UMF
UNIVERSITATEA DE MEDICINĂ ȘI FARMACIE
IULIU HAȚIEGANU
CLUJ-NAPOCA

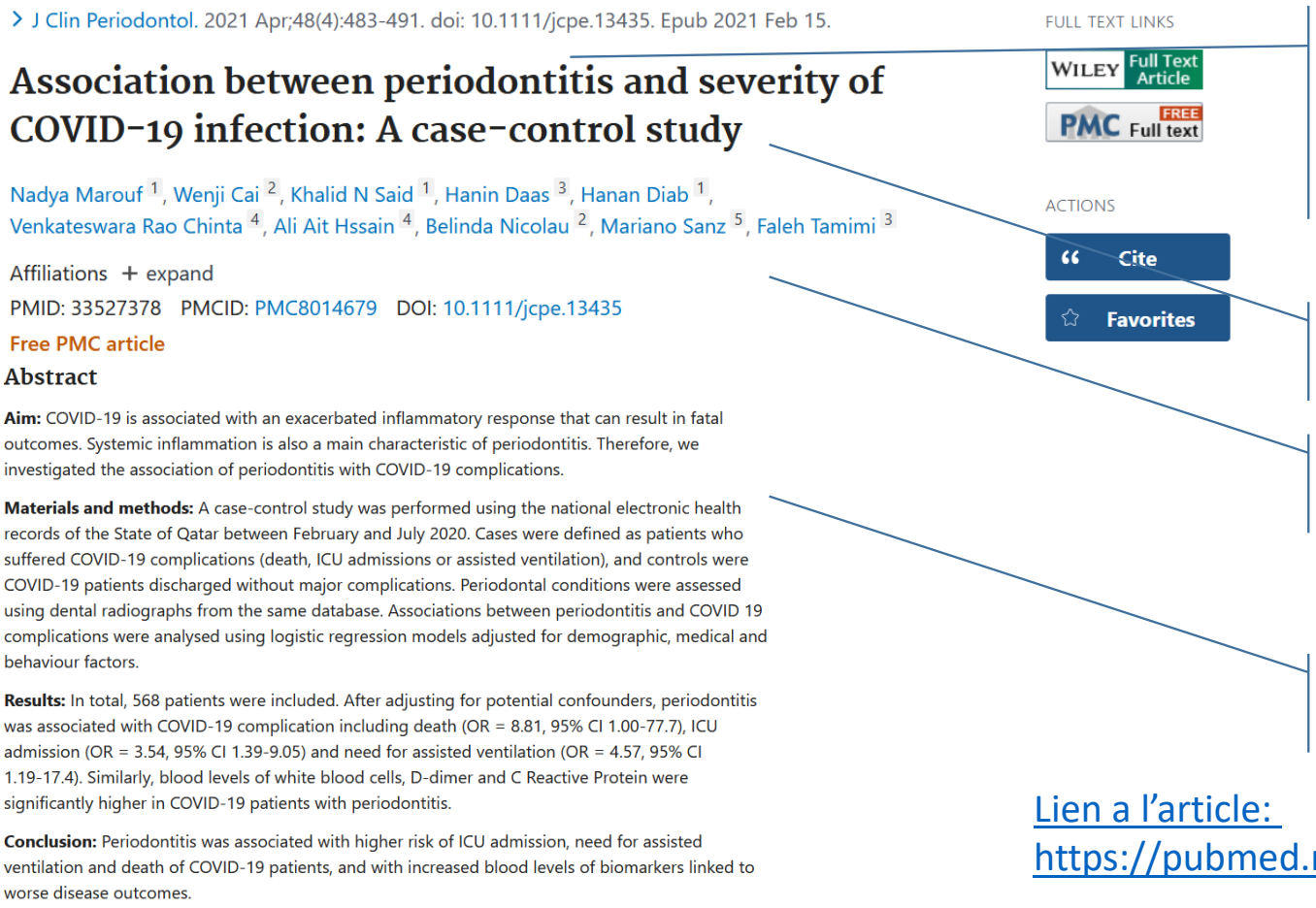
Statistiques Descriptives (3)-
Description d'une variable quantitative

Description d'une variable quantitative (Continues, Discrètes)

- *Measures de tendence centrale*
  - Moyenne arithmétique
  - Médiane
  - Mode
- *Measures de dispersion*
  - Variances
  - Deviation standard
  - Coefficient de variation
  - Error standard
- *Measures de localisation & asymétrie/aplatissement*
  - Asymétrie (Skewness)
  - Coefficient d'aplatissement (Kurtosis)
  - Quartiles
  - Percentiles
- *Graphiques*
  - Histograms
  - Box-plots
  - Graphique avec des Moyennes et barres d'erreur (*engl. means error plot*)

Exemples d'articles scientifiques (littérature scientifique en dentisterie)

E1.



> J Clin Periodontol. 2021 Apr;48(4):483-491. doi: 10.1111/jcpe.13435. Epub 2021 Feb 15.

Association between periodontitis and severity of COVID-19 infection: A case-control study

Nadya Marouf [1], Wenji Cai [2], Khalid N Said [1], Hanin Daas [3], Hanan Diab [1], Venkateswara Rao Chinta [4], Ali Ait Hssain [4], Belinda Nicolau [2], Mariano Sanz [5], Faleh Tamimi [3]

Affiliations + expand

PMID: 33527378 PMCID: PMC8014679 DOI: 10.1111/jcpe.13435

Free PMC article

Abstract

Aim: COVID-19 is associated with an exacerbated inflammatory response that can result in fatal outcomes. Systemic inflammation is also a main characteristic of periodontitis. Therefore, we investigated the association of periodontitis with COVID-19 complications.

Materials and methods: A case-control study was performed using the national electronic health records of the State of Qatar between February and July 2020. Cases were defined as patients who suffered COVID-19 complications (death, ICU admissions or assisted ventilation), and controls were COVID-19 patients discharged without major complications. Periodontal conditions were assessed using dental radiographs from the same database. Associations between periodontitis and COVID 19 complications were analysed using logistic regression models adjusted for demographic, medical and behaviour factors.

Results: In total, 568 patients were included. After adjusting for potential confounders, periodontitis was associated with COVID-19 complication including death (OR = 8.81, 95% CI 1.00-77.7), ICU admission (OR = 3.54, 95% CI 1.39-9.05) and need for assisted ventilation (OR = 4.57, 95% CI 1.19-17.4). Similarly, blood levels of white blood cells, D-dimer and C Reactive Protein were significantly higher in COVID-19 patients with periodontitis.

Conclusion: Periodontitis was associated with higher risk of ICU admission, need for assisted ventilation and death of COVID-19 patients, and with increased blood levels of biomarkers linked to worse disease outcomes.

FULL TEXT LINKS

WILEY Full Text Article

PMC FREE Full text

ACTIONS

Cite

Favorites

où et quand l'article scientifique a été publié : nom du journal, volume numéro/ numéro, année

titre de l'article de recherche

les auteurs

résumé d'article scientifique (ou abstract)

Lien a l'article: https://pubmed.ncbi.nlm.nih.gov/33527378/

Exemples d'articles scientifiques:
Description des variables par des statistiques descriptives

TABLE 1 Selected characteristics of the cases and controls

	Controls	Cases			
	COVID 19 patients without complications (n = 528)	All complications (N = 40)	Death N = 14 (%)	ICU admission N = 36 (%)	Assisted ventilation N = 20 (%)
Sex					
Male	290 (54.9)	20 (50.0)	7 (50.0)	17 (47.2)	10 (50.0)
Female	238 (45.1)	20 (50.0)	7 (50.0)	19 (52.8)	10 (50.0)
Age					
Mean, years (SD)	41.5 (14.1)	53.6 (15.0)	56.6 (17.6)	52.8 (15.4)	53.3 (15.7)
Smoker					
Never	460 (87.1)	29 (72.5)	8 (57.1)	28 (77.8)	15 (75.0)
Past/current	68 (12.9)	11 (27.5)	6 (42.9)	8 (22.2)	5 (25.0)
Diabetes					
Yes	147 (27.8)	17 (42.5)	8 (57.1)	20 (55.6)	12 (60.0)
No	381 (42.9)	23 (57.5)	6 (42.9)	16 (44.4)	8 (40.0)
Comorbidity					
None	314 (59.5)	5 (12.5)	0 (0)	5 (13.9)	3 (15.0)
One comorbidity	103 (19.5)	11 (27.5)	4 (28.6)	10 (27.8)	5 (25.0)
Two comorbidities	111 (21.0)	24 (60.0)	10 (71.4)	21 (58.3)	12 (60.0)

Les variables

Les catégories/ modalités de la variable

Fréquence absolue (l'effectif de la catégorie)

Fréquence relative (%)

Moyenne arithmétique

l' écart-type

Exemples d'articles scientifiques (littérature scientifique en dentisterie)

TABLE 3 Associations between periodontal condition and COVID-19 complications

	Controls (n = 528)	Cases: All COVID complications (n = 40)	
Periodontal condition			**Unadjusted OR (95% CI)**
Stage 0–1	303 (57.4)	7 (17.5)	1
Stage 2–4	225 (42.8)	33 (82.5)	6.34 (2.79–14.61)
		Cases: death (n = 14)	
Stage 0–1	303 (57.4)	1 (7.1)	1
Stage 2–4	225 (42.8)	13 (92.9)	17.5 (2.27–134.8)
		Cases: ICU admission (n = 36)	
Stage 0–1	303 (57.4)	7 (19.4)	1
Stage 2–4	225 (42.8)	29 (80.6)	5.57 (2.40–12.9)
		Cases: need for assisted ventilation (n = 20)	
Stage 0–1	303 (57.4)	3 (15.8)	1
Stage 2–4	225 (42.8)	17 (85.0)	7.31 (2.21–26.3)

Exemples d'articles scientifiques:
Description des variables par des graphiques

E2.

Braz Oral Res. 2021 Mar 12;35:e048. doi: 10.1590/1807-3107bor-2021.vol35.0048. eCollection 2021.

Knowledge, stress levels, and clinical practice modifications of Turkish dentists due to COVID-19: a survey study

Ayca Sarialioglu Gungor [1], Nazmiye Donmez [1], Yesim Sesen Uslu [2]

Affiliations + expand

Abstract

Dentists are exposed to the highest risk of occupational respiratory and droplet infections by working face-to-face with patients. The aim of this study was to investigate the knowledge of symptoms and modes of transmission of COVID-19, stress levels and clinical practice modifications of Turkish dentists during the COVID-19 pandemic. An online survey (15 questions) was sent to Turkish dentists from May 5 to 12 May, 2020. The survey comprised questions about dentists' demographic characteristics, their knowledge about COVID-19, stress levels and the measures taken in dental clinics against COVID-19. This study included a total of 1,095 Turkish dentists. The data were expressed as frequency with percentage values for overall variables. Dentists were most familiar with high fever among the symptoms of COVID-19 (99.4%) and 99.2% of them reported that COVID-19 was transmitted with eye, mouth and nasal mucosa contact on surfaces contaminated with the droplets of infected persons. While the stress levels of females were higher than males, the stress levels of dentists with more than 20 years of professional experience were found to be lower. Regarding the precautions to be taken as a preventive measure when working again, 86.6% of the dentists took precautions by increasing daily patient care intervals and only 38.4% of the dentists wore an N95 mask. During this pandemic, knowing the conditions about when the treatments can be applied and the precautions to be taken will shed light on dentistry staff. Current recommendations of national authorities about the coronavirus should be followed.

Exemples d'articles scientifiques:
Description des variables par des graphiques

Les réponses aux les questions pour quantifier le niveau du stress utilisent une échelle en 10 points où 0 =« pas du tout stressant »,..., 5 =« modérément stressant » et 10 = "extrêmement stressant"

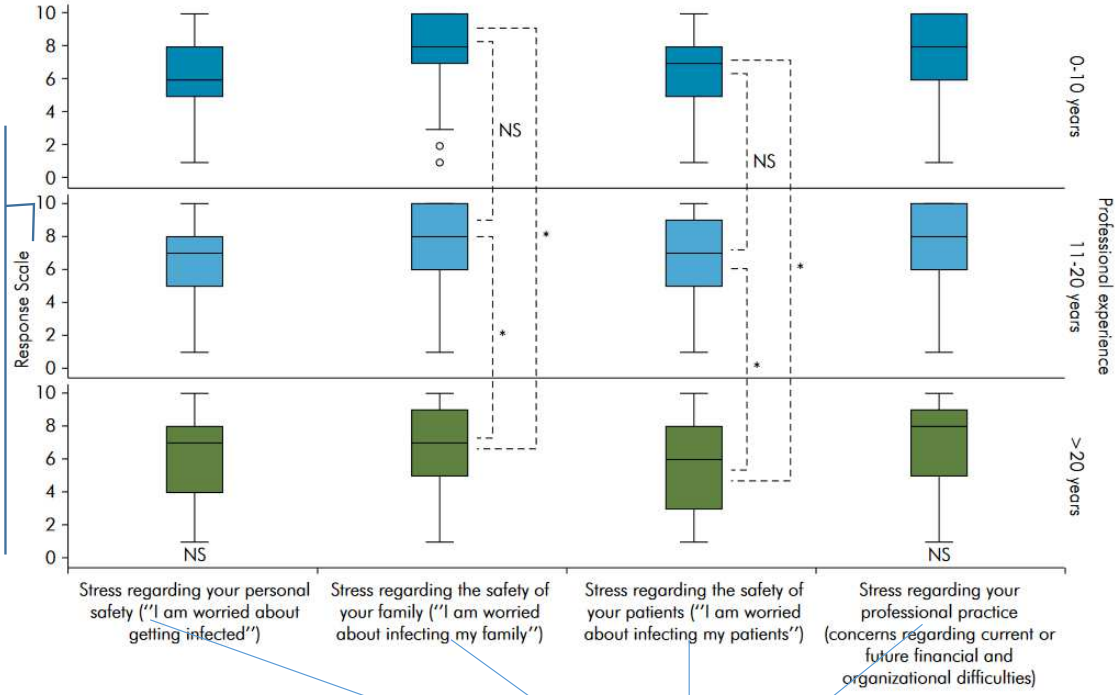


Figure 3. Box-plot representing the statistical relation of professional experience and stress scale responses. [Data represented as median (horizontal bars inside the box) and range (Y-error bars). * = represents significant difference ($p < 0.05$). NS = Non-significant ($p > 0.05$).

Les variables (les questions du formulaire enligne)

Mesures (statistiques) descriptives de tendance centrale

- Moyenne arithmétique
- Mediane
- Mode
- Valeur centrale
- Moyenne pondérée
- Moyenne géométrique

Statistiques observées sur l'échantillon≈ paramètre de population

Échantillon versus Population: Statistiques vs. Paramètres

Population → paramètre

Moyenne arithmétique calculée sur une population

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

x_i = la valeur de la variable quantitative pour le i -ème patient
N=taille de la population

Échantillon→ statistique

Moyenne arithmétique calculée sur un échantillon

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

x_i = la valeur de la variable quantitative pour le i -ème patient
n=taille de l' échantillon

Mesures de tendance centrale: MOYENNE

EXEMPLE:
Quel est le nombre moyenne de dentes sans des caries chez les 10 patients diabétiques ?

Nombre de dents sans caries
23
27
25
24
20
24
20
30
28
23

Moyenne arithmetique= $\bar{X}$ =(23+27+25+…+23)/10=25

Moyenne arithmétique (Ma)

Avantages

- ✓Toute valeur de la série est prise en compte dans le calcul de la moyenne
- ✓la somme des écart à la moyenne est nulle:

$$\sum_{i=1}^{n}(x_i - \bar{x}) = 0$$

Désavantages

- ✗peut être influencée par les valeurs extrêmes
- ✗ modification d'une seule valeur de la série de données va influencer (changer) la moyenne
- ✗n'a de sens que pour une variable quantitative

Médiane (Me): définition et calcul

- **Médiane:** la valeur qui partage la série des individus en 2 groupes d'effectifs égaux
- **Comment trouver/determiner le médiane?**

  - triez / ordonnez les données en ascendant
  - regardez la taille de l'échantillon (« *n* »)

$$Me = \begin{cases} x_{\frac{n+1}{2}} & \text{si n est impaire,} \\ \dfrac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2} & \text{si n est paire.} \end{cases}$$

Médiane (Me):

Avantages

- ✓ variables quantitatives et qualitatives ordinales
- ✓ peu sensible aux valeurs extrêmes contrairement à la moyenne

Désavantages

- ✘ elle ne se prête pas à des opérations algébriques
- ✘ ne s'applique pas aux variables qualitatives nominales

Exemple de calcul: moyenne et médiane

	Variable quantitative continue	Variable quantitative continue ~ discrète	Variable qualitative dichotomique	Variable qualitative ordinale
Patient	**Taille (cm)**	**Age (années)**	**Sexe**	**Parodontite**
1	170	50	F	légère
2	160	45	F	modérée
3	187	38	H	modérée
4	172	25	H	légère
5	157	65	F	sévère
6	175	56	H	sévère

Variable: Âge

Moyenne arithmétique (Ma) =?

Ma=(50 + 45 + 38 + 25 + 65 + 56) / 6 = 46,5 ans

Médiane (Me)=?

Taille de l'échantillon: nombre paire = 6 =>

Me = (45+50) / 2= 47,5

Paire (2, 4, …)

$$Me = \frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2}$$

MODE: définition et calcul

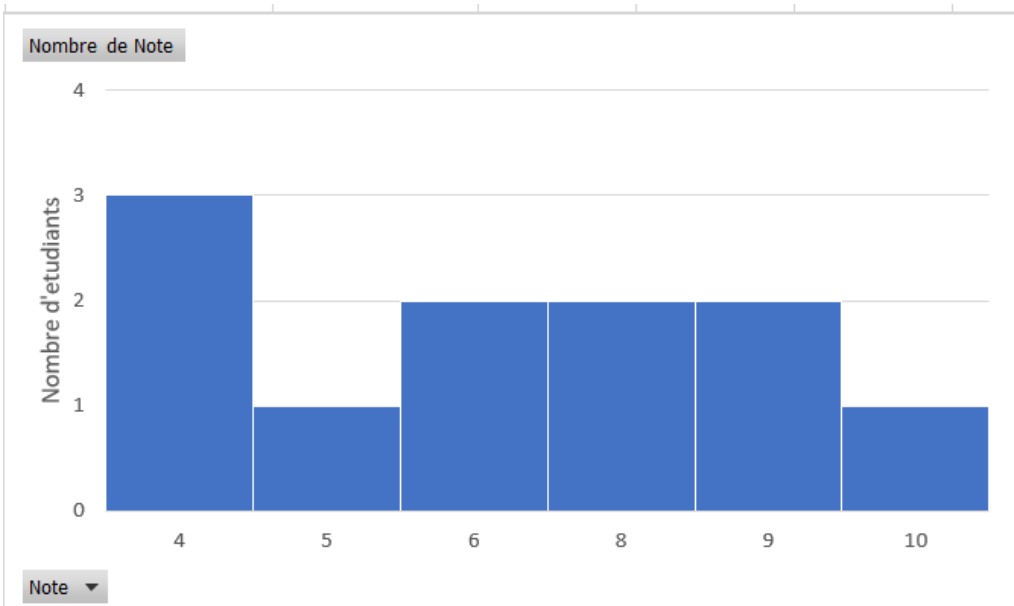
- la valeur (donnée) de la série statistique ayant la fréquence la plus élevée (la valeur de la série qui revient le plus souvent)
- il n'y a pas de formule mathématique pour le calculer
- une série statistique de données peut posséder plusieurs valeurs modales.

MODE: définition et calcul

Les notes obtenues à l'examen pratique de Biostat par un groupe de 11 étudiants:

4, 9, 5, 8, 6, 4, 9, 10, 8, 6, 5, 4

- Mode: 4 → série unimodale

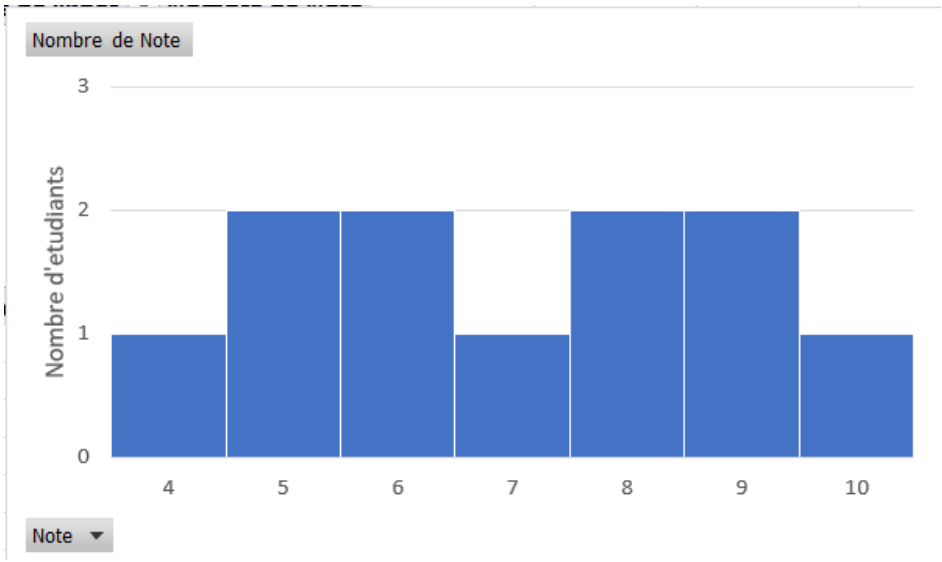


MODE: exemple

Les notes obtenues à l'examen pratique de Biostat par un groupe de 11 étudiants MDFR:

4, 9, 5, 8, 6, 7, 9, 10, 8, 6, 5

- Mode: 5, 6, 8, 9
- → série multimodale (plurimodale)



MODE (Mo): propriétés

Avantages

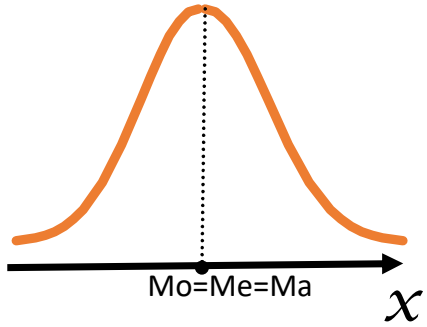
- ✓ faible sensibilité aux valeurs extrêmes de la série.
- ✓ peut être un indicator d'une série de données hétérogène
- ✓ si les données sont hétérogènes (série bimodale), il vaut mieux deux valeurs modales qu'une médiane

Désavantages

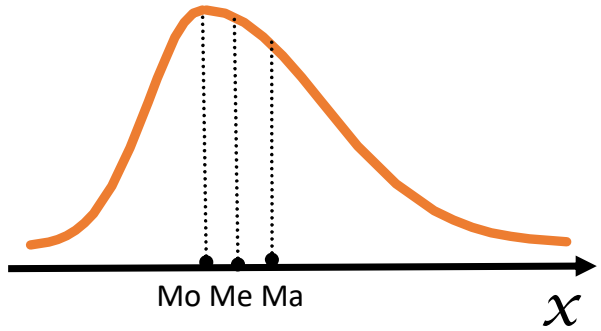
- ✗ ne se prête pas aux calculs (Transformer l'échelle de mesure de la série de données statistiques: X" = C·X, C=constante)

Positions relatives de la moyenne, médiane et mode

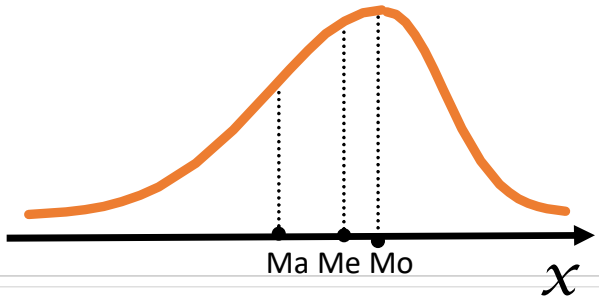
- Distribution **symétrique:**
  Mode ≈Moyenne ≈ Médiane



- Distribution **asymétrique (étalée) à droite:**
  Mode < Médiane < Moyenne



- Distribution **asymétrique à gauche:**
  Mode >Médiane > Moyenne



D’autre type de moyennes

- la moyenne géométrique : transformation de la variable quantitative sur une autre
- echelle (par exemple: echelle logarithmique)
- moyenne pondérée (si les poids sont egaux: moyenne arithmetique)

Mesures de tendance centrale: Valeur centrale

- **Valeur centrale**= moyenne aritmetique entre la valeur maximale et minimale
- **Comment trouver la valeur centrale?**

Valeur centrale=(minimum+maximum)/2

Mesures de tendance centrale: Valeur centrale

Avantages

- ✓ rarement utilisé dans l'analyse statistique

Désavantages

- ✘ l'efficacité réduite parce que tient compte des valeurs minimum et maximum
- ✘ Peut être influencer par des valeurs aberrantes
- ✘ Le manque de représentativité: prendre la même valeur pour toute série ayant des valeurs minimales et maximales identiques

MESURES DE DISPERSION

- Montrent si les valeurs sont plus ou moins proches autour de la moyenne (ou un autre « centre » - médiane) de la série de données
- Quantifient le taux de variabilité des données autour de la valeur centrale
- **Mesures:**
    - amplitude
    - écart interquartile
    - variance
    - écart-type
    - coefficient de variation
    - erreur standard

MESURES DE DISPERSION: la variance

La variance: mesure la dispersion des données autour de la moyenne

La variance (d'une population): moyenne des carrés des écarts des valeurs à la moyenne de la population.

- notation: sigma^2

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

ou

μ est la moyenne d'une population

x_i sont les valeurs de la variable quantitative

N est la taille de la population

MESURES DE DISPERSION: la variance

La variance (de l'échantillon)=somme des carrés des écarts des valeurs à la moyenne de l'échantillon,

- est la variance descriptive
- utilisé pour décrire les échantillons
- sous-estimer la variance de la population
- notation s^2 (en minuscule)

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

ou

$\bar{x}$ est la moyenne d'un echantillon

x_i sont les valeurs de la variable quantitative

n est la taille de l' échantillon

MESURES DE DISPERSION: la variance

- la variance d' échantillonnage (variance d'un échantillon optimisée pour approcher aux mieux la variance de la population):

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

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

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

MESURES DE DISPERSION: la variance

Avantages

- ✓ positive
- ✓ grande variance => grande dispersion autour de la moyenne

Désavantages

- ✘ sensible aux valeurs extrêmes
- ✘ unités de la variance = le carré des unités de la variable

MESURES DE DISPERSION: l'écart type

- **L'écart type (la deviation standard, DS):** 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 – majuscule)

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

- la déviation standard d'une population est:

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

MESURES DE DISPERSION: l'écart type

Avantages

- ✓ Mêmes unités de mesure que la variable elle-même.
- ✓ Positive
- ✓ 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 dispersion autour de la moyenne

Désavantages

- ✗ Peut être utilise dans le cas ou les valeurs de la variable suivent une distribution symétrique et unimodale.

L'écart type : exemple de calcul

	Age(*X*)	$X_i-\bar{X}$ ou i=1,…,10	$(X_i-\bar{X})^2$
X1	26	0.5	0.25
X2	25	-0.5	0.25
X3	28	2.5	6.25
X4	24	-1.5	2.25
X5	26	0.5	0.25
X6	27	1.5	2.25
X7	22	-3.5	12.25
X8	30	4.5	20.25
X9	25	-0.5	0.25
X10	22	-3.5	12.25
Somme	255	**0**	**56.5**
$\bar{X}$	**25.5**	$(X_i-\bar{X})^2/(n-1)$	**6.28**
		St.Dev.	**2.51**

les écarts des valeurs à la moyenne

Le carré des écarts

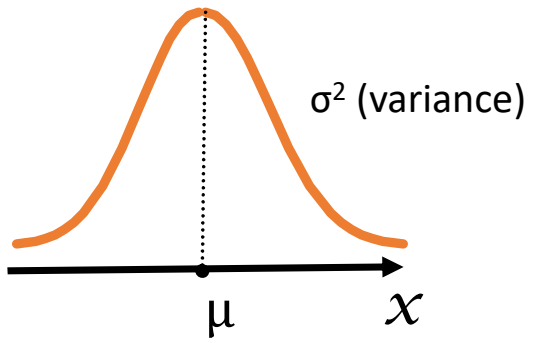
La somme des carrés

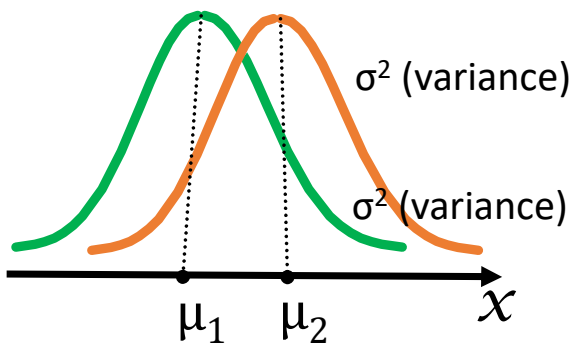
La somme des carrés divisé par la taille de l'echantillon -1

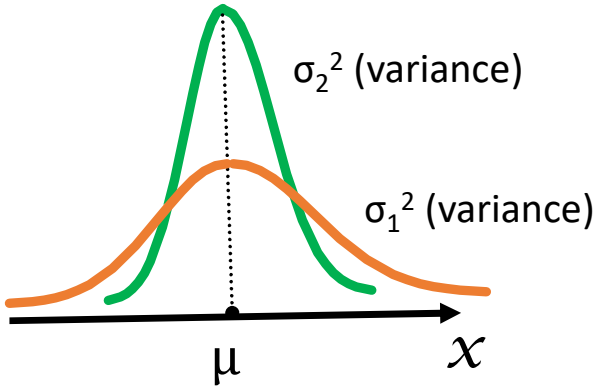
L' ecart-type= la racine carré

Loi Normale (ou loi de Gauss)

- en forme de cloche (unimodale);
- symétrique par rapport à sa moyenne ;
- décalée vers la droite si la moyenne est augmentée (en supposant la variance constante);
- s'aplatit à mesure que la variance augmente mais devient plus pointue quand la variance diminue (pour la moyenne fixe).
- la moyenne et la médiane (et le mode) d'une distribution Normale sont égales ;

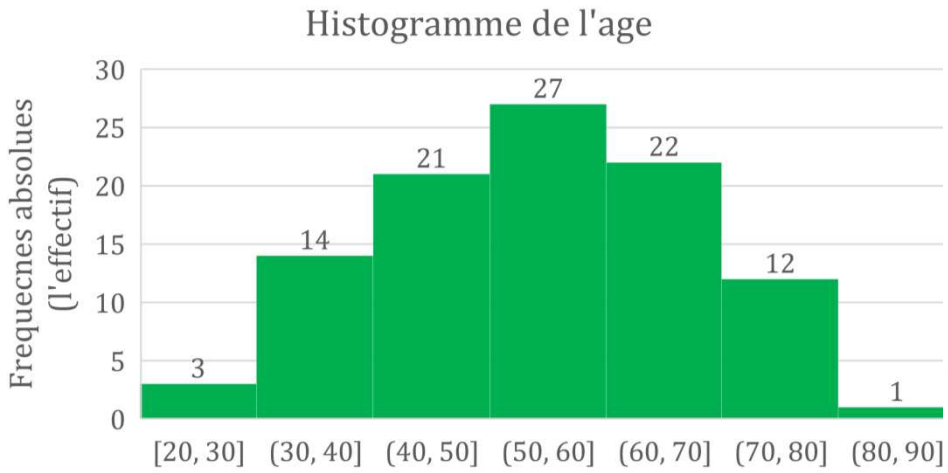




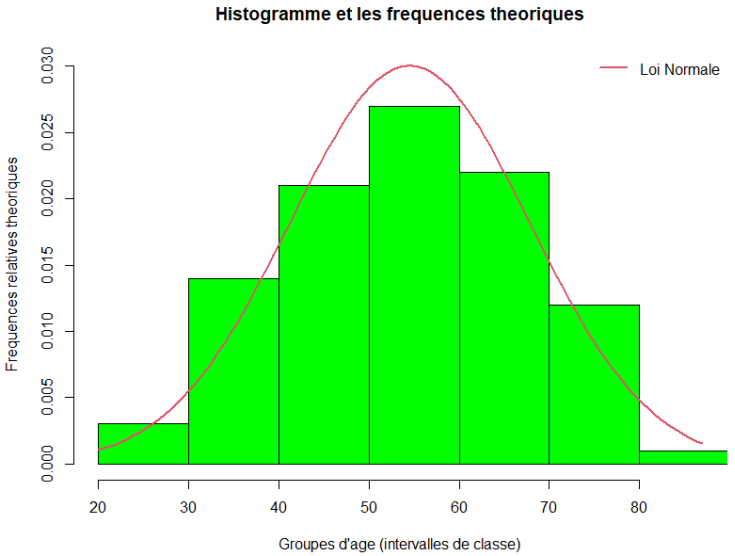


Loi Normale: exemple

	A	B
1	Id_patient	Age (ans)
2	1	59
3	2	57
4	3	63
5	4	48
6	5	70
7	6	72
8	7	39
9	8	45
10	9	64
11	10	34
12	11	55
13	12	53
14	13	63
15	14	55
16	15	71
17	16	68
18	17	48
19	18	37
20	19	54
21	20	56
22	21	52
23	22	68
24	23	68
25	24	87
26	25	34
27	26	48



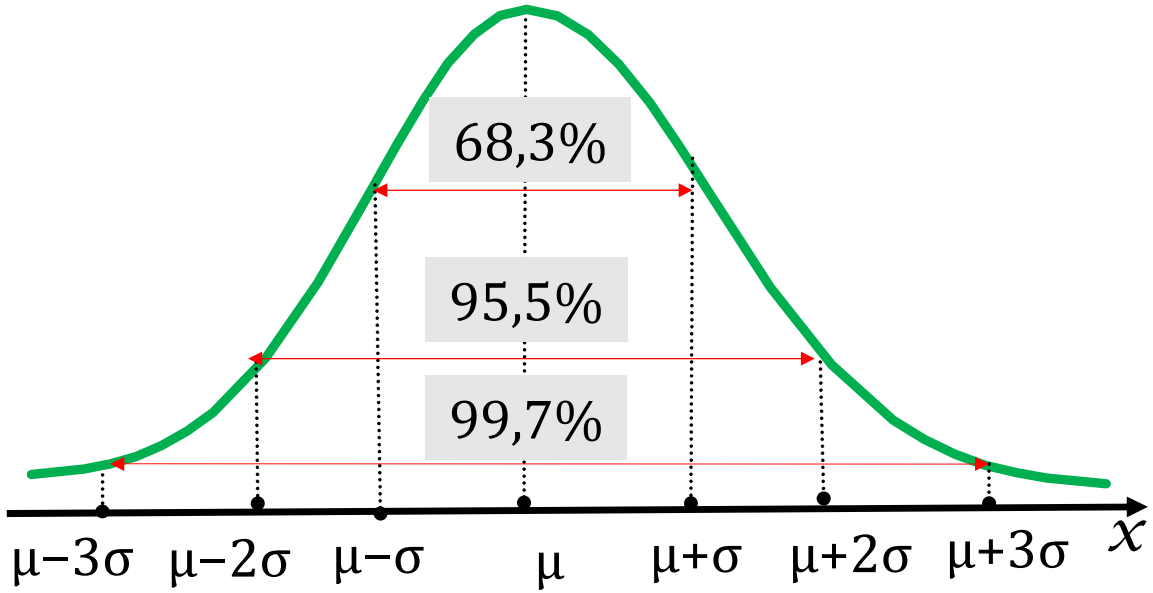
Age (ans)	
Moyenne arithmétique	54.53
Erreur standard	1.34
Mediane	55
Mode	48
Deviation standard	13.35
Variance	178.23
Kurtosis	-0.21
Skewness	-0.24
Amplitude	67
Minimum	20
Maximum	87
Nombre de cas	100



Loi Normale

Pour une série des données qui suit une distribution Normale (Gaussienne):

- Dans l'intervalle **[μ-σ; μ+σ]** on trouve **~68,3 %** de la population.
- Dans l'intervalle **[μ-2σ; μ+2σ]** on trouve **~95,5 %** de la population.
- Dans l'intervalle **[μ-3σ; μ+3σ]** on trouve **~99,7 %** de la population.



MESURES DE DISPERSION: le coefficient de variation

- Il n'a pas une unité de mesure

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

ou

$\bar{x}$ est la moyenne d'un echantillon
S est la déviation standard d' échantillonnage
n est la taille de l' échantillon

MESURES DE DISPERSION: le coefficient de variation

Avantages

- ✓ Il n'a pas unité de mesure
- ✓ Positive
- ✓ Permet la comparaison de la dispersion de deux variables ayant différentes unités de mesure.
- ✓ peut être exprimée en pourcentage

Désavantages

- ✗ n'a de sens que pour des variables quantitatives

MESURES DE DISPERSION: le coefficient de variation

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

Valeur du coefficient de variation	interprétation
CV < 10%	Série de données homogène
10% ≤ CV <20%	Série de données relativement homogène
20% ≤ CV <30%	Série de données relativement hétérogène
CV ≥ 30%	Série de données hétérogène

MESURES DE DISPERSION: l'erreur standard

- **L' erreur standard**

$$ES = \frac{S}{\sqrt{n}}$$

ou

$\bar{x}$ est la moyenne d'un echantillon

S est la déviation standard d' échantillonnage

n est la taille de l' échantillon

- utilisé
  - pour déterminer l'intervalle de confiance de la moyenne d'une population
  - dans la statistique inférentielle

Mesures de symétrie: le coefficient d'asymétrie (skewness en anglais)

Coefficient d'asymétrie (α_3):

- degré d'asymétrie d'une distribution
- la direction de cette asymétrie (positive ou négative);

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

$\boldsymbol{\alpha_3} \approx 0$ => une distribution symétrique.

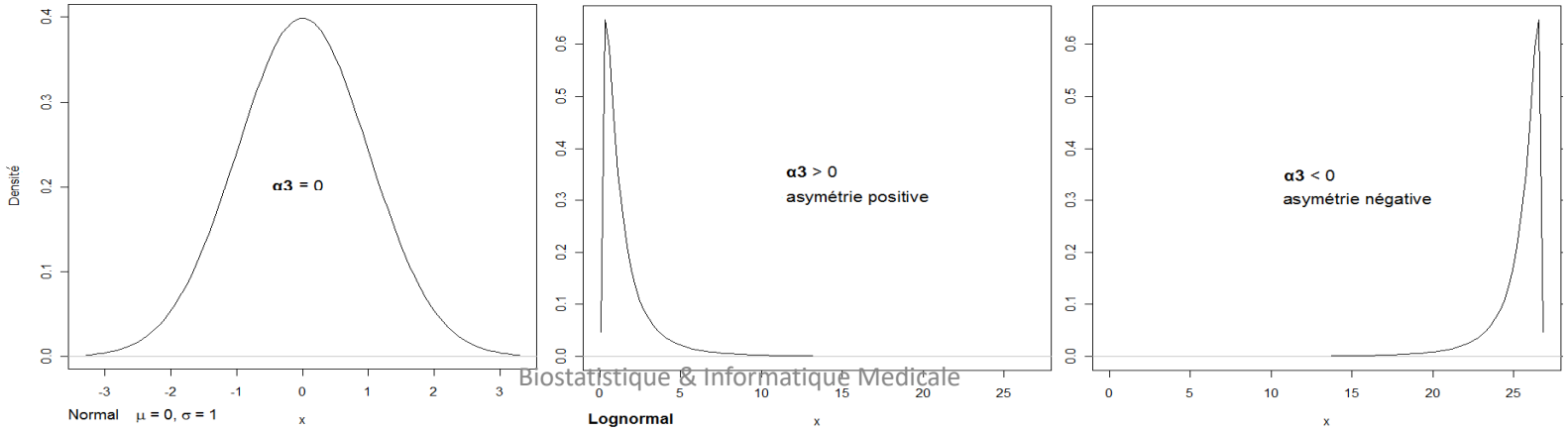
$\boldsymbol{\alpha_3} > 0$ => distribution est plus allongée vers la droite – asymétrie positive

$\boldsymbol{\alpha_3} < 0$ => distribution est plus allongée vers la gauche – asymétrie négative

$\alpha_3 \in$ [0,5; 0,5]: distribution approximative symétrique

$\alpha_3 \in$ [1; -0,5[ou (0,5; – 1]: distribution avec une asymétrie modérée

$\alpha_3 <$ -1 ou >1: asymétrie importante

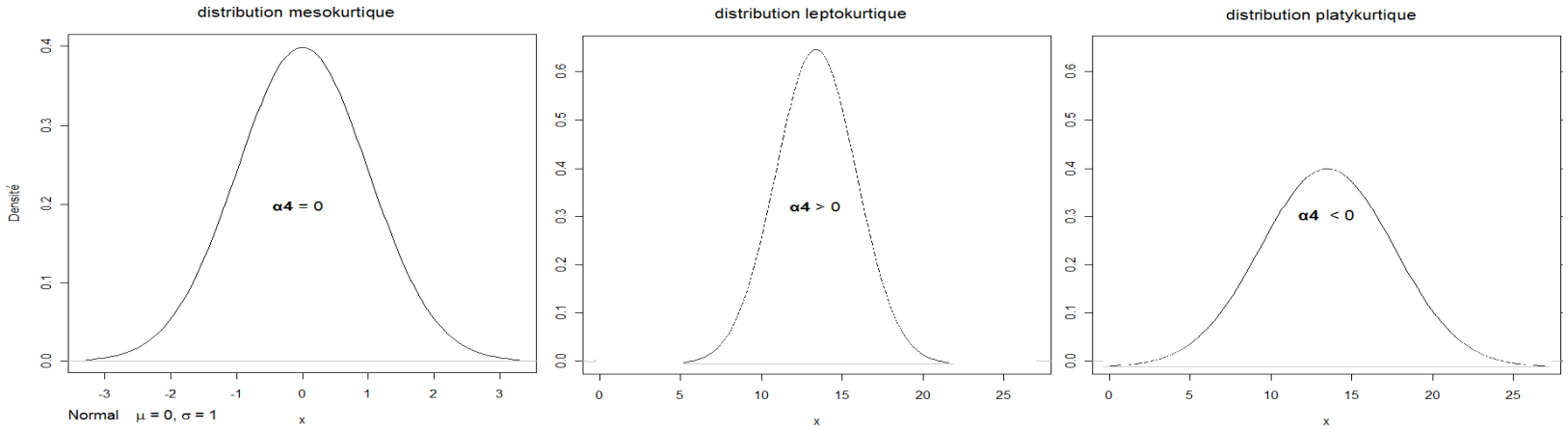


Mesures d'aplatissement: Le coefficient d'aplatissement (Kurtosis en anglais)

Le coefficient d'aplatissement (α_4):

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

- l'angle de la courbe du milieu d'une distribution
- par rapport a une distribution normale (gaussienne)
- **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



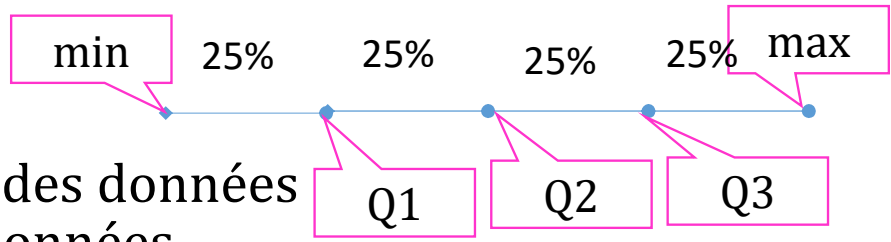
MESURES DE POSITION: quantiles et percentiles

- **Les quantiles** (Q_1,Q_2,...,Q_{q-1}): valeurs remarquables qui partagent la série de données ordonnées en q sous-ensembles (groupes) consécutifs égaux.
- **le quantile d'ordre α (0<α<1)** est la valeur x_α telle qu'une proportion α (%) des données soit plus petite que x_α
- **les plus utilises quantiles: les quartiles, les quintiles, les deciles, les centiles**
- **le** $p^{ième}$ **percentile**:
  - divise la série en deux sous-ensembles (tell qu'au plus p% des valeurs sont en-dessous d'elle et au plus (100-p)% sont au-dessus)
  - est le quantile d'ordre **α=p/100**

MESURES DE POSITION: quantiles & percentiles

•**Les quartiles (Q_1, Q_2, Q_3): divisent la série de données en quatre groupes** ayant la même proportion des données **.**

- le 1er quartile (Q_1) sépare les 25% inférieurs des données
  - •25% des valeurs sont ≤ Q1, 75% sont ≥Q1
- le 2e quartile est la médiane de la série (50%)
  - •50% des valeurs sont ≤ Q2/ médiane, 50% sont ≥Q2/ médiane
- le 3e quartile sépare les 75% inférieurs des données
  - •75% des valeurs sont ≤ Q3, 25% sont ≥Q3



• **Les quintiles (V_1, V_2, V_3, V_4):**

- le 1er quintile (V_1) sépare les 20% inférieurs des données
- le 2e quintile sépare les 40% inférieurs des données
- le 3e quintile sépare les 60% inférieurs des données
- le 4e quintile sépare les 80% inférieurs des données

MESURES DE POSITION: déciles & centiles

- **Les déciles (D_1,..., D_9):**
  - le 1er décile sépare les 10% inférieurs des données
  - le 2e décile sépare les 20% inférieurs des données
  - ...
  - le 9e décile sépare les 90% inférieurs des données

- **Les centiles (C_1,...,C_{99}): permettent de scinder la série de données en 100 sous-ensembles égaux)**
  - le 1er centile (C_1) sépare le 1% inférieurs des données
  - ....
  - le 99e centile (C_{99}) sépare le 99% inférieurs des données

MESURES DE DISPERSION: l'écart/l'intervalle interquartile

- **Intervalle interquartile IQR: [Q1, Q3]**
  - C'est l'intervalle entre le première quartile (Q1) et le 3eme quartile (Q3)
  - D'habitude il est mis après la médiane dans les résultats d'articles scientifiques
    - Ex: Cholestérol mg/dl (médiane, IQR): 189 [162; 211]

  - **Ecart interquartile EQR: Q3-Q1**
  - Parfois certain chercheurs montre seulement la différence entre la quartile 3 et la quartile 1 (étendu ou écart inter quartile EQR=Q3-Q1) au lieu des deux valeurs
    - Ex: Cholestérol mg/dl (médiane, EQR): 189 (49)
  - La meilleure représentation est si on montre les quartiles (Q1, Q3), pas seulement le EQR

Description des variables quantitatives par des graphiques

- **Description d'une seule variable *quantitative***
  - **Histograme, polygone des fréquences, graphique des quantiles**
    - Pour evaluer la forme de la distribution
  - graphique des moyennes – pour des variables normalement distribuées
  - **Graphique par des Colonnes avec barre d' erreur –** pour une variable normalement distribuée
  - **box and wiskers (boite à moustaches) –** pour une variable non normalement distribuée

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

- **La relation entre deux variables**
  - ***Qualitative***
    - **Column** (Clustered Column/ Stacked Column/ 100% Stacked column),
    - ou **Bar** (Clustered Bar / Stacked Bar / 100% Stacked Bar)
  - ***Quantitative***
    - **Scatter** (nouage des points)

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

- **La relation entre deux variables**
  - ***Une variable quantitative en fonction d'une variable qualitative***
    - **Si les données sont normalement distribuées**
      - Graphique des moyennes (avec deviation standard)
    - **Si les données ne sont pas normalement distribuées**
      - Graphique box plot ou wiskers/boite a moustaches

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

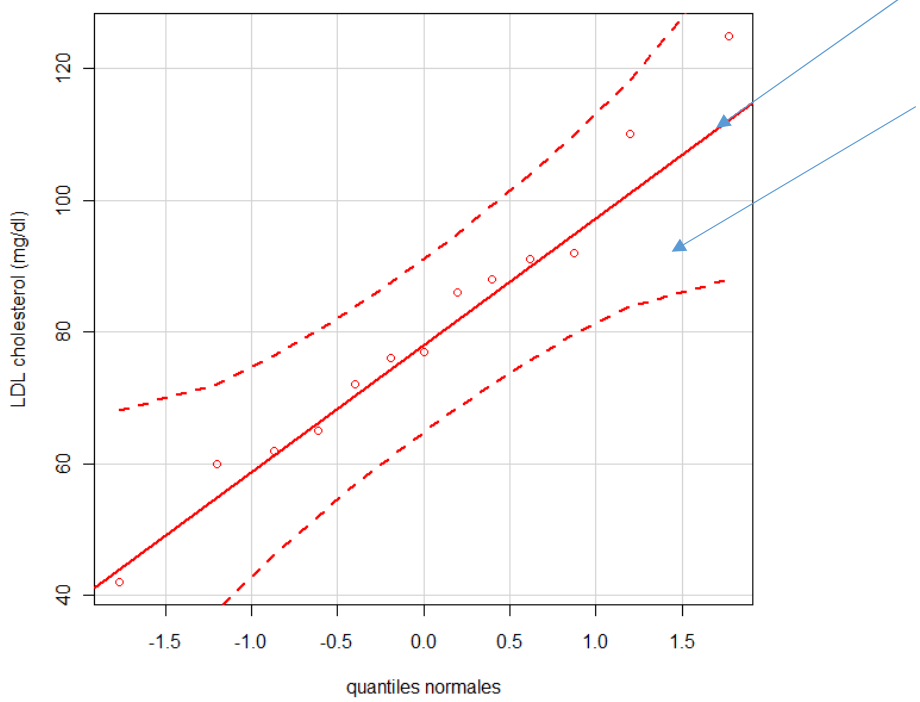
- **L'évolution dans le temps d'une variable qualitative ou quantitative**
  - **Line** (Ligne)
- **La relation entre trois variables quantitatives**
  - **Bubble** (nouage des spheres)
  - Nouage des points tridimensionel

Cas d'une variable quantitative: 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, mieux que l'histogramme, et le polygone des fréquences

Cas d'une variable quantitative: Graphique des quantiles

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



- représente graphiquement les quantiles de deux distributions (la distribution observée des données et la distribution théorique (Normale))
- La droite continue rouge = la droite selon laquelle les points devraient s'aligner en cas de concordance parfaite entre la distribution observée et la distribution Normale
- bande (zone) de confiance à 95% (surface entre les courbes pointillées rouges) tient compte de la variabilité aléatoire d'un échantillon à l'autre: 95% des points doivent, en principe, se trouver à l'intérieur de la zone de confiance
- si presque tous les points les points sont compris dans la bande (zone) de confiance à 95%, le graphique indique que la distribution de données observee semble s'ajuster a une distribution normale.

Excel, EpiInfo:

✓**-n'existe pas**

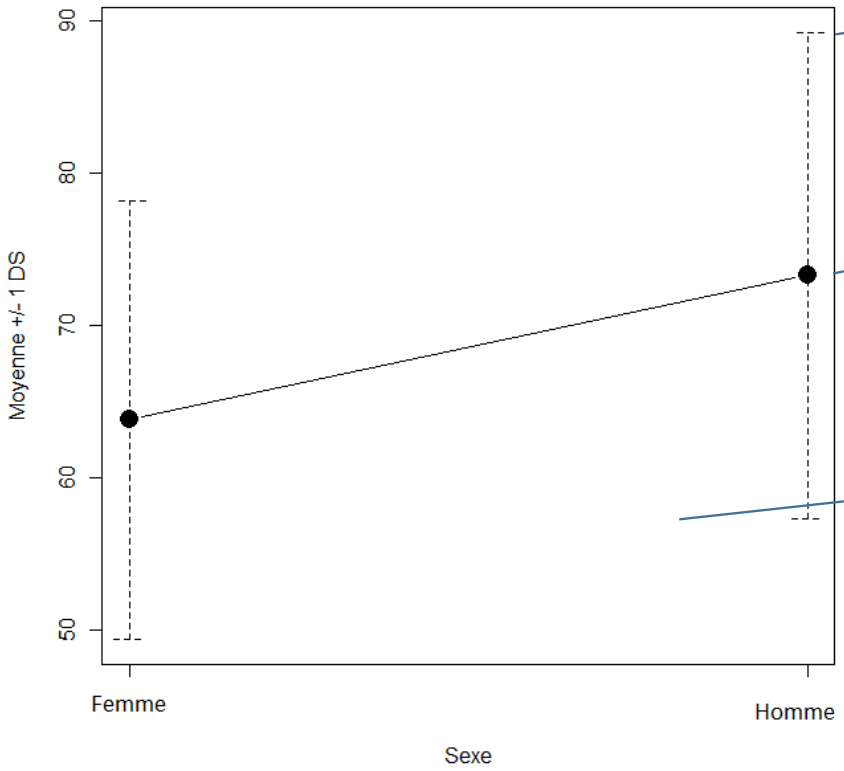
R (R Commander):

✓**KMggplot2/Q-Q plot…**

✓**ou Graphs/Quantile-comparison plot…**

Une variable quantitative: graphique des indicateurs: moyennes et déviations standard (pour la représentation des données a distribution normale)

- Le poids âpres la dialyse en fonction du sexe



Moyenne + écart type

Moyenne

Moyenne – écart type

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**

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

- la boîte à moustache(le box-plot/box and wiskers) = graphique qui permet de visualiser la distribution d'une variable quantitative
- pour la représentation des données qui **ne suivent pas la distribution Normale**
- On la construit de la manière suivante :

  - on trace une boîte de longueur Q3 − Q1
  - on partage la boîte par un trait à la position de la médiane
  - on trace la moustache gauche/inferieur de longueur

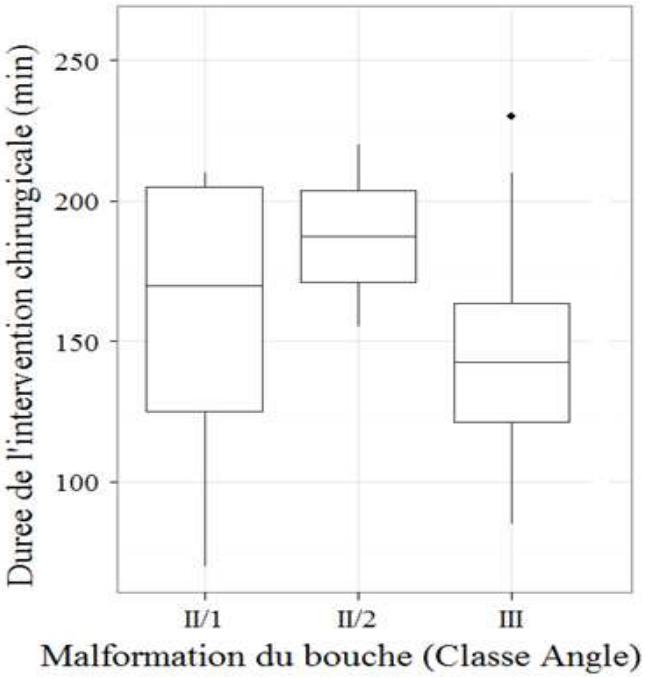
$$\min\left(Q_1 - X_{\min}, 1.5 * \left(Q_3 - Q_1\right)\right)$$

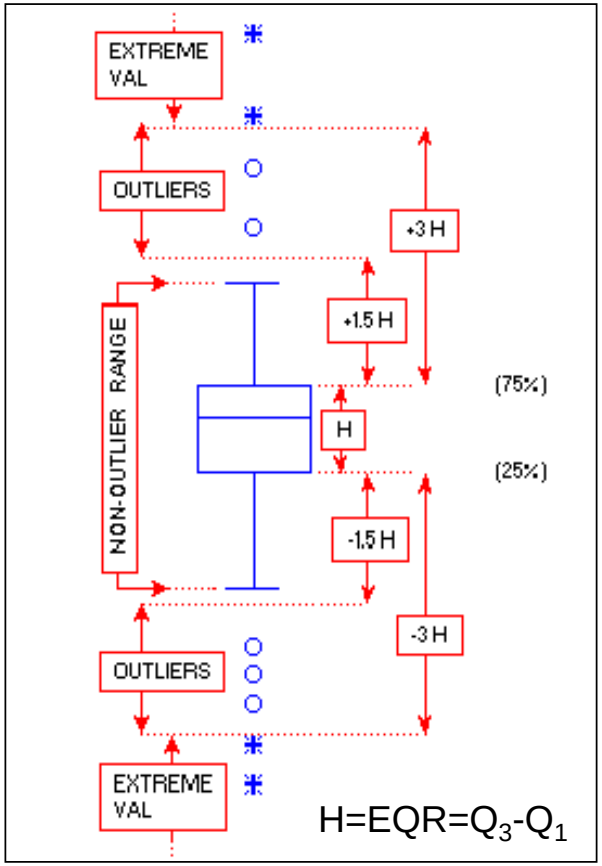
  - on trace la moustache droite/supérieur de longueur

$$\min\left(X_{\max} + Q_3, 1.5 * \left(Q_3 - Q_1\right)\right)$$

  - si certains individus sont en dehors des moustaches, on les représente par des * (les valeurs extrêmes=valeurs inhabituelles inferieurs à $Q_1 - 3\times EQR$ ou supérieurs à $Q_3 + 3\times EQR$) et ° (outliers = valeurs inferieurs à $Q_1 - 1{,}5\times EQR$ ou supérieurs à $Q_3 + 1{,}5\times EQR$) –voir le graphique

Une variable quantitative: box-plot (pour la représentation des données a distribution non normale)

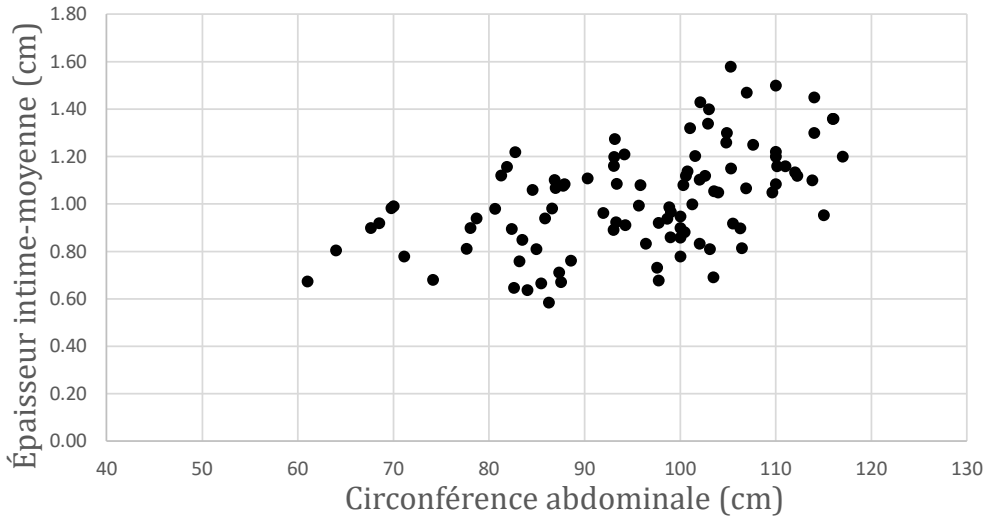




R (R Commander):
KMggplot2/Box plot/Error bar plot…
ou, Graph/Box plot…

Deux variables quantitatives: graphique par nuage de points

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



Relation de proportionnalité

Excel:

✓**INSERT CHART: type: XY SCATTER**

R (R Commander):

✓**KMggplot2/Scatter plot…**

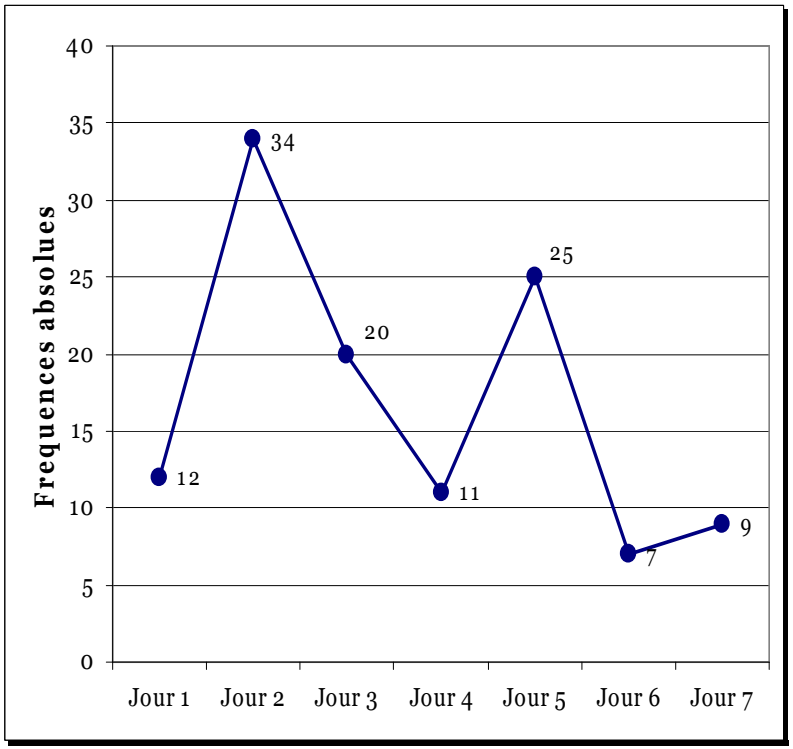
✓**ou Graphs/Scatterplot…**

EpiInfo:

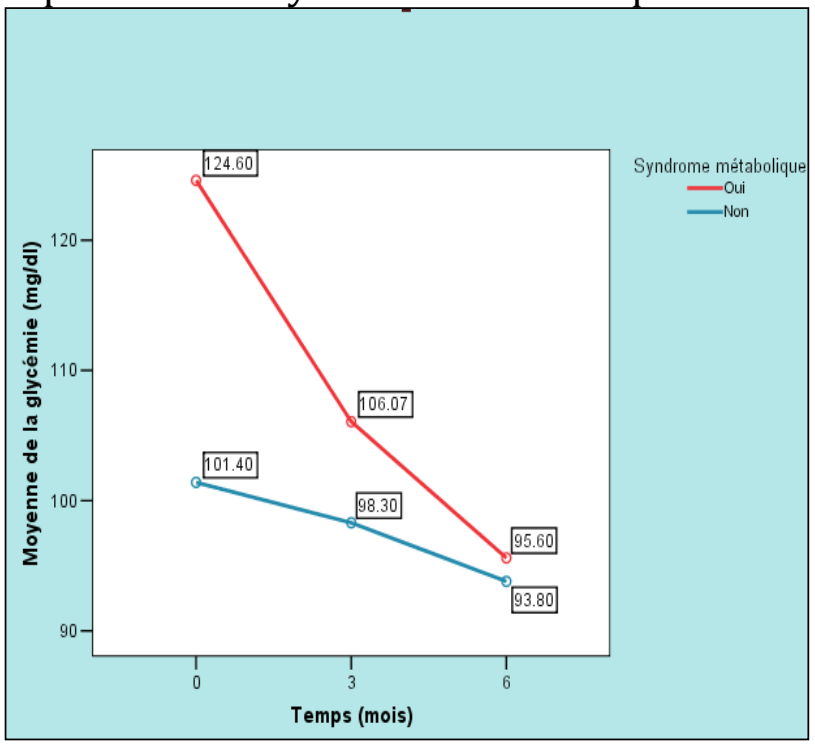
✓**STATISTICS/GRAPH/SCATTER XY**

L'evolution d' une variable quantitative dans le temps: Graphique linéaire

L'évolution de nombre des poussées hypertensifs pendant un semaine



L'évolution du moyenne de la glycémie en fonction de la présence du syndrome métabolique



Excel:
✓**INSERT CHART/LINE**

EpiInfo:
✓**STATISTICS/GRAPH/ LINE**

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

Ce qu'on a appris…

- Statistique descriptive pour les **variables quantitatives**:
  - Mesures de centralité
  - Mesures de position
  - 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

Exemple de questions

E1. L'âge des 30 patients hospitalisée atteints de la COVID-19 a été observé. Les suivantes statistiques ont été calculée : moyenne=41,2 ans, médiane=41,1 ans, écart type descriptif=10 ans, coefficient d'asymétrie= -0,30, coefficient d'aplatissement= -0,4, coefficient de variation = 0.25.

Lesquelles des réponses suivantes sont correctes?

A. les données semble être approximative normalement distribuées
B. Approximative (environ) 68% des malades sont âgés entre 47.2 et 75.2 ans
C. la distribution de l'Age est un peu aplatée
D. la distribution de l'Age a une queue vers la gauche
E. les données de l'Age sont relativement homogènes

R1: A, B, C, D

Solution de l'exercice:

- Pour etablir si les valeurs de l'Age suivent la loi Normale, une des methodes connue sera basé sur les informations concernant le coefficient d'asymétrie et le coefficient d'aplatissement ainsi que la moyenne et la médiane. Si les données sont normalement distribuées, la moyenne et la médiane doivent être proches l'une de l'autre. mais à quelle distance/écart ?? Cela dépend du contexte et il est difficile d'avoir un moyen uniforme de vérifier; voici une des raisons pour lesquelles nous regardons aussi a des autres coefficients (coefficient d'asymétrie et le coefficient d'aplatissement). Si l'un des deux coefficients est en dehors de l'intervalle [-1, 1], on considère qu'il y a des suggestions que les données ne sont pas normalement distribuées. Si les deux coefficients sont compris dans l'intervalle [-1, 1], on considère qu'il y a des suggestions que les données sont normalement distribuées. Compte tenu de tout cela, la réponse (A) est correcte.
- Si les donnes sont normalement distribuées, on sait que dans l'intervalle [moyenne - 1 × écart type, moyenne + 1 × écart type] on trouve 68% des données (voir le diapo 33). 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'aplatissement suggère une distribution des données avec une 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 0,25, donc 25%, qui indique que les données sont relativement hétérogènes, ainsi (E) est faux .

Exemple de questions

E2. Nombre d'implants dentaires par patient des 40 personnes âgées a été observé. Les suivantes statistiques ont été calculée : le deuxième quartile=4, le première quartile =1, le troisième quartile = 5, le minimum=0, le maximum 7.

Lesquelles des réponses suivantes sont correctes?

A. l' écart interquartile est égale a 4

B. le percentile 75% = 5

C. le 5e décile est égale a la 4

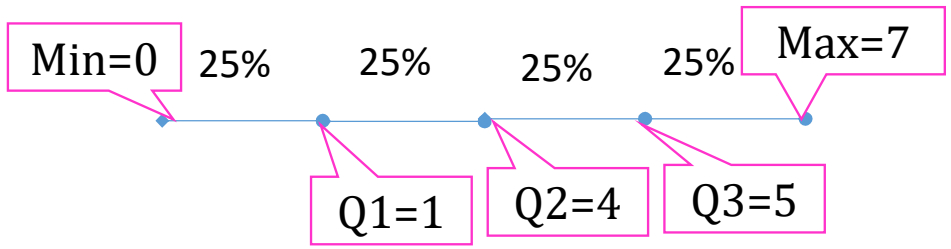
D. l'amplitude est égale à 7

E. dans une graphique boite à moustaches pour le Nombre d'implants dentaires par patient, la ligne inferieure de la boite correspond à 4

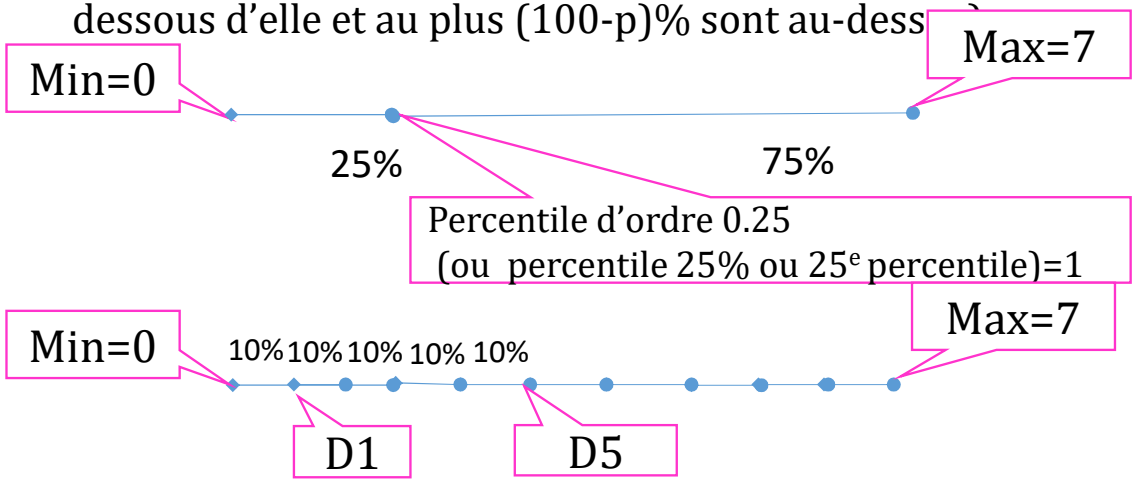
R2: A, B, C, D

Solution de l'exercice:

- Les quartiles:



•**le** p^{ième} **percentile**: divise la série en deux sous-ensembles (tell qu'au plus p% des valeurs sont en-dessous d'elle et au plus (100-p)% sont au-dess



Exemple de questions

E3. Lesquelles des réponses suivantes sont correctes :

A. une bonne graphique pour le Niveau de la intelligence (réduite, normale, supérieure) est la graphique ligne

B. une bonne graphique pour les Types des différents médicaments utilisées dans un cabinet dentaire est le graphique camembert

C. une bonne graphique pour la relation entre le Poids (kg) et la Circonférence abdominale (cm) est la graphique ligne

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

E. la présence des Réactions secondaires (vrai/faux) au vaccin anti-COVID-19 peut être bien représente avec un graphique boite a moustache

R3: B, D

Solution de l'exercice:

A. Ici, la variable **Niveau de la intelligence** (réduite, normale, supérieure) est une variable qualitative ordinale, et elle n'est pas représenté dans le temps. Si dans l' énonce du problème on observe l' évolution dans le temps d'une variable qualitative, ou quantitative, on peut utiliser une graphique ligne, donc la réponse A est fausse

B. la variable **Types des différents médicaments utilisées** dans un cabinet dentaire est une variable qualitative donc pour la répartition des catégories d'une seule variable qualitative, le camembert sera un bon graphique d'être utilise donc la réponse B est correct

C. les variables **Poids** (kg) et la **Circonférence abdominale** (cm) sont des variable quantitatives donc pour représenter graphiquement leur relation, le nouage de points sera un bon graphique d'être utilise donc la réponse C est fausse

D. le graphique boite a moustaches (ou box plot), nous aide a vérifier s'il existe une asymétrie autour de la valeur médiane (en regardant a la distance aux quartiles, ou a la longueur des moustaches); l'existence de la symétrie n'implique pas l'existence la normalité mais une distribution Normale este forcement symétrique; l' histogramme nous permet de visualiser la distribution empirique de données et de regarder si la distribution de données semble s'ajuster a une distribution normale (voir le diapo 32) donc la réponse D est correct

E. Ici, la variable Réactions secondaires (vrai/faux) au vaccin anti-COVID-19 est une variable qualitative dichotomique pour laquelle on peut utiliser un graphique camembert, bar, ou colonne donc la réponse E est incorrect

Exemple de questions

E4. Lesquelles des réponses suivantes sont correctes :

A. une bonne graphique pour la Longueur d'un implant dentaire (mm) est le graphique des moyennes

B. une bonne graphique pour l'évolution dans le temps de la Hauteur (cm) est la graphique ligne

C. une bonne graphique pour l'évolution dans le temps du Nombre des interventions d'ostéointégration d'implants dans la mâchoire par jour est le graphique ligne

D. pour la variable Température corporelle le graphique boite a moustache est meilleur que le graphique des quantiles pour évaluer la normalité

E. la présence de la Douleur dentaire (présente/absente) peut être bien représente par une histogramme

R4: A, B, C

Solution de l'exercice:

A. Ici, la variable Longueur d'un implant dentaire (mm) est une variable quantitative continue, et on peut utiliser une graphique contenant des moyennes (voir le diapo 51) donc la réponse A est correcte

B. la variable Hauteur (cm) est une quantitative continue donc pour la répartition des ses valeurs pendent d'une période du temps le graphique par des lignes (voir le diapo 55 – premier graphique) sera un bon graphique d'être utilise donc la réponse B est correct

C. les mêmes explications que pour le point B

D. le graphique quantile-quantile (voir le diapo 49) nous offre une représentation graphique plus précise pour comparer la distribution empirique (observée) des données et la distribution Normale (les quantiles = valeurs qui divisent la série de données en intervalles/parties contenant le même effectif; médiane = quantile d'ordre 0,5, le premier quartile (Q1) = quantile d'ordre 0,25);

E. Ici, la variable Douleur dentaire (présente/absente) est une variable qualitative dichotomique pour laquelle on peut utiliser un graphique camembert, bar, ou colonne donc la réponse E est incorrect

Exemple de questions

E5. Lesquelles des réponses suivantes sont correctes :

A. une bonne graphique pour la Longueur d'un implant dentaire (mm) est le graphique histogramme

B. une bonne graphique pour l'évolution dans le temps de la Hauteur (cm) est la graphique ligne

C. une bonne graphique pour l'évolution dans le temps du Poids (kg) est la graphique colonne

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

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

R5: A, B, D, E

Exemple de questions

E6. Lesquelles des réponses suivantes sont correctes, concernant le graphique a coté:

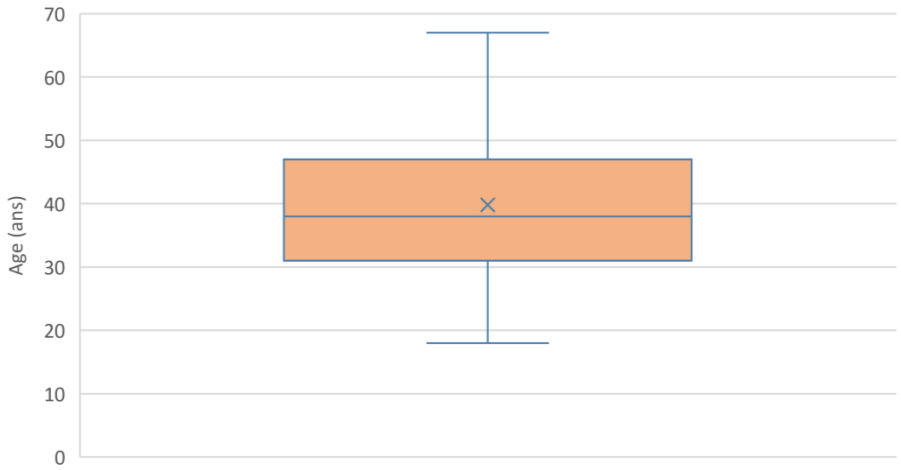
A. les données semblent être normalement distribuées

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 la même que la différence entre la médiane et la quartile 1

E. le coefficient d'asymétrie est plus probable < 0



R6: B, C

Solution de l'exercice:

A. 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) ne sont pas également espacés de la médiane (la ligne horizontale à l'intérieur de la boîte), et les deux moustaches n'ont pas la même longueur; il y a donc une suggestion que les données ne sont pas normalement distribuées, et la réponse (A) est incorrecte.

- Les réponses (B) est correcte, car nous pouvons observer l'asymétrie dans le graphique; la distribution de l'Age est positivement asymétrique, car la portion droite de la boîte et la moustache droite sont plus longues qu'à gauche de la médiane (nombreux patients ont l'âge <l'âge moyenne) ainsi, le coefficient d'asymétrie est plus probablement > 0 (donc la réponse E est fausse)
- Le quartile 2 est la médiane, et nous pouvons voir/déduire sur le graphique la distance/écart entre la médiane, et les deux quartiles (1 et 3); l' écart n'est pas la même donc la réponse (D) est incorrecte

MERCI POUR VOTRE ATTENTION!