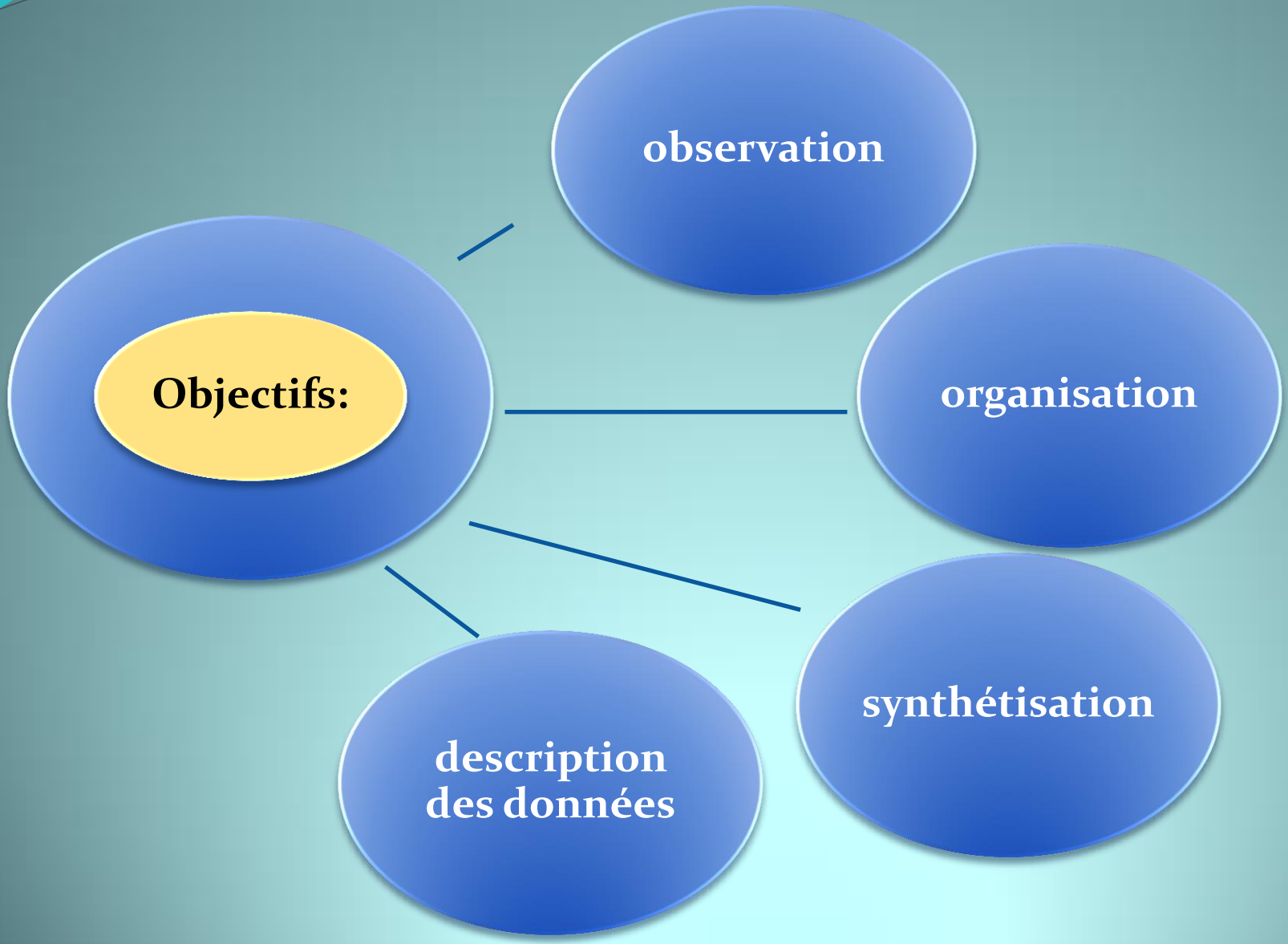
Statistique descriptive

Indicateurs synthétiques de distributions statistiques

STATISTIQUE DESCRIPTIVE



1. **Population**
2. **Echantillon**
3. **Caractéristique observée**

Tableau d'enregistrement primaire

Distribution de fréquence

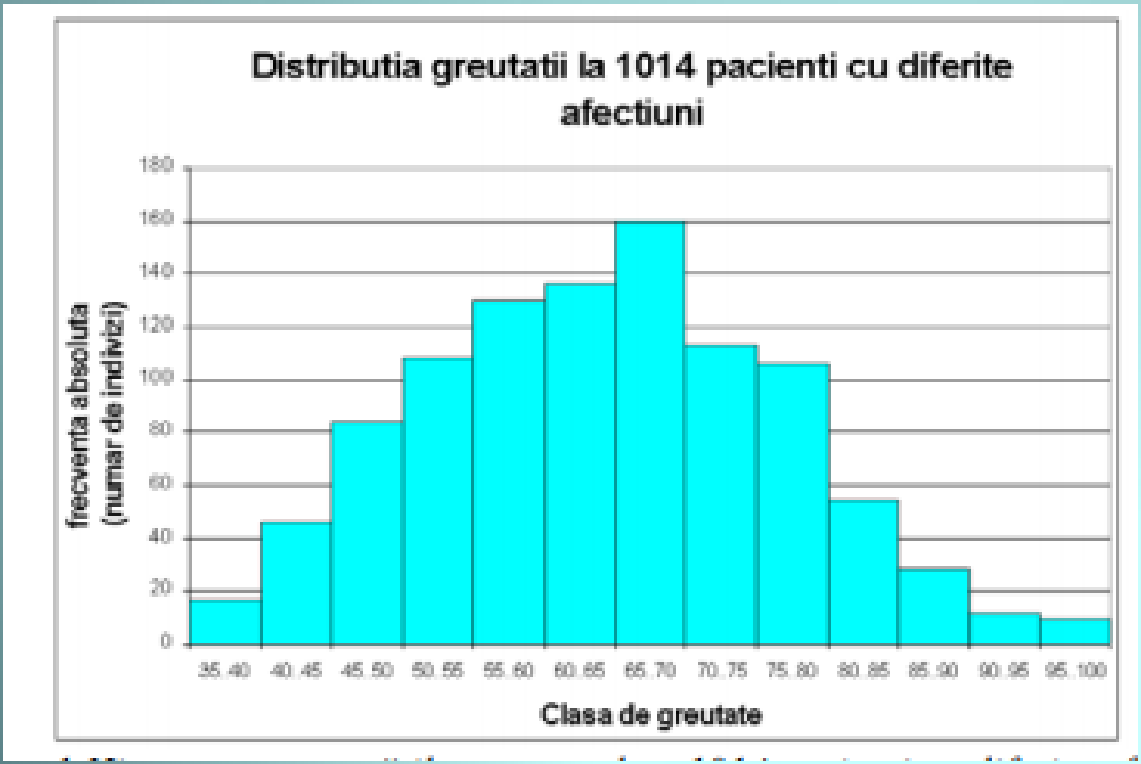
Histogramme!!!!!

Indicateurs synthétiques

Que dit l'histogramme?

C'est un graphique qui donne des informations sur la distribution des valeurs dans une série de valeurs

Poids corporel chez 1014 patients atteints de maladies différentes: sur des classes de 5 kg en 5 kg (http://www.umfcv.ro/files/b/i/Biostatistica%20MG%20-%20Cursul%20IV.pdf)



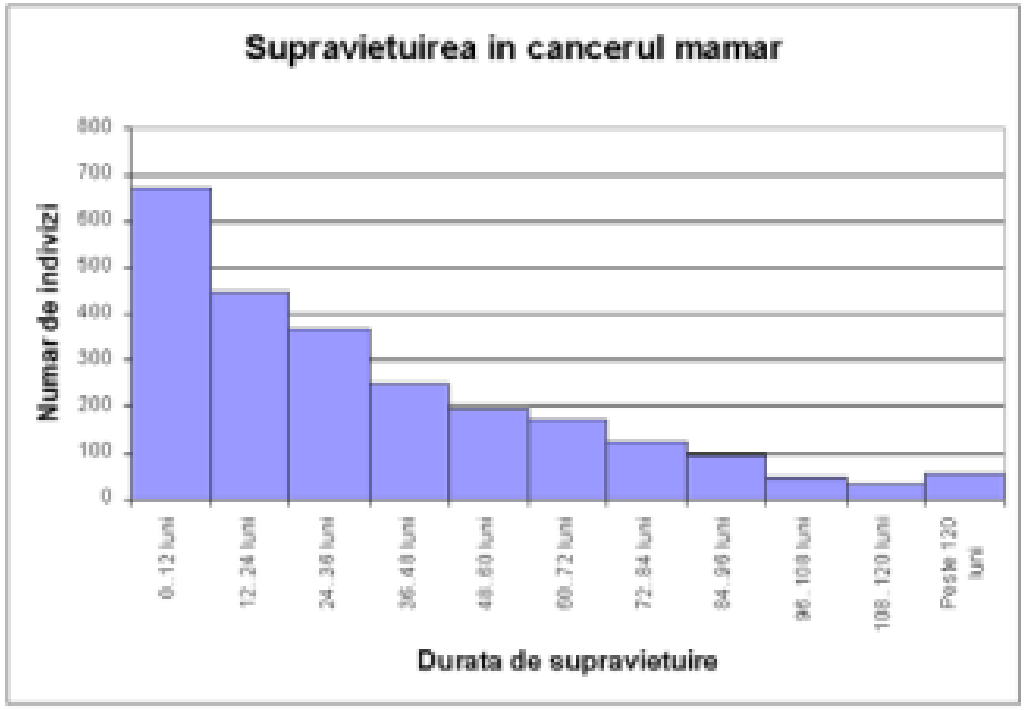
La façon dont les barres grandissent est différente de la façon dont elles diminuent.

- **asymétrie - asymétrie droite.**

Si les individus auprès desquels les données ont été recueillies étaient normaux, l'histogramme aurait une apparence plus symétrique.

Tabelul 2 Situaţia supravieţuirilor în cazurile de cancer mamar. Gruparea în clase de 12 luni

Nr.Crt	Perioada	Nr.cazuri	Procent %	Procent Cumulat %
1	0..12 luni	672	27.36	27.36
2	12..24 luni	446	18.16	45.52
3	24..36 luni	368	15.00	60.52
4	36..48 luni	249	10.14	70.66
5	48..60 luni	196	8.00	78.66
6	60..72 luni	172	7.00	85.66
7	72..84 luni	126	5.13	90.79
8	84..96 luni	98	4.00	94.79
9	96..108 luni	45	1.83	96.62
10	108..120 luni	31	1.26	97.88
11	Peste 120 luni	52	2.12	100.00



En règle générale, il est bon de noter que:

• Plus on perd d'informations, plus le nombre de classes est faible. Les histogrammes de 2 à 4 classes ne sont pas recommandés.

• Un trop grand nombre de classes conduit à dissimuler l'essence par des aspects insignifiants.

Il est recommandé:

- **Pour quelques dizaines de valeurs, choisissez un maximum de 6 à 8 classes**
- **Pour quelques centaines de valeurs, choisissez entre 10 et 15 classes**
- **Pour quelques milliers de valeurs, choisissez plus de 15 classes**

Un histogramme est l'information d'une série de valeurs avec perte d'information.

Plus on perd d'informations, moins il y a de classes.

Plus on perd d'informations, plus les classes sont longues.

Non recommandé

- Utiliser plus de 20 à 30 classes que dans des cas particuliers dans des études portant sur plusieurs milliers de cas.
- Utilisez moins de 4-6 classes.
- Utilisez des histogrammes si nous n’avons pas au moins quelques dizaines de valeurs.

Par exemple, pour une série de 15 valeurs, aucun histogramme n'est créé

INDICATEURS STATISTIQUES
Statistiques descriptives - objectifs

- **Comment sont présentées les valeurs d'une distribution?**
  - À quel point sont-ils proches l'un de l'autre?
  - En quoi sont-ils différents les uns des autres?
- **Y a-t-il des valeurs qui représentent la distribution entière?**

Que sont-ils?

Les indicateurs synthétiques sont des descripteurs numériques qui condensent en une valeur unique une certaine caractéristique de toute une distribution de valeurs.

Catégories d'indicateurs

1. Indicateurs de la tendance centrale

- valeurs représentatives typiques qui décrivent la distribution dans son intégralité

2. Indicateurs de diffusion

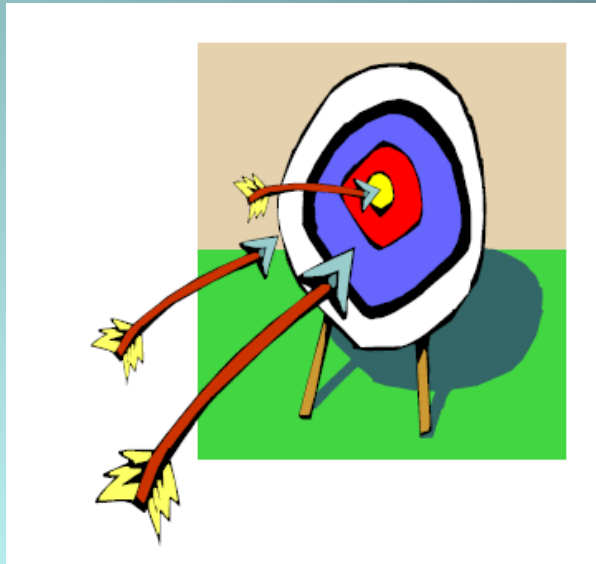
- décrivent la fonction de diffusion de la valeur de distribution

3. Indicateurs de la forme de distribution

- se réfère à la forme de la représentation graphique de la distribution

Indicateurs de tendance centrale

1. valeur moyenne;
2. valeur médiane;
3. valeur dominante (module);
4. Quartiles.



Indicateurs de tendance centrale - MEDIA

La moyenne est à la fois la mesure la plus importante et la plus populaire de la tendance centrale d'une distribution.

La moyenne de sondage ***(Sample Mean)*** **est un indicateur qui caractérise un échantillon (une population) du point de vue d'une fonction étudiée.**

La moyenne de la population ***(Population Mean)*** **est la moyenne des nombres dans une population numérique.**

Cette valeur est un paramètre de la population, par opposition à la moyenne calculée à partir d'un échantillon, qui n'est qu'une estimation du paramètre.

Moyenne arithmétique

Moyenne arithmétique pondérée

Médias géométriques

Moyenne arithmétique (m)

Elle est calculée en faisant la somme de toutes les valeurs observées de la série de données, divisée par le nombre d'observations.

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Notes communes:

- μ (*miu*), **lorsqu'elle est la moyenne de toute la population de référence**
- $(\bar{\chi}$ *barat)* *ou m*, **lors du calcul d'un échantillon (le plus fréquent)**

Moyenne arithmétique

Exemple: Pour distribution: 5,8,3,2,5,4

$$m = \frac{\sum X}{N} = \frac{5+8+3+2+5+4}{6} = \frac{26}{6} = 4{,}50$$

Exemple: Pour distribution: 5,8,3,3,3,2,4,2,3,5,4

2	2
3	4
4	2
5	2
8	1

$$m = \frac{2*2+3*4+4*2+5*2+8*1}{2+4+2+2+1} = \frac{42}{11} = 3{,}81$$

Détermination de la moyenne pour les données groupées

Intervalle (i)	Centre i (x)	Fréquence (fi)	x * f
3-5	4	2	8
6-8	7	3	21
9-11	10	5	50
12-14	13	5	65
15-17	16	9	144
18-20	19	10	190
21-23	22	6	132
24-26	25	4	100
27-29	28	2	56
30-32	31	1	31

$$m = \frac{\sum x_i \cdot f_i}{\sum f_i} = \frac{31 \cdot 1 + 28 \cdot 2 + \cdots + 4 \cdot 2}{1 + 2 + 4 + \cdots 2} = \frac{797}{47} = 16{,}96$$

Propriétés de la moyenne arithmétique

Ajouter / supprimer une constante à chaque valeur de distribution augmente / diminue la moyenne de cette valeur

Multiplier / diviser chaque valeur de distribution avec une constante, multiplie/divise le média par cette constante

La somme de l'écart par rapport à la moyenne est toujours zéro

La somme du carré des écarts par rapport à la moyenne sera toujours inférieur à la somme des carrés des écarts par rapport à tout autre point de distribution

Propriétés de la moyenne

variabila	constantă	+	*	abateri medie
5	2	7	10	0,50
8	2	10	16	3,50
3	2	5	6	-1,50
2	2	4	4	-2,50
5	2	7	10	0,50
4	2	6	8	-0,50
m=4.5		**m=6.5**	**m=9**	**suma=0 media=?**

Moyenne arithmétique pondérée

Exemple: pour la distribution: 5,8,3,3,3,2,4,2,3,5,4

2	2
3	4
4	2
5	2
8	1

$$m = \frac{\sum (X * f)}{\sum f} = \frac{5*2+8*1+3*4+2*2+4*2}{2+1+4+2+2} = \frac{43}{11} = 3{,}90$$

Moyenne géométrique

Il est utilisé dans le cas de distributions de fréquence représentant un caractère avec un taux de croissance uniforme (tel que celui de la division cellulaire), ou pour rechercher des valeurs intermédiaires qui entraînent un rythme plus géométrique (donc en se multipliant) que arithmétique (donc en s'ajoutant)

$$x_{geom} = \sqrt[n]{x_1 * x_2 * x_3 * \ldots * x_n} = \sqrt[n]{\prod x_i}$$

Moyenne géométrique

Ex1: Après une expérience, 10 cas positifs ont été trouvés le premier jour et 1 000 cas positifs le troisième jour. Quel est la moyenne?

$$\bar{x} = \frac{10 + 1000}{2} = 1010 : 2 = 505$$

$$x_{geom} = \sqrt[2]{10 * 1000} = \sqrt[2]{10000} = 100$$

MODULE (Mo) ou valeur dominante

Définition:

Le MODULE ou la valeur DOMINANTE est la valeur ou la classe de la plage de la caractéristique ayant la fréquence d'occurrence la plus élevée.

Elle est obtenue en faisant le tableau de fréquences (simple ou en cluster) et elle est la valeur à laquelle correspond la fréquence absolue la plus élevée.

- Distributions unimodales *(**5**8325**5**4 Mo=5)*
- Distributions bimodales *(**5**832**225**4 Mo=5; =2)*
- Distributions multimodales*(**588**3**225**4 Mo=5; =2; =8)*

Exemple:

Dans la série de valeurs 5,8,3,2,5,4, Mo=5 (apparait la plupart du temps)

x	n
2	1
3	1
4	1
5	2
8	1

Pour les données en cluster, la plage ayant la fréquence la plus élevée est recherchée.

Intervalle	Fréquence	Intervalle	Fréquence
3-5 (4)	2	18-20 (19)	10
6-8 (7)	3	21-23 (22)	6
9-11 (10)	5	24-26 (25)	4
12-14 (11)	5	27-29 (28)	2
15-17 (16)	9	30-32 (31)	1

Dans notre cas, cette plage est comprise entre 18 et 20 dans laquelle 10 valeurs sont présentes.

La valeur modale est égale à la valeur trouvée au centre de cette plage, dans ce cas Mo = 19.

Caractéristiques du module:

- **considérer uniquement les mesures les plus représentatives;**
- **nécessite d'ordonner les données**
- **correspond à un ou plusieurs éléments de la série (en cas d'égale fréquence).**

Médiane (Me)

- La médiane d'une série statistique ordonnée est la valeur qui divise la chaîne de valeurs ordonnée de la variable en deux parties, chaque partie contenant le même nombre de valeurs. Elle est notée avec Me et a

  - 50% de ses valeurs au dessus

  -et 50% de ses valeurs en dessous

- Si le nombre d'observations est impair, -Me est la valeur du milieu suivant leur ordination.
- Si le nombre d'observations est pair, -Me est calculée en tant que moyenne arithmétique des valeurs situées au milieu de la série statistique ordonnée.

5,8,3,2,5,4, →2,3,4,5,5,8→**Me=4,5**

Série statistique **Série statistique ordonnée**

Comment est-ce déterminé?

1. La série est ordonnée en ordre ascendant.
   La valeur moyenne est déterminée.
2. Pour les distributions avec un nombre impair de valeurs, Me est la valeur respective.
3. Dans le cas des distributions paires, Me est calculé comme la moyenne des deux valeurs situées au milieu de la distribution

Médiane

une valeur médiane n'existe réellement que si le nombre n est impair, alors qu'il existe en fait un individu moyen (le [n + 1] / 2 ème) dont la valeur est médiane.

Si n est pair, nous prenons les individus de rang n / 2 et n / 2 + 1

EXEMPLE:

Dans la série de valeurs 5,8,3,2,5,4, en ordre croissant (2,3,4,5,5,8), Me = 4,5 (valeur moyenne des valeurs 4 et 5 au milieu d'une distribution paire.

Si la distribution avait 5 valeurs (en excluant 2, par exemple), Me = 5

Modul, Mediana și Media vizează același lucru, tendința centrală…

…și totuși…

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
60	45	22	20	16	16	16	15	14	14	14	12	12	12	12

N=15

ΣX=300

media=20
ΣX/N=300/15

mediana=15
(N+1)/2=8

modul=12
(valoarea cea mai frecventă)

Quartiles

- **Le quartile est la valeur pour laquelle nous avons un quart des valeurs des séries inférieures et supérieures.**
- **La médiane est un indicateur de la tendance centrale, il s'agit de la valeur moyenne d'une série de valeurs.**

Définition:

Le quartile Q1 est cette valeur dans une série de valeurs pour laquelle 25% des valeurs de la série sont inférieures à Q1 et 75% lui sont supérieures.

Définition:

Le quartile Q3 est cette valeur dans une série de valeurs, pour laquelle 75% des valeurs de la série sont inférieures à Q3 et 25% supérieures.

INDICATEURS DE LA TENDENCE CENTRALE

étroitement liée au niveau de mesure des variables

Avantages

- **Module**
  - Facile à calculer (insignifiant pour le moment)
  - Peut être utilisé pour n'importe quelle échelle
  - C'est le seul indicateur pour les échelles nominales
- **Médiane**
- Peut être utilisé sur l'échelle ordinale et d'intervalle / rapport
- **Moyenne**
  - Reflète les valeurs de la distribution entière
  - Elle ne peut être calculée que pour des variables mesurables sur des échelles intervalle et rapport

Inconvénients

Module

- Généralement incertain, en particulier dans les petits échantillons, quand il peut changer radicalement à un changement mineur de valeur;

Médiane

- Elle peut ne pas correspondre à une valeur réelle (N pair);
- Elle est moins sûre en extrapolant l'échantillon à la population;

Moyenne

- Cela ne correspond généralement pas à une valeur réelle;
- Cela conduit à des interprétations erronées sur les distributions asymétriques
- Peut être sévèrement affectée par des scores extrêmes;

Indicateurs de la tendance centrale. (Résumé)

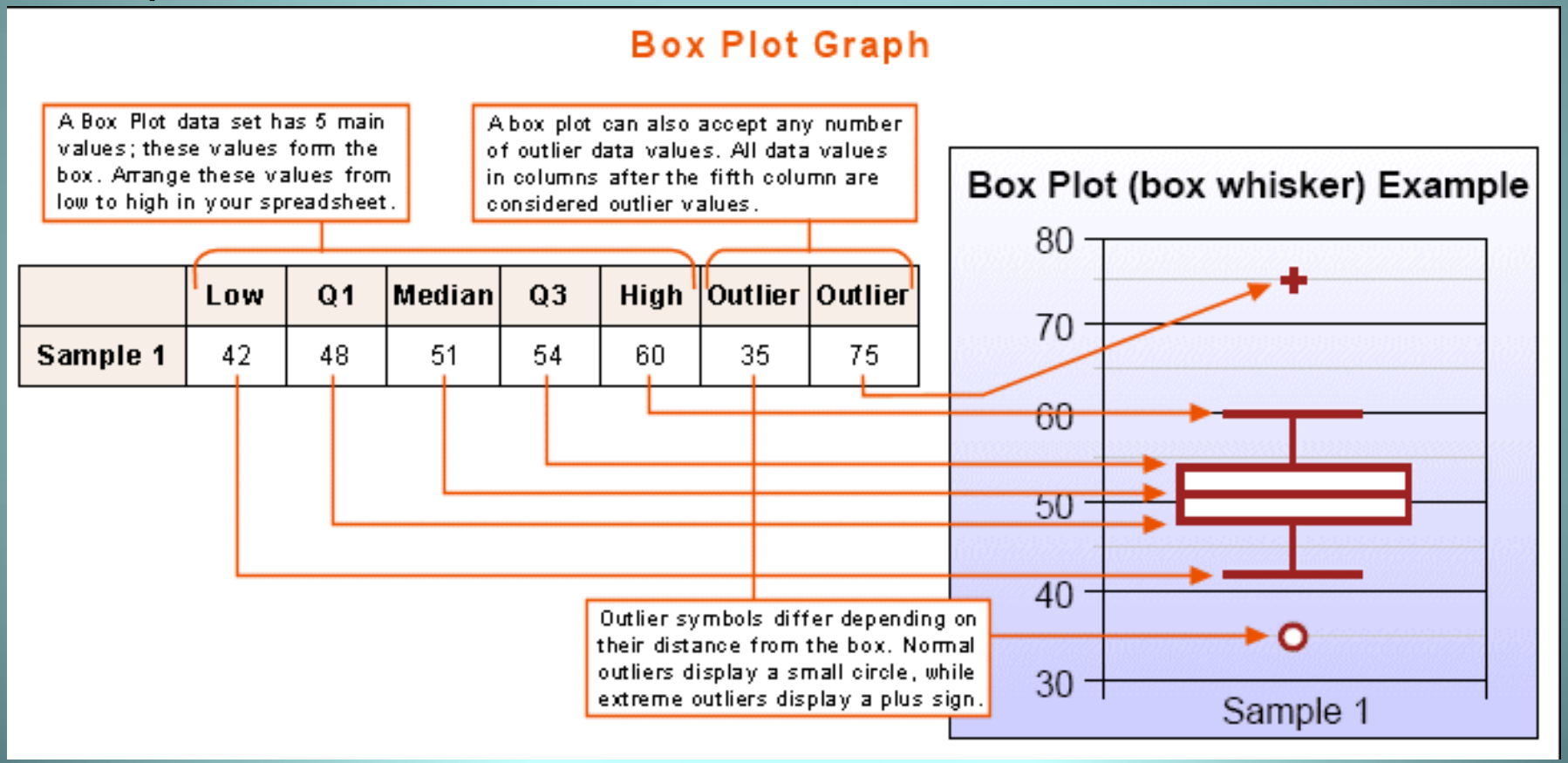
- Les indicateurs les plus importants de la tendance centrale sont la moyenne, la médiane et le module.
- La moyenne indique la tendance centrale lorsque la série de valeurs est répartie symétriquement autour de celle-ci et lorsque les valeurs ne présentent pas une dispersion trop grande.
- Dans le cas de valeurs distribuées très asymétriquement, la tendance centrale n'est plus indiquée par la moyenne, mais par la médiane.
- Le module est un indicateur de la tendance centrale de la série unimodale, c'est-à-dire lorsqu'il n'y a qu'un seul maximum dans le tableau de fréquences. Si nous avons une série multimodale, il perd sa qualité en tant qu'indicateur de la tendance centrale.

Graphiques en boîte à moustache (diagramme en boîte)

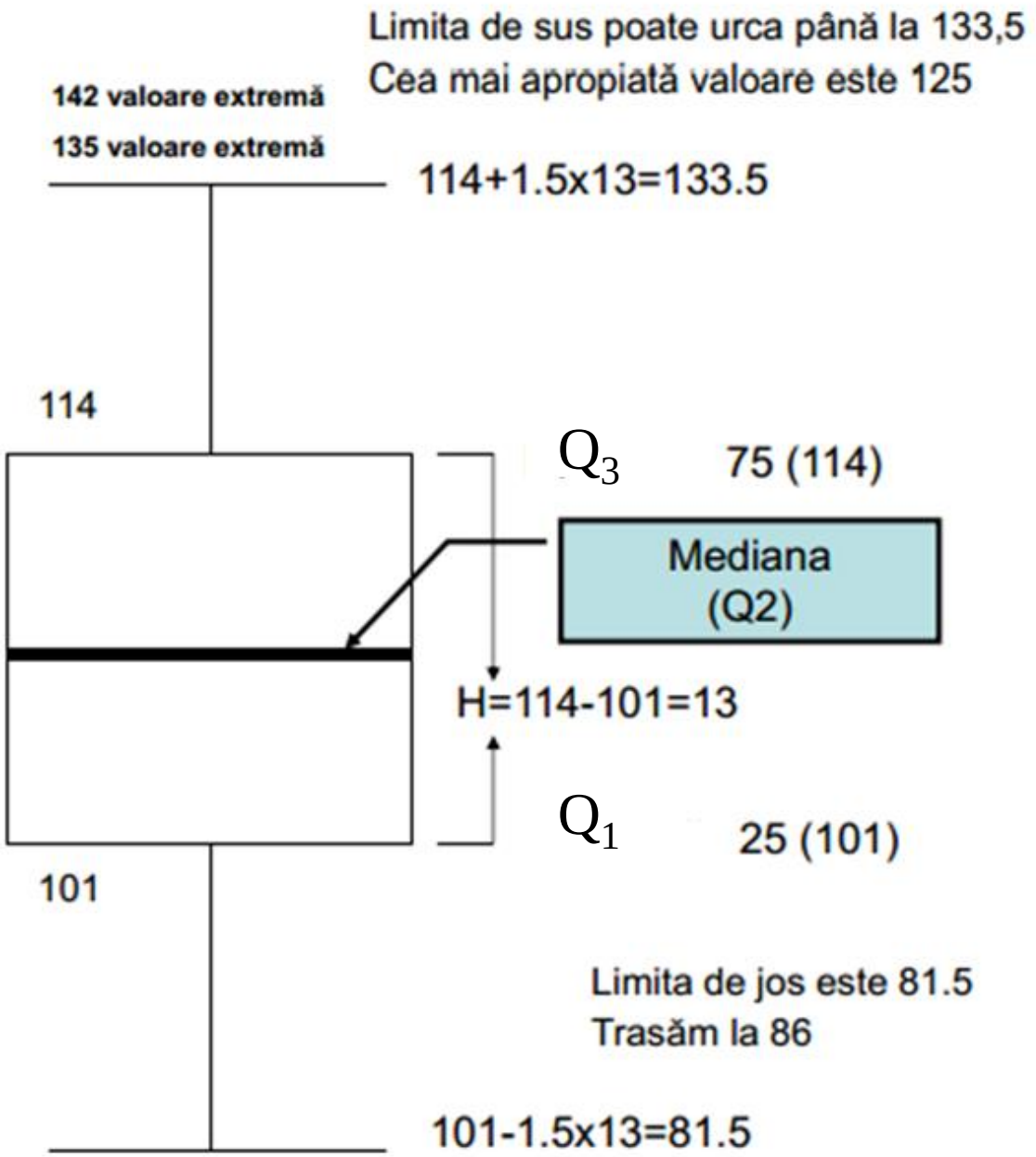
Graphiques en boîte à moustache (diagramme en boîte)

- sont assez communs dans les statistiques et les mesures de qualité.
- ont cinq valeurs principales: minimum, Q1, Mediane, Q3 et Maximum

Exemple: 35, 42, 48, 50, 51, 53, 54, 60, 75



Valori OI	fa	fr%	frc% (rang percentil)
142	1	1,9	100,0
135	1	1,9	98,1
125	1	1,9	96,2
124	1	1,9	94,2
123	1	1,9	92,3
121	1	1,9	90,4
118	2	3,8	88,5
117	1	1,9	84,6
116	2	3,8	82,7
115	2	3,8	78,8
114	3	5,8	75,0
113	2	3,8	69,2
112	1	1,9	65,4
110	1	1,9	63,5
109	4	7,7	61,5
108	3	5,8	53,8
107	3	5,8	48,1
106	2	3,8	42,3
105	1	1,9	38,5
104	1	1,9	36,5
102	3	5,8	34,6
101	4	7,7	28,8
98	1	1,9	21,2
97	2	3,8	19,2
96	1	1,9	15,4
94	1	1,9	13,5
92	2	3,8	11,5
91	1	1,9	7,7
89	1	1,9	5,8
87	1	1,9	3,8
86	1	1,9	1,9
	N=52	100,0	



Population size: 52
Median: 108
Minimum: 86
Maximum: 142
First quartile: 101
Third quartile: 114.75
Interquartile Range: 13.75
Outliers: 86 87 142 135

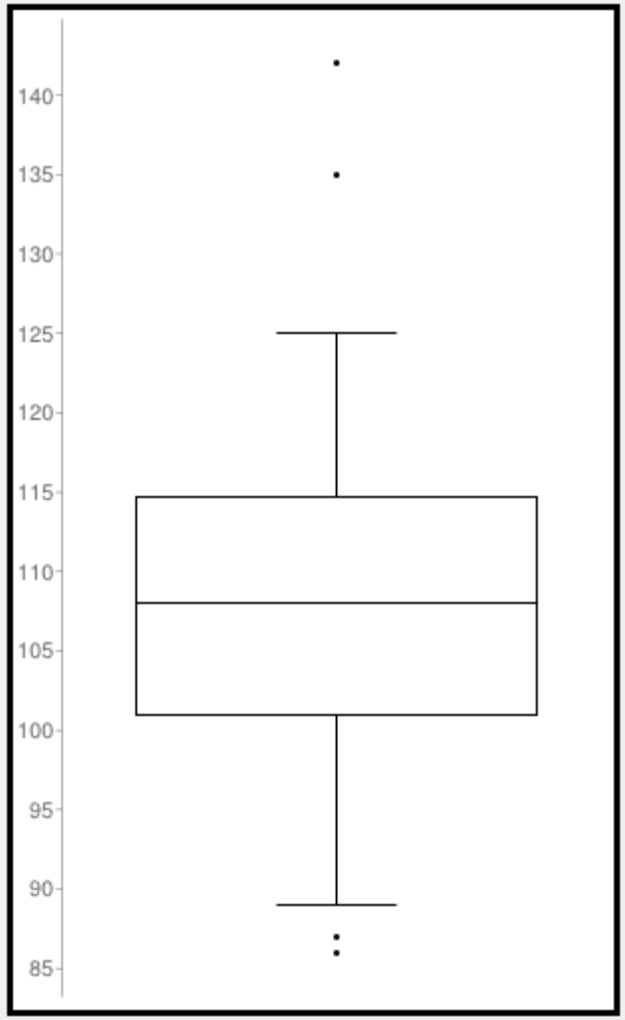
Arithmetic mean (μ): 107.8076923076

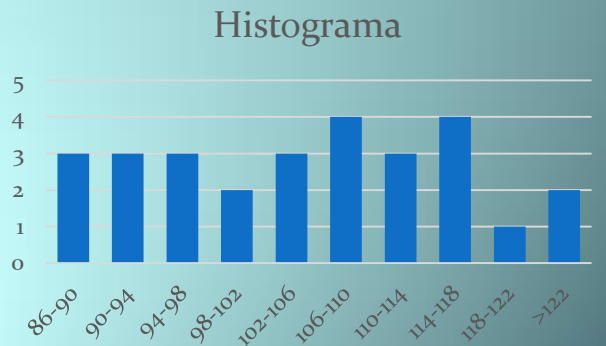
Median: 108

Modes: 101 109

86,87,89,91,92,92,94,96,97,97,98,101,
101,101,101,102,102,102,104,105,106,
106,107,107,107,108,108,108,109,109,
109,109,110,112,113,113,114,114,114,
115,115,116,116,117,118,118,121,123,
124,125,135,142

124	1
123	1
121	1
118	2
117	1
116	2
115	2
114	3
113	2
112	1
110	1
109	4
108	3
107	3
106	2
105	1
104	1
102	3
101	4
98	1
97	2
96	1
94	1
92	2
91	1
89	1
87	1
86	1





http://www.alcula.com/calculators/statistics/box-plot/

Traiter les valeurs extrêmes ou aberrantes (outlier)

- Déterminer la nature des valeurs extrêmes:
  - erreurs d'enregistrement (dactylographie);
  - erreurs de mesure;
  - résultats influencés par des anomalies des conditions expérimentales.
  - l'échantillon a été extrait d'une population asymétrique
  - ces valeurs font parties d'une autre population de valeurs
  - échantillon trop petit
- En les traitant de l'une des manières possibles:
  - suppression (s'il y a des erreurs irrécupérables);
  - correction (si possible)