UMFT
Universitatea de
Medicină și Farmacie
„Victor Babeș"
din Timișoara

INFORMATIQUE
Médicale et
BIOSTATISTIQUE

Estimation statistique. Estimation de la moyenne

Cours 5

Dr. Mirela FRANDES

UMFT
Universitatea de Medicină și Farmacie „Victor Babeș” din Timișoara

Statistique descriptive

- ensemble des méthodes visant à résumer l’information contenue dans un ensemble de données à l’aide de
  - tableaux
  - graphiques
  - indicateurs numériques (statistiques)

UMFT
Universitatea de Medicină și Farmacie „Victor Babeș” din Timișoara

Méthodes de présentation des données

- Tableaux:
  - une variable
  - deux ou plus de variables
- Diagrammes
  - une variable
  - deux/trois variables

UMFT
Universitatea de Medicină și Farmacie „Victor Babeș” din Timișoara

Tableaux

- distribution de la fréquence **d’une variable**:
  - qualitative / quantitative discrète;
  - quantitative continue;

- distribution de la fréquence pour **deux variables** qualitative
  - (tableau de contingence / tableau croisé);
  - permet de voir un relation entre les deux variables

- des **indicateurs/statistiques**
  - Moyenne, déviation standard
  - Médiane, quartile 1, 3
  - permet de décrire un variable ou de comparer des groups

UMFT
Universitatea de Medicină și Farmacie „Victor Babeș” din Timișoara

Comment peut-on obtenir la distribution de fréquences ?

- On regroupe les données par **classes** dans un **tableau** indiquant la répartition des individus selon le **caractère** (variable) étudie

- **Variable qualitative ou quantitative discrète**
  - Classe=une valeur de la variable
- **Variable quantitative**
  - Classe=intervalle ou classe de valeurs

UMFT
Universitatea de Medicină și Farmacie „Victor Babeș” din Timișoara

Une variable qualitative nominale: distribution de fréquences

- On montre les fréquences absolues et fréquences relatives

Causes de décès	*Nombre de patients*	*Pourcentage (%)*
L'asphyxie à la naissance	527	26.30
Blessures obstétricales	92	4.59
Pneumonie	181	9.03
Malformations congénitales	598	29.84
Autres causes	606	30.24
Total	**2004**	**100.00**

- La distribution des causes de décès chez les nouveaux nées

UMFT
Universitatea de Medicină și Farmacie „Victor Babeș" din Timișoara

Une variable qualitative nominale: distribution de fréquences

- On montre les fréquences absolues et relatives

Diagnostique	Nombre de patients	Pourcentage (%)
Intrusion procesus alveolaire	1	3,70
laterognatie du mandibule	1	3,70
maxilar etroit	18	66,67
oclusion ouverte	1	3,70
protrusion mentoniere	2	7,41
Retruzion mentoniere	4	14,81
Total	27	100,00

- La distribution des diagnostiques chirurgicales dentaires

Distribution de fréquences

Que faut-il indiquer pour chaque classe ?	*Définitions*
fréquence **absolue** (ni)	**nombre** d'individus de la classe
fréquence **relative** (fi)	**proportion** de sujets de la population ou de l'échantillon appartenant a une certaine classe ;
fréquence **absolue** cumulée croissante de la classe (xi)	**effectif** d'individus dans la population (ou l'échantillon) pour lesquelles le caractère étudié peut prendre une valeur inférieur ou égale a xi
fréquence **absolue** cumulée décroissante de la classe (xi)	**effectif** d'individus dans la population (ou l'échantillon) pour lesquelles le caractère étudié peut prendre une valeur supérieur ou égale a xi
fréquence **relative** cumulée croissante de la classe (xi)	**proportion** d'individus dans la population (ou l'échantillon) pour lesquelles le caractère étudié peut prendre une valeur inférieur ou égale a xi
fréquence **relative** cumulée décroissante de la classe (xi)	**proportion** d'individus dans la population (ou l'échantillon) pour lesquelles le caractère étudié peut prendre une valeur supérieur ou égale a xi

UMFT
Universitatea de Medicină și Farmacie „Victor Babeș" din Timișoara

Variables Qualitatives

Tableau de fréquences

X (Variable)	Frequence absolue (ni)	Fréquence relative (%) (fi) $f_i = \frac{n_i}{n} x100$	Fréquence absolue cumulée croissante (Ni)	Fréquence relative cumulée croissante (Fi) (%)
x1	n1	f1	N1 =n1	F1= f1
x2	n2	f2	N2 = n1 + n2	F2 = f1 + f2
…	…	…	…	…
Xk	nk	fk	Nk = n	Fk = 100
Total	n	100		

Excel:
COUNTIF
DATA-PIVOT TABLE

EpiInfo:
STATISTICS/FREQUENCIES

UMFT
Universitatea de Medicină și Farmacie „Victor Babeș" din Timișoara

Deux variables qualitative

tableau de contingence / tableau croise

- Montre la relation entre deux variables (e.g., Obésité et Diabète)

		Diabète		
		Oui	Non	**Total**
Obésité	Oui	24	54	**78**
	Non	2	10	**12**
	Total	**26**	**64**	**90**

		Diabète		
		Oui %	Non %	**Total %**
Obésité	Oui	30,77	69,23	**100.0**
	Non	16,67	83,33	**100.0**
	Total%	**28,89**	**1,11**	**100.0**

Excel:
DATA-PIVOT TABLE

EpiInfo:
STATISTICS/TABLES

Une variable quantitative

tableau de classes de fréquences (exemple)

Classes du poids du nouveau nees (g)	Frequences absolues	Frequences relatives (%)	Fréquences relatives cumulées croissante (%)
<=2500	14	2.31	2.31
(2500,2700]	31	5.12	7.43
(2700,2900]	76	12.54	19.97
(2900,3200]	170	28,05	48,02
(3200,3400]	97	16,01	64,03
(3400,3600]	94	15,51	79,54
(3600,3800]	62	10,23	89,77
(3800,4000]	31	5,12	94,88
>4000	31	5,12	100,00
Total	**606**	**100,00**	

Une variable quantitative

Tableau de classes de fréquences

X (Variable)	Frequence absolue (ni)	Fréquence relative (%) (fi) $f_i = \frac{n_i}{n} x100$	Fréquence absolue cumulée croissante (Ni)	Fréquence relative cumulée croissante (Fi) (%)
(a1, a2]	n1	f1	N1 =n1	F1= f1
(a2, a3]	n2	f2	N2 = n1 + n2	F2 = f1 + f2
…	…	…	…	…
$(a_k, a_{k+1}]$	nk	fk	Nk = n	Fk = 100
Total	n	100		

- en général, on va travailler avec des classes de même amplitude

EXCEL:
TOOLS-DATAANALYSYS-HISTOGRAMME

EpiInfo:
STATISTICS/FREQUENCIES

UMFT
Universitatea de Medicină și Farmacie „Victor Babeș" din Timișoara

Variable quantitative

Tableau des indicateurs

- Permet de décrire un variable ou de comparer différents groups en ce qui concerne les valeurs d'une variable
- On utilise le plus fréquent:
  - moyenne +/- déviation standard
  - médiane [quartile 1;3]
- Les caractéristiques du cholestérol chez les sujets avec Gaucher et les témoins

Caractéristique	Group	
	cas (n=13)	témoin (n=10)
Cholestérol total (mg/dl) (Moyenne+/-1DS)	128,46+/-36,14	145,50+/-20,79
HDL cholestérol (mg/dl) (Moyenne+/-1DS)	24,53+/-6,24	54,00+/-8,68
LDL cholestérol (mg/dl) (Médiane [quartile 1;3])	75 [55; 84]	82 [68; 97,25]

Diagrammes

Variable qualitative	Variable quantitative
Pour décrire une variable -La diagramme à **secteurs**; (camembert/pie) -La diagramme en **barres (colonnes)**; -La diagrammes en **bandeaux**;	**Pour décrire une variable** **-L'histogramme** ; -Le **polygone** des effectifs (ou des fréquences); -La courbe cumulative (ou polygone des fréquences cumulées); Les graphiques **des indicateurs**: -La graphique **boîte à moustaches** (le box-plot); -Le graphique du **moyenne et déviation standard** -Le graphique des **quantiles** -Le graphique **linéaire**
Relation entre deux variables: - La diagramme en barres ou en bandeaux, pour chaque catégorie du deuxième variable, mais sur le même graphique ou deux graphiques a cote	**Relation entre deux variables:** - Le graphique « nouage des points »

UMFT
Universitatea de Medicină și Farmacie „Victor Babeș” din Timișoara

Principes de faire les graphiques

- toute représentation graphique doit avoir:
  - titre clair, concis et précis;
  - définitions des axes, sans abréviations (a l’exception des unités de mesure);
  - unités de mesure;
  - légende (s’il faut);
  - toute symboles, des chiffres ou lettres utilisées dans la figure doivent être expliqué clairement dans la légende.

UMFT
Universitatea de Medicină și Farmacie „Victor Babeș” din Timișoara

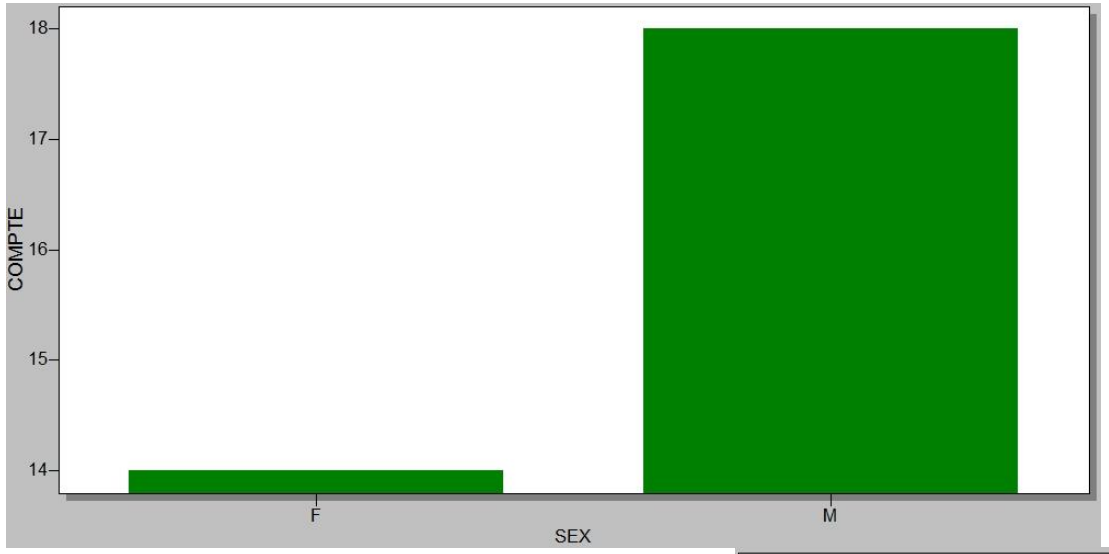
Une variable qualitative

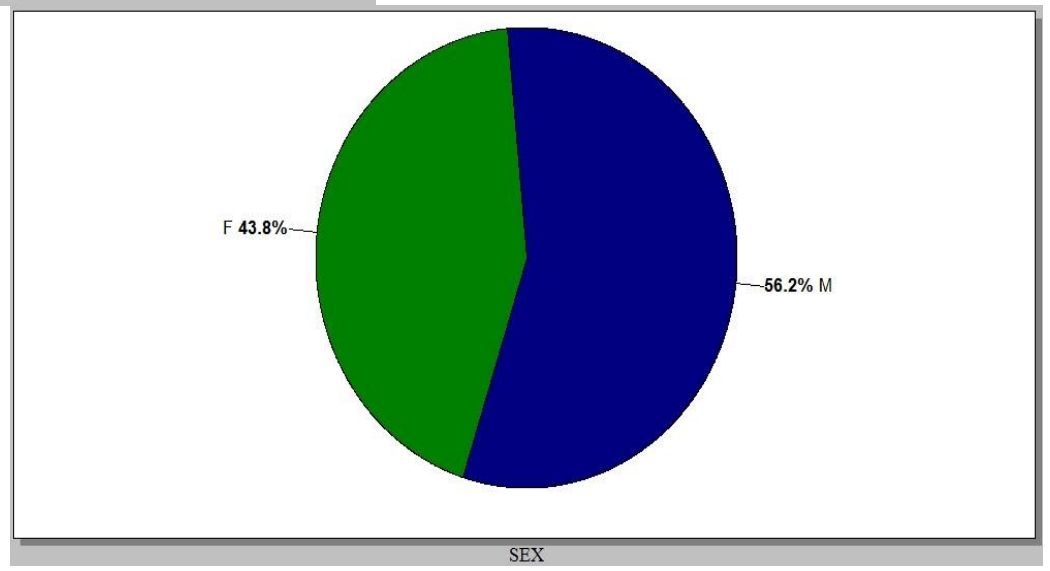
Diagrammes

- Représentent des aires proportionnelles aux fréquences des catégories (classes) de la variable statistique (absolues/relatifs)
- Types
  - Diagrammes à secteurs circulaires
  - Diagrammes en colonnes/barres
  - Diagrammes en bandeaux

- Conseilles:
  - Pourcentages sont meilleures que les valeurs absolues mais les valeurs absolues sont très importantes aussi

Une variable qualitative:
Diagramme en barres et camembert/sectorielle





Excel:
INSERT/GRAPH/BAR ou PIE

EpiInfo:
STATISTICS/GRAPH/BAR ou PIE

UMFT
Universitatea de Medicină și Farmacie „Victor Babeș" din Timișoara

Une variable quantitative continue

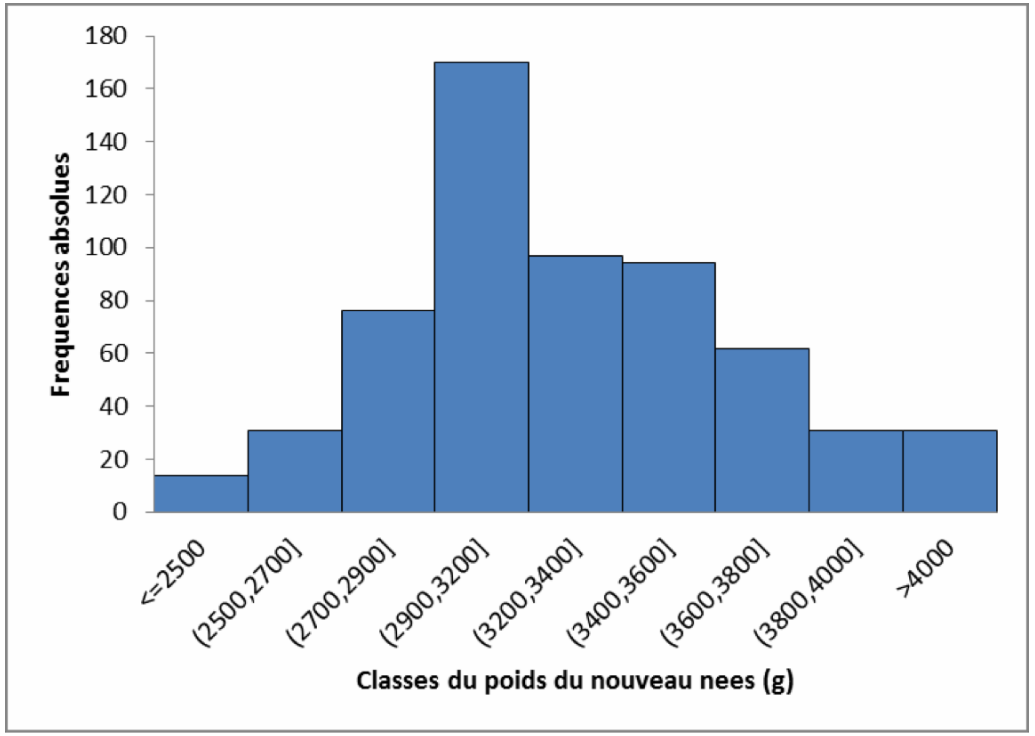
Histogramme = rectangles juxtaposes

- chacune des bases est égale a l'intervalle de chaque classe
- la hauteur est la fréquence absolue/relative
- l'aire de chaque rectangle est proportionnelle
- aux effectifs ou aux fréquences de la classe correspondante;

Polygone des effectifs (ou des fréquences):

- on joint les points milieu du sommet des rectangles adjacents
- par un segment de droite
- le polygone est fermé aux deux bouts en le prolongeant sur l'axe horizontal

Une variable quantitative continue:
Histogramme



Excel:
TOOLS/DATA ANALYSIS/HISTOGRAM

EpiInfo:
STATISTICS/GRAPH/ HISTOGRAM

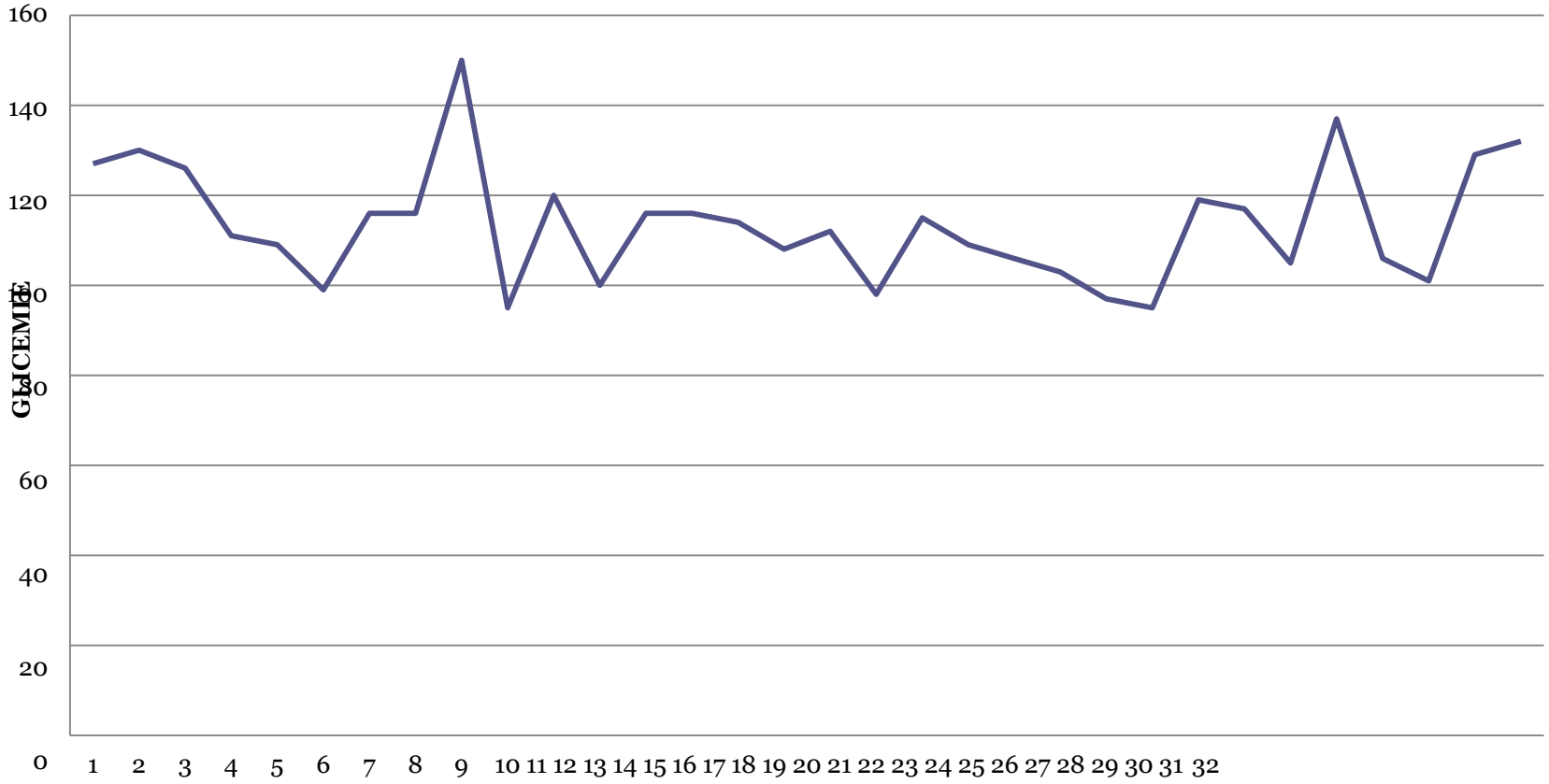
UMFT
Universitatea de Medicină și Farmacie „Victor Babeș" din Timișoara

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

L’évolution d’une variable quantitative: graphique linéaire



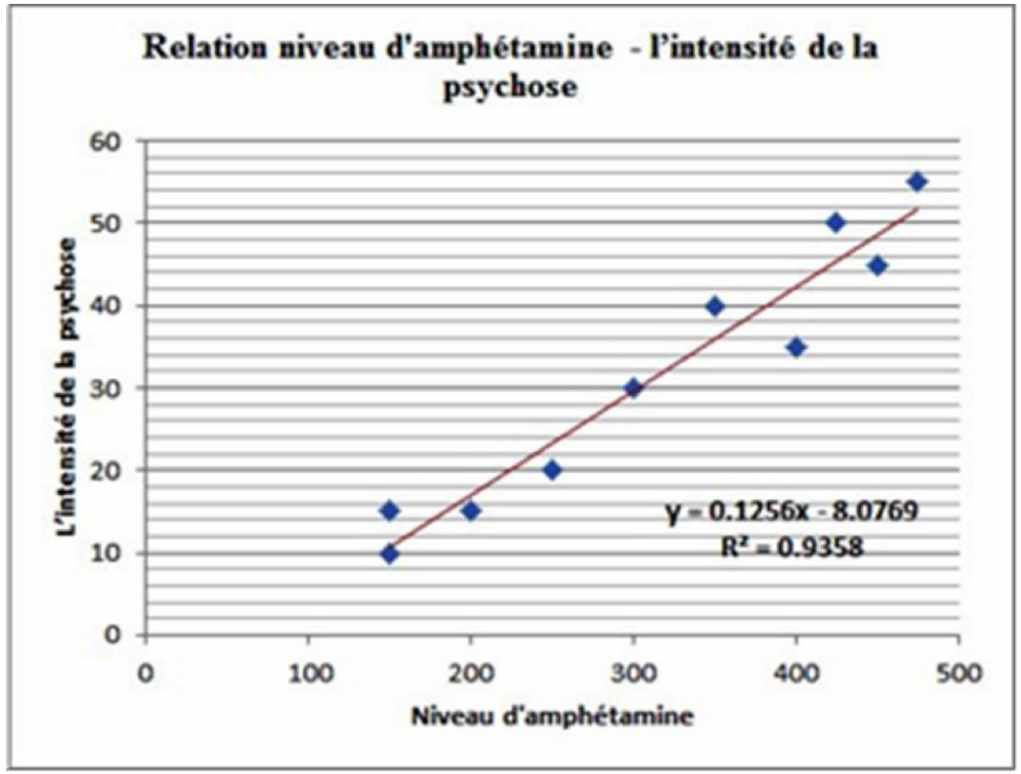
Excel:
INSERT CHART: type: LINE

EpiInfo:
STATISTICS/GRAPH/LINE

Deux variables quantitatives:

Diagramme par un nuage de points

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



Excel:
INSERT CHART: type: XY SCATTER

EpiInfo:
STATISTICS/GRAPH/ SCATTER XY

Estimation statistique

UMFT
Universitatea de Medicină și Farmacie „Victor Babeș” din Timișoara

Inférence statistique

- **Statistique** = but d’étudier **les populations**
- **Échantillon** (sélection) – un ensemble finit extrait d’une population
- **INFÉRENCE STATISTIQUE** = extension de les propriétés déterminé sur un échantillon aux population entière =

échantillon → population

UMFT
Universitatea de Medicină și Farmacie „Victor Babeș” din Timișoara

Inférence statistique

- Pour construire un **échantillon représentative** il faut sélecter:
  - ses éléments d’une manière aléatoire
  - un nombre suffisant d’ éléments
- Interpréter:
  - **indicateurs de tendance centrale**
    - Moyenne
    - Médiane
    - mode
  - **indicateurs de dispersion**
    - amplitude
    - écart-type
    - variance

UMFT
Universitatea de Medicină și Farmacie „Victor Babeș" din Timișoara

Inférence statistique

- Échantillon - caractéristiques sont notée par des lettres latines (m, s)

- Y échantillons de taille *n (une population de taille N)*
  – caractéristiques sont notée par des lettres grecques (μ, σ)

UMFT
Universitatea de Medicină și Farmacie „Victor Babeș” din Timișoara

La déviation standard

- L’échantillon $$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$$

- La population $$\sigma = \sqrt{\frac{\sum (x_i - \mu)^2}{n}}$$

UMFT
Universitatea de Medicină și Farmacie „Victor Babeș” din Timișoara

Inférence statistique

échantillon → population

- Les paramètres d’échantillon = déterminer
- Les paramètres de la population= estimer

- Estimer la moyenne n’est pas possible de dire **« quelle valeur »** mais **« dans quelle intervalle »**

Example

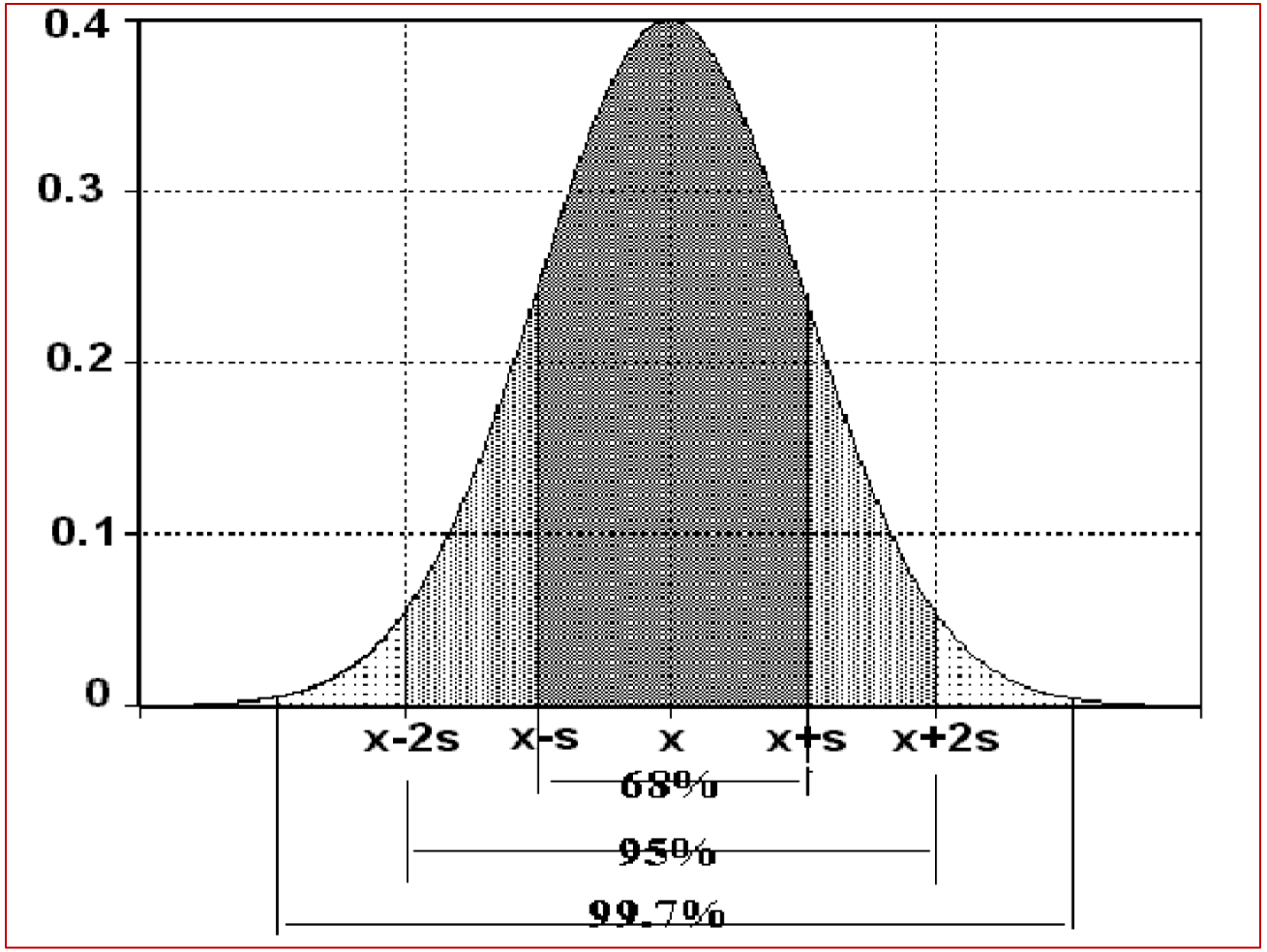
- 25 enfants de 10 ans
  - m = 137 cm (moyenne)
  - s =5cm (deviation standard)
  - 68% (d'enfants) ➔(137 -5, 137+5)
    (132cm, 142cm)

$$x_i \in (\bar{x} - s, \bar{x} + s);\ p = 68\%$$

$$x_i \in (\bar{x} - 2s, \bar{x} + 2s);\ p = 95\%$$

$$x_i \in (\bar{x} - 3s, \bar{x} + 3s);\ p = 99.7\%$$

Distribution normale de Gauss



Example

- Les enfants de 10 ans – Timisoara (population)
  - 25 enfants
  - m = 137 cm (moyenne)
  - s = 5 cm (déviation standard)

- Extraire toutes les échantillons possibles d'enfants de 10 ans

La moyenne de la population

- Pour une population de dimension *N*, on a les moyennes

$$\bar{x}_1 + \bar{x}_2 + \ldots\ldots \bar{x}_T$$

$$\mu = \frac{1}{T} \sum_{j=1}^{T} \bar{X}_j$$

- La distribution des moyennes des échantillons sont situé sur une courbe de Gauss (échantillons de taille > 30)

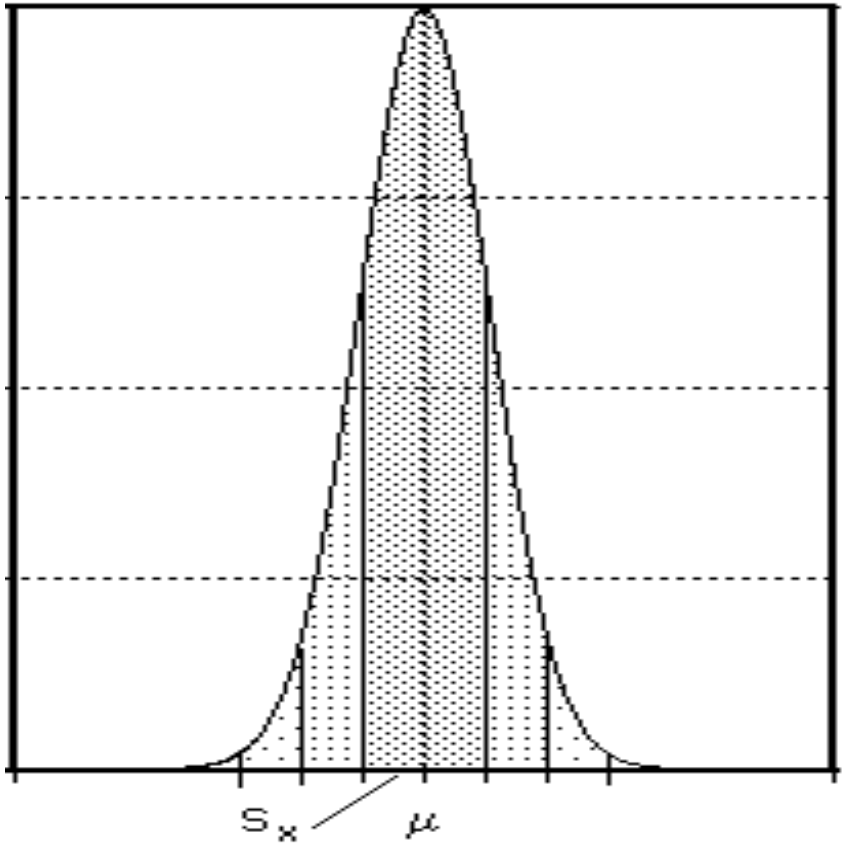
1. Les variations de moyennes (des échantillons) sont dans une intervalle plus aigue que les valeurs individuelles
2. La valeur autour de laquelle les variations sont symétriques est la moyenne de la population

UMFT
Universitatea de Medicină și Farmacie „Victor Babeș" din Timișoara

Distribution des moyennes des échantillons

- Caractérisée par :
- μ – moyenne de la population
- $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{N}}$ erreur standard de la moyenne

(N = volume de la population
σ = deviation standard)



Erreur standard de la moyenne

- Déviation standard de l'échantillon ***s***

  - Dispersion les valeurs individuelles tout autour de la moyenne de l' échantillon

- Erreur standard de la moyenne:

  - Dispersion de la moyenne de l'échantillons autour de la **moyenne de la population**

    - Echantillons grands $S_x = \frac{s}{\sqrt{n}}$

    - Echantillons petites $S_x = \frac{s}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$

Estimation statistique – distribution des milieux des échantillons

Echantillon

$$x_i \in (\bar{x} - s, \bar{x} + s); p = 68\%$$

$$x_i \in (\bar{x} - 2s, \bar{x} + 2s); p = 95\%$$

$$x_i \in (\bar{x} - 3s, \bar{x} + 3s); p = 99.7\%$$

p = **probabilité** pour un individu d'avoir la taille dans l'intervalle

Population

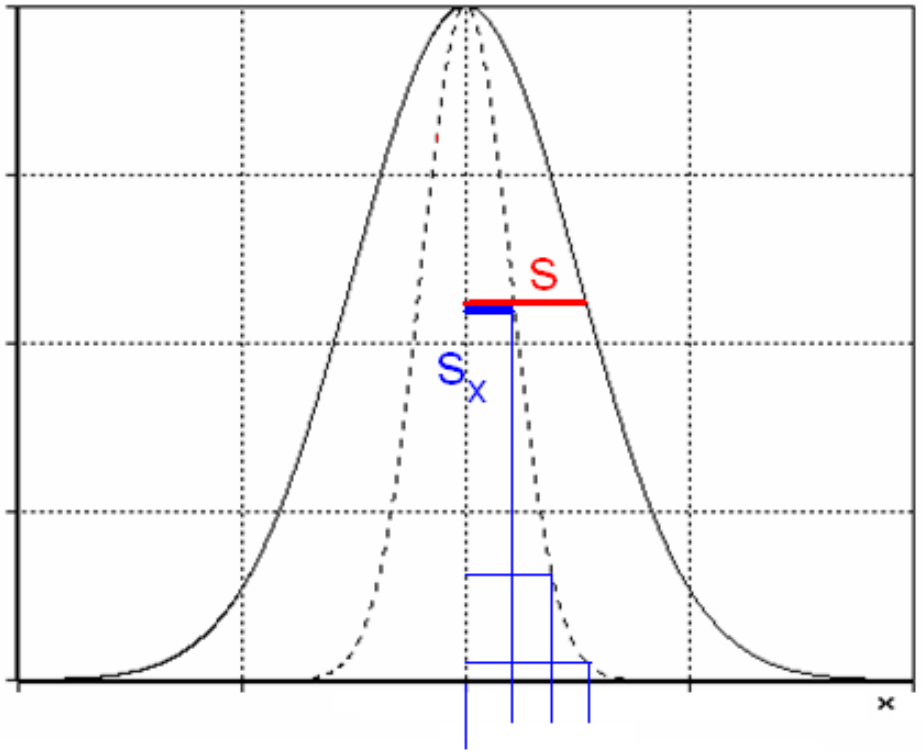
$$\mu \in (\bar{x} - s_x, \bar{x} + s_x); p = 68\%$$

$$\mu \in (\bar{x} - 2s_x, \bar{x} + 2s_x); p = 95\%$$

$$\mu \in (\bar{x} - 3s_x, \bar{x} + 3s_x); p = 99.7\%$$

p = **probabilité** pour la moyenne de la population d'etre dans l'intervalle

Estimation



S = déviation standard pour les valeurs individuelles

Sx = l'erreur standard de la moyenne (pour les moyennes des échantillons)

Inférence statistique

échantillon → population

- Les paramètres d’échantillon = déterminer
- Les paramètres de la population= estimer

- L’estimation d’intervalle de confiance

UMFT
Universitatea de Medicină și Farmacie „Victor Babeș" din Timișoara

L'intervalle de confiance

- L'estimation d' intervalle est exprimé par: (moyenne +/-la précision)
- La **qualité d' estimation** de l'intervalle dépend de la largeur de l'intervalle:
  - Intervalle étroite
    - la moyenne situé avec précision
    - la probabilité faible
  - Intervalle large
    - la confiance augmenté/grande ➔ la probabilité grande
    - la moyenne situé imprécise

Estimation de la moyenne pour des échantillons

- Échantillons grandes > 30
- Échantillons petits ≤ 30

Échantillons grandes (>30)

- Distribution de moyennes d'échantillon est normale
- Niveau de confiance = 95%
- Z = 1.96
- La précision de la localisation de la moyenne de la population (μ) dépend de S_x
- S_x dépend de la déviation standard et de la dimension de l'échantillon

$$\mu \in (\bar{x} - z^* s_x, \bar{x} + z^* s_x)$$

Example

- 144 individus
- m =4.84 pour VEMS (volume maximale expirateur dans 1 sec)
- s = 0.36
- Dans quel intervalle ont peut estimer de trouver la moyenne de la population avec un niveau de confiance de 98%?

- Erreur standard de la moyenne $S_x = \frac{S}{\sqrt{n}} = \frac{0.36}{\sqrt{144}} = 0.03$

- $1 - \alpha = 98\%$, Z=2.33
- $\mu \in (4.84 - 0.03\ x\ 2.33;\ 4.84 + 0.03\ x\ 2.33)$
- $\mu \in (4.84 - 0.07;\ 4.84 + 0.07)$
- $\mu \in (4.77; 4.91)$

Example

- Conclusion:

- On a une confiance de 98% que la vrai moyenne de VEMS pour les tous les individus est entre 4.77 et 4.91, ce qui signifie que la probabilité pour que la moyenne de VEMS sont a l'exterieur de cet intervalle est au dessous 2%

Distribution normale

	Valeurs						
Niveau de confiance **(1-α)**	0.68	0.90	0,95	0,954	0,98	0,99	0.997
Seuil de signification **α**	0.32	0.10	0.05	0.046	0.02	0.01	0.003
Z	1	1.65	1.96	2.00	2.33	2.58	3.00

Échantillons petites (≤ 30)

- N'avons pas une distribution normale
- Distribution **t**
- La distribution dépend du niveau de confiance (1-α) et de la dimension de l'échantillon (n)

- La distribution **t** dépend de la dimension de l' échantillon, étant caractérisée par un paramètre ϑ - le nombre de degré de liberté

$$\mu \in (x - t_{Z,\vartheta}\bar{s}_x, x + t_{Z,\vartheta}\bar{s}_x)$$

Example

- Echantillon n=16
- m=4.84
- s=0.36
- 1 - α = 98%
- υ =n-1=16-1=15
- $S_m = \frac{s}{\sqrt{n}} = \frac{0.36}{\sqrt{16}} = 0.09$
- $t_{\alpha/2,\vartheta} = 2.6$
- $\mu \in (4.84 - 0.09x2.6;\ 4.84 + 0.09x2.6)$
- $\mu \in (4.84 - 0.23;\ 4.84 + 0.23)$
- $\mu \in (4.6; 5.08)$

Estimation des proportions

- **Proportion de la classe:** $p_i = \frac{N_i}{N} * 100$

- L' écart type (la déviation standard) de la proportion:

$$s_p = \sqrt{p_i q_i / N} \quad q_i = 1 - p_i$$

- p_i = probabilité appartenir a la classe
- q_i = probabilité de n'appartenir pas a la classe

$$\hat{p} \in \left(p - z_{\alpha/2}.s_p ; p + z_{\alpha/2}.s_p\right)$$

Example

- Echantillon 80 individus
- 21 ont la groupe sanguine A
- Quelle est la proportion de la groupe sanguine Adans la population étudiée avec un niveau de confidence 95%?

- $\text{p}=\frac{24}{80}x100 = 30\%$
- $S_p = \sqrt{\frac{0.3x0.7}{80}} = 0.0512 \cong 5.12\%$
- $p \in (30 - 1.96\ x\ 5.12; 30 + 1.96\ x\ 5.12)$
- $p \in (30 - 10; 30 + 10)$
- $p \in (20\%, 30\%)$

Estimation des différences

A. Entre les moyennes

$$S_d = \sqrt{\frac{S_a^2}{n_A} + \frac{S_b^2}{n_B}}$$ la déviation standard

$$\hat{d}_x \in \left(\bar{d} - z_{\alpha/2}.s_d ; \bar{d} + z_{\alpha/2}.s_d \right)$$

$$d_{\bar{X}} = \bar{X}_a - \bar{X}_b$$

B. Entre les pourcents

$$d_p = p_2 - p_1$$

$$Sp = \sqrt{\frac{p_1(1-p_1)}{n_1 - 1} + \frac{p_2(1-p_2)}{n_2 - 1}}$$

$$\hat{d}_p \in \left(\bar{d} - z_{\alpha/2}.s_{pd} ; \bar{d} + z_{\alpha/2}.s_{pd} \right)$$