

Statistiques
descriptives

v.a. continues

v.a. discrètes

Intervalle de
confiance

v.a. continues

v.a. discrètes

Les statistiques descriptives et les intervalles de confiance

Yohann.Foucher@univ-nantes.fr

Equipe d'Accueil 4275 "Biostatistique, recherche clinique et mesures subjectives en santé", Université de Nantes

Master 2 - Cours #2



UNIVERSITÉ DE NANTES



itun

institut
transplantation
urologie
néphrologie
INSERM - UNIV NANTES

Statistiques
descriptives

v.a. continues

v.a. discrètes

Intervalle de
confiance

v.a. continues

v.a. discrètes

1. Statistiques descriptives

v.a. continues

v.a. discrètes

2. Intervalle de confiance

v.a. continues

v.a. discrètes

Statistiques descriptives

v.a. continues

v.a. discrètes

Intervalle de confiance

v.a. continues

v.a. discrètes

1. Statistiques descriptives

v.a. continues

v.a. discrètes

2. Intervalle de confiance

v.a. continues

v.a. discrètes

Deux grandes catégories d'indicateurs

Statistiques
descriptives

v.a. continues

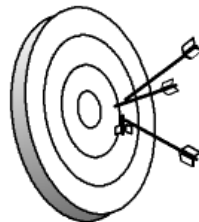
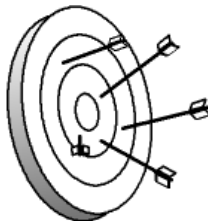
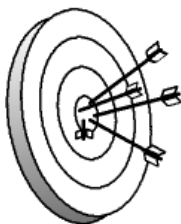
v.a. discrètes

Intervalle de
confiance

v.a. continues

v.a. discrètes

- Les statistiques dites de localisation.
- Les statistiques dites de dispersion.



** Auteur : Christophe Dang Ngoc Chan*

Statistiques
descriptives

v.a. continues

v.a. discrètes

Intervalle de
confiance

v.a. continues

v.a. discrètes

1. Statistiques descriptives

v.a. continues

v.a. discrètes

2. Intervalle de confiance

v.a. continues

v.a. discrètes

- **La moyenne** est l'indicateur de localisation le plus fréquemment utilisé.
Soit X cette v.a. dont on observe N expériences : $x_1, x_2, \dots, x_N$.

$$\bar{x} = \frac{\sum_{i=1}^N x_i}{N} = \frac{x_1 + x_2 + \dots + x_N}{N}$$

- **La médiane** est la valeur qui partage l'échantillon en deux groupes de même effectif.
- Pour le calcul, ordonner les valeurs observées :

1.1, 7.8, 9.9, 3.3, 5.4

↓

1.1, 3.3, 5.4, 7.8, 9.9

Les indicateurs de localisation

Statistiques
descriptives

v.a. continues
v.a. discrètes

Intervalle de
confiance

v.a. continues
v.a. discrètes

- Le 1er quartile (Q_1) est la valeur qui identifie le quart des données inférieures.
- Le 2nd quartile est aussi la médiane...
- Le 3ème quartile (Q_3) est la valeur qui identifie le quart des données supérieures.
- Le mode est la valeur la plus représentée.
- Le minimum est la valeur la plus faible.
- Le maximum est la valeur la plus forte.

- **La variance** mesure la variabilité autour de la moyenne.

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

- **L'écart-type** respecte la même unité que la variance (plus facile à interpréter)

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

- **L'intervalle interquartiles (IQ)** est la distance entre le 1er et le 3ème quartile.
- **L'étendue** est la distance entre le minimum et la maximum.

Statistiques
descriptives

v.a. continues
v.a. discrètes

Intervalle de
confiance

v.a. continues
v.a. discrètes

- Grands échantillons :
 - 1 moyenne ($\pm$ écart-type)
 - 2 moyenne (minimum-maximum)
- Petits échantillons :
 - 1 médiane (intervalle interquartiles)
 - 2 médiane (minimum-maximum)

Quelques conventions

Statistiques
descriptives

v.a. continues

v.a. discrètes

Intervalle de
confiance

v.a. continues

v.a. discrètes

- Grands échantillons :
 - 1 moyenne ($\pm$ écart-type)
 - 2 moyenne (minimum-maximum)
- Petits échantillons :
 - 1 médiane (intervalle interquartiles)
 - 2 médiane (minimum-maximum)

Problèmes liés à la moyenne :

- Sensible aux valeurs extrêmes (en particulier pour les petits échantillons).
- N'a de sens que pour des lois symétriques et unimodales.

Les règles à respecter pour un graphique

Statistiques
descriptives

v.a. continues

v.a. discrètes

Intervalle de
confiance

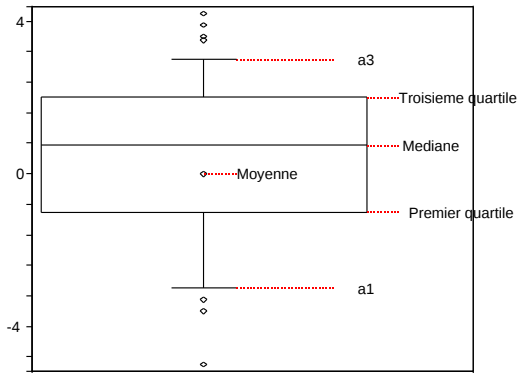
v.a. continues

v.a. discrètes

- Libellés des deux axes.
- Unités des deux axes.
- Légende explicite.
- Titre.
- Eviter les couleurs ou la 3D sauf nécessité.

La boîte à moustache

- a_1 est la plus petite valeur supérieure à $Q_1 - 1.5 \times IQ$
- a_3 est la plus grande valeur inférieure à $Q_3 + 1.5 \times IQ$
- Les valeurs en dehors de ces bornes sont "extrêmes"



* Auteur : Hélène Guérin

La boîte à moustache pour comparer deux groupes

Statistiques
descriptives

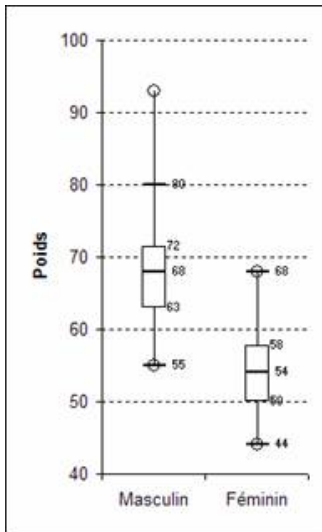
v.a. continues

v.a. discrètes

Intervalle de
confiance

v.a. continues

v.a. discrètes



La boîte à moustache pour décrire une évolution

Statistiques
descriptives

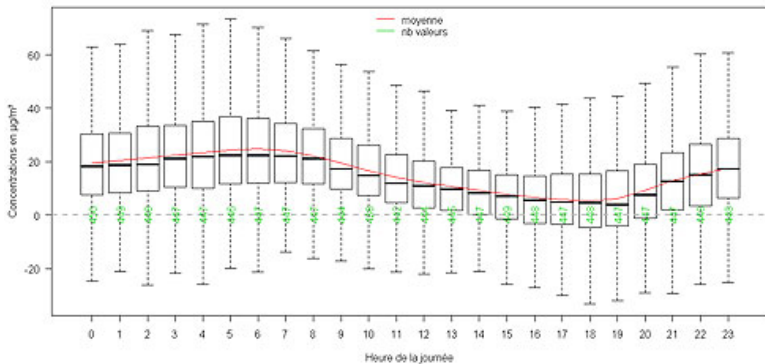
v.a. continues

v.a. discrètes

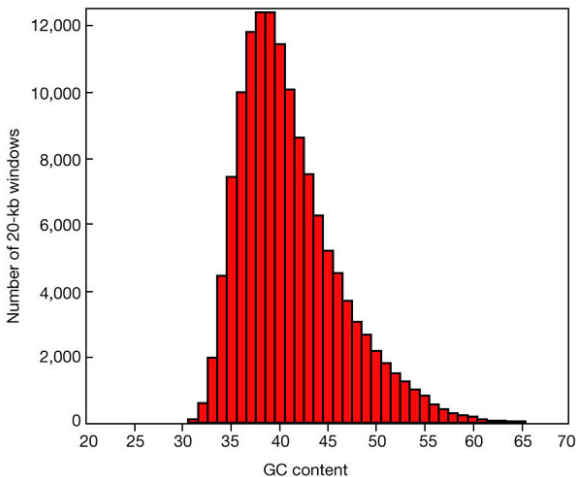
Intervalle de
confiance

v.a. continues

v.a. discrètes



L'histogramme (les rectangles sont contigus)



Statistiques
descriptives

v.a. continues

v.a. discrètes

Intervalle de
confiance

v.a. continues

v.a. discrètes

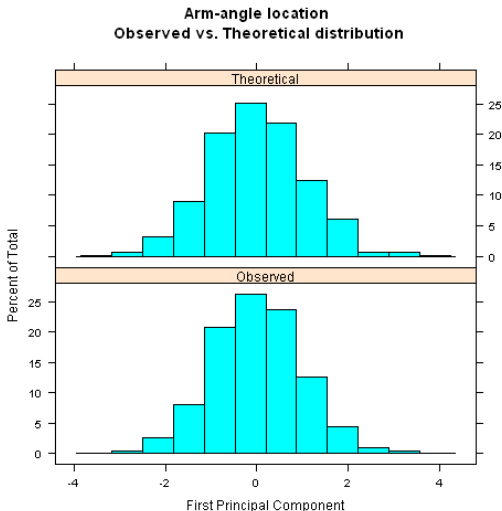
L'histogramme pour comparer deux groupes

Statistiques
descriptives

v.a. continues
v.a. discrètes

Intervalle de
confiance

v.a. continues
v.a. discrètes



Statistiques
descriptives

v.a. continues

v.a. discrètes

Intervalle de
confiance

v.a. continues

v.a. discrètes

1. Statistiques descriptives

v.a. continues

v.a. discrètes

2. Intervalle de confiance

v.a. continues

v.a. discrètes

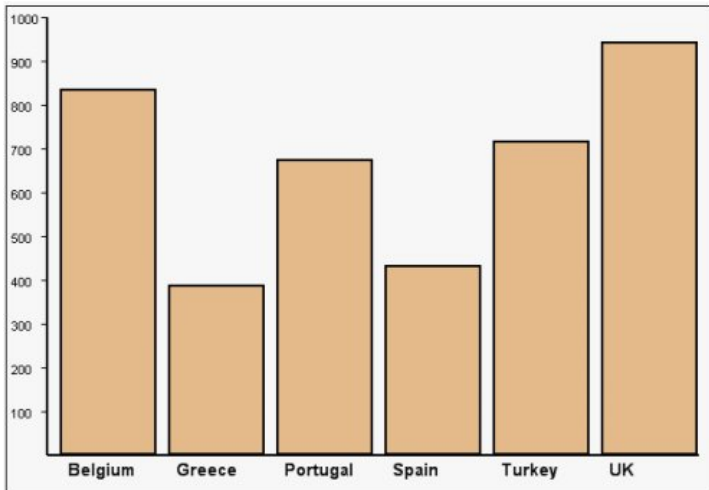
On s'intéresse à un v.a. aléatoire discrète X à p modalités à partir d'un échantillon de taille N :

Modalités	x_1	x_2	...	x_p
Effectifs	N_1	N_2	...	N_p
Fréquences	$f_1 = N_1/N$	$f_2 = N_2/N$	...	$f_p = N_p/N$

Deux représentations couramment utilisées :

- le graphique en bâtons
- le camembert
- le graphique en araignée

Graphique en bâtons



Statistiques
descriptives

v.a. continues

v.a. discrètes

Intervalle de
confiance

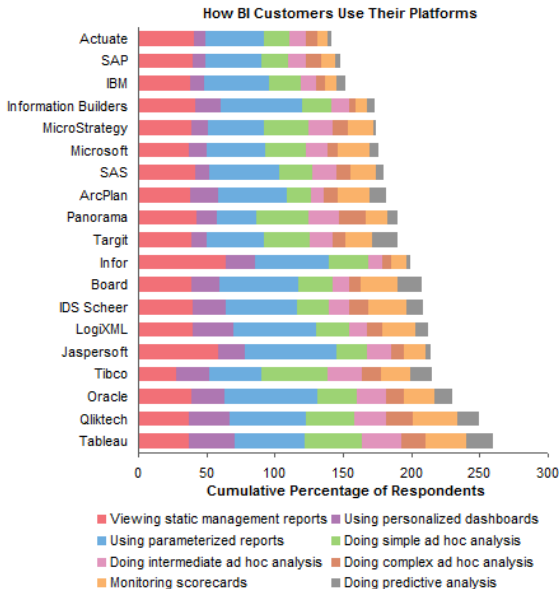
v.a. continues

v.a. discrètes

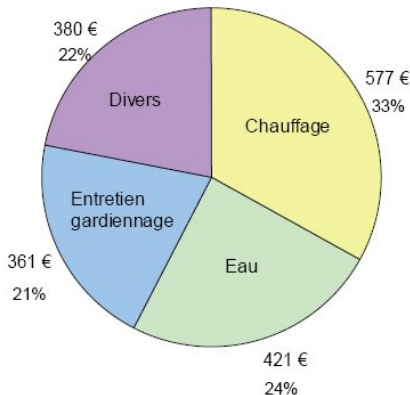
Graphique en bâtons

Statistiques
descriptives
v.a. continues
v.a. discrètes

Intervalle de
confiance
v.a. continues
v.a. discrètes



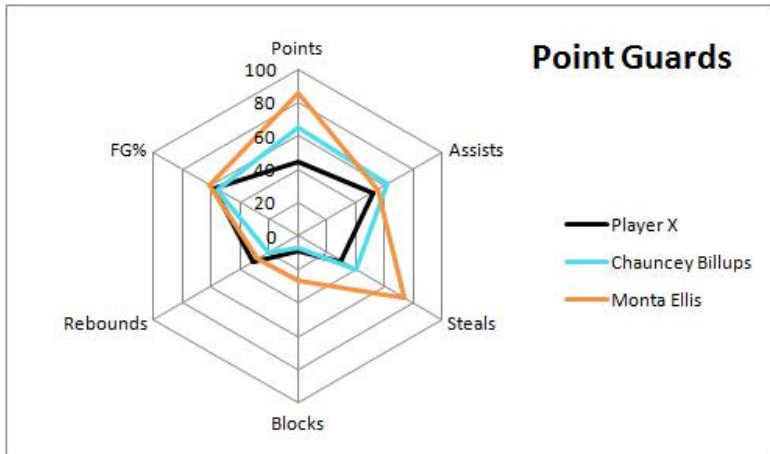
Exemple de répartition des postes de charges en 2006
(résidences équipées d'un chauffage, d'une production d'eau
chaude sanitaire collective et d'un ascenseur)



Source : étude sur les charges locatives en 2006.

Statistiques
descriptives
v.a. continues
v.a. discrètes

Intervalle de
confiance
v.a. continues
v.a. discrètes



Les tableaux de contingence

- On s'intéresse à 2 v.a. aléatoires discrètes.
- Exemple : Tests diagnostiques.

		Test diagnostique		Total
		T^+	T^-	
Statut de référence	M^+	a	b	$a + b$
	M^-	c	d	$c + d$
Total		$a + c$	$b + d$	N

- Le statut de référence est défini par le **gold standard**.
- Deux types d'erreurs :
 - b faux négatifs : sujets ayant un test négatif alors qu'ils sont malades.
 - c faux positifs : sujets ayant un test positif alors qu'ils sont sains.

Statistiques
descriptives

v.a. continues

v.a. discrètes

Intervalle de
confiance

v.a. continues

v.a. discrètes

1. Statistiques descriptives

v.a. continues

v.a. discrètes

2. Intervalle de confiance

v.a. continues

v.a. discrètes

Statistiques
descriptives

v.a. continues

v.a. discrètes

Intervalle de
confiance

v.a. continues

v.a. discrètes

1. Statistiques descriptives

v.a. continues

v.a. discrètes

2. Intervalle de confiance

v.a. continues

v.a. discrètes

- Soit un échantillon de taille N et $x_1, x_2, \dots, x_N$ les observations d'une variable aléatoire continue.
- La moyenne de la population (μ) est estimée par $\bar{x}$.
- Si $N > 30$ (TCL), alors on sait aussi que

$$\bar{X} \sim \mathcal{N}(\mu, s/\sqrt{N})$$

- En centrant et réduisant, on obtient :

$$\frac{\bar{X} - \mu}{s/\sqrt{N}} \sim \mathcal{N}(0, 1)$$

- Pour obtenir l'intervalle de confiance à 95% de la moyenne estimée, on cherche l'intervalle à l'intérieur duquel on a 95% de chance de retrouver la moyenne théorique de la population :

$$Pr(-1.96 < \frac{\bar{X} - \mu}{s/\sqrt{N}} < 1.96) = 0.95$$

Intervalle de confiance

Statistiques
descriptives

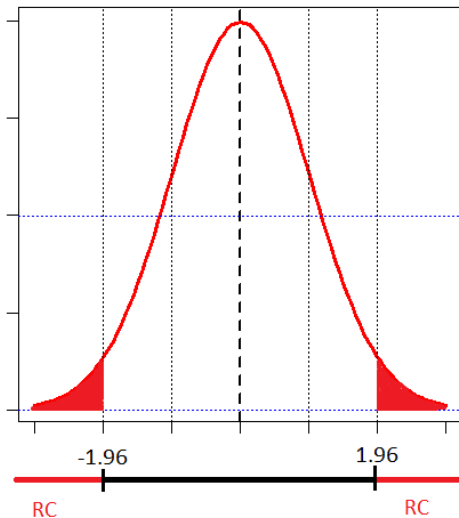
v.a. continues

v.a. discrètes

Intervalle de
confiance

v.a. continues

v.a. discrètes



Intervalle de confiance

Statistiques
descriptives

v.a. continues

v.a. discrètes

Intervalle de
confiance

v.a. continues

v.a. discrètes

$$Pr\left(-1.96 < \frac{\bar{X} - \mu}{s/\sqrt{N}} < 1.96\right) = 0.95$$

$\Leftrightarrow$

$$Pr\left(\bar{X} - 1.96 \frac{s}{\sqrt{N}} < \mu < \bar{X} + 1.96 \frac{s}{\sqrt{N}}\right) = 0.95$$

$\Leftrightarrow$

$$IC_{95\%}(\mu) = \left[\bar{x} \pm 1.96 \frac{s}{\sqrt{N}}\right]$$

- Exemple

On cherche à estimer l'âge moyen au diagnostic d'un cancer du sein. On observe une moyenne $\bar{x}$ de 58 ans et l'écart-type estimé de l'âge est à 30 ans (échantillons de 100 femmes). **Quelle est l'intervalle de confiance de la moyenne ?**

$$\begin{aligned} IC_{95\%}(\mu) &= \left[58 \pm 1.96 \frac{30}{\sqrt{100}} \right] \\ &= [52.12; 63.88] \end{aligned}$$

Si on réalise 100 échantillons comme celui-ci, on attend 95 moyennes estimées comprises entre 52 et 63 ans.

Statistiques
descriptives

v.a. continues

v.a. discrètes

Intervalle de
confiance

v.a. continues

v.a. discrètes

1. Statistiques descriptives

v.a. continues

v.a. discrètes

2. Intervalle de confiance

v.a. continues

v.a. discrètes

- Soit un échantillon de taille N et $x_1, x_2, \dots, x_i, \dots, x_N$ les observations d'une variable aléatoire discrète $X = \{0, 1\}$.
- La proportion p de $\{X = 1\}$ de la population est estimée par la fréquence $f = \sum_i x_i / N$.
- Comme précédemment grâce au théorème central limite, nous pouvons utiliser les propriétés de la loi normale.
- Si $N > 30$ $Np > 5$ et $N(1 - p) > 5$ (TCL), alors on sait aussi que

$$F \sim \mathcal{N}\left(p, \sqrt{\frac{p(1-p)}{N}}\right)$$

- En respectant les mêmes développements que précédemment, on démontre :

$$IC_{95\%}(p) = \left[f \pm 1.96 \sqrt{\frac{f(1-f)}{N}} \right]$$