ASSISTANCE PUBLIQUE HÔPITAUX DE PARIS
Direction de la Recherche Clinique et de Innovation

Center for Methodology and Modeling
Malaria – Public Health – Infectious Diseases.
West-Africa
C2M

Biostatistiques

Partie 2 : Résumer les données et probabilités en statistiques

DU EPIDEMIOLOGIE DE TERRAIN - IFRISSE

Kankoé SALLAH MD, PhD

kankoe.sallah@univ-amu.fr
kankoe@skml.fr

Apr 2020

Résumer les données

OBJECTIF MAJEUR

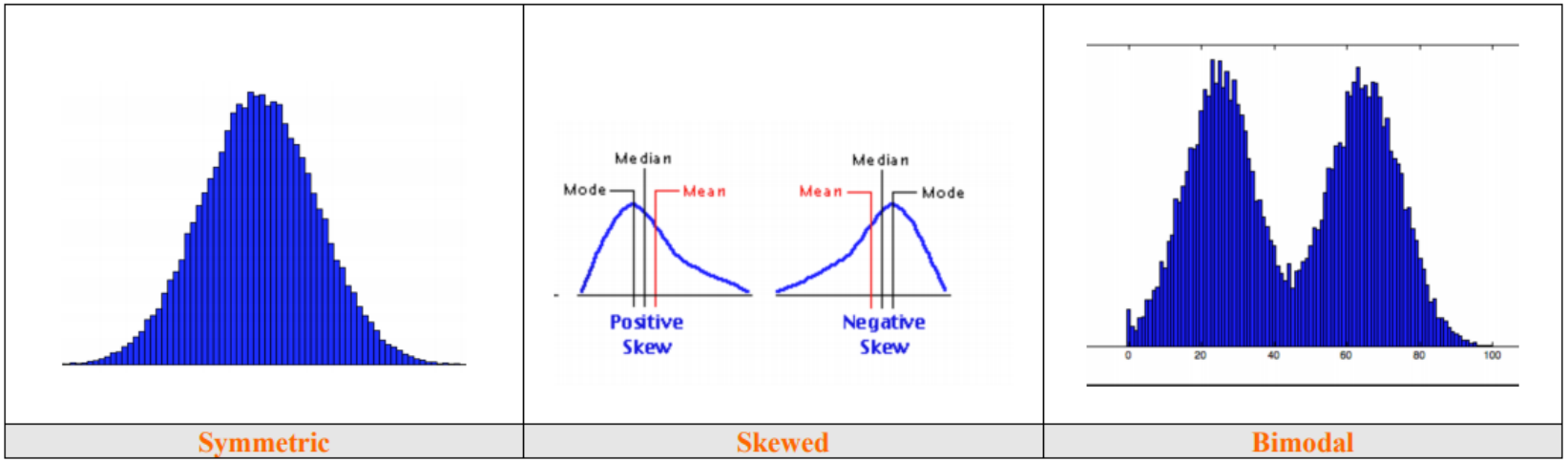
Utiliser les principales commandes R pour décrire les données, comprendre les notions de probabilités, sensibilité, spécificité, valeurs prédictives, odds ratios et risques relatifs

Résumer les données

	Variable catégorielle		Variable numérique	
Type de variable	Nominale	Ordinale	Discrete	Continue
Méthodes graphiques	Bar chart (diag. en barres) Pie chart (diag. en camembert)	Bar chart Pie chart	Bar chart, Pie chart, Dot diagram, Scatter plot (2 variables), Stem-Leaf, Histogram, Box Plot, Quantile-Quantile Plot	Dot diagram, Scatter plot (2 vars), Stem-Leaf, Histogram, Box Plot, Quantile-Quantile Plot
Synthèse numérique	Frequency Relative Frequency	Frequency Relative Frequency Cumulative Frequency	Frequency Relative Frequency Cumulative Frequency means, variances, percentiles	Means (moyennes), variances, Percentiles

```
hist(), barplot(), pie(), dotchart(), plot(),
boxplot(), qqplot(), qqnorm(), qqline()
```

Résumer les données



Mode. Valeur ou modalité la plus représentée de la série. Non influencée par les valeurs extrêmes.

Moyenne. Moyenne arithmétique des valeurs de la série. Influencée par les valeurs extrêmes. `mean(data$age)`

Médiane. Valeur divisant la série en 2 sous-échantillons de même taille. Non influencée par les valeurs extrêmes. `median(data$age)`

Résumer les données

Moyenne `mean(data$age)`

$$\text{Mean} = \frac{\text{sum of values}}{\text{sample size}} = \frac{\Sigma\ (\text{values})}{\text{n}}$$

$$\text{Grouped mean} = \frac{\sum(\text{data value})(\text{frequency of data value})}{\sum(\text{frequencies})}$$

Paramètres de tendance centrale

Médiane `median(data$age)`

If the sample size n is ODD

$$\text{median} = \frac{\text{n}+1}{2}\text{th largest value}$$

If the sample size n is EVEN

$$\text{median} = \text{average of}\left(\left[\frac{\text{n}}{2}\right]th, \left[\frac{\text{n}+2}{2}\right]th\right)\text{values}$$

Résumer les données

Variance.

$$\sigma^2 = \frac{\sum_{i=1}^{N}(X_i-\mu)^2}{N} \qquad S^2 = \frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{(n-1)}$$

$$\text{Sample Standard Deviation (S or SD)} = \sqrt{\frac{\sum(X-\bar{X})^2}{n-1}}$$

```
var(data$age)
sd(data$age)
```

Paramètres de dispersion

Médiane des déviations absolues à la médiane

$$\text{MADM} = \text{median of}\left[\ |X_i - \text{median of}\ \{X_1,\ldots,X_n\}\ |\ \right]$$

Erreur standard sur la moyenne $SE(\bar{X}) = \frac{SD}{\sqrt{n}}$

```
sd(data$age)/sqrt(length(age))
```

Résumer les données

Coefficient de variation = $\frac{\textit{Déviation standard}}{\textit{Moyenne}}$

$$\xi = \frac{\sigma}{\mu}$$

$$cv = \hat{\xi} = \frac{S}{\overline{X}}$$

```
CV <- function(x){ (sd(x)/mean(x))*100 }
CV(data$age)
```

Paramètres de dispersion

Etendue: Max-Min

```
max(data$age) - min(data$age)
```

Distance interquartile (interquartile range)

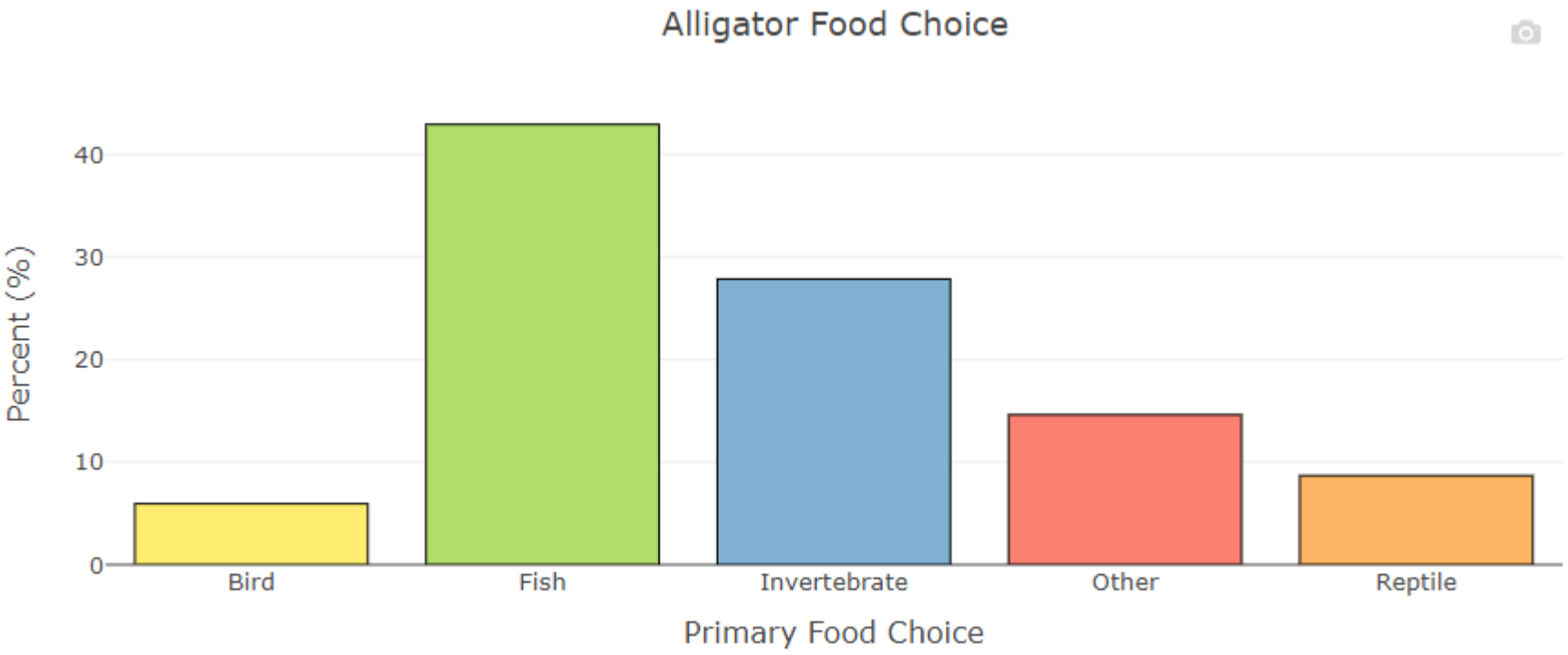
$$IQR = \text{Interquartile Range} = [P_{75} - P_{25}]$$

```
IQR(data$age)
quantile(data$age, prob=0.25)
quantile(data$age, prob=0.50)
quantile(data$age, prob=0.75)
quantile(data$age, prob=0.75)-quantile(data$age, prob=0.25)
```

Percentiles : valeurs divisant la série en sous-ensembles dont la largeur est égale à 1/1000 ème de l'étendue

Visualiser les données

Exemple. Diagrammes en barres



Bases de probabilités

- **Une variable** peut prendre différentes valeurs dépendamment de la chance.
- On appelle **évènement** une réalisation de la variable.
- On appelle probabilité la **proportion attendue** de réalisation d'un évènement dans une population

- **L'espace de probabilité** représente l'ensemble des réalisations possibles

- **Evènements disjoints :** ne peuvent se produire simultanément.

- **Evènements indépendants :** la réalisation de l'un n'a pas d'incidence sur la réalisation de l'autre.

- **2 évènements disjoints de probabilités non nulles ne sont jamais indépendants**

Bases de probabilités

- **L'évènement complémentaire** de E est la non réalisation de E

- **L'union des évènements A et B est l'évènement :** A se réalise ou B se réalise ou A et B se réalisent simultanément. Il est noté AUB

- **L'intersection des évènements A et B est l'évènement :** A se réalise et B se réalise. Il est noté A∩B

- **Probabilité conjointe de 2 évènements indépendants**

$$P(A \cap B) = P(A) \times P(B)$$

- **De façon générale**

$$P(A \cup B) = P(A) + P(B) - P(A) \cap P(B)$$

Bases de probabilités

- **Probabilités conditionnelles.** La réalisation de A a une incidence sur a réalisation de B

$$P(A \cap B) \neq P(A) \times P(B)$$

$$P(A \cap B) = P(A) \times P(B \mid A)$$

Théorème de Bayes

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)}$$

Attention : ici, il y a une notion d'ordre dans les réalisations

$$P(A \cap B \cap C) = P(A) \times P(B \mid A) \times P(C \mid B)$$

Bases de probabilités

- **Théorème des probabilités totales.** Si A1, A2…An représentent une partition de A alors

$$P(E) = P(A_1) \times P(E \mid A_1) + P(A_2) \times P(E \mid A_2) + \ldots + P(A_n) \times P(E \mid A_n)$$

Or

$$P(Ai \mid E) = \frac{P(E \mid A_i) \times P(A_i)}{P(E)}$$

D'où

$$P(Ai \mid E) = \frac{P(E \mid A_i) \times P(A_i)}{P(A_1) \times P(E \mid A_1) + P(A_2) \times P(E \mid A_2) + \ldots + P(A_n) \times P(E \mid A_n)}$$

Probabilités en épidémiologie

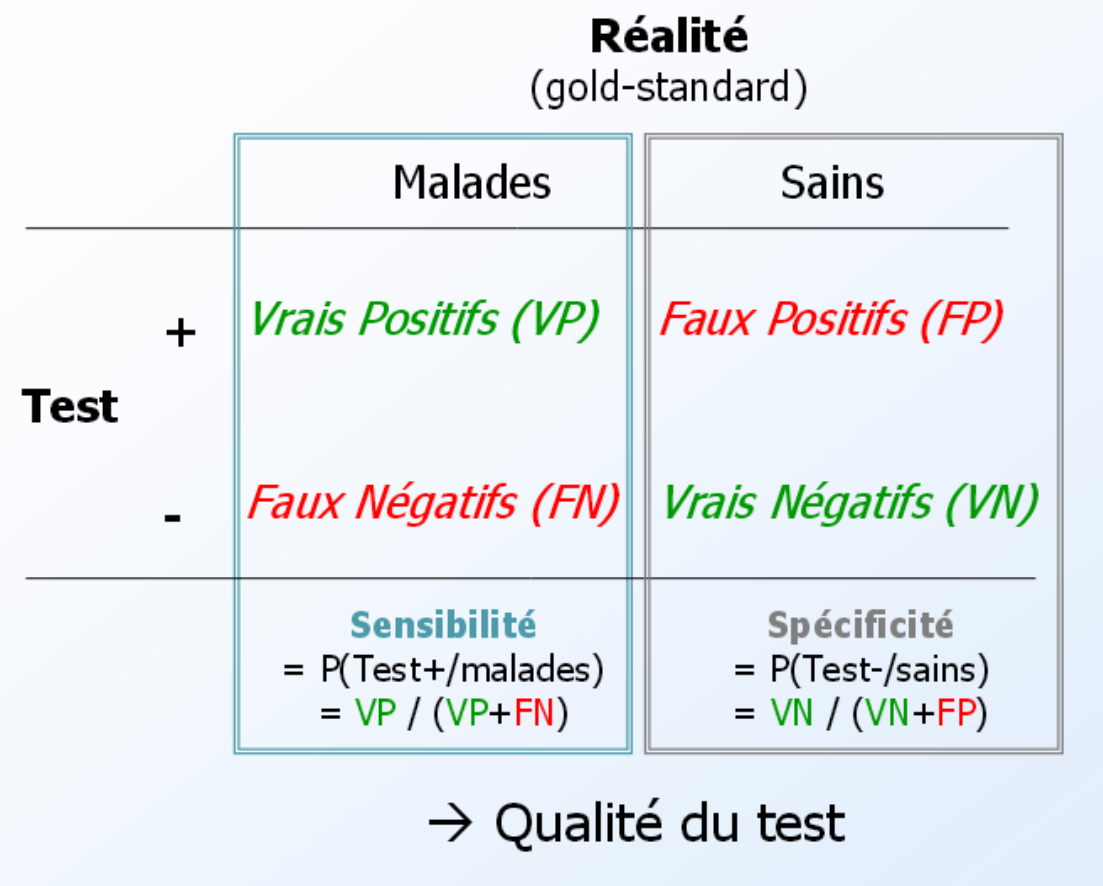
- **Prévalence d'une maladie:** Proportion d'individus ayant une maladie dans le population en un instant *t*

- **Incidence:** Effectif de nouveaux cas par unité de temps (Nx cas / Δt)

- **Sensibilité:** $$\text{sensitivity} = \Pr[+\text{test}|\text{disease}] = \frac{\Pr[\text{"+test" AND "disease"}]}{\Pr[\text{disease}]}$$

- **Spécificité:** $$\text{specificity} = \Pr[-\text{test}|\text{ NO disease}] = \frac{\Pr[\text{"-test" AND "NO disease"}]}{\Pr[\text{NO disease}]}$$

Probabilités en épidémiologie



Probabilités en épidémiologie

Valeurs prédictives

Test	Réalité (gold-standard) : Malades	Réalité (gold-standard) : Sains	
+	*Vrais Positifs (VP)*	*Faux Positifs (FP)*	**VPP** = P(Malades/Test+) = VP / (VP+FP)
-	*Faux Négatifs (FN)*	*Vrais Négatifs (VN)*	**VPN** = P(Sains/Test-) = VN / (VN+FN)

→ Apport du test pour un patient donné

Probabilités en épidémiologie

Odds ratios, Risques relatifs

Maladie → Exposition ↓	M+ Malades	M- Non Malades	Total
E+ Exposés	O_{11}	O_{10}	L_1
E- Exposés	O_{01}	O_{00}	L_0
Total	C_1	C_0	n

$$\psi = Rapport \text{ de côtes} = \frac{Côte \text{ de l'exposition chez les malades}}{Côte \text{ de l'exposition chez les non malades}} = \frac{\frac{O_{11}}{O_{01}}}{\frac{O_{10}}{O_{00}}} = \frac{O_{11}}{O_{01}} \times \frac{O_{00}}{O_{10}} = \frac{O_{11}O_{00}}{O_{10}O_{01}}$$

Plus l'exposition est spécifique des malades, plus Ψ est grand

Probabilités en épidémiologie

Odds ratios, Risques relatifs

Maladie → Exposition ↓	M+ Malades	M- Non Malades	Total
E+ Exposés	O_{11}	O_{10}	L_1
E- Exposés	O_{01}	O_{00}	L_0
Total	C_1	C_0	n

Variance OR $$S^2_{\hat{\psi}} = \frac{1}{O_{11}} + \frac{1}{O_{00}} + \frac{1}{O_{10}} + \frac{1}{O_{01}}$$

Probabilités en épidémiologie

Odds ratios, Risques relatifs

Exposition ↓ Maladie →	M+	M-	Total
E+	O_{11}	O_{10}	L_1
E-	O_{01}	O_{00}	L_0
Total	C_1	C_0	n

- Risque relatif

Valable si échantillon représentatif

$$RR = \frac{\text{Risque chez les exposés}}{\text{Risque chez les non exposés}} = \frac{\frac{O_{11}}{O_{11}+O_{10}}}{\frac{O_{01}}{O_{01}+O_{00}}} = \frac{R_1}{R_0}$$

- Odds ratio

$$OR = \psi = Rapport \text{ de côtes} = \frac{Côte \text{ de l'exposition chez les malades}}{Côte \text{ de l'exposition chez les non malades}} = \frac{\frac{O_{11}}{O_{01}}}{\frac{O_{10}}{O_{00}}} = \frac{O_{11}}{O_{01}} \times \frac{O_{00}}{O_{10}} = \frac{O_{11}O_{00}}{O_{10}O_{01}}$$

On démontre : $OR = RR \times \frac{1-R_0}{1-R_1}$

Si maladie rare [R_0<<1 et R_1<<1] alors OR≈RR

Merci
Fin de la Partie 2 : résumer les données et probabilités en épidémiologie

A suivre : Lois de probabilités et tests d'hypothèses