STATISTIQUE DESCRIPTIVE 2

Daniel-Corneliu Leucuța

1

Objectifs

- Statistique descriptive
  - Donnes quantitatives:
    - Mesures de dispersion
      - dispersion, amplitude, écart interquartiles, variance, écart-type, coefficient de variation,
    - Asymétrie
    - Aplatissement
    - Tableaux et graphiques – continuation
      - Graphique des moyennes, des quantiles, boite a moustaches, ligne, nouage des points
    - La technique du chois des graphiques
    - L' évaluation de la normalité des données

2

MESURES DE DISPERSION :

Montrent si les valeurs sont plus ou moins proches autour de la moyenne (ou un autre « centre » - médiane) de l' échantillon

Quantifient le taux de variabilité des données autour d'une mesure de centralité

Exemples:

- amplitude
- écart interquartile
- variance
- écart-type
- coefficient de variation

3

MESURES DE DISPERSION

L'amplitude (l'entendue) (range – en anglais): la différence entre la plus grande et la plus petite observation.

= Xmax - Xmin

Avantages	Désavantages
✓**simple** a calculer ✓**mêmes unités** de mesure que la variable elle-même. ✓utilise pour décrire **l'âge** des sujets au début des études (mais on préfère de montrer l'interval minimun-maximun (range – en anglais, eg: 43-82) que de montrer l'amplitude!!!)	× elle donne un information concernant seulement « **deux** » **valeurs** dans la série des données ×cette mesure n'est pas très utile ×elle este **très sensible** a les valeur **extrêmes** ×n'indique **pas claire**ment comment les données **varient** autour de la moyenne

4

MESURES DE DISPERSION

L' écart interquartile (interquartile range / IQR – en anglais): la différence entre la quartile 3 et 1

=Q3 – Q1

au moins 50% des données dans la série des valeurs se retrouvent dedans cet intervalle.

5

MESURES DE DISPERSION

Groupe 1 Hauteur (cm)	Groupe 2 Hauteur (cm)
150	168
160	172
175	176
180	181
195	183

Moyennes

Amplitude	**45**=195-150	**15** =183-168
Q1	155	170
Q2 (median)	175	176
Q3	187.5	182
écart inter-quartile=Q3-Q1	**32.5**=187.5-155	**12** =182-170

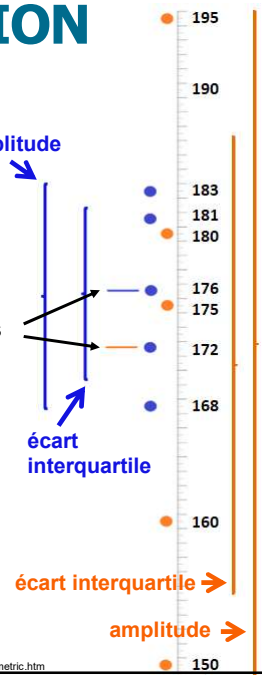


Image du centimetre: http://thehomeschoolzone.com/FREEBasics/FREEBasicRulerMeasurements-centimeterandmillimetermetric.htm

6

Propriétés de la moyenne des écarts de la moyenne/ médiane :

L' écart interquartile (interquartile range / IQR – en anglais): la différence entre la quartile 3 et 1

=Q3 – Q1

au moins 50% des données dans la série des valeurs se retrouvent dedans cet intervalle.

Avantages	Désavantages
✓**simple** a calculer ✓elle n'est **pas si sensible** au valeurs extrêmes comme l' amplitude ✓utilise pour décrire les **variables** quantitatives **non normale distribuées** (a suivre)(mais on préfère de montrer les quartiles au lieux de IQR) ✓**mêmes unités** de mesure que la variable elle-même.	×elle donne un information concernant seulement un **nombre réduit** («~ 4 ») des **valeurs** dans la série des données

7

MESURES DE DISPERSION

Pour mesurer la dispersion on est intéresse de la distance entre les valeurs et la valeur au « centre » de la série. Et on fait la moyenne des ces écarts. Ils utilisent toutes les valeurs de la série des donnes.

La moyenne des écarts a la moyenne

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

Si on n'ignore pas les signes elle est égale a 0 – n'apporte d'informations => on utilise la valeur absolue

Notations: x_i – valeur pour le sujet i, $\bar{x}$ - moyenne de l' échantillon, n – nombre des sujets dans l' échantillon, $\sum_{i=1}^{n}()$ - la somme des toutes les valeurs depuis la première jusqu'a n

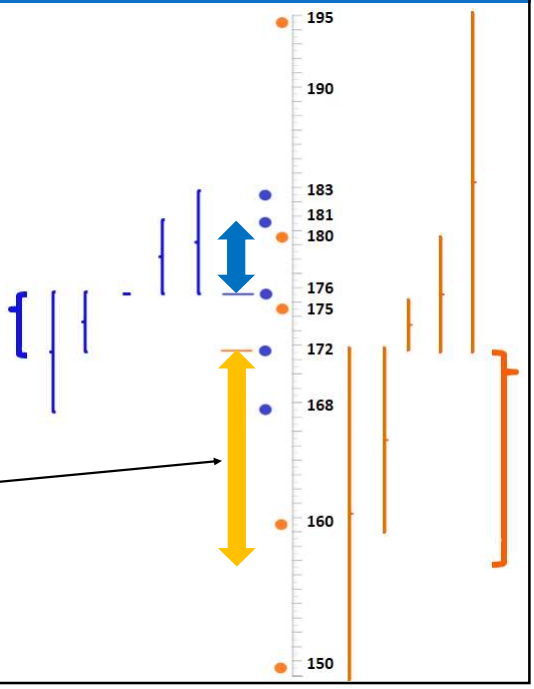
8

MESURES DE DISPERSION

- **La moyenne des écarts de la moyenne**

= (|195 – 172| +
|180 – 172| +
|175 – 172| +
|160 – 172| +
|150 – 172|) / 5
= (22 + 12 + 3 + 8 + 23) /5
= **14 cm**

(la moyenne est égale a 172)



9

MESURES DE DISPERSION

Groupe 1			Groupe 2		
Hauteur	x-m	|x-m|	Hauteur	x-m	|x-m|
150	-22	22	168	-8	8
160	-12	12	172	-4	4
175	3	3	176	0	0
180	8	8	181	5	5
195	23	23	183	7	7
Moyenne	172			176	
somme(x-m)/5	0		somme(x-m)/5	0	
somme|x-m|/5		14	somme|x-m|/5		4.8

10

Propriétés de la moyenne des écarts de la moyenne :

Avantages	Désavantages
✓utilise **toutes les valeurs** de la série des données ✓**mêmes unités** de mesure que la variable elle-même. ✓**positive**	×elle est **sensible** a les valeurs **extrêmes** ×on **ne peut** pas faire des **calculs** arithmétiques !!

11

MESURES DE DISPERSION

Mais la valeur absolue n'a pas des bons propriétés mathématiques, donc, on fait des carrées des écarts. Puis on fait la somme des carrées, et enfin la moyenne des carrées.

La variance (variance – en anglais)**:** mesure la dispersion des données autour de la moyenne

La variance d'une population
- notation sigma^2

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

La variance d'un échantillon:
- est la variance descriptive
- utilise pour décrire les échantillons
- sous-estime la variance de la population
- notation s^2 minuscule

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

Notations: μ - moyenne de la population, σ^2 – variance de la population, σ – écart type ou déviation standard de la population, n – nombre des sujets dans l'échantillon, N – nombre des sujets dans la population, X_i – valeur pour le sujet I, s (minuscule) ^{2} - variance descriptive ou de l' échantillon, s (minuscule) – écart type ou déviation standard descriptive ou de l' échantillon, $\bar{x}$ - moyenne de l' échantillon, $\sum_{i=1}^{n}()$ - la somme des toutes les valeurs depuis la première jusqu'a n

12

MESURES DE DISPERSION

- la variance d' échantillonnage d'un échantillon:

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

$$S^2 = \frac{n}{n-1} s^2$$

- corrige l'erreur faite par la formule précédente
- utilise dans la statistique inférentielle (analytique)
- notation S majuscule ^{2}

- Notations: S (majuscule) ^{2} - variance d' échantillonnage, x_i – valeur pour le sujet I, s (majuscule) – écart type ou déviation standard de l' échantillonnage, $\bar{x}$ - moyenne de l' échantillon, n – nombre des sujets dans l'échantillon, $\sum_{i=1}^{n}()$ - la somme des toutes les valeurs depuis la première jusqu'a n

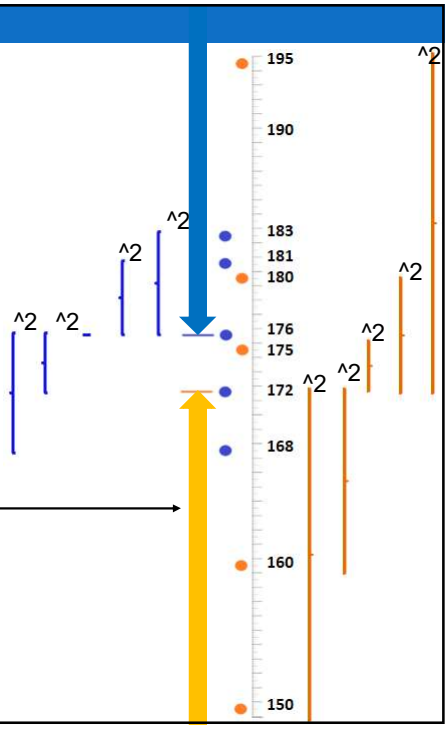
13

MESURES DE DISPERSION

- **La variance**

= ((195 – 172)^2 +
(180 – 172)^2 +
(175 – 172)^2 +
(160 – 172)^2 +
(150 – 172)^2) / 5
=(22^2 + 12^2 + 3^2 + 8^2 + 23^2)/5
= **246 cm^2**

(la moyenne est égale a 172)



14

Propriétés de la variance :

Avantages	Désavantages
✓utilise **toutes les valeurs** de la série des données ✓**positive** ✓on **peut faire des calculs** arithmétiques	×les unités de la variance = le **carré des unités** de la variable ×grande variance => grande dispersion autour de la moyenne ×=> **difficile de comprendre** et de faire des comparaisons ×elle est **sensible** a les valeurs **extrêmes**

15

Variance

	Groupe 1				Groupe 2		
	Hauteur		x-m	(x-m)²	Hauteur	x-m	(x-m)²
	150		-22	484	168	-8	64
	160		-12	144	172	-4	16
	175		3	9	176	0	0
	180		8	64	181	5	25
	195		23	529	183	7	49
Moyenne	172				176		
		somme des carrées ou somme des écarts		1230	cm²		154
s²		**variance descriptive**		**246**	**cm²**		**30.8**
S²		**variance d'échantillonnage**		**307.5**	**cm²**		**38.5**

16

MESURES DE DISPERSION

- **L'écart type (la déviation standard - DS) (standard deviation / SD – en anglais)**
- Est la racine carrée de la variance
- La déviation standard (notation s petit) d'un échantillon est défini comme suit:

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

- La déviation standard d'échantillonage (notation S – s grand)

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

- la déviation standard populationnelle est (notation σ - sigma):

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{N}}$$

Notations: μ - moyenne de la population, σ – écart type ou déviation standard de la population, n – nombre des sujets dans l'échantillon, N – nombre des sujets dans la population, X_i – valeur pour le sujet I, s (minuscule) ^{2} - variance descriptive ou de l' échantillon, s (minuscule) – écart type ou déviation standard descriptive ou de l' échantillon, S (majuscule) ^{2} - variance d' échantillonnage, x_i – valeur pour le sujet I, s (majuscule) – écart type ou déviation standard de l' échantillonnage, $\bar{x}$ - moyenne de l' échantillon, $\sum_{i=1}^{n}()$ - la somme des toutes les valeurs depuis la première jusqu'a n

17

MESURES DE DISPERSION

- **L'écart type**

(déviation standard)

=Racine carre de

((195 – 172)^2 +

(180 – 172)^2 +

(175 – 172)^2 +

(160 – 172)^2 +

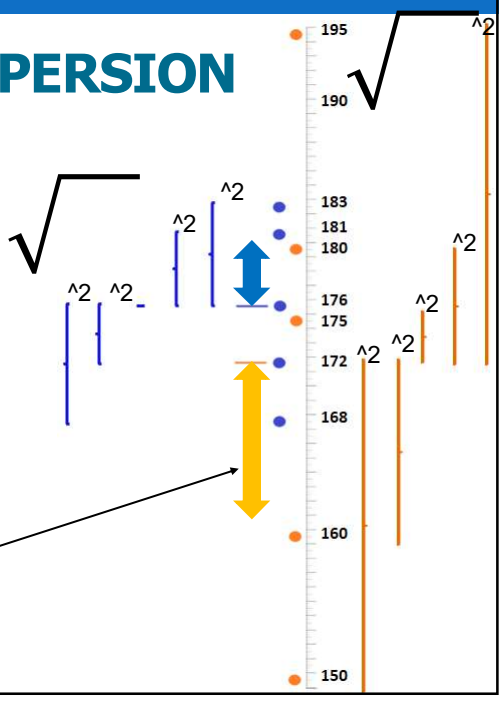
(150 – 172)^2) / 5

=√(22^2 + 12^2 + 3^2 + 8^2 + 23^2)/5

= √246

= 15.68 cm

(la moyenne est égale a 172)



18

L' écart type (la déviation standard)

	Groupe 1 Hauteur		x-m	$(x-m)^2$	Groupe 2 Hauteur	x-m	$(x-m)^2$
	150		-22	484	168	-8	64
	160		-12	144	172	-4	16
	175		3	9	176	0	0
	180		8	64	181	5	25
	195		23	529	183	7	49
Moyenne	172				176		
		somme des carrées ou somme des écarts		1230	cm^2		154
	s^2	variance descriptive		246	cm^2		30.8
	S^2	variance d'échantillonnage		307.5	cm^2		38.5
	s	**écart type descriptive**		**15.68**	**cm**		**5.55**
	S	**écart type d'échantillonnage**		**17.54**	**cm**		**6.20**

19

Propriétés de l'écart type :

Avantages

- ✓ utilise **toutes les valeurs** de la série des données
- ✓ **mêmes unités** de mesure que la variable elle-même.
- ✓ **positive**
- ✓ on **peut faire des calculs** arithmétiques
- ✓ si on multiplie les valeurs d'une série statistique avec une constante, l' écart type se multiplie avec la même constante
- ✓ si on ajoute une valeur a chaque valeur d'une série statistique, l' écart type ne se modifie pas

Désavantages

- × Elle est **sensible** a les valeurs **extrêmes**

20

Propriétés de l'écart type :

•Pour la plupart des distributions uni modales:

- Au moins 50% des données se trouvent à 1 écart-type de la moyenne.
- Au moins 75% à 95% des données se situent à 2 écarts-types de la moyenne.
- La quasi-totalité des données se situent à 3 écarts-types de la moyenne.

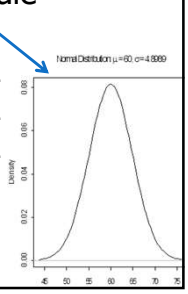
Pour une série des données avec un distribution normale (gaussienne):

- Dans l'intervalle m +/- **1** DS on trouve **~68**,3 **%** de la population.
- Dans l'intervalle m +/- **2** DS on trouve **~95**,5 **%** de la population.
- Dans l'intervalle m +/- **3** DS on trouve **~99**,7 **%** de la population.

(voir le cours avec les variables aléatoires)

m - moyenne

DS = déviation standard, ou écart type



21

MESURES DE DISPERSION

Le coefficient de variation

quel pourcentage est la déviation standard par rapport a la moyenne

$$CV = \frac{s}{x} \times 100$$

Avantages	Désavantages
✓utilise **toutes les valeurs** de la série des données ✓permet la **comparaison** de la **dispersion** de **deux variables** qui ont des moyennes différentes ✓il n'a **pas une unité** de mesure	×beaucoup des fois dans le domaine médical on est intéressée de l' utilité clinique et on a besoin de l' unité de mesure

Notations: s (minuscule) – écart type ou déviation standard descriptive ou de l' échantillon, $\bar{x}$ - moyenne de l' échantillon,

22

MESURES DE DISPERSION

L' interprétation du coefficient de variation d'une série des données

Coefficient de variation	Interprétation
CV < 10% (dev. std. très petite)	homogène
10% ≤ CV <20%	relativement homogène
20% ≤ CV <30%	relativement hétérogène
CV ≥ 30% (dev. std. très grande)	hétérogène

23

Le coefficient de variation

	Groupe 1				Groupe 2		
	Hauteur		x-m	$(x-m)^2$	Hauteur	x-m	$(x-m)^2$
	150		-22	484	168	-8	64
	160		-12	144	172	-4	16
	175		3	9	176	0	0
	180		8	64	181	5	25
	195		23	529	183	7	49
Moyenne	172				176		
		variance		246	cm^2		30.8
		ecart type		15.68	cm		5.55
		Coefficient de variation		**0.09**	-		**0.03**
		=15,68/172=		**9%**	=5,55/176		**3%**
		Interpretation du CV		homogène			homogène

24

Comparatif des mesures de dispersion

Mesure de dispersion	Groupe 1	Groupe 2	Unite de mesure	Problèmes
amplitude	45.0	15.0	cm	Sensible a v. extrêmes, dépend de seulement 2 valeurs
écart interquartile	32.5	12.0	cm	Dépend de seulement quelques valeurs
moyenne des écarts de la moyenne	14.0	4.8	cm	Les propriétés mathématiques du valeur absolue les rends difficile a être utilise
variance descriptive	246.0	30.8	cm^2	S'exprime dans le carrée de l'unité de mesure, sensible au valeurs extrêmes
écart type descriptive	15.7	5.6	cm	Sensible au valeurs extrêmes
coefficient de variation	0,09	0,03	-	N'a pas d'unité de mesure

25

Exemples dans des articles

Observez l'utilisation du: moyenne, médiane, quartiles, %, nombre des sujets, dans les résultats des articles scientifiques:

Characteristics	Treatment Group (N=257)	Control Group (N=257)
Age, mean (sd), years	56.7 (10.5)	57.9 (9.6)
Female, No. (%)	114 (44.4)	123 (47.9)
Smoking history, No. (%)		
Never smoked	129 (50.2)	144 (56.0)
Former	89 (34.6)	86 (33.5)
Current	39 (15.2)	27 (10.5)
Diabetes factors, mean (SD)		
HbA1c, %	7.8 (0.65)	7.8 (0.60)
Fasting glucose, mg/dL , median	150	147
IQR (quartile 1 - quartile 3)	(125 – 174)	(122 – 172)
Duration of diabetes, years	12.3 (8.2)	11.3 (8.4)
Anthropometrics, mean (SD)		
Weight, kg	99.5 (24.3)	97.5 (21.7)
BMI, kg/m^2	34.7 (7.5)	34.2 (6.7)
Blood pressure[c], mean(sd), mm Hg		
Systolic	133.1 (20.7)	135.1 (20.4)
Diastolic	78.8 (12.3)	78.8 (10.9)

Pour le poids (weight) – le fait qu'ils montre la moyenne et la déviation standard signifie que les données sont normale distribuées (on suppose q'ils ont vérifiée ca mais ils ne montre pas cette étape).
La moyenne de 99,5 nous **indique** ou est le centre des valeurs du poids dans le group avec traitement
La **déviation standard (SD – standard deviation)** de 24,3 nous indique que dans l' intervalle moyenne de 99,5 moins 24,3 (1DS) et moyenne plus un déviation standard, il y a approximative 68% des poids des sujets

Depuis: Engebretson SP, Hyman LG, Michalowicz BS, Schoenfeld ER, Gelato MC, Hou W, Seaquist ER, Reddy MS, Lewis CE, Oates TW, Tripathy D, Katancik JA, Orlander PR, Paquette DW, Hanson NQ, Tsai MY. **The effect of nonsurgical periodontal therapy on hemoglobin A1c levels in persons with type 2 diabetes and chronic periodontitis: a randomized clinical trial.** JAMA. 2013 Dec 18;310(23):2523-32.

26

Exemples dans des articles

Observez l'utilisation du: moyenne, médiane, quartiles, %, nombre des sujets, dans les résultats des articles scientifiques:

Characteristics	Treatment Group (N=257)	Control Group (N=257)
Age, mean (sd), years	56.7 (10.5)	57.9 (9.6)
Female, No. (%)	114 (44.4)	123 (47.9)
Smoking history, No. (%)		
Never smoked	129 (50.2)	144 (56.0)
Former	89 (34.6)	86 (33.5)
Current	39 (15.2)	27 (10.5)
Diabetes factors, mean (SD)		
HbA1c, %	7.8 (0.65)	7.8 (0.60)
Fasting glucose, mg/dL , median	150	147
IQR (quartile 1 - quartile 3)	(125 – 174)	(122 – 172)
Duration of diabetes, years	12.3 (8.2)	11.3 (8.4)
Anthropometrics, mean (SD)		
Weight, kg	99.5 (24.3)	97.5 (21.7)
BMI, kg/m^2	34.7 (7.5)	34.2 (6.7)
Blood pressure[c], mean(sd), mm Hg		
Systolic	133.1 (20.7)	135.1 (20.4)
Diastolic	78.8 (12.3)	78.8 (10.9)

Depuis: Engebretson SP, Hyman LG, Michalowicz BS, Schoenfeld ER, Gelato MC, Hou W, Seaquist ER, Reddy MS, Lewis CE, Oates TW, Tripathy D, Katancik JA, Orlander PR, Paquette DW, Hanson NQ, Tsai MY. **The effect of nonsurgical periodontal therapy on hemoglobin A1c levels in persons with type 2 diabetes and chronic periodontitis: a randomized clinical trial.** JAMA. 2013 Dec 18;310(23):2523-32.

Pour la glucose a jeun (Fasting glucose) – le fait qu'ils montre la médiane et les quartiles signifie que les données ne sont pas normale distribuées (on suppose q'ils ont vérifiée ca mais ils ne montre pas cette étape). **La médiane** de 150 nous **indique** que au moins 50% des sujets ont les valeurs de glucose <= 150 (ou 150 divise la série des données dans deux parties égales) dans le group avec traitement Les quartiles **(IQR – interquartile range)** de 125-174 nous indique que dans l' intervalle interquartile, il y a approximative au moins 50% des valeurs de la glucose des sujets. Aussi on sait que au moins 25% des sujets on des valeurs <=125, et au moins 75% des sujets on des valeurs <=174 dans le group avec traitement

27

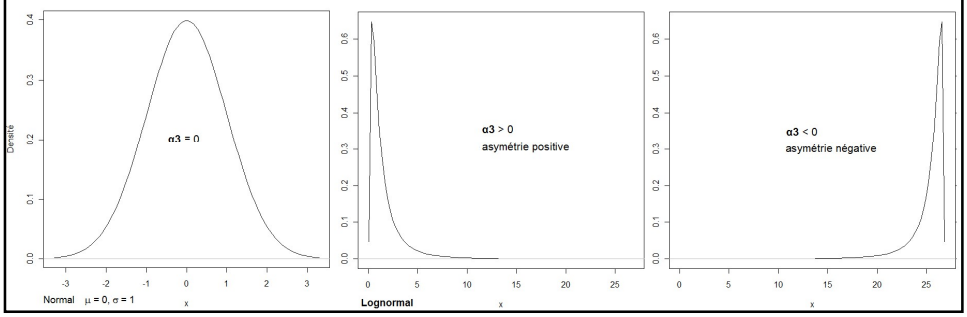
Mesures de symétrie

Coefficient d'asymétrie (α_3) (skewness – en anglais):
degré d'asymétrie d'une distribution
la direction de cette asymétrie (positive ou négative);

$$\alpha_3 = \frac{1}{s^3} \frac{\sum_{i=1}^{n} (x_i - \overline{X})^3}{n}$$

α3 ≈ 0 => une distribution symétrique.
α3 > 0 => distribution est plus allongée vers la droite – asymétrie positive
α3 < 0 => distribution est plus allongée vers la gauche – asymétrie négative

α3 (-1 – 1) suggestion que la distribution est normale
α3 < -1 ou >1 suggestion que la distribution n'est pas normale



28

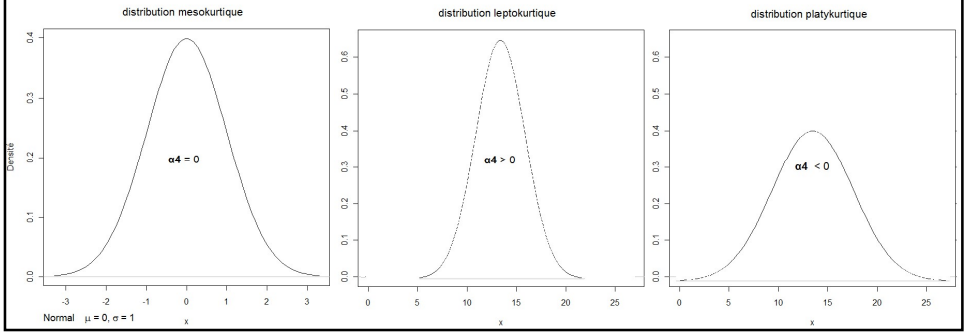
Le coefficient d'aplatissement

Le coefficient d'excèss d'aplatissement (α_4) (excess kurtosis – en anglais):
l'angle de la courbe du milieu d'une distribution
par rapport a une distribution normale (gaussienne)

$$\alpha_4 = \frac{1}{s^4}\frac{\sum_{i=1}^{n}(x_i - \overline{X})^4}{n} - 3$$

- **α4** ≈ 0 => l'angle normal => distribution mesokurtique
- **α4** > 0 => l'angle aigu => distribution leptokurtique - centre élevée
- **α4** < 0 => la pente aplaté => distribution platykurtique – centre plus bas

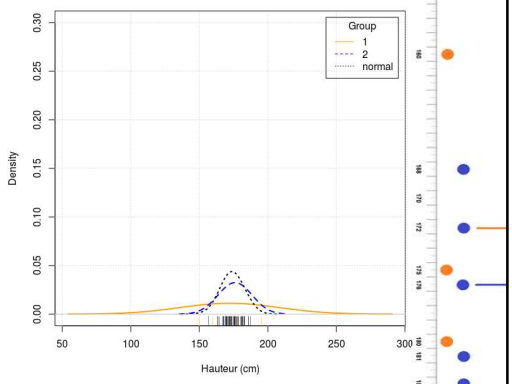
α4 (-1 – 1) suggestion de distribution normale
α4 < -1 ou >1 suggestion que la distribution n'est pas normale



29

Asymétrie, aplatissement

Groupe 1	Groupe 2
Hauteur	Hauteur
150	168
160	172
175	176
180	181
195	183



c. asymétrie	**0.0254997**	**-0.188374**
c. aplatissement	**-0.874479**	**-1.77804**

30

TABLEAU RÉCAPITULATIF

Mesures de tendance centrale:	**Mesures de dispersion:**
✓ **Moyenne** ✓utilisée fréquemment ✓ **Médiane** ✓utilisée fréquemment ✓ Mode	✓ Amplitude (entendue) (utilisée rarement) ✓ Intervalle interquartile (utilisée rarement, idéal de montrer les quartiles) ✓moyenne des écarts de la moyenne (utile didactique pour comprendre la déviation standard) ✓Variance (utilise fréquemment, mais dans l'appareil mathématiques au delà de l'article) ✓ **Déviation standard** (**écart-type**) ✓Utilisée fréquemment ✓ Coefficient de variation
Mesures de symétrie/aplatissement:	**Mesures de position:**
✓ Coefficient **d'asymétrie** (skewness) ✓ Coefficient **d'aplatissement** (Kurtosis) ✓ne sont pas montrées dans les articles mais ils peuvent être utilisée pour évaluation de la normalité des données	✓ **Quartiles** ✓utilisée fréquemment ✓on préfère montrer les quartiles que l'intervalle interquartile car elles sont plus informatives! ✓ Déciles ✓ Percentiles

31

Continuation des graphiques utilises dans les articles scientifiques médicales

- Graphiques pour des variables quantitatives
  - Pour évaluer la forme de la distribution
    - Histogramme
    - Graphique des quantiles
  - Pour des variables normale distribuées
    - Graphique des moyennes
    - Graphique colonnes avec barre d' erreur
  - Pour des variables non-normale distribuées
    - Graphique boite a moustaches
    - Graphique de bande de gigue
    - Graphique en essaim
- Graphiques pour la relation entre deux variables quantitative
  - Graphique nuage des points
- Graphiques pour montrer l'évolution dans le temps des variables qualitatives ou quantitatives
  - Graphique ligne

32

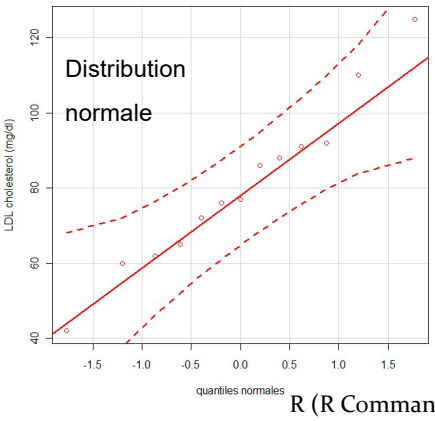
Une variable quantitative: Graphique des quantiles

- (quantile-quantile plot / QQ plot – en anglais)
- Permet de comparer deux distributions
- On peut comparer la distributions de la série des données observées (les points) avec un distribution théorétique (normale – la ligne)
  - Si les points sont sur la ligne – distribution approximative normale
  - Si les points s'éloigne de la ligne – distribution non normale
- La meilleur façon d'évaluer la normalité des données, mieux que l'histogramme

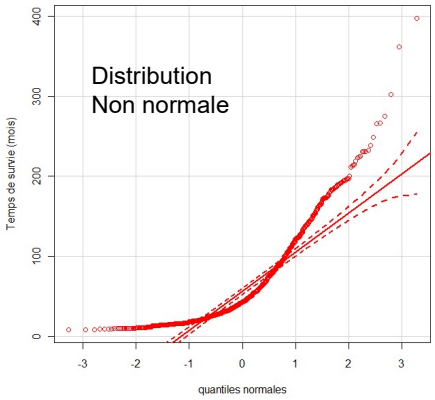
33

Une variable quantitative: Graphique des quantiles

La distribution du LDL cholestérol chez un group des sujets avec la maladie Gaucher



Le temps de survie pour un group des sujets dialysée



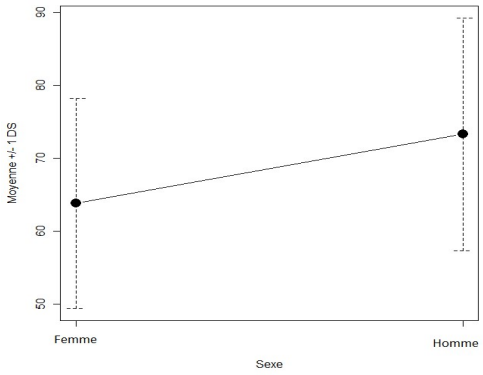
Excel, EpiInfo:
✓-n'existe pas

R (R Commander):
✓KMggplot2/Q-Q plot…
✓ou Graphs/Quantile-comparison plot…

34

Pour des variables **quantitatives**: graphique des indicateurs: moyennes et déviations standard (pour la représentation des données a distribution **normale**)

- Le poids âpres la dialyse en fonction du sexe



Pour une **seule variable quantitative** on va avoir un seule représentation graphique – un point avec des barres d'erreur.
Si on montre **la relation** entre **une variable quantitative** et **une variable qualitative**, sur un axe on a la représentation du variable quantitative (Moyenne, déviation standard) et sur l'autre axe on a les groups. Ici, on observe la relation entre le sexe (deux groups) et le poids.

R (R Commander):
✓**KMggplot2/Box plot …**
✓**ou Graphs/Plot of means…**

EpiInfo:
✓**STATISTICS/GRAPH/BOX-WHISKER – Mean-1SD-2SD**

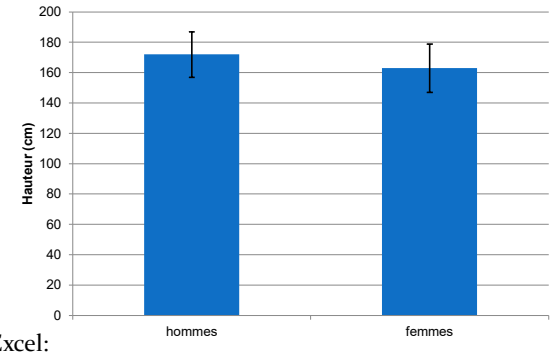
Excel:
✓**INSERT CHART: type: STOCK – HIGH-LOW-CLOSE**

35

Pour des variables **quantitatives**: graphique des indicateurs: moyennes et déviations standard (pour la représentation des données a distribution **normale**)

- L'hauteur moyenne (déviation standard) en fonction du sexe dans l' échantillon étudiée

Graphique Colonnes avec barre d' erreur



Excel:
✓**INSERT CHART: type: Columns avec Error Bars**

Pour une **seule variable quantitative** on va avoir un seule représentation graphique – un colonne avec des barres d'erreur.
Si on montre **la relation** entre **une variable quantitative** et **une variable qualitative**, sur un axe on a la représentation du variable quantitative (Moyenne, déviation standard) et sur l'autre axe on a les groups. Ici, on observe la relation entre le sexe (deux groups) et l'hauteur.

36

37

Une variable quantitative: graphiques des indicateurs (de position: quartiles)

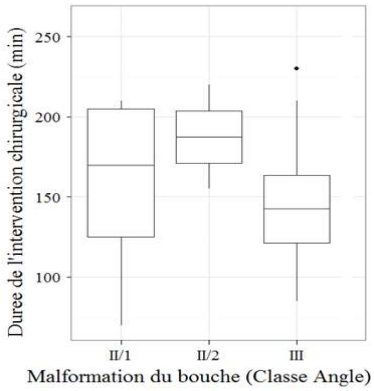
- **la boîte à moustache** (le box-plot / box and wiskers – en anglais) = graphique **utile** pour **visualiser la distribution d'une variable quantitative**
- pour la représentation des données avec **distribution non normale**
- On la **construit** de la manière suivante :

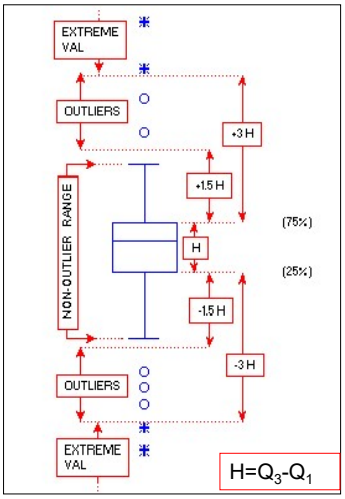
- on trace une boîte de longueur Q_3-Q_1
- on partage la boîte par un trait à la position de la médiane
- on trace la moustache de gauche/inferieur de longueur le point le plus éloignée jusqu'à -1,5*(Q_3-Q_1) $\min(Q_1 - X_{\min}, 1.5*(Q_3 - Q_1))$
- on trace la moustache de droite/supérieur de longueur le point le plus éloignée jusqu'à 1,5*(Q_3-Q_1) $\min(X_{\max} + Q_3, 1.5*(Q_3 - Q_1))$
- si certains individus sont en dehors des moustaches, on les représente par des * (valeurs extrêmes > 3*(Q_3-Q_1) et ° (valeurs aberrantes [en anglais outliers] > 1,5*(Q_3-Q_1) et < 3*(Q_3-Q_1)). On peut nommer tout valeur même extrême comme valeur aberrante

- Les moustaches **peuvent être construits avec autres techniques** (ex percentile 2,5, et 97,5, ou le minimum et le maximum

37

Une variable **quantitative**: boite a moustaches (en anglais: box-plot/box and wiskers) (pour la représentation des données a distribution **non normale**)





R (R Commander):
KMggplot2/Box plot/Error bar plot…
ou, Graph/Box plot…

38

Une variable quantitative: Graphique boite a moustaches (en anglais box and wiskers/box plot)

- Avec le graphique boite a moustaches on peut déduire l'asymétrie de la distribution
- a) approximative symétrique (qui est une suggestion de normalité des données)

Suggestions: la distance entre la Q3 (la ligne supérieure du boite, pour le graphique en position verticale, ou la ligne droite de la boite pour le graphique en position horizontale) et la médiane (la ligne a l'intérieur du boite) approximative égale avec la distance entre la médiane et Q1 (la ligne inferieure du boite, pour le graphique en position verticale, ou la ligne gauche de la boite pour le graphique en position horizontale); **et** les moustaches égales.

- b) une asymétrie a droite / positive (qui est une suggestion de non normalité des données)

Suggestions: la distance entre la Q3 (la ligne supérieure du boite, pour le graphique en position verticale, ou la ligne droite de la boite pour le graphique en position horizontale) et la médiane (la ligne a l'intérieur du boite) plus grande que la distance entre la médiane et Q1 (la ligne inferieure du boite, pour le graphique en position verticale, ou la ligne gauche de la boite pour le graphique en position horizontale); ou la moustache en haut est plus grande que la moustache en bas (pour le graphique en position verticale); ou la moustache a droite est plus grande que la moustache a gauche (pour le graphique en position horizontale); ou les deux.

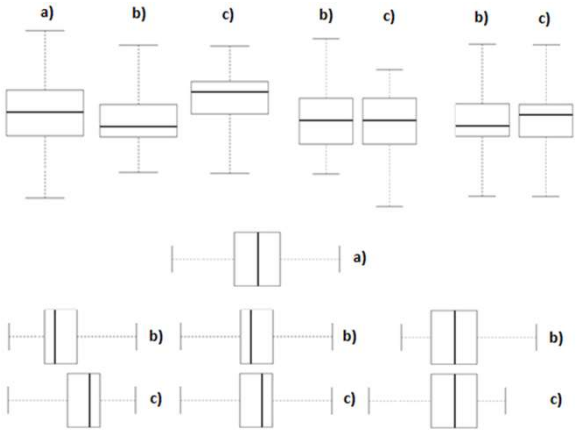
- c) une asymétrie a gauche / négative (qui est une suggestion de non normalité des données)

Suggestions: la distance entre la Q3 (la ligne supérieure du boite, pour le graphique en position verticale, ou la ligne droite de la boite pour le graphique en position horizontale) et la médiane (la ligne a l'intérieur du boite) plus petite que la distance entre la médiane et Q1 (la ligne inferieure du boite, pour le graphique en position verticale, ou la ligne gauche de la boite pour le graphique en position horizontale); ou la moustache en haut est plus petite que la moustache en bas (pour le graphique en position verticale); ou la moustache a droite est plus petite que la moustache a gauche (pour le graphique en position horizontale); ou les deux.

39

Une variable quantitative: Graphique boite a moustaches (en anglais box and wiskers/box plot)

- Avec le graphique boite a moustaches on peut déduire l'asymétrie de la distribution
- a) approximative symétrique – suggestion de normalité
- b) une asymétrie a droite / positive – suggestion de non normalité
- c) une asymétrie a gauche / négative – suggestion de non normalité



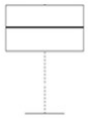
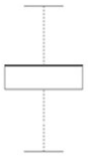
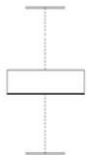
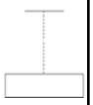
EpiInfo:
✓**STATISTICS/GRAPH/BOX-WHISKER – Median-25%-10%**

Excel:
✓**INSERT CHART: type: STOCK – OPEN-HIGH-LOW-CLOSE**

40

Une variable quantitative: Graphique boite a moustaches (en anglais box and wiskers/box plot)

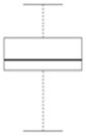
- Autres situations identifiables avec les boites a moustaches

La moustache supérieure est identique avec la quartile 3 (asymétrie a gauche, suggestion de non normalité) 	La moustache inferieure est identique avec la quartile 1 (asymétrie a droite, suggestion de non normalité) 
La quartile 3 est égale a la médiane (asymétrie a gauche, suggestion de non normalité) 	La quartile 1 est égale a la médiane (asymétrie a droite, suggestion de non normalité) 
La moustache supérieure est identique avec la quartile 3 et avec la médiane (asymétrie a gauche, suggestion de non normalité) 	La moustache inferieure est identique avec la quartile 1 et avec la médiane (asymétrie a droite, suggestion de non normalité) 

41

Une variable quantitative: Graphique boite a moustaches (en anglais box and wiskers/box plot)

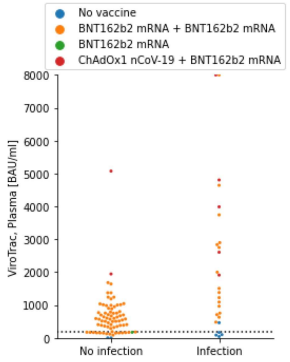
- Autres situations identifiables avec les boites a moustaches

La quartile 3 est égale a la médiane et la quartile 1 (suggestion symétrie, mais le fait que q1=q2=q3 – suggestion de non normalité) 	La moustache supérieure est égale avec la quartile 3 et la médiane et la quartile 1 (suggestion asymétrie a gauche, suggestion de non normalité) 
Situation avec asymétrie indécidable (la moustache supérieure est plus petite que la moustache inferieure [indique asymétrie a gauche], mais la distance entre la Q3 et la médiane est plus grande que la distance entre la médiane et Q1 [indique asymétrie a droite]). Toutefois - suggestion de non normalité 	La moustache supérieure est égale avec la quartile 3 et la médiane et la quartile 1 et la moustache inferieure (suggestion symétrie, mais le fait que q1=q2=q3 – suggestion de non normalité)

42

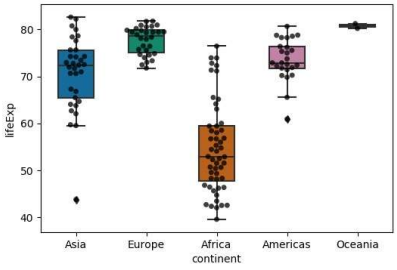
Une variable quantitative: Graphique en essaim (en anglais swarmplot)

- graphique utile pour **visualiser la distribution d'une variable quantitative**
- pour la représentation des données avec un **distribution non normale**
- **Utilisée d'habitude quand le nombre des données est réduit**
- Fréquemment il est superposée sur un graphique boite a moustaches
- **Toutes les observations** sont représentées sur le graphique, et **elle ne se superposent pas**



Exemple niveau anticorps anti SARS-COV-2

https://doi.org/10.1101/2021.09.17.21263729



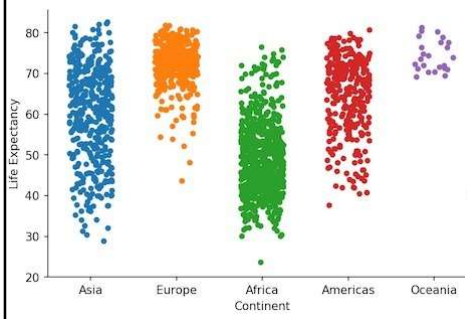
Exemple graphique espérance de vie en function du continent de bande de gigue avec graphique box and wiskers

https://cmdlinetips.com/2018/03/how-to-make-boxplots-in-python-with-pandas-and-seaborn/

43

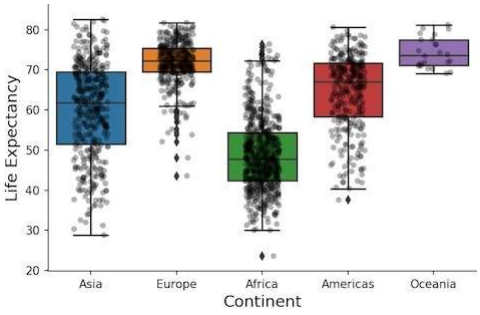
Une variable quantitative: Graphique de bande de gigue (en anglais jitter strip chart)

- graphique utile pour **visualiser la distribution d'une variable quantitative**
- pour la représentation des données avec un **distribution non normale**
- **Utilisée d'habitude quand le nombre des données est important**
- Fréquemment il est superposée sur un graphique boite a moustaches
- **Toutes les observations** sont représentées sur le graphique, et **mais elle peuvent se superposer**



Exemple espérance de vie en function du continent

https://cmdlinetips.com/2019/03/catplot-in-seaborn-python/

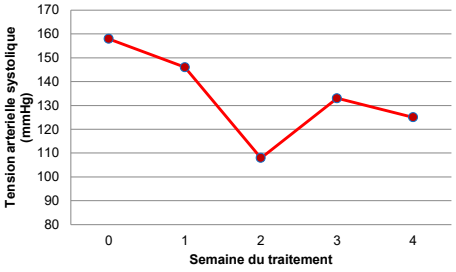


Exemple graphique de bande de gigue avec graphique box and wiskers,

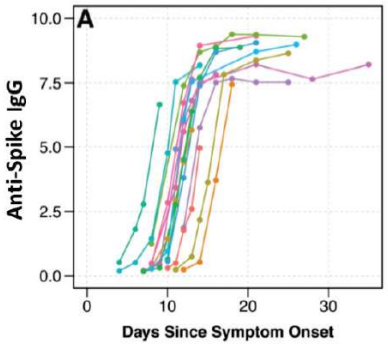
44

L'evolution d' une variable quantitative, ou qualitative, ou même évolution des statistiques (moyennes, fréquences, pourcentages) dans le temps: graphique linéaire

L'évolution de la tension artérielle systolique pendant le traitement



L'évolution du niveau des anticorps anti-spike IgG en fonction du jour depuis le debout des symptômes



Excel:
✓**INSERT CHART: type: LINE**

R (R Commander):
✓**KMggplot2/Line chart…**
✓**ou Graphs/Line graph…**

EpiInfo:
✓**STATISTICS/GRAPH/LINE**

45

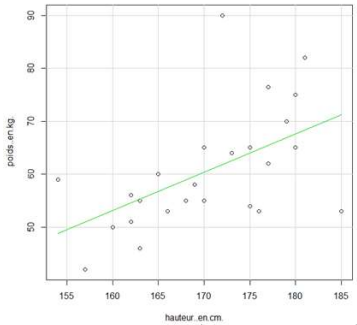
46

Deux variables quantitatives: diagramme par un nuage de points

(scatter plot – en anglais)

Le nouage des points peut être utilise aussi pour la relation entre deux variables qualitatives ordinales, ou pour la relation entre une variable quantitative, et un variable qualitative ordinale

Montre la relation direct/inverse proportionnelle, linéaire ou pas.



relation direct proportionnelle, linéaire.

Excel:
✓**INSERT CHART: type: XY SCATTER**

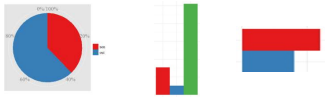
R (R Commander):
✓**KMggplot2/Scatter plot…**
✓**ou Graphs/Scatterplot…**

EpiInfo:
✓**STATISTICS/GRAPH/SCATTER XY**

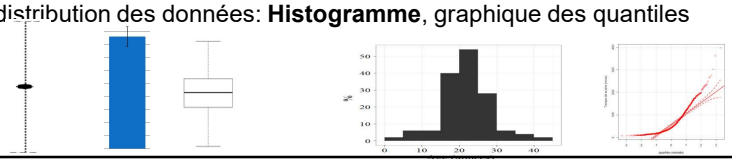
46

Le chois du type du graphique en fonction des types des variables et but

- Pour faire la **choix**, comptez **combien des variables** sont et quel **est leur type.**
- **Description <u>d'une seule</u> variable**
  - ***Qualitative***
    - camembert (sectoriel – **Pie**)
    - **Colonne** (si les noms des catégories ne sont pas très longues) – plus claire
    - **Bar** (si les noms des catégories sont très longues) – plus claire



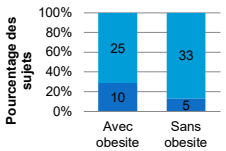
  - ***Quantitative***
    - montrées dans des articles: graphique des moyennes, colonnes avec barres d'erreur, box and wiskers (boite a moustaches),
    - d'habitude moins montrées, mais faites avant écrire l'article pour évaluer la distribution des données: **Histogramme**, graphique des quantiles



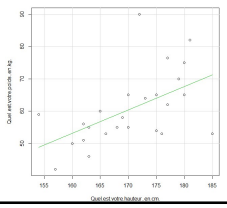
47

Le chois du type du graphique en fonction des types des variables et but

- **La relation entre <u>deux variables</u>**
  - ***Qualitatives***
    - **Colonne** (Clustered Column/ Stacked Column/ 100% Stacked column), ou **Bar** (Clustered Bar / Stacked Bar / 100% Stacked Bar)



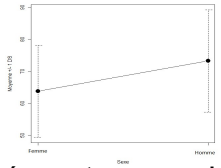
  - ***Quantitatives***
    - **Scatter** (nouage des points)

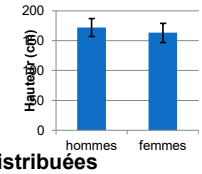


48

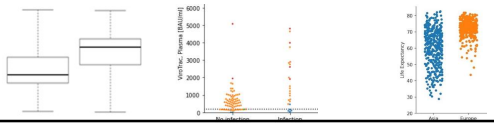
Le chois du type du graphique en fonction des types des variables et but

- **La relation entre deux variables**
  - ***Une variable quantitative en fonction d'une variable qualitative***
    - **Si les données sont normale distribuées**
      - Graphique **des moyennes** (avec déviation standard)
      - Graphique **colonnes avec barre d' erreur**





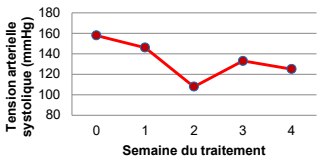
    - **Si les données sont non normale distribuées**
      - Graphique **boite a moustaches** (box plot ou box and wiskers plot, en anglais)
      - Ou graphique en essaim (swarm plot), ou graphique de bande de gigue (jitter band plot)



49

Le chois du type du graphique en fonction des types des variables et but

- **L'évolution dans le temps d'une variable qualitative ou quantitative**
  - **Line** (Ligne)

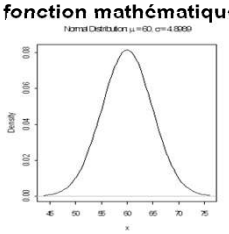


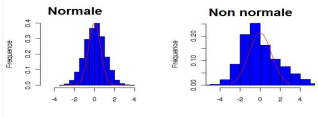
- **La relation entre trois variables quantitatives**
  - **Bubble** (nouage des sphères)
  - Nouage des points tridimensionnel
- **Une variable qualitative en fonction des intervalles d'une variable quantitative**
  - **Area** (Surface)

50

Vérification de la condition de normalité des données

On considère que les données d'un variable quantitative sont normale distribuées si les fréquences des valeurs de ce variable sont similaires a les fréquences théoriques générée par une fonction mathématique. (on va le voir dans les cours suivants)





Utilité:

- Importante pour choisir les statistiques descriptives
  - Ex: pour les données normale distribuées on utilise la moyenne et la déviation standard,
  - Ex: pour des données non normale distribuées on utilise la médiane et les quartiles
- Importante pour choisir les techniques statistiques analytiques/ inferentielles
  - Ex: appliquer des test paramétriques, avec condition de normalité:
    - Test Z pour les moyennes
    - Test t (Student)
    - Test ANOVA

51

Vérification de la condition de normalité des données

Modalités de vérification (ici, <u>conditions de normalité</u>:

- des graphiques (les meilleures) modalités)
  - Histogramme (symétrique, comme un chapeau)
  - Boite a moustaches (symétrique autour de la médiane)
  - Le graphique des quantiles (voir diapositive suivant)
  - *L'ordre de qualité entre ces techniques: Le graphique des quantiles – le meilleur, puis la histogramme, et après la boite a moustache – le moins bon.*
- des statistiques descriptives (pas très fiables)
  - Si la moyenne est ~= médiane
  - Si le coefficient de l'aplatissement ~=0 / appartient a [-1, 1] (exces kurtosis)
  - Si le coefficient de symétrie ~=0 / appartient a [-1, 1] (skewness)
  - *L'ordre de qualité entre ces techniques: Les qu'efficients d' asymétrie et aplatissement – le meilleur, puis l' égalité entre moyenne et la médiane – le moins bon.*
- des tests de normalité: (ne sont pas très recommandées)
  - Test de Kolmogorov-Smirnov ($p < 0,05$ – non normale, $p > 0,05$ normale)
  - Test de Shapiro-Wilk ($p < 0,05$ – non normale, $p > 0,05$ normale)

52

Vérification de la condition de normalité des données

Modalités de vérification (ici, conditions de normalité):

Des tests de normalité: (ne sont pas très recommandées)

On va voir comment utiliser les tests statistiques dans les cours suivantes

- Test de Kolmogorov-Smirnov
- Test de Shapiro-Wilk

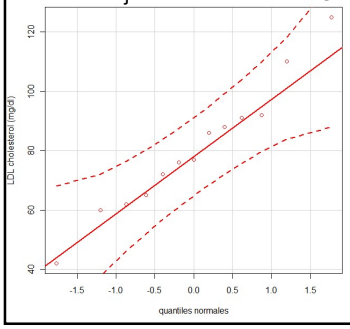
- **Les hypotheses** des tests statistiques de normalité:
  - L' hypothèse **nulle**: H_0
    - **Il n'y a pas** un différence statistiquement significative entre la distribution observée et la distribution normale (théorique)
    - Les données observées sont normale distribuées
  - L' hypothèse **alternative**: H_1 (négation du H_0)
    - **Il y a** un différence statistiquement significative entre la distribution observée et la distribution normale (théorique)
    - Les données observées ne sont pas normale distribuées
  - **Decision** a l'aide de la valeur du p
    - Si **p-value ≤ alpha (=0,05) => on rejet H_0 et on accepte H_1**
      - Les données observées ne sont pas normale distribuées
    - Si **p-value > alpha (=0,05) => on ne peut pas rejeter H_0**
      - **On ne peut pas dire** que les données observées ne sont pas normale distribuées
      - Jusqu'a on va trouver d'autres informations, on peut utiliser des statistiques qui considère les données normale distribuées

53

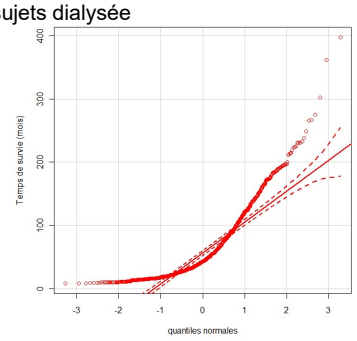
Vérification de la condition de normalité des données

- Graphique des quantiles – permet de comparer deux distributions
  - On peut comparer la distributions de la série des données observées (les points) avec un distribution théorétique normale (la ligne)
  - Si les points sont sur la ligne – distribution approximative normale
  - Si les points s'éloigne de la ligne – distribution non normale
- La meilleur façon d'évaluer la normalité des données

La distribution du LDL cholestérol chez un group des sujets avec la maladie Gaucher



Le temps de survie pour un group des sujets dialysée



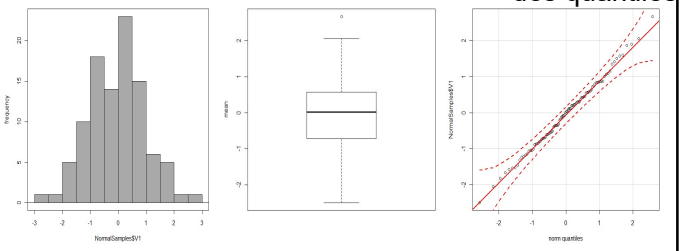
54

Comparaison des données normale/non normale distribuées

- **Normale**

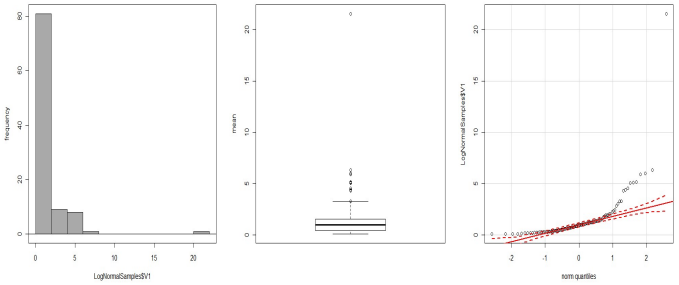
moyenne ~= médiane
(=-0,03 =0,015)
c. asymétrie = 0,11
apartient a [-1, 1], ~= 0
c. ex. aplatissement = -0,09
apartient a [-1, 1] ~= 0
Shapiro-Wilk test
p= 0,99 > 0,05

Histogramme | Boite a moustaches | Le graphique des quantiles



- **Non normale**

moyenne <> médiane
(=1.57 =0,98)
c. asymétrie = 5,59
> 1, <> 0
c. ex. aplatissement = 40,63
> 1, <> 0
Shapiro-Wilk test
p=~0 < 0,05



55

Exemples des questions pour l'examen

1) L'âge des 30 malades a été observé. Les suivantes statistiques ont été calculée : moyenne=61,2 ans, médiane=61,1 ans, écart type descriptif=14 ans, coefficient d'asymétrie= -0,30, coefficient d'excès d'aplatissement= -0,4, coefficient de variation = 0.23. Lesquelles des réponses suivantes sont correctes :

a) les données semble être approximative normale distribuées

b) approximative 68% des données sont compris entre 47.2 et 75.2 ans

c) la distribution des données est un peu aplatée

d) la distribution a une queue vers la gauche

e) les données sont relativement homogènes

Réponse: a, b, c, d

Pour trouver des indices sur la normalité en ayant des informations sur le coefficient d'asymétrie et l'excès d'aplatissement ainsi que la moyenne et la médiane, nous préférons les informations des deux premiers coefficients. Si les données sont normalement distribuées, la moyenne et la médiane doivent être proches l'une de l'autre; sinon, ils devraient être plus éloignés. Mais combien est assez proche? Cela dépend du contexte et il est difficile d'avoir un moyen uniforme de vérifier; ceci étant une des raisons de préférer les autres coefficients. Si l'un des deux coefficients (coefficient d'asymétrie ou d'excès d'aplatissement) est en dehors de l'intervalle [-1, 1], nous considérons que nous avons des suggestions que les données ne sont pas normalement distribuées. Si les deux coefficients sont à l'intérieur de l'intervalle [-1, 1], nous considérons que nous avons des suggestions que les données sont normalement distribuées. Compte tenu de tout cela, la réponse (A) est correcte. Si les donnes sont normale distribuees, on sait que entre la moyenne +/- 1 écart type, on trouve 68% des données. Ici 47,2 = moyenne (61,2) – 1 écart type (14), et 75.2 = moyenne (61,2) + 1 écart type (14), donc, la réponse (B) est correct. Une valeur négative pour un coefficient d'excès d'aplatissement suggère une distribution des données avec un tendance d'être aplatée; par conséquent, (C) est correct. La valeur négative du coefficient d'asymétrie suggère une asymétrie vers la gauche; ainsi, (D) est correct. Pour savoir si les données sont homogènes ou non, on vérifie le coefficient de variation. (CV < 10% - homogène; 10% ≤ CV <20% - relativement homogène; 20% ≤ CV <30% - relativement hétérogène; CV ≥ 30% - hétérogène). Ici, CV est 0,23, donc 23%, qui indique que les données sont relativement hétérogènes, ainsi (E) est faux.

56

Exemples des questions pour l'examen

2) Le niveau du cholestérol des 30 malades a été observé. Les suivantes statistiques ont été calculée : moyenne=200 mg/dL, médiane=202 mg/dL, écart type descriptif=30 mg/dL, coefficient d'asymétrie=0,37, coefficient d'excès d'aplatissement=0,7. Lesquelles des réponses suivantes sont correctes :

a) les données semble être approximative normale distribuées

b) approximative 99% des données sont compris entre 110 et 290 mg/dl

c) la distribution des données a une tendance vers un angle aigu

d) la distribution a une queue vers la gauche

e) les données sont hétérogènes

Réponse: a, b, c

Voir l'explication du diapo precedent. Les deux coefficients, le coefficient d'asymétrie et celui d'excès d'aplatissement sont à l'intérieur de l'intervalle [-1, 1], nous considérons que nous avons des suggestions que les données sont normalement distribuées, ainsi la réponse (A) est correcte. Si les donnes sont normale distribuees, on sait que entre la moyenne +/- 3 déviations standard, on trouve 99% des données. Ici 110 = moyenne (200) – 3 * déviation standard(30), et 290 = moyenne (200) + 3 * déviation standard (30), donc, la réponse (B) est correct. Une valeur positive pour un coefficient d'excès d'aplatissement suggère une distribution des données avec un tendance vers un angle aigu; par conséquent, (C) est correct. La valeur positive du coefficient d'asymétrie suggère une asymétrie vers la droite; ainsi, (D) est faux. Pour savoir si les données sont homogènes ou non, on vérifie le coefficient de variation. On le calcule comme écart type / moyenne * 100 = 15 %. (CV < 10% - homogène; 10% ≤ CV <20% - relativement homogène; 20% ≤ CV <30% - relativement hétérogène; CV ≥ 30% - hétérogène). Ici, CV est 15%, qui indique que les données sont relativement homogènes, ainsi (E) est faux.

57

Exemples des questions pour l'examen

3) Le corps du fémur des 40 malades a été observé. Les suivantes statistiques ont été calculée : le deuxième quartile=49 cm, le première quartile =41 cm, le troisième quartile =56cm, le minimum=39 cm, le maximum 59 cm. Lesquelles des réponses suivantes sont correctes :

a) l'intervalle interquartile à la dimension de 15 cm

La quartile 3 (56) – la quartile 1 (41) = 15, le réponse (A) est correcte.

b) le percentile 75= 56 cm

La quartile 3 (56), est égale a la percentile 75, ainsi, le réponse (B) est correcte.

c) le percentile 25= 49 cm

La quartile 1 (41), est égale a la percentile 25, ainsi, le réponse (C) est faux.

d) l'amplitude est égale à 20 cm

Le maximum (59) – le minimum (39) = 20, le réponse (D) est correcte.

e) dans une graphique boite à moustaches (en position vertical) la ligne inferieure de la boite correspond à 49 cm

Dans une graphique boite à moustaches (en position vertical) la ligne inferieure de la boite correspond a la quartile 1 (41 cm), donc, le réponse (E) est faux.

Réponse: a, b, d

58

Exemples des questions pour l'examen

4) Le corps du fémur des 40 malades a été observé. Les suivantes statistiques ont été calculée : médiane=45 cm, le première quartile=44 cm, le troisième quartile=55cm, le minimum=43 cm, le maximum 68 cm. Lesquelles des réponses suivantes sont correctes :

a) les données semble être normale distribuées

b) un bonne chois du graphique pour représenter ces données est le graphique histogramme

c) un bonne chois du graphique pour représenter ces données est la graphique ligne

d) le coefficient d'asymétrie est plus probable < 0

e) la distribution semble avoir une queue vers la droite

Réponse: b, e

Pour trouver des indices sur la normalité en ayant des informations sur le 25e centile, le 75e centile, ou sur le minimum et le maximum, on peut vérifier s'il existe des suggestions d'une asymétrie autour de la valeur médiane: en cas de symétrie autour de la médiane, on soupçonnerait la normalité des données; sinon, nous soupçonnerions la non-normalité des données. Par exemple, la distance entre le 75e percentile (le troisième quartile) et la médiane est de 55 cm - 45 cm = 10 cm, tandis que la distance entre la médiane et le 25e percentile (le premier quartile) est de 45 cm - 44 cm = 1 cm; il y a donc une asymétrie vers la droite (la distance de la médiane au troisième quartile est plus grande que la distance au premier quartile), ce qui suggère une non-normalité. De plus, la distance entre le maximum et la médiane est de 68 cm - 45 cm = 23 cm, tandis que la distance entre la médiane et le minimum est de 45 cm - 43 cm = 2 cm; ainsi, il y a une asymétrie vers la droite (la distance de la médiane au maximum est supérieure à la distance au minimum), ce qui suggère une non-normalité. Cela dit, nous avons des indices suggérant une non-normalité, avec un biais à droite, et donc (A) et (D) sont faux, tandis que et (E) est correct. Pour les données normale distribuées on peut utiliser toujours des histogrammes (d'habitude pendant l'analyse, mais rarement dans les résultats des études), donc (B) est correct. Dans le cas de données qui ne sont pas normalement distribuées, le graphique recommandé est le graphique à boîte et moustaches. Les graphiques ligne sont utilisés pour les variables qualitatives, ou quantitatives, mais pour montrer l'évolution dans le temps (qu'on n'a pas ici), ainsi (C) est faux.

59

Exemples des questions pour l'examen

5) Lesquelles des réponses suivantes sont correctes :

a) une bonne graphique pour le niveau de la intelligence (réduite, normale, supérieure) est la graphique ligne

Ici, la variable est une variable qualitative ordinale (il y a plus des trois catégories, avec ordre), mais elle n'est pas représenté dans le temps, donc le réponse (A) est faux. Si dans l' énonce on observe l' évolution dans le temps d'une variable qualitative, ou quantitative, on peut utiliser une graphique ligne.

b) une bonne graphique pour les types des différents médicaments (aspirine, paracétamol, ibuprofène) utilisées est le graphique camembert

Ici, la variable est une variable qualitative nominale (on ne peut pas ordonner les médicaments, et il y a plus des deux catégories), pour laquelle on peut utiliser un graphique camembert, bar, ou colonne, donc le réponse (B) est correcte.

c) une bonne graphique pour la relation entre le poids et la circonférence abdominale est la graphique ligne

Ici, les deux variables sont des variables quantitatives continues (on peut mesurer ces variables avec précision, au niveau des décimales ou plus). Un bonne graphique pour la relation entre deux variables quantitatives (ou entre deux variables qualitatives ordinales, ou entre une variable quantitative, et un variable qualitative ordinale), est un graphique nouage des points (scatter), ainsi le réponse (C) est faux.

d) pour la variable longueur du os cubitus (cm) le graphique histogramme est plus bonne que la boite a moustache pour évaluer la normalité

Ici, la variable est une variable quantitative continue (on peut mesurer cette variable avec précision, au niveau des décimales ou plus), pour laquelle on peut utiliser en ordre descendante de qualité: le graphique des quantiles, puis l'histogramme, puis la boite a moustaches, donc le réponse (B) est correcte.

e) la présence des réactions secondaires (vrai/faux) peut être bien représente avec un graphique boite a moustache

Ici, la variable est une variable qualitative dichotomique (on a seulement deux catégories), pour laquelle (mais en général pour toute variable qualitative) on peut utiliser un graphique camembert, bar, ou colonne, donc le réponse (E) est faux.

Réponse: b, d

60

Exemples des questions pour l'examen

6) Lesquelles des réponses suivantes sont correctes :

a) une bonne graphique pour la longueur du muscle triceps (cm) (normale distribuée) est le graphique des moyennes

Ici, la variable est une variable quantitative continue (on peut mesurer cette variable avec précision, au niveau des décimales ou plus), et ils indiquent que elle est normale distribuées, situation dans laquelle une bon graphique a utiliser est le graphique des moyennes, ou le graphique colonne avec barres d'erreur, donc le réponse est vraie.

b) une bonne graphique pour l'évolution dans le temps de la hauteur (cm) est la graphique ligne

Ici, la variable est une variable quantitative continue (on peut mesurer cette variable avec précision, au niveau des décimales ou plus), et elle est représenté dans le temps. Si dans l' énonce on observe l' évolution dans le temps d'une variable qualitative, ou quantitative, on peut utiliser une graphique ligne, donc le réponse est correcte.

c) une bonne graphique pour l'évolution dans le temps du nombre des interventions chirurgicales par jour est le graphique ligne

Ici, la variable est une variable quantitative discrète (on ne peut pas avoir une demie d'intervention chirurgicale), et elle est représenté dans le temps. Si dans l' énonce on observe l' évolution dans le temps d'une variable qualitative, ou quantitative, on peut utiliser une graphique ligne, donc le réponse est correcte.

d) pour la variable température le graphique boite a moustache est meilleur que le graphique des quantiles pour évaluer la normalité

Ici, la variable est une variable quantitative continue (on peut mesurer cette variable avec précision, au niveau des décimales ou plus), pour laquelle on peut utiliser en ordre descendante de qualité: le graphique des quantiles, puis l'histogramme, puis la boite a moustaches, donc le réponse est faux.

e) la présence des migraines (présent/absent) peut être bien représente avec un graphique histogramme

Ici, la variable est une variable qualitative dichotomique (il y a seulement deux catégories), pour laquelle (mais en général pour toute variable qualitative) on peut utiliser un graphique camembert, bar, ou colonne, donc le réponse est faux.

Réponse: a, b, c

61

Exemples des questions pour l'examen

7) Lesquelles des réponses suivantes sont correctes :

a) une bonne graphique pour le niveau de anxiété (mesurée sur une échelle numérique avec des valeurs de 0 à 100) est le graphique des quantiles

Ici, la variable est une variable quantitative (c'est écrite qu'elle est numérique, mais on sait pas si elle est continue ou discrète), mais ils ne précisent pas la normalité des données situation dans laquelle n'importe quel graphique entre les suivants: un graphique boite a moustaches, une histogramme, et une graphique des quantiles, donc le réponse est vraie.

b) une bonne graphique pour les types des sports pratiqués (volleyball, football, patinage) est la graphique colonne

Ici, la variable est une variable qualitative nominale (on ne peut pas ordonner les types du spot, et il y a plus des deux catégories), pour laquelle on peut utiliser un graphique camembert, bar, ou colonne, donc le réponse est correcte.

c) une bonne graphique pour la relation entre le diamètre du fémur (cm) et la longueur du fémur (cm) est le graphique nouage des points

Ici, les deux variables sont des variables quantitatives continues (on peut mesurer ces variables avec précision, au niveau des décimales ou plus). Un bonne graphique pour la relation entre deux variables quantitatives (ou entre deux variables qualitatives ordinales, ou entre une variable quantitative, et un variable qualitative ordinale), est un graphique nouage des points (scatter), ainsi le réponse (C) est correcte.

d) pour la variable vélocité du sang dans un vaisseau (cm/s), le graphique des quantiles est meilleur que la boite a moustache pour évaluer la normalité

Ici, la variable est une variable quantitative continue (on peut mesurer cette variable avec précision, au niveau des décimales ou plus), pour laquelle on peut utiliser en ordre descendante de qualité: le graphique des quantiles, puis l'histogramme, puis la boite a moustaches, donc le réponse est correcte.

e) la présence d'un bebe (oui/non) peut être bien représente avec un graphique des quantiles

Ici, la variable est une variable qualitative dichotomique (on a seulement deux catégories), pour laquelle (mais en général pour toute variable qualitative) on peut utiliser un graphique camembert, bar, ou colonne, donc le réponse est faux.

Réponse: a, b, c, d

62

Exemples des questions pour l'examen

8) Lesquelles des réponses suivantes sont correctes :

a) une bonne graphique pour la longueur du muscle biceps (cm) est le graphique histogramme

Ici, la variable est une variable quantitative continue (on peut mesurer cette variable avec précision, au niveau des décimales ou plus), mais ils ne précisent pas la normalité des données situation dans laquelle n'importe quel graphique entre les suivants: un graphique boite a moustaches, une histogramme, et une graphique des quantiles, donc le réponse est vraie.

b) une bonne graphique pour l'évolution dans le temps de la hauteur (cm) est la graphique ligne

Ici, la variable est une variable quantitative continue (on peut mesurer cette variable avec précision, au niveau des décimales ou plus), et elle est représenté dans le temps. Si dans l' énonce on observe l' évolution dans le temps d'une variable qualitative, ou quantitative, on peut utiliser une graphique ligne, donc le réponse est correcte.

c) une bonne graphique pour l'évolution dans le temps du poids (kg) est la graphique camembert

Ici, la variable est une variable qualitative quantitative continue (on peut mesurer cette variable avec précision, au niveau des décimales ou plus), mais ils ne précisent pas la normalité des données situation dans laquelle n'importe quel graphique entre les suivants: un graphique boite a moustaches, une histogramme, et une graphique des quantiles, donc le réponse est faux.

d) pour la variable diamètre de la carotide (mm) le graphique des quantiles est meilleur que l'histogramme pour évaluer la normalité

Ici, la variable est une variable quantitative continue (on peut mesurer cette variable avec précision, au niveau des décimales ou plus), pour laquelle on peut utiliser en ordre descendante de qualité: le graphique des quantiles, puis l'histogramme, puis la boite a moustaches, donc le réponse est correcte.

e) la présence d'une allergie aux antibiotiques (vrai/faux) peut être bien représente avec un graphique camembert

Ici, la variable est une variable qualitative dichotomique (on a seulement deux catégories), pour laquelle (mais en général pour toute variable qualitative) on peut utiliser un graphique camembert, bar, ou colonne, donc le réponse est faux.

Réponse: a, b, d, e

63

Exemples des questions pour l'examen

9) Lesquelles des réponses suivantes sont correctes, concernant le graphique a cote:

a) les données semble être normale distribuées

b) le coefficient d'asymétrie est plus probable < 0

c) la distribution semble avoir une queue vers la droite

d) la différence entre la quartile 3 et quartile 2 est très différente de la différence entre la médiane et la quartile 1

e) le coefficient d'asymétrie est plus probable proche 0

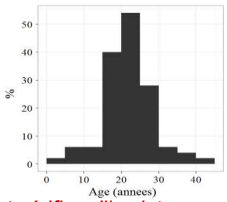
Réponse: a, e

Pour trouver des indices de normalité en ayant un box plot, on peut vérifier s'il existe une asymétrie autour de la valeur médiane (soit la distance aux quartiles, soit la longueur des moustaches): dans le cas d'une symétrie autour de la médiane, on soupçonnerait normalité des données; sinon, nous soupçonnerions la non-normalité des données. Ici, les deux quartiles (le premier quartile est la ligne du bas de la boîte, tandis que le troisième quartile est la ligne du haut de la boîte) sont également espacés de la médiane (la ligne horizontale à l'intérieur de la boîte), et les deux moustaches ont la même longueur ; il y a donc une suggestion que les données sont normalement distribuées, et la réponse (A) est correcte. Les réponses (B) et (C) sont incorrectes, tandis que la réponse (E) est correcte car nous n'avons pas pu déceler d'asymétrie dans le graphique; ainsi, le coefficient d'asymétrie est plus vraisemblablement proche de 0. Le quartile 2 est, en fait, la médiane, et nous pouvons voir la distance entre la médiane, et les deux quartiles (1 et 3) sont égaux; ainsi, la réponse (D) est incorrecte.

64

Exemples des questions pour l'examen

10) Lesquelles des réponses suivantes sont correctes, concernant le graphique a cote :



a) les données semble être normale distribuées
b) le coefficient d'asymétrie est plus probable < 0
c) la distribution semble avoir une queue vers la droite
d) de coefficient d' aplatissement est plus probable > 0
e) le coefficient d'asymétrie est plus probable proche a 0

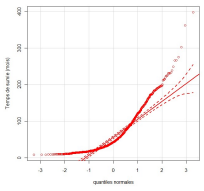
Réponse: a, d, e

Pour trouver des indices de normalité en ayant un histogramme, on peut vérifier s'il existe une asymétrie et une image très grossièrement en forme de cloche (au moins pour avoir une distribution unimodale): en cas de symétrie autour de la moyenne, on soupçonnerait la normalité des données; sinon, nous soupçonnerions la non-normalité des données. De plus, une distribution qui n'est pas unimodale suggère la non-normalité des données. Ici, la distribution est unimodale et symétrique; ainsi, nous avons une suggestion de normalité des données, donc la réponse (A) est correcte. Les réponses (B) et (C) sont incorrectes, tandis que la réponse (E) est correcte car nous n'avons pas pu déceler d'asymétrie dans le graphique; ainsi, le coefficient d'asymétrie est plus vraisemblablement proche de 0. Puisque la distribution est plus vraisemblablement étroite et haute (avec une pointe plus nette), il y a une suggestion d'une distribution leptokurtique, lorsque le kurtosis est supérieur à 0; ainsi, la réponse (D) est correcte.

65

Exemples des questions pour l'examen

11) Lesquelles des réponses suivantes sont correctes, concernant le graphique a cote:



a) les données semble être normale distribuées
b) la ligne represente la distribution normale
c) les données ne semble pas être normale distribuées
d) la ligne represente la distribution non normale
e) les points represente la distribution normale

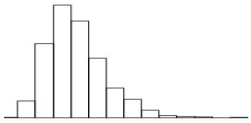
Réponse: b, c

Pour trouver des indices sur la normalité lors d'un graphique quantile-quantile, on peut vérifier si les points sont proches de la ligne diagonale: dans le cas où les points sont proches de la diagonale, on soupçonnerait la normalité des données; sinon, nous soupçonnerions la non-normalité des données. Ici, les points sont clairement éloignés de la ligne diagonale; ainsi, les données ne semblent pas être normalement distribuées, donc la réponse (A) est incorrecte, tandis que la réponse (C) est correcte. La ligne représente la distributions normale, tandis que les points représentent les points de données, donc la réponse (B) est correcte, tandis que les réponses (D) et (E) sont incorrectes.

66

Exemples des questions pour l'examen

12) Lesquelles des réponses suivantes sont correctes, concernant le graphique a cote :



a) les données semble être normale distribuées
b) le coefficient d'asymétrie est plus probable > 0
c) la distribution semble avoir une queue vers la droite
d) la médiane est plutôt plus petite que la moyenne
e) la différence entre la quartile 3 et quartile 2 est plutôt plus grande que la différence entre la médiane et la quartile 1

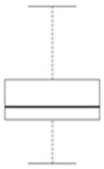
Réponse: b, c, d, e

Voir la justification de la question 10. Pour les histogrammes ayant des colonnes éloignées (de la partie centrale de la distribution) vers la droite, représentent une distribution avec une queue a la droite et suggère une valeur positive pour le coefficient d'asymétrie. D'autre part, avoir des colonnes distantes (de la partie centrale de la distribution) vers la gauche représente une distribution avec une queue gauche et suggère une valeur négative pour le coefficient d'asymétrie. Ici, la distribution est asymétrique, donc la réponse (A) est correcte. Les réponses (B) et (C) sont correctes car il existe une distribution à queue droite. Dans le cas d'une distribution avec asymétrie a la droite, la moyenne est supérieure à la médiane; ainsi, la réponse (D) est correcte. Dans le cas d'une distribution à queue gauche, la moyenne est inférieure à la médiane. La différence entre le quartile 3 et la médiane (quartile 2) est plutôt plus grande que la différence entre la médiane et le quartile 1, dans le cas d'une distribution avec une queue a la droite; ainsi, la réponse (E) est correcte. En revanche, la différence entre le quartile 3 et la médiane (quartile 2) est plutôt inférieure à la différence entre la médiane et le quartile 1, en cas de distribution à queue gauche

67

Exemples des questions pour l'examen

13) Lesquelles des réponses suivantes sont correctes, concernant le graphique a cote:



a) les données semble être normale distribuées
b) la distribution semble avoir une queue vers la droite
c) les données ne semble pas être normale distribuées
d) la distribution semble avoir une queue vers la gauche
e) Le coefficient d'asymétrie est plus probable > 0

Réponse: b,c,e

Explication:

b,d) on observe une inégalité entre la distance entre quartile 3 – médiane (plus grande) par rapport a la distance entre médiane – quartile 1, et aussi on observe une inégalité entre les moustaches (en haut il est plus grande que en bas) => asymétrie a droite, donc une queue vers la droite.

a,c) Si les données ont une asymétrie ca indique que les données ne semble pas normale distribuées.

e) Une coefficient d'asymétrie plus grand que 0 est associée a une queue vers la droite.

68

Exemples des questions pour l'examen

14) Lesquelles des réponses suivantes sont correctes, concernant le graphique ici:

a) Le graphique 1, 2, 3, semble suggérer une distribution unimodale

b) Le graphique 5 semble suggérer une distribution bimodale

c) Le graphique 6 semble suggérer une distribution multimodale

d) Le graphique 4 semble suggérer une distribution uniforme

e) Le coefficient d'asymétrie pour les données représentées dans le graphique 3 et 5 est plus probable > 0

f) Le coefficient d'asymétrie pour les données représentées dans le graphique 2 est plus probable < 0

g) Le coefficient d'asymétrie pour les données représentées dans le graphique 1 est plus probable proche a 0

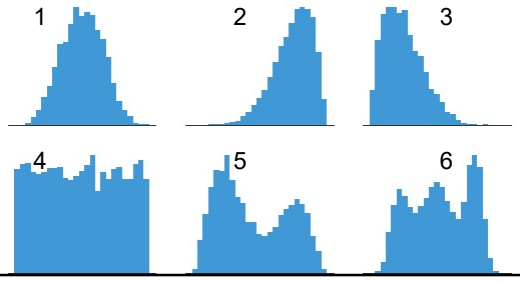
h) Le graphique 2 semble suggérer une distribution avec une queue a la droite

Réponse: a, b, c, d, e, f, g

Explication:

a) On observe que il y a une seul mode (valeur plus fréquente) dans les graphiques 1, 2, 3

b) On observe qu'il y a deux mode, une mode principal et une mode secondaire, et reflet l'existence possible des deux populations



https://chartio.com/learn/charts/histogram-complete-guide/

69

Exemples des questions pour l'examen

14) Explication:

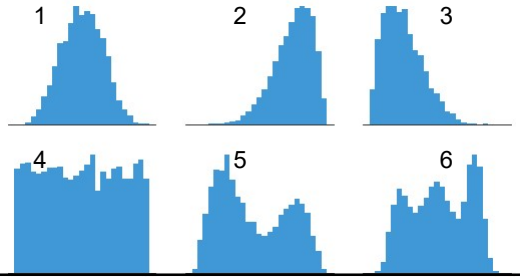
c) On observe qu'il y a trois modes, et reflet l'existence possible des trois populations

d) On observe que n'importe la valeur sur l'axe horizontale, sur l'axe verticale on a presque la même fréquence. Ca indique une possible distribution uniforme (chaque valeur a la même fréquence)

On observe que les distributions 3 et 5 ont une queue a la droite (certaines valeurs sont plus distancées du centre des données vers la droite), ca suggère une asymétrie a la droite. Le coefficient d' asymétrie a des valeurs plus grande que 0 si il y a une asymétrie a la droite => réponse e) est correcte, et réponse h) est incorrecte

On observe que les distributions 2 ont une queue a la gauche (certaines valeurs sont plus distancées du centre des données vers la droite), ca suggère une asymétrie a la gauche. Le coefficient d' asymétrie a des valeurs plus petites que 0 si il y a une asymétrie a la gauche => réponse f) est correcte, et réponse h) est incorrecte

On observe que la distributions 1 est une distribution symétrique autour du centre. Le coefficient d' asymétrie a des valeurs proches a 0 si il y a distribution symétrique => réponse g) est correcte



https://chartio.com/learn/charts/histogram-complete-guide/

70

Exemples des questions pour l'examen

15) Lesquelles des réponses suivantes sont correctes, concernant le tableau suivant qui présente la statistique descriptive d'un étude médical (on considère que les auteurs de l'étude ont bien choisi la modalité de représenter les données):

Caractéristique	Traitement	Control
Age (années), moyenne (DS)	24 (10)	26 (10)
Femmes n (%)	102 (51)	104 (52)
Triglycérides (mg/dL), médiane (IQR)	150 (130-190)	160 (140-210)

a) on suppose que les auteurs ont évalue la normalité des données quantitatives

b) les auteurs ont choisi de montrer la moyenne et l'écart type pour la variable âge parce que l' âge est normale distribuée

c) les auteurs ont choisi de montrer la médiane et les quartiles pour la variable Triglycérides parce que les Triglycérides ne suivent pas une distribution normale

d) la distribution des Triglycérides pour le group Traitement semble avoir une queue vers la droite

e) le coefficient d'asymétrie des Triglycérides pour le group Control est plus probable > 0

f) le coefficient d'asymétrie de l' âge pour le group Traitement est plus probable proche a 0

g) entre 14 ans et 34 ans il y a ~68% des valeurs de l'âge observées dans l' étude dans le groupe qui a reçu le traitement

h) entre 130 mg/dL et 190 mg/dL il y a au moins 50% des valeurs des Triglycérides observées dans l' étude pour le group Traitement

i) entre 14 ans et 34 ans il y a ~95% des valeurs de l'âge observées dans l' étude dans le groupe qui a reçu le traitement

j) c'est mieux de montrer les quartiles que de montrer l'intervalle interquartiles

71

Exemples des questions pour l'examen

15)

Caractéristique	Traitement	Control
Age (années), moyenne (DS)	24 (10)	26 (10)
Femmes n (%)	102 (51)	104 (52)
Triglycérides (mg/dL), médiane (IQR)	150 (130-190)	160 (140-210)

k) Un bon graphique pour comparer le traitement avec le contrôle concernant l'âge est : graphique des moyennes

l) Un bon graphique pour comparer le traitement avec le contrôle concernant l'âge est : graphique des colonnes avec des barres d'erreur

m) Des bons graphiques pour comparer le traitement avec le contrôle concernant le sexe sont : graphique des colonnes ou graphique des barres

n) Un bon graphique pour comparer le traitement avec le contrôle concernant le sexe sont: graphique des barres

o) Un bon graphique pour comparer le traitement avec le contrôle concernant les niveaux des triglycérides est : graphique boite a moustaches

p) Un bon graphique pour comparer le traitement avec le contrôle concernant les niveaux des triglycérides est : graphique en essaim

q) Un bon graphique pour comparer le traitement avec le contrôle concernant les niveaux des triglycérides est : graphique en bande en gigue

r) Un bon graphique pour évaluer la relation entre le traitement et les niveaux des triglycérides est: graphique boite a moustaches

s) Un bon graphique pour évaluer la distribution des triglycérides dans le groupe traitement: graphique boite a moustaches

t) Un bon graphique pour évaluer la distribution du sexe dans le groupe traitement: graphique camembert

u) L' écart type d'âge est l' écart type descriptif

Réponse: a, b, c, d, e, f, g, h, j, k, l, m, n, o, p, q, r, s, t, u

72

Exemples des questions pour l'examen

Explications:

a) Dans l'énonce c'est écrit « on considère que les auteurs de l'étude ont bien choisi la modalité de représenter les données « donc ca implique la vérification de normalité normalité des données quantitatives

b) les auteurs doit montrer la moyenne et l'écart type pour les variables avec une normale distribuée

c) les auteurs ont choisi de montrer la médiane et les quartiles les variables avec qui ne suivent pas une distribution normale

d) la distance entre quartile 3 (190) – médiane (150) = 40, est plus grande par rapport a la distance entre médiane (150) – quartile 1 (130) = 20, donc ca indique une asymétrie a la droite => la distribution semble avoir une queue vers la droite

e) la distance entre quartile 3 (210) – médiane (160) = 50, est plus grande par rapport a la distance entre médiane (160) – quartile 1 (140) = 20, donc ca indique une asymétrie a la droite => la distribution semble avoir une queue vers la droite => le coefficient d'asymétrie des Triglycérides pour le group Control est plus probable > 0

f) du au fait que l'article montre la moyenne et la déviation standard, c'est une indication que les données sont normale distribuées. Et pour cette situation le coefficient d'asymétrie est plus probable proche a 0

g) Dans l'intervalle entre la moyenne (24) moins une déviation standard (24 – 10 = 14), et la moyenne (24) plus une déviation standard (24 + 10 = 34), il y a ~68% des valeurs de l'âge observées dans l' étude dans le groupe qui a reçu le traitement, du au fait que si les données sont normale distribuées ils ont cette propriété.

h) entre la première quartile (130 mg/dL – 25% des valeurs des Triglycérides sont inferieures ou égale a 130) et la troisième quartile (190 mg/dL – 75% des valeurs des Triglycérides sont inferieures ou égale a 190) il y a au moins 50% des valeurs des Triglycérides observées dans l' étude pour le group Traitement

i) voir g)

j) c'est mieux de montrer les quartiles que de montrer l'intervalle interquartiles – voir la théorie

u) Les tableaux qui présent les caractéristiques des sujets montre l' écart type descriptive. Elle ne montre pas l' écart type d' échantillonnage, qui est utilise pour la statistique inferentielle, analytique

73

Exemples des questions pour l'examen

16) Le niveau des triglycérides a été observe dans un échantillon et le test de Shapiro Wilk a trouve un valeur du p=0,01 pour ceux qui ont reçu un traitement pour réduire les triglycérides, et le test de Kolmogorov Smirnov a trouve un valeur du p=0,58 pour ceux qui ont reçu un placebo. Lesquelles des réponses suivantes sont correctes:

a) les données semble être non normale distribuées pour le group qui ont reçu un traitement pour réduire les triglycérides

b) pour le group qui ont reçu un placebo il n'y a pas de suggestion qu'elle ne sont pas normale distribuées

c) Le premier test statistique dans l' énonce a un résultat statistiquement significative

d) l'hypothèse nulle du test de normalité pour ceux qui on reçu le placebo est: L' hypothèse nulle: Il n'y a pas un différence statistiquement significative entre la distribution observée et la distribution normale (théorique)

e) l'hypothèse alternative du test de normalité pour ceux qui on reçu le placebo est: Les données observées ne sont pas normale distribuées

Réponse: b, c, d, e

Pour évaluer la normalité on peut utiliser des tests de normalité comme (Shapiro Wilk, ou Kolmogorov Smirnov). L' hypothèse nulle: H0 Il n'y a pas un différence statistiquement significative entre la distribution observée et la distribution normale (théorique) ou: Les données observées sont normale distribuées. L' hypothèse alternative: H1 (négation du H0): Il y a un différence statistiquement significative entre la distribution observée et la distribution normale (théorique) ou:

Les données observées ne sont pas normale distribuées. => Donc les réponses d) et e) sont correctes

La décision a l'aide de la valeur du p pour ces tests est: Si p-value ≤ alpha (=0,05) => on rejet H0 et on accepte H1, ou Si p-value > alpha (=0,05) => on ne peut pas rejeter H0. => Donc les réponses b) [p=0.58 > 0,05] et a) et c) [p=0,01 < 0,05] sont correctes

74

Ce qu'on a appris

- Statistique descriptive
  - Donnes quantitatives:
    - Mesures de dispersion
      - dispersion, amplitude, écart interquartiles, variance, écart-type, coefficient de variation,
    - Asymétrie
    - Aplatissement
    - Tableaux et graphiques – continuation
      - Graphique des moyennes, des quantiles, boite a moustaches, ligne, nouage des points
    - La technique du chois des graphiques
    - L' évaluation de la normalité des données

75

MERCI!

76

To do: tabel cu toate tehnicile de reprezentare grafica
Explicatii pt ex 14 – cu grafice

77