université PARIS-SACLAY | FACULTÉ DES SCIENCES D'ORSAY

Entrées d'un modèle

Marc Girondot, Université Paris Saclay
marc.girondot@universite-paris-saclay.fr

1

Plan

- 1/ Les deux composantes d'un modèle
- 2/ Incertitude sur des valeurs empiriques
  - L'écart-type, l'erreur standard et autres
- 3/ Incertitude sur les paramètres du modèle
  - L'erreur standard et la distribution par MCMC
- 4/ Impact des paramètres sur les sorties d'un modèle
  - L'élasticité: « One at the time »
  - La sensibilité par décomposition de la variance

2

Notions à retenir

- Robustesse
- Ecart-type, erreur standard, quantiles
- Matrice hessienne
- MCMC par méthode bayésienne et algorithme de Metropolis-Hastings
- Elasticité
- Sensibilité, Indice de Sobol

3

Partie I

LES DEUX COMPOSANTES D'UN MODÈLE

4

Qu'est ce qu'un modèle ?

- Un modèle est constitué de deux parties:
  - Une formalisation de connaissances
  - Une paramétrisation de ces connaissances
- La formalisation des connaissances peut être verbale ou par des équations.
  - Il n'y a pas de différence fondamentale entre un modèle verbal ("Il fait plus chaud quand il y a une couverture nuageuse plus faible") ou un modèle décrit par des équations:
    $T_1 > T_2$ si $N_1 < N_2$.
  - L'utilisation d'équations permet cependant souvent d'être plus concis et précis.

5

Exemple

- Prenons comme exemple la croissance d'une population sous la forme d'un modèle logistique:

$$N_t = \frac{N_0 K}{N_0 + e^{-rt}\left(K - N_0\right)}$$

- Qui est la solution de l'équation différentielle:

$$\frac{dN_t}{dt} = rN_t\left(1 - \frac{Nt}{K}\right)$$

6

Croissance logistique

- Le modèle est donc constitué:
  - de l'équation qui correspond à une formalisation de connaissances;
  - des valeurs des paramètres N_0, r et K qui décrivent un comportement particulier du modèle.

7

La croissance logistique

Lambert Adolphe Jacques Quetelet né à Gand le 22 février 1796 et mort à Bruxelles le 17 février 1874.

- P. F. Verhulst est élève de Adolphe Quételet, un des fondateurs des statistiques.
- Quételet connait bien les travaux de Malthus et la croissance géométrique des populations.
  - Il faut noter que Malthus discute déjà des mécanismes écologiques qui doivent freiner la croissance. Mais il ne fournit pas d'équation.
- Quételet propose à Verhulst de trouver une fonction de freinage en s'inspirant de la résistance de l'air en aérodynamique:
  - La résistance aérodynamique s'écrit : $Ra = 1/2 \cdot \mu \cdot k \cdot v^2$ où μ représente la masse volumique de l'air, k un coefficient dépendant de la surface frontale du véhicule et de sa résistance aérodynamique et v sa vitesse.

Pierre-François Verhulst, né le 28 octobre 1804 et mort le 15 février 1849 à Bruxelles

8

Lambert Adolphe Jacques Quetelet né à Gand le 22 février 1796 et mort à Bruxelles le 17 février 1874.

L'origine du modèle

- « j'ai tenté depuis longtemps de déterminer par l'analyse, la loi probable de la population ; mais j'ai abandonné ce genre de recherches parce que les données de l'observation sont trop peu nombreuses pour que les formules puissent être vérifiées, de manière à ne laisser aucun doute sur leur exactitude » (*in* Quételet, 1850)

9

Les fonctions de freinage

$$\frac{dp}{dt} = (m\ p)\ -\ \varphi(p)$$

Verhulst indique ainsi avoir testé successivement quatre fonctions retardatrices :

$\varphi(p) = n\,.\,p^2$ $\varphi(p) = n\,.\,p^3$

$\varphi(p) = n\,.\,p^4$ $\varphi(p) = n\,.\log(p)$

Il choisit la première car c'est la plus simple et il n'avait pas d'argument pour choisir les autres !

10

Croissance logistique

- La relation mécaniste entre un modèle démographique et l'équation logistique n'a été démontrée qu'en 2002 dans:

ELSEVIER

Ecological Modelling 150 (2002) 55–81

ECOLOGICAL MODELLING

www.elsevier.com/locate/ecolmodel

Techniques of spatially explicit individual-based models: construction, simulation, and mean-field analysis

Luděk Berec *

Department of Theoretical Biology, Institute of Entomology, Academy of Sciences of the Czech Republic, Faculty of Biological Sciences, University of South Bohemia, Branišovská 31, 370 05 České Budějovice, Czech Republic

Received 23 February 2001; received in revised form 24 August 2001; accepted 28 October 2001

11

Formalisation de connaissance

- La formalisation des connaissances sous la forme d'un modèle peut être elle-même sujette à une part d'incertitude.
- Cependant souvent cette incertitude n'est pas prise en compte. On considèrera alors ce modèle comme décrivant au mieux les processus que l'on cherche à modéliser.
  - C'est clairement trop optimiste.

12

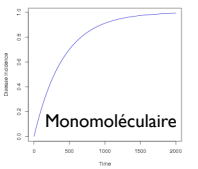
Une solution

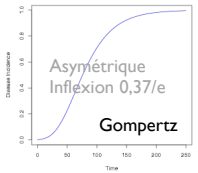
Richards F.J. 1959. A flexible growth function for empirical use. Journal of Experimental Botany 29:290-300.

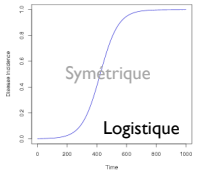
$$g = a(1 + b\exp(-rt))^{1/(1-k)}$$

$$g = \begin{cases} a(1+be^{-rt}) & \text{Monomolecular curve when } k = 0 \\ ae^{-be^{-rt}} & \text{Gompertz's curve when } k = 1 \\ \dfrac{a}{1+be^{-rt}} & \text{Autocatalytic or logistic curve when } k = 2. \end{cases}$$

Façon astucieuse de tester différentes formulation de modèles qu'on peut changer avec un paramètre.







13

Paramètres

- L'autre composante du modèle est celle pour laquelle on considère le plus souvent l'incertitude, il s'agit des paramètres du modèle.
- Les valeurs de ces paramètres peuvent être déterminées expérimentalement et elles sont donc entachées d'incertitude.

14

Partie II

INCERTITUDE SUR DES VALEURS EMPIRIQUES (ISSUES DE L'OBSERVATION)

15

Paramètres mesurés empiriquement

- Certains paramètres peuvent être directement mesurés sur le terrain.
- Il convient cependant d'effectuer une série de mesures pour capturer la variabilité sur ce paramètre.

Comment décrire cette variabilité ?

16

Les mesures de la dispersion

- Maximum et minimum
- L'écart-type
- L'erreur-standard
- Intervalle interquantile

- Toutes ces mesures sont justes mais ne représentent pas la même chose ; il faut donc bien savoir ce que l'on cherche à représenter.

17

Différence principale

- Ce qui différencie principalement ces mesures de dispersion, c'est leur comportement par rapport à des valeurs extrêmes ou aberrantes ainsi que la symétrie de l'estimateur autour de la moyenne.

18

Robustesse d'une statistique

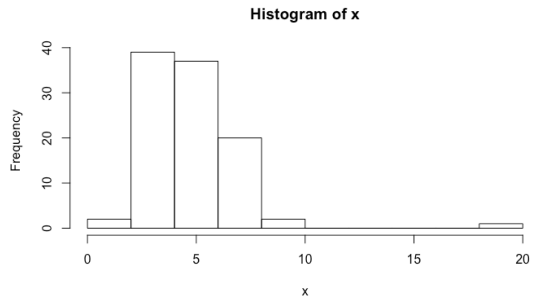
- La robustesse d'une statistique est sa capacité à supporter des violations d'hypothèses.
- Dans notre cas, si on veut mesurer la dispersion de nos observations, l'hypothèse est que les points soient tous tirés d'une même distribution.
- Mais si un point sort clairement du lot, il faut que la statistique décrivant la distribution ne dévie pas trop à cause de ce point.

19

Exemple

```
> x <- c(rgamma(100, shape=9, scale=0.5), 20)
```

- `hist(x)`



20

Minimum et maximum

- Les minimum et maximum sont des statistiques très sensibles aux valeurs extrêmes. Ce sont les statistiques les moins robustes pour décrire la dispersion.

```
> x <- rgamma(100, shape=9, scale=0.5)
> x2 <- c(x, 20)
> range(x)
[1] 1.558270    9.499228
> range(x2)
[1] 1.558270    20.00000
```

21

Ecart-type - *standard deviation*

- L'écart-type mesure la dispersion des valeurs.
- Dans une loi normale, 95% des observations sont situées dans l'intervalle moyenne ± 1,96 SD.

```
> sd(x)
[1] 1.542434
> sd(x2)
[1] 2.179
> 100*(sd(x2)-sd(x))/sd(x)
[1] 41.27029
```

22

L'écart-type (2)

- L'écart-type est très sensible à la présence d'une valeur extrême puisqu'il augmente de 41% seulement à cause d'une valeur.

Attention: Les observations sont situées dans l'intervalle moyenne ± 1.96 SD seulement dans le cas de la loi normale.

23

L'erreur standard - *standard error*

- L'erreur standard mesure la dispersion de la moyenne de nos observations. Il se calcule comme:

$$SE = \frac{SD}{\sqrt{N}}$$

```
> sd(x)/sqrt(length(x))
[1] 0.1542434
> sd(x2)/sqrt(length(x2))
[1] 0.2168186
```

24

L'erreur standard (2)

- Mais l'erreur standard reste très sensible à une observation extrême puisqu'il augmente aussi de 41% sur la base de l'inclusion d'une seule valeur extrême.

```
> sd(x)/sqrt(length(x))
[1] 0.1542434
> sd(x2)/sqrt(length(x2))
[1] 0.2168186
```

25

Masquer les erreurs

- On a comme relation : $SE = \frac{SD}{\sqrt{N}}$

« Donc SE est toujours plus faible que SD; et donc autant présenter SE que SD sur les barres d'erreur, cela cachera qu'on a mal travaillé ! »

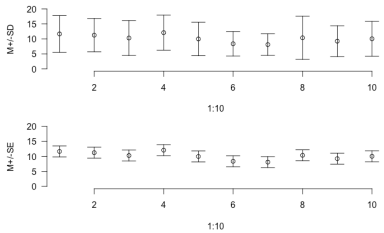
26

Représenter SE ou SD

```
M <- NULL
SD <- NULL
SE <- NULL

for (i in 1:10) {
  x <- runif(10, 5, 15)
  M <- c(M, mean(x))
  SD <- c(SD, sd(x))
  SE <- c(sd(x)/sqrt(length(x)))
}

library(HelpersMG)
par(mar=c(4, 4, 1, 1))
layout(mat = matrix(1:2, nrow=2))
plot_errbar(1:10, M, errbar.y = deux*SD, bty="n", las=1, ylab="M+/-SD", xlab="1:10",
ylim=c(0, 20))
plot_errbar(1:10, M, errbar.y = deux*SE, bty="n", las=1, ylab="M+/-SE", xlab="1:10",
ylim=c(0, 20))
```



27

Barres d'erreur

- En fait le terme « barre d'erreurs » serait à proscrire. Ces barres mesurent la dispersion d'un résultat, pas forcément une erreur.
- La source de dispersion des valeurs est le plus souvent due à une variabilité naturelle. Bien sûr parfois la variabilité peut être générée par la mesure ou bien réellement à des erreurs de mesure, mais c'est rare.
- Vouloir masquer la variabilité naturelle est stupide puisque justement c'est cette variabilité naturelle qui est utile pour répondre à des questions intéressantes (ou pas) !

28

SD *vs* SE

- SD est une mesure de la dispersion des observations.
- SE est une mesure de la dispersion des moyennes, c'est à dire où se trouve la moyenne. La moyenne peut être considérée déjà comme un modèle.

- **SD et SE ne sont donc pas interchangeables ; selon ce qu'on veut montrer sur un graphique, il faut choisir l'un ou l'autre.**

29

Expression of TRPV4 gene

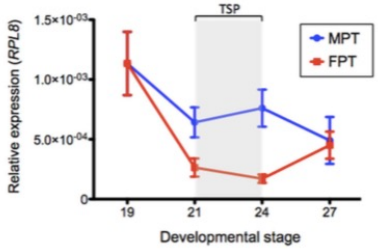
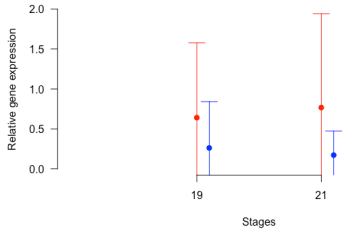


Figure 1. Developmental expression profile of American alligator TRP channels in gonad during sexual development. (*A*) The mRNA levels of various thermosensitive TRP channels were assessed in gonads at the onset of TSP (stage 21) incubated under MPT and FPT conditions. Gene expressions of 5 AmTRP ion channels (AmTRPV2, AmTRPV4, AmTRPA1, AmTRPM3, AmTRPM8) were observed in varying expression levels. (*B*) Quantitative RT-PCR analysis was performed for AmTRPV4 at various key sexual developmental stages including bipotential (stage 19; n = 13), sex determination (stage 21; n = 14, 14), sex differentiation (stage 24; n = 14, 15), and pre-hatching (stage 27; n = 14, 15) stages at both FPT and MPT temperature conditions respectively; ± SEM. Temperature sensitive period is indicated in gray.

30

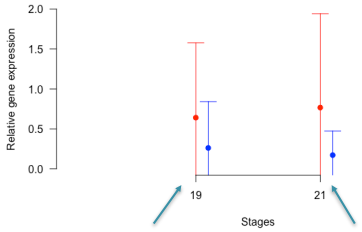
Expression of TRPV4 gene



- The error bars show ±SEM, then the dispersion at individual level is $(14)^{-2}$=3.7 times higher.
- Here is graph showing ±2 SD which is the correct metric to show that individuals at MPT and FPT could exhibit different expression level.

31

Expression of TRPV4 gene



- Notez qu'en utilisant ± 2 SD on trouve que l'expression relative pourrait être négative, cela n'a aucun sens.
- Cela est dû au fait que la distribution de l'expression relative (un rapport) n'est pas normale alors que quand on fait ± 2 SD, on suppose implicitement qu'elle est normale.

32

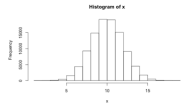
Le nombre magique 2

- On a tous entendu que 95% des points sont situés dans l'intervalle moyenne ± 2 SD.
- Déjà, on sait tous que ce n'est pas 2, mais 1,96:
- > (deux <- qnorm(p=0.975, 0, 1))
- [1] 1.959964
- Si on prend 2, cela fait:
- > pnorm(2, 0, 1)-pnorm(-2, 0, 1)
- [1] 0.9544997
- Pas si différent si l'effectif n'est pas trop important.

33

Distribution des observations

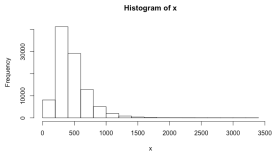
- > x <- rnorm(100000, 10, 2)
- > hist(x)
- > mean(x)
- [1] 10.01199
- > sqrt(sum((x-mean(x))^2)/length(x))
- [1] 1.997326
- > sd(x)
- [1] 1.997336
- > sum(x<(mean(x)-deux*sd(x)))/length(x)
- [1] 0.02401
- > sum(x>(mean(x)+deux*sd(x)))/length(x)
- [1] 0.02531



34

Distribution des observations

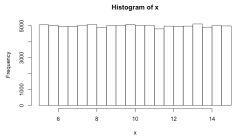
- > x <- rlnorm(100000, 6, 0.5)
- > hist(x)
- > mean(x)
- [1] 457.4219
- > sd(x)
- [1] 245.5845
- > sum(x<(mean(x)-deux*sd(x)))/length(x)
- [1] 0
- > sum(x>(mean(x)+deux*sd(x)))/length(x)
- [1] 0.04643



35

Distribution des observations

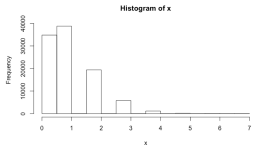
- > x <- runif(100000, 5, 15)
- > hist(x)
- > mean(x)
- [1] 9.994812
- > sd(x)
- [1] 2.890294
- > sum(x<(mean(x)-deux*sd(x)))/length(x)
- [1] 0
- > sum(x>(mean(x)+deux*sd(x)))/length(x)
- [1] 0



36

Distribution des observations

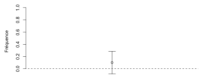
- > x <- rbinom(100000, 10, 0.1)
- > hist(x)
- > mean(x)
- [1] 1.00094
- > sd(x)
- [1] 0.9483818
- > sum(x<(mean(x)-deux*sd(x)))/length(x)
- [1] 0
- > sum(x>(mean(x)+deux*sd(x)))/length(x)
- [1] 0.07043



37

Conclusion

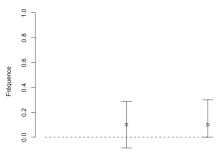
- 95% des observations ne sont pas dans l'intervalle ±1.96 SD sauf dans un cas très particulier: la loi normale.
- Cela peut conduire à des situations à la limite de l'absurde:
- x <- x/10
- mean(x)
- sd(x)
- library(HelpersMG)
- plot_errbar(1, mean(x), errbar.y = deux*sd(x), bty="n",
- ylim=c(-0.1, 1), xlab="", ylab="Fréquence", xaxt="n")
- segments(x0=0, x1=2, y0=0, y1=0, lty=2)



38

Retour aux fréquences

```
x <- rbinom(100000, 10, 0.1)
x <- x/10
mean(x)
sd(x)
l <- quantile(x = x, probs = c(0.025, 0.975))
library(HelpersMG)
plot_errbar(c(1, 2), c(mean(x), mean(x)), errbar.y.plus =  c(deux*sd(x),
    l[2]-mean(x)),  errbar.y.minus = c(deux*sd(x), mean(x)-l[1]),
    bty="n", ylim=c(-0.1, 1), xlab="", ylab="Fréquence", xaxt="n",
    xlim=c(0, 3))
segments(x0=0, x1=2, y0=0, y1=0, lty=2)
```



39

Les quantiles

- Un quantile q est la valeur pour laquelle une fréquence q des observations sont situées en dessous du quantile.
- On utilisera les quantiles 0.025 et 0.975 pour définir les bornes qui incluent 95% des valeurs (0.975-0.025=0.95).

```
> quantile(x, probs=c(0.025, 0.975))
   2.5%    97.5%
2.273944 7.963572
> quantile(x2, probs=c(0.025, 0.975))
   2.5%    97.5%
2.276463 8.151774
```

40

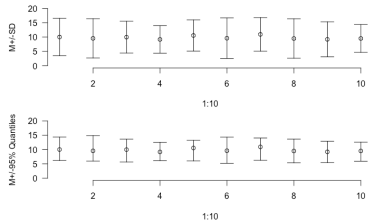
Les quantiles (2)

- Les quantiles sont des statistiques très robustes notamment aux points extrêmes mais aussi à l'asymétrie de la distribution.

41

Les quantiles 0,025 et 0,975

- M <- NULL; SD <- NULL; SE <- NULL; QM <- NULL; QP <- NULL
- for (i in 1:10) {
- x <- runif(10, 5, 15)
- M <- c(M, mean(x))
- l <- quantile(x = x, probs = c(0.025, 0.975))
- QM <- c(QM, l[1])
- QP <- c(QP, l[2])
- SD <- c(SD, sd(x))
- SE <- c(sd(x)/sqrt(length(x)))
- }



- layout(mat = matrix(1:2, nrow=2))
- plot_errbar(1:10, M, errbar.y = deux*SD, bty="n", las=1, ylab="M+/-SD", xlab="1:10", ylim=c(0, 20))
- plot_errbar(1:10, M, y.plus = QP, y.minus = QM, bty="n", las=1, ylab="M+/-95% Quantiles", xlab="1:10", ylim=c(0, 20))

42

Conclusion

- Pour mesurer la dispersion des points, nous utiliserons les quantiles 0.025 et 0.975.
- Pour mesurer la dispersion de la moyenne nous utiliserons l'erreur standard sachant qu'il y a 95% de chance que la vraie moyenne soit entre moyenne ± 1.96 SE

43

D'où vient cette affirmation ?

- Le théorème central limite (Laplace, 1809) établit la convergence en loi de la somme d'une suite de variables aléatoires vers la loi normale.
- La moyenne est une somme divisée par une constante (N), donc le théorème central limite s'applique mais que dit-il en clair: Définition - On dit que la suite $(X_n)_{n\geq 1}$ *converge en loi* vers X si, pour toute fonction φ *continue bornée* sur E, à valeurs dans R:

23 mars 1749, Beaumont-en-Auge
5 mars 1827, Paris

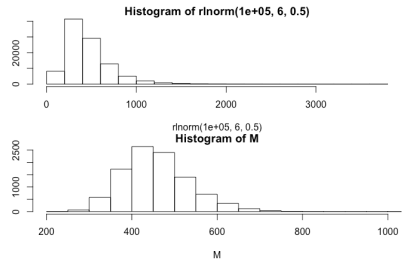
$$\lim_n \mathbb{E}\left[\varphi(X_n)\right] = \mathbb{E}\left[\varphi(X)\right].$$

Pierre-Simon Laplace, « Mémoire sur les approximations des formules qui sont fonctions de très-grands nombres, et sur leur application aux probabilités », Mémoires de la Classe des sciences mathématiques et physiques de l'Institut de France, 1809, p. 353-415

44

Distribution de la moyenne

- M <- NULL; SD
- for (i in 1:10000
- x <- rlnorm(1(
- M <- c(M, mea
- SD <- c(SD, sd
- SE <- c(sd(x)/s
- }



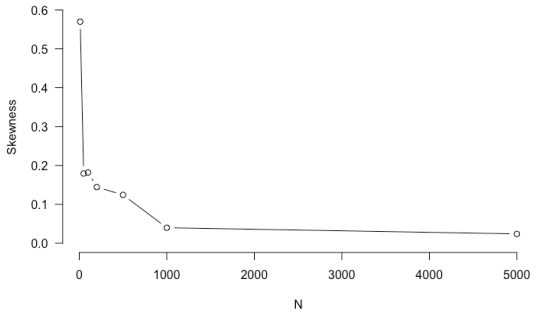
- layout(mat = matrix(1:2, nrow=2))
- hist(rlnorm(100000, 6, 0.5))
- hist(M, xlim=c(200, 1000))

```
> mean(((M-mean(M))/sd(M))^3)
[1] 0.5624483
> library(e1071)
> skewness(M, type=3)
[1] 0.5624483
```

45

Quel n est nécessaire ?

- > plot(c(10, 50, 100, 200, 500, 1000, 5000), S, type="b", bty="n", las=1, xlab="N", ylab="Skewness",ylim=c(0,0.6))
- > S
- [1] 0.56986114 0.17927412 0.18199925 0.14438459 0.12433417 0.03980372 0.02376757



46

La quête du graal

- Le théorème central limite nous dit que les estimateurs sont distribués normalement mais on ne l'atteint réellement que quand un nombre très grand d'observations utilisés pour estimer les distributions.
- Qu'en est-il pour les observations elles-mêmes ?
- Revenons aux caractéristiques de la loi normale:
- Non bornée
- Symétrique

- Est-ce réaliste: non !
- La loi normale admet des valeurs négatives alors que la plupart des métriques utilisée en biologie n'en admettent pas.

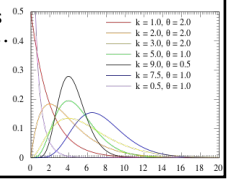
47

La quête du graal

Donc pourquoi continuer à utiliser une distribution que l'on sait non-appropriée ?

- On a d'autres distributions qui sont flexibles et plus appropriées:
- - La binomiale pour les proportions
- - La binomiale négative pour les comptages
- - La gamma pour les valeurs ≥0
- - La lognormale pour les valeurs >0

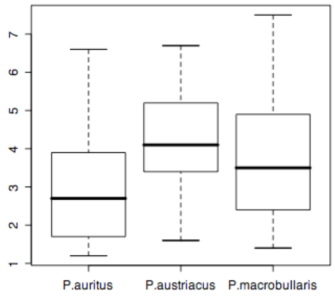
- Bien sûr on peut s'en sortir en parlant de tests robustes et continuer à utiliser la loi normale…



48

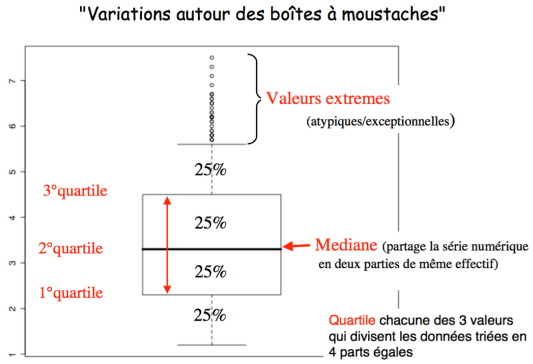
Autre façon de représenter une distribution univariée

```
boxplot(d[,2]~d[,1])
```



49

Les moustaches: les 4 quarts et un peu plus

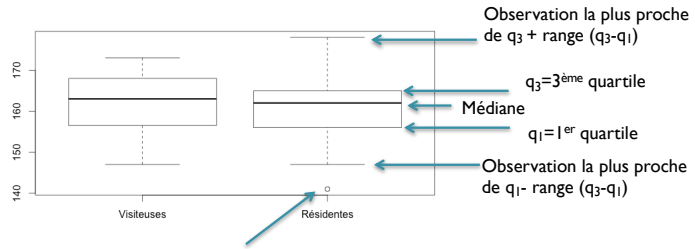


50

Les tiers et les quarts

51

La taille des moustaches



« Outliers » range = 1,5 par défaut

- La taille des moustaches est dépendante de l'échantillon observé. Ce n'est pas un quantile.
- Lorsque la distribution est normale et infinie, 95,45% des observations sont entre les moustaches,
- Donc la taille des moustaches n'a pas d'interprétation simple probabiliste.

McGill, R., Tukey, J.W., Larsen, W.A., 1978. Variations of Box Plots. The American Statistician 32, 12-16.

52