une fonction définie sur l'ensemble des éventualités.

Loi de probabilité

Une v.a. est décrite par sa loi de probabilité qui mesure, pour chaque valeur possible de X, la probabilité pour que la v.a. X prenne cette valeur.

ex : loi uniforme, loi normale, loi binomiale, etc.

Exemple

La mesure de la température par un thermomètre i peut être modélisée par une v.a. X_i suivant une même loi de probabilité F. L'observation x_i d'une température s'appelle une **réalisation** de la variable aléatoire X_i.

Estimation de la moyenne

La température affichée par un thermomètre i peut être modélisée par une v.a. X_i suivant une loi de probabilité F telle $E[F] =$ température de la pièce.

"Connaître" la température de la pièce
revient à
estimer l'espérance (moyenne) de la loi de probabilité F.

Rappel $E[X] = \sum_k P(X = k).k$

Estimation de la moyenne

La température affichée par un thermomètre i peut être modélisée par une v.a. X_i suivant une loi de probabilité F telle $E[F] =$ température de la pièce.

"Connaître" la température de la pièce
revient à
estimer l'espérance (moyenne) de la loi de probabilité F.

Rappel $E[X] = \sum_k P(X = k).k$

Estimation à partir des observations x_i

Espérance $\mu = E[F]$ de la loi F estimée par la moyenne empirique :

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

$\hat{\mu}$ est une estimation de la moyenne μ

Estimation de la moyenne

La température affichée par un thermomètre i peut être modélisée par une v.a. X_i suivant une loi de probabilité F telle $E[F] =$ température de la pièce.

"Connaître" la température de la pièce
revient à
estimer l'espérance (moyenne) de la loi de probabilité F.

Rappel $E[X] = \sum_k P(X = k).k$

Estimation à partir des observations x_i

Espérance $\mu = E[F]$ de la loi F estimée par la moyenne empirique :

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

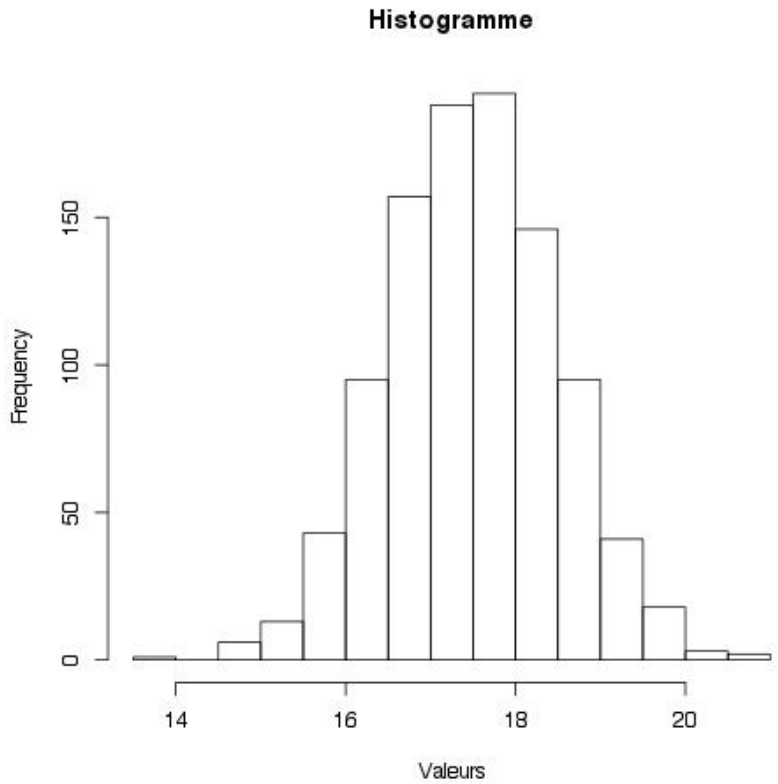
$\hat{\mu}$ est une estimation de la moyenne μ

Théorie des sondages, loi des grands nombres ($\hat{\mu}_n$ converge vers $E[X]$)

Décrire une distribution

Distribution

Une **distribution** de probabilité est l'énumération de tous les "cas" possibles avec leur probabilité respective.



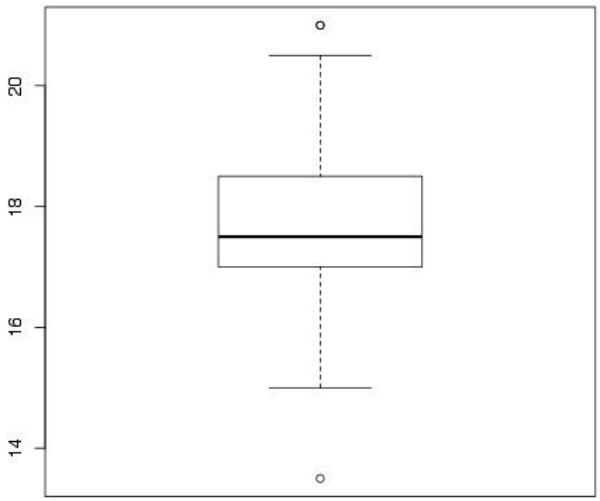
Histogramme

Une distribution peut-être estimée à l'aide d'un **histogramme** *i.e.* la fréquence d'apparition de chaque observation.

Décrire une distribution

Histogramme décrit par le minimum, le maximum, la médiane, les quartiles, la moyenne, l'écart-type, etc.

Boite à moustache :



- Cercles : données aberrantes (outliers),
- De bas en haut : min, premier quartile, médiane, troisième quartile, max.

Estimateur de la variance

Variance et écart-type (standard deviation)

Echantillon de n observations $\{x_1, \ldots, x_n\}$

Variance :

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

Ecart-type :

$$\sigma = \sqrt{\sigma^2}$$

Estimateur de la variance

Variance et écart-type (standard deviation)

Echantillon de n observations $\{x_1, \ldots, x_n\}$

Variance :

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

Ecart-type :

$$\sigma = \sqrt{\sigma^2}$$

Estimateur (non-biaisé) de la variance d'une loi

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

Loi normale

Loi normale ou loi gaussienne

- Loi adaptée à modéliser de nombreux phénomènes naturels aléatoires
- Complètement caractérisée par la moyenne μ et la variance σ^2
- Densité :

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

- Notation :

$$X \sim \mathcal{N}(\mu, \sigma^2)$$

Intervalle de Confiance

Principe

IC est un intervalle (calculé à partir d'observations) utilisé pour indiquer la qualité d'une estimation, la marge d'erreur d'une estimation.

Intervalle de Confiance pour l'estimation de la moyenne

Définition

L'intervalle de confiance *IC* de niveau α est intervalle de valeurs qui a $\alpha\%$ chance de contenir la moyenne à estimer.

Pour l'estimation de la moyenne à partir de *n* observations :

$$IC = \left[\hat{\mu} + t_\alpha \frac{\hat{\sigma}}{\sqrt{n}}, \hat{\mu} - t_\alpha \frac{\hat{\sigma}}{\sqrt{n}}\right]$$

où :

- $\hat{\mu}$ est la moyenne observée
- $\hat{\sigma}$ est l'écart-type observé
- t_α est un réel qui dépend du niveau de confiance : pour $\alpha = 95\%$, $t_\alpha = 1.96$

Logiciel R

Historique

- `R` est un logiciel libre (projet GNU), open source
- Langage et environnement dédié au traitement statistique et à la représentation graphique des données
- Clone du logiciel $S+$ (qui n'est pas libre)
- Disponible sous toutes les plateformes : `http://www.r-project.org`
- Enormement de librairies disponibles (très pointues), projet très dynamique
- Langage basé sur le calcul matriciel de la même famille que Matlab, Scilab

Logiciel R

Aide

Documentation

Très nombreuses documentations (livres, site web, forums, etc.)
Un site parmi d'autres :
`http://www.duclert.org/Aide-memoire-R/Le-langage/Introduction.php`

Aide en ligne

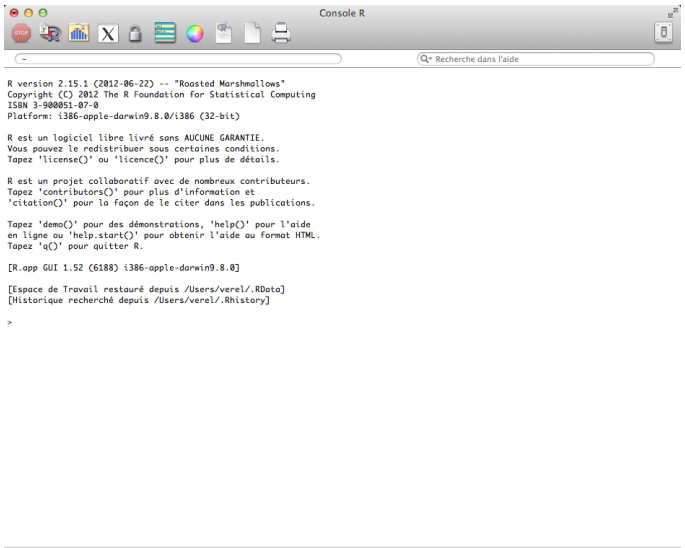
Pour avoir de l'aide sur une fonction :

```
help(mean)
```

ou

```
?mean
```

Environnement



- Les commandes peuvent être exécutées directement en les frappant au niveau du prompt
- Possibilité d'écrire des scripts (extension `.R`), lecture : `source("blabla.R", echo = T)`
- Pour quitter la session : `q()`
- L'historique peut être enregistrer pour ne pas perdre les données (fichier `.Rhistory`)

Calculatrice

- le "top level" peut être utilisé comme une calculatrice :
  ```
  sqrt(3 * 3 + 4 * 4)
  sin(pi / 2)
  ```
- Notion de variable :
  ```
  a <- 2
  a <- a + 1
  ```
- Lister l'ensemble des variables de l'environnement :
  ```
  ls()
  ```
- Effacer une variable de l'environnement :
  ```
  rm(a)
  ```

Vecteurs

- Objets de base de `R`
- Suite d'éléments de même type :
  - numérique (réelles ou complexes),
  - booléen (`TRUE, FALSE`)
  - alphanumérique (chaine de caractères)

```
c(elt1, elt2, elt3, ...)
```

Vecteurs

Création

- En extension (c pour combine) :
  `c(100, 20, 50)`
  vecteur composé des 3 entiers 100, 20 et 50.
- Par "répétition" d'un élément avec `rep` :
  `rep("a", 20)`
  vecteur composé de 20 éléments égaux au caractère "a"
- Par description d'une suite d'entiers consécutifs avec " :" :
  `1:100`
  vecteur composé des 100 premiers entiers
- Par description d'une suite d'entiers consécutif avec `seq` :
  `seq(11, 31, 2)`
  vecteur composé des nombres impairs entre 11 et 31.

Vecteurs

Longueur

- La longueur d'un vecteur est donné par `length` :
  `length(rep(1, 100))` donne 100

Concaténation

- L'ajout d'éléments s'effectue dela manière suivante :
  ```
  u <- c(100, 20, 50)
  u <- c(u, 500)
  ```
  donne : 100, 20, 50, 500

Tirage sans remise

- `sample` effectue un tirage sans remise des éléments :
  `sample(1:100, 10)`
  donne un vecteur constitué de 10 valeurs aléatoirement sélectionnées sans remise parmi le vecteur 1:100

Vecteurs

Lecture / ecriture

- Pour accèder à un élément donnée :
  `u[2]`
  élément d'indice 2 du vecteur u
- Attention le premier élément est d'indice 1 !
  `u <- c(100, 20, 50)`
  `u[1]` a pour valeur 100
- Sous-vecteur :
  `u[2:50]`
  éléments de l'indice 2 à l'indice 50.
  `u[c(2, 4, 6)]`
  éléments d'indice 2, 4 et 6.
- Sous-vecteur avec sélection :
  `u[u > 0]`
  vecteur composé des éléments strictement positifs de u

Opération sur les vecteurs

- Les opérations sont rapides et ne nécessite pas de boucles :
  `u + v`
  somme du vecteur u et v
- Opération sur chaque élément :
  `u * 2`
  multiplie tous les éléments par 2
- Application de fonction :
  `sin(u)`
  applique la fonction sinus à tous les éléments
- Tests :
  `u <- c(100, 20, 50)`
  `u > 30`
  donne le vecteur dont les éléments sont le résultat du test
  `TRUE FALSE TRUE`

Nom des éléments

- Par défaut, les éléments sont indexés par des entiers
- Il est possible de donner des noms aux éléments :
  `u <- c(100, 20, 50)`
  `names(u) <- c("un", "deux", "trois")`
  `u["un"]` a pour valeur 100
- Connaitre les noms des éléments :
  `names(u)`
  `"un", "deux", "trois"`

Matrices

- Objets de base de `R`
- Tableau à 2 dimensions d'éléments de même type :
  - numérique (réelles ou complexes),
  - booléen (`TRUE, FALSE`)
  - alphanumérique (chaine de caractères)

Matrices

Création

- En "colonne" :

```
A <- matrix(c(1:6), ncol = 2)
```

```
     [,1] [,2]
[1,]    1    4
[2,]    2    5
[3,]    3    6
```

- En "ligne" :

```
A <- matrix(c(1:6), ncol = 2, byrow=TRUE)
```

```
     [,1] [,2]
[1,]    1    2
[2,]    3    4
[3,]    5    6
```

Matrices

Lecture / écriture

```
A <- matrix(c(1:6), ncols = 2)
```

- Notation matricielle usuelle :
  `A[1,2]` a pour valeur 4
- une ligne entière :
  `A[1,]`
  la première ligne
- une colonne entière :
  `A[,2]`
  la deuxième colonne

Matrices

Opérations

- Opérations terme à terme : `A + B`
  `A * B`
- Opérations matricielles usuelles :
  `A %*% B`

Listes

Liste

- Suite d'objets (vecteurs, matrices, etc.) qui ne sont pas nécessairement du même type
- Utilisées par de nombreuses fonctions en résultat

Listes

Création

- Utilisation du mot clé `list` :

```
maSuperListe <- list(je = c("pense", "suis"),
data = matrix(c(1:6) ncols = 2))
```

Listes

Création

- Utilisation du mot clé `list` :
  `maSuperListe <- list(je = c("pense", "suis"), data = matrix(c(1:6) ncols = 2))`

Lecture / écriture

- Utilisation du signe $:
  `maSuperListe$je` donne l'élément `je` de la liste
- Utilisation du double crochet :
  `maSuperListe[[1]]` donne le premier élément de la liste (= `maSuperListe$je`)

Les `data.frame`

`data.frame`

- Liste de plusieurs vecteurs de même longueur
- Proche de la notion de tableau

Les `data.frame` : Création

Fichier csv

- Les fichiers d'extension `.csv` (comma-separated values) ont pour convention d'être des fichiers textes où les données sont enregistrées sous forme de tableau (ligne / colonne)
- Sur une même ligne, les données sont séparées par un caractère (espace, virgule, etc.)

```
initial final
155.209786 118.799252
367.658596 118.795867
405.252902 118.795936
...
```

Les `data.frame` : Création

Création à partir d'un fichier

- Les data frame peuvent se construire à partir d'un fichier csv :

```
frame <- read.table("data.csv", sep = " ",
header = TRUE)
```

  - le premier argument est le nom du fichier
  - `sep` indique le caractère de séparation des données
  - `header` indique si la première ligne contient le nom des colonnes

Les `data.frame` : Création

Création à partir d'un fichier

- Les data frame peuvent se construire à partir d'un fichier csv :
  ```
  frame <- read.table("data.csv", sep = " ",
  header = TRUE)
  ```
  - le premier argument est le nom du fichier
  - `sep` indique le caractère de séparation des données
  - `header` indique si la première ligne contient le nom des colonnes

Lecture / écriture

Pour obtenir les vecteurs constituant le data frame :

- Utilisation du signe $:
  `frame$initial`
- Utilisation du double crochet :
  `frame[[1]]`

Les fonctions

Fonction définie par l'utlisateur

```
maFonction <- function(arg1, arg2) {
    bloc d'instructions
}
```

- Le dernier résultat du bloc d'instruction est la valeur finale de la fonction

Exemple

```
plusUn <- function(x) {
    x + 1
}
```

Branchements et Itérations

si ... alors ... sinon ...

```
if (x < 0)
   - x
else
   x
```

Itération "pour"

```
for(name in c("un", "deux", "trois"))
   print(name)
```

Itération "tant que"

```
i <- 0
while(i < 100) {
   print(i)
   i <- i + 1
}
```

Statistique descriptive

- Moyenne :
  `mean(u)`
- Variance :
  `var(u)`
- Ecart-type :
  `sd(u)`
- Minimum et maximum :
  `min(u)`
  `max(u)`
- Quartiles :
  `quantile(u)`
- Résumé des principales statistiques :
  `summary(u)`

Observation d'une distribution empirique

Histogramme

```
hist(u)
```

Boite à moustache

- Unique :
  `boxplot(u)`
- Plusieurs vecteurs :
  `boxplot(list(u, v))`

Loi normale

Loi normale ou loi gaussienne

- Loi adaptée à modéliser de nombreux phénomènes naturels aléatoires
- Complètement caractérisée par la moyenne μ et la variance σ^2
- Densité :

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

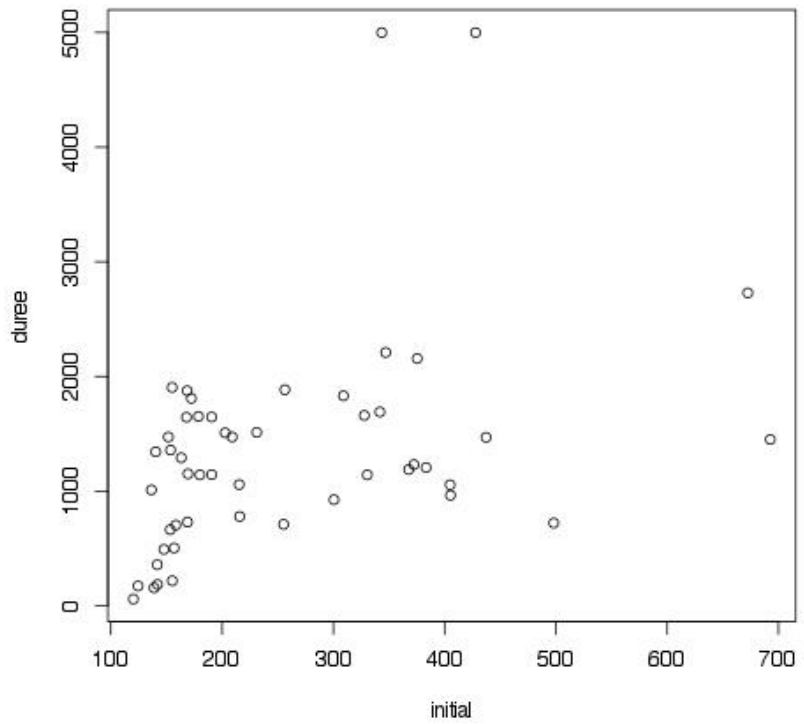
- Notation :

$$X \sim \mathcal{N}(\mu, \sigma^2)$$

en R

- Obtenir des réalisations :
  `rnorm(1000, 0, 1)`
  1000 réalisations d'une loi normale de moyenne nulle et de variance 0.

Données bivariées



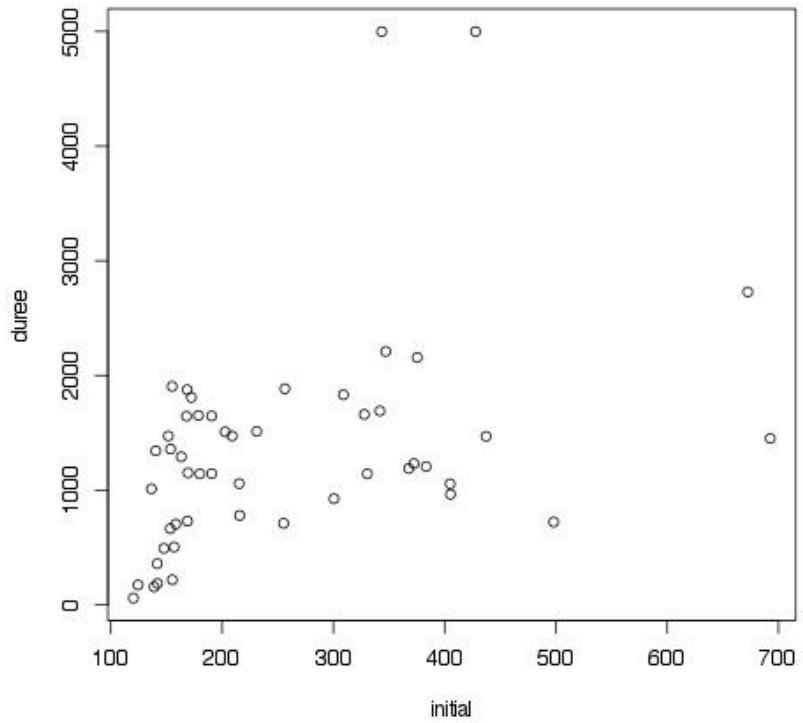
Définition

Ensemble de couples de données $\{(x_1, y_1), (x_2, y_2), \ldots\}$

en R

- Si `x` et `y` sont des vecteurs de même longueur (abscisses et ordonnées) :
  `plot(x,y)`
- Si `frame` est un data.frame (`initial` et `duree` sont des colonnes) :
  `plot(duree ~ initial, frame)`

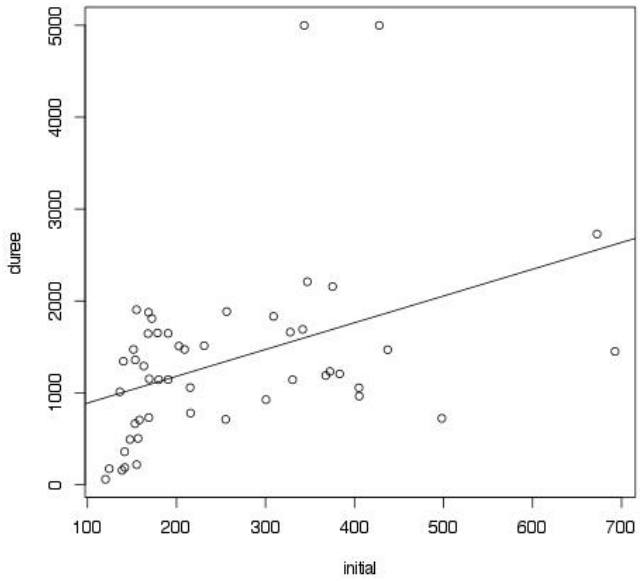
Statistique inférentielle



Questions

Existe-t-il une relation entre les variables x et y ?
Laquelle ?

Modèle linéaire



Modèle linéaire

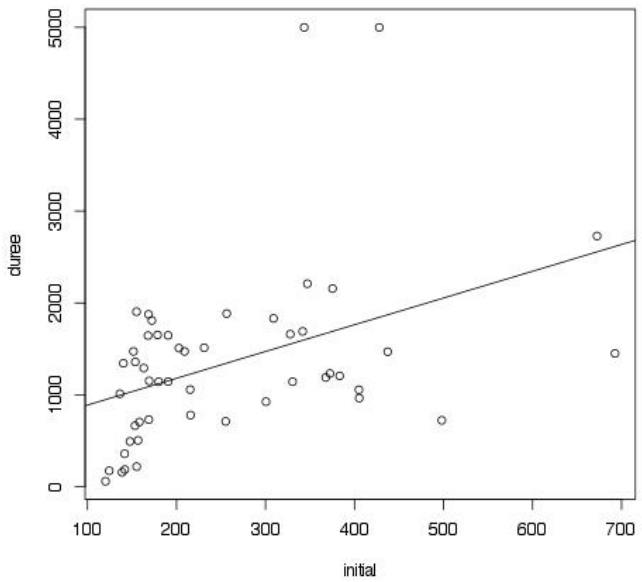
Le nuage de point est remplacé par une droite :

$$y = a.x + b$$

On peut alors :

- Expliquer la relation entre les variables
- Prédir les valeurs

Modèlisation et erreur



Erreur quadratique moyenne (MSE)

Erreur du modèle est la distance entre observations et prédictions :

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(f(x_i) - y_i)^2$$

D'autres mesures de distance sont possibles...

Modèlisation et erreur

Ajustement du modèle

$$f(x) = a.x + b$$

- Choisir les paramètres du modèle pour minimiser l'erreur
- Choisir a et b pour minimiser l'erreur MSE
- L'erreur est alors égale au coefficient de corrélation (au signe prêt) :

$$\rho = \frac{cov(x, y)}{\sigma_x \sigma_y}$$

Modèle linéaire en R

Fonction lm

- Utiliser la fonction `lm` (linear model) :
  - Avec data.frame :
    `lin <- lm(y ~ x, frame)`
  - Avec vecteurs de même longueur :
    `lin <- lm(x, y)`
- Pour avoir le descriptif du modèle :
  `summary(lin)`
- En particulier, on peut lire la p-value du modèle :
  Si p-value < 0.05, le modèle linéaire est significatif
- Pour avoir les coefficients a et b :
  `lin$coefficients`

Corrélation R

- Utiliser la fonction `cor` :
  - Avec data.frame :
    `cor(frame)`
    On obtient une matrice de corrélation entre toutes les variables du data.frame
  - Avec vecteurs de même longueur :
    `cor(x, y)`
- Si le coefficient est négatif, les variables sont anti-corrélées (variation inverse)
- Un fort coefficient est supérieur à 0.8 ou 0.9 (en valeur absolue)
- Un faible coefficient est compris entre -0.2 ou 0.2

La vie de tous les jours

Petit situation presque fictive

Ma grand-mère et mon père sont enrhumés. Mon père prend le médicament *touxClaire*, ma grand-mère prend le médicament *douceNuit*.
Au bout de 3 jours,

- ma grand-mère est guérie
- mon père tousse toujours

Qu'en déduire ?

La vie de tous les jours

Petit situation presque fictive

Ma grand-mère et mon père sont enrhumés. Mon père prend le médicament *touxClaire*, ma grand-mère prend le médicament *douceNuit*.
Au bout de 3 jours,

- ma grand-mère est guérie
- mon père tousse toujours

Qu'en déduire ?

Statistiquement parlant : RIEN.

La vie de tous les jours

Petit situation presque fictive

Ma grand-mère et mon père sont enrhumés. Mon père prend le médicament *touxClaire*, ma grand-mère prend le médicament *douceNuit*.
Au bout de 3 jours,

- ma grand-mère est guérie
- mon père tousse toujours

Qu'en déduire ?

Statistiquement parlant : RIEN.

Ne jamais rien déduire d'une seule observation
ou d'un faible (?) nombre d'observations

Mon dé fait toujours 6 ! ... Enfin une fois

Différence observée

La question

La différence observée est-elle la conséquence :

- d'une différence significative entre les phénomènes aléatoires
- de fluctuations aléatoires, de l'incertitude

cf. Tableau

Principe d'un test

- Formuler une hypothèse H_0 (hypothèse nulle)
- Calculer la probabilité d'obtenir les observations si l'hypothèse est vraie : **p-value**
- Si la p-value est plus petite qu'un seuil,
  alors on rejète l'hypothèse nulle H_0
  - En général (simplement raison d'usage) le seuil est 5% : p-value < 0.05
  - On peut fixer d'autres seuils, par ex. 1% : plus forte présomption contre l'hypothèse nulle qu'à 5%

Principe d'un test

- Formuler une hypothèse H_0 (hypothèse nulle)
- Calculer la probabilité d'obtenir les observations si l'hypothèse est vraie : **p-value**
- Si la p-value est plus petite qu'un seuil,
  alors on rejète l'hypothèse nulle H_0
  - En général (simplement raison d'usage) le seuil est 5% : p-value < 0.05
  - On peut fixer d'autres seuils, par ex. 1% : plus forte présomption contre l'hypothèse nulle qu'à 5%

- De nombreux tests ont été développé
- Pour tester si les observations suivent une loi donnée, pour comparer des moyennes, pour comparer des taux, si les observations sont issues d'une même distribution, etc.

Tester de normalité

- Hypothèse nulle H_0 :
  "Les observations (iid) suivent une loi normale"
- Test de Kolmogorov-Smirnov :
  `ks.test(x, "pnorm", 0, 1)`
  test de normalité pour la normale de moyenne nulle et de variance 1
- Test de Shapiro-Wilk :
  `shapiro.test(x)`

Obtenir la p-value

```
test <- shapiro.test(x)
test$p.value
```

Remarque : `qqnorm(x)` permet de représenter les observations "contre" une loi normale

Comparaison de moyennes

Quelques tests classiques

Hypothèse nulle H_0 :

"Les moyennes observées sont égales"

2 possibilités :

- Les distributions suivent des lois normales :
  - Test de Fisher exact (plus précis, effectif faible) :
    `fisher.text(mat)`
    où `mat` est une matrice à 2 colonnes contenant les 2 séries d'observations
  - Test t de student :
    `t.test(v1, v2)`
- Les distributions ne suivent pas des lois normales (tests non-paramétriques) :
  - Test de Mann-Whitney (ou Mann-Whitney-Wilcoxon) :
    `wilcox.test(v1, v2)`