LP05 – PRESENTATION DES DONNEES STATISTIQUES (1)

Objectifs:

I. **Présentation des données par des tableaux**

- **Réalisation du tableau d' évidence primaire**
  Ce tableau contient des valeurs d' observation distinctes x_i du caractère étudié.
- **Réalisation du tableau ou de la distribution de fréquence**
  La distribution de fréquences est une liste des valeurs (catégories) possibles d'une variable, accompagnées par le nombre d' observations qui prennent les valeurs respectives (qui se trouvent dans chacune de catégories respectives).
- **Calcul:**
  Fréquence absolue (fa)
  Fréquence cumulée (fc)
  Fréquence relative (fr)
  Fréquence relative cumulée frc
  Fréquence relative en pourcentage fr%
  Fréquence relative en pourcentage cumulée fr%c

II. **Présentation graphique des données**

- Construire un graphique
- Lire et interpréter un graphique

1. **Graphique barre**
   On l'utilise au moment où on désire représenter une variable „discrète" (qui présente des valeurs entières).
2. **Graphique galette (PIE)**
   Îl est utilisé dans les situations où les valeurs sont „partie d'un tout"
   Les graphiques barre et galette sont utilisés pour montrer les dimensions relatives de données.
3. **Graphique ligne**
   Un graphique ligne est utilisé pour représenter graphiquement des variables quantitatives continues.
   Ce type de graphique est créé par la connexion d'une série de points unis par une ligne qui montre les changements dans le temps.
   A ces types de graphiques, l'axe Ox représente l'axe du temps, et l'axe Oy représente la caractéristique étudiée
4. **Histogramme**
   C'est la représentation graphique de la distribution de fréquence. Les deux axes représentent : l'axe Ox – la caractéristique étudiée, l'axe Oy la fréquence d' apparition de la caractéristique étudiée. Il est semblable au graphique barre. Les données observées sont groupées et ordonnées de manière croissante.
5. **Le nuage de points (scater plot)**
   Ce type de graphique montre la relation entre deux sets de données.

Problème:

Un médecin a réalisé une étude pour identifier le risque cardiaque. A cause de ressources limitées on a fait recours au choix aléatoire d'un échantillon de 30 patients.

Les données suivantes ont été recueillies chez chaque patient : sexe, consommation affirmative d' alcool (oui/non), habitude de fumer (oui/non, affirmatif), âge (ans, fiche du patient).

Les mesurages suivants ont été réalisés pour chaque patient : poids (kg), taille (cm), TAS (tension artérielle systolique, mmHg), TAD (tension artérielle diastolique, mmHg). Les déterminations biochimiques ont été réalisées après une prise de sang : glycémie (mg/dl) et cholestérol (mg/dl).

Les données ont été collectées et sont présentées dans le tableau ci-dessous :

Sexe	Alcool	Fumer	Age (ans)	Poids (kg)	Taille (cm)	TAS (mmHg)	TAD (mmHg)	Glycémie(mg/dL)	Cholestérol (mg/dL)
F	non	non	27	53	162	110	80	75	168
M	non	non	41	106	176	130	70	92	343
M	non	non	67	91	170	170	100	77	229
F	non	non	60	107	168	190	120	128	157
M	non	non	26	84	174	130	90	81	161
M	non	non	46	119	182	170	90	138	192
M	non	oui	34	86	180	110	70	88	218
M	non	non	31	80	178	110	70	72	159
M	oui	oui	45	82	179	100	70	71	272
M	non	non	35	100	172	130	90	80	195
F	non	non	64	74	162	130	100	91	220
F	non	non	64	78	154	170	100	94	246
F	non	non	34	55	152	110	80	90	147
M	non	oui	35	57	173	110	70	90	157
M	oui	non	41	89	172	120	90	96	175
F	non	non	49	95	163	130	100	83	257
F	non	non	64	78	154	160	110	88	223
M	non	non	43	79	180	120	80	92	184
M	non	non	58	96	178	145	85	79	245
M	non	non	44	64	155	150	100	75	242
F	non	non	45	51	152	120	80	92	162
M	non	oui	25	71	177	130	80	96	215
F	oui	non	62	76	158	130	80	88	293
F	non	oui	39	81	158	100	70	72	197
M	non	non	26	75	176	120	80	77	219
F	non	non	41	70	162	120	80	86	225
M	non	non	30	86	173	120	80	74	178
F	non	non	49	76	165	130	80	80	152
M	non	oui	24	88	178	120	70	90	154
M	non	oui	27	88	182	130	90	88	216

Exigences:

1. Copiez le tableau dans un fichier Microsoft Excel.
2. Réalisez les tableaux de fréquence pour les variables TAS et TAD.
3. Complétez les tableaux avec les fréquences observées (f_a) avec 5 (cinq) colonnes de plus où on calcule :

   Fréquence cumulée (fc)
   Fréquence relative (fr)
   Fréquence relative cumulée frc
   Fréquence relative en pourcentage fr%

Fréquence relative en pourcentage cumulée fr%c

4. Représentez graphiquement pour les variables TAS et TAD:
   4.1. Sous la forme de colonnes (histogramme) les fréquences absolues
   4.2. Graphique ligne fréquences cumulées
   4.3. Graphique pie (galette) fréquences relatives
5. Réalisez les tableaux de fréquence pour les variables qualitatives (Sexe, Alcool, Habitude de fumer) utilisant Tableau croise dynamique.
6. Réalisez les histogrammes pour ces variables utilisant Pivot Chart.

Indications :

1. Copiez le tableau dans un fichier Microsoft Excel. [Copy - Paste] [File – Save As]

On considère que c'est le **tableau d' évidence primaire**.
Dans les études biostatistiques, le tableau est le principal mode de présentation des données statistiques. C'est pourquoi les tableaux sont construits de telle manière qu'ils permettent la réalisation d'une analyse correcte.
A la réalisation des tableaux on tiendra compte de :

- le tableau doit porter un titre qui soit concis et lié au sujet ;
- les lignes et les colonnes qui indiquent la nature des données soient étiquetées de manière simple et précise;
- les unités de mesure des données sont incluses;
- les sources d' information sont précisées;
- il est préférable que des lignes ou colonnes contenant des moyennes ou totaux existent;
- le formatage des tableaux doit être suggestif.

2. Réalisation des tableaux de fréquence

Les tableaux de fréquence sont des tableaux à 2 colonnes (x_i n_i) où x_i est le caractère recherché et n_i sa fréquence d' apparition ou la fréquence absolue **(f_a)**.
La fréquence absolue d'une valeur x d'une série statistique S est le nombre de répétition de la valeur x dans la série S. Donc, le total des fréquences absolues de toutes les valeurs distinctes d'une série statistique est égal à la taille ou au volume de la série.
Les tableaux de fréquence peuvent être obtenus par plusieurs méthodes:

1. Par numération effective
2. En Excel par numération à l'aide de la fonction NB ou NB.SI
3. En Excel utilisant la fonction FREQUENCE
4. En Excel utilisant l'option HISTOGRAMME de Data Analisys

Dans le travail pratique LP05 on utilise la méthode 1 et 2.

N	O
TAS (mmHg)	
100	=COUNTIF($L$2:$L$31;N2)
110	=COUNTIF($L$2:$L$31;N3)
120	=COUNTIF($L$2:$L$31;N4)
130	=COUNTIF($L$2:$L$31;N5)
145	=COUNTIF($L$2:$L$31;N6)
150	=COUNTIF($L$2:$L$31;N7)
160	=COUNTIF($L$2:$L$31;N8)
170	=COUNTIF($L$2:$L$31;N9)
190	=COUNTIF($L$2:$L$31;N10)

Par la méthode 1, pour élaborer ces tableaux on procède de façon suivante :

- La colonne qui contient la variable est copiée du tableau de la feuille de calcul dans une autre colonne;
- On ordonne de manière croissante;
- On élabore le tableau de fréquence (x_i n_i), chaque valeur distincte x_{i1} (voir footnote) est écrite dans une cellule de

[1] Valeurs distinctes x_i peuvent être obtenues utilisant l'option Remove duplicates du menu Data

la feuille de calcul et devant elle, dans la cellule voisine, sa fréquence d' apparition. Par exemple, pour la variable TAS, la valeur 100 apparait 2 fois, la valeur 110 5 fois etc.

Par la méthode 2 on se propose que Excel compte (COUNT) le nombre de valeurs sélectées ou (COUNTIF) le nombre de valeurs égales à la valeur introduite à *Criteria* qui se trouvent dans le domaine sélecté à *Range*.

La syntaxe de la fonction COUNT

La fonction NB a les arguments :

- ***valeur1*** *Obligatoire. Le premier élément, référence de cellule ou la zone dont on désire compter les nombres.*
- ***Valeur 2, ...*** *Optionnel. Jusqu'à 255 éléments supplémentaires, références de cellules ou zones ou on désire compter les nombres.*

Par exemple, on peut introduire la formule pour compter les nombres de la zone A1:A20: ***=NB(A1:A20)****.*

Dans cet exemple, si cinq des cellules de la zone contiennent des nombres, le résultat est 5.

La syntaxe de la fonction NB.SI (zone, critères)

zone *(obligatoire)*

Le groupe de cellules qu'on désire compter. Zone peut contenir des nombres, matrices, une zone nommée ou des références qui contiennent des nombres. Les valeurs texte et non complétées sont ignorées.

critères *(obligatoires)*

Un nombre, une expression, une référence de cellule ou une ligne de texte qui détermine les cellules qui seront comptées.

Par exemple:

- *=NB.S (A2:A5;"pommes") compte le nombre de cellules de la zone A2:A5 qui contient le texte "pommes"*
- *=NB.SI(A2:A5;A4) compte le nombre de cellules de la zone A2:A5 qui contient la valeur qui se trouve dans la cellule A4*

3. Suite à la résolution du point 2, on obtient les fréquences absolues ou les fréquences d'apparition (f_a) . Leur somme représente le nombre total d' observations (30).

	A	B	C	D	E	F	G
1	TAS (mmHg)	fa	fc	fr	frc	fr%	frc%
2	100	2	=B2	=B2/$B$12	=D2	=D2	=F2
3	110	5	=C2+B3	=B3/$B$12	=E2+D3	=D3	=G2+F3
4	120	7	=C3+B4	=B4/$B$12	=E3+D4	=D4	=G3+F4
5	130	9	=C4+B5	=B5/$B$12	=E4+D5	=D5	=G4+F5
6	145	1	=C5+B6	=B6/$B$12	=E5+D6	=D6	=G5+F6
7	150	1	=C6+B7	=B7/$B$12	=E6+D7	=D7	=G6+F7
8	160	1	=C7+B8	=B8/$B$12	=E7+D8	=D8	=G7+F8
9	170	3	=C8+B9	=B9/$B$12	=E8+D9	=D9	=G8+F9
10	190	1	=C9+B10	=B10/$B$12	=E9+D10	=D10	=G9+F10
11		0					
12		30					

Pour la variable TAS et TAD, le tableau des fréquences est complété par cinq colonnes où on va calculer : les fréquences cumulées, relatives, relatives cumulées, relatives en pourcentage et relatives cumulées en pourcentage.

- **Les fréquences cumulées (f_c)** sont calculées par la cumulation de la fréquence suivante au total de la fréquence antérieure. (2; 2+5=7; 7+7=14; etc.) (On introduit la formule de calcul)
  - La fréquence absolue cumulée croissante d'une valeur x dans une série statistique S est la somme des fréquences absolues des valeurs de la série inférieure ou égale à x.
  - La fréquence absolue cumulée décroissante d'une valeur x dans une série statistique S est la somme des fréquences absolues des valeurs de la série supérieure ou égale à x.
- **Les fréquences relatives (f_r)** représentent le pourcentage de chaque fréquence dans le total n (30). (On introduit la formule de calcul) La fréquence relative d'une valeur x

dans une série statistique S est le rapport entre la fréquence absolue de la valeur x et la taille (volume) de la série. D'habitude la fréquence relative est présentée en pourcentage.

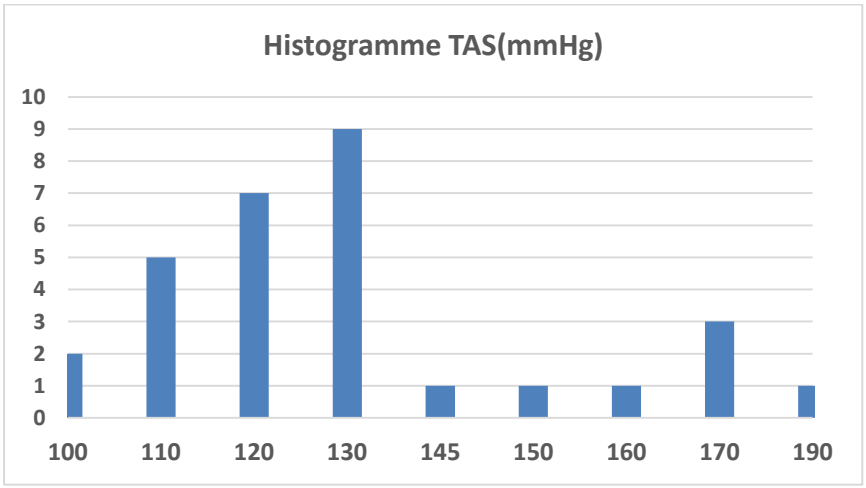
- **Les fréquences relatives cumulées (f_{rc})** sont calculées de la même manière que les fréquences absolues cumulées (voir ci-dessus) (On introduit la formule de calcul)
    - La fréquence relative cumulée croissante d'une valeur x dans une série statistique S est le rapport entre la fréquence absolue cumulée croissante de la valeur x et la taille (volume) de la série.
    - La fréquence relative cumulée décroissante d'une valeur x dans une série statistique S est le rapport entre la fréquence absolue cumulée décroissante de la valeur x et la taille (volume) de la série.
- **Les fréquences relatives en pourcentage** sont calculées conformément à la formule (voir cours) ou par formatage des données (copiez les fréquences relatives, sélectez et choisissez Pourcentage dans le menu Général) (on introduit la formule de calcul)

Les formules de calcul sont introduites pour la première ligne du tableau , puis on les copie avec Fill down.
Les fréquences en pourcentage sont formatées avec Pourcentage

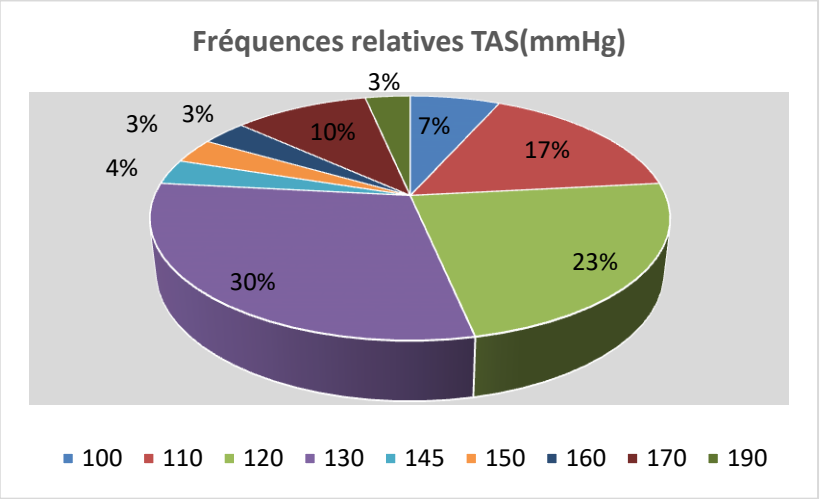
4. La représentation graphique des fréquences absolues (fa)

La représentation graphique des fréquences absolues s'appelle HISTOGRAMME.
Dans un **histogramme**, sur l'**axe X** (axe horizontale) sont représentés les intervalles de valeurs et sur l'**axe Y** (axe verticale) sont représentées les valeurs des fréquences correspondantes aux intervalles de valeurs.
Pour réaliser l'histogramme on peut utiliser le crayon et le papier (manuellement) ou les fonctions EXCEL.
Attention!!!! Le graphique doit porter un titre, sur l'axe Ox apparaitront les valeurs de la caractéristique (100,110, etc) et sur l'axe Oy la fréquence d' apparition (2, 5 etc)
Le graphique peut être réalisé par 2 méthodes (voir le travail LP04):

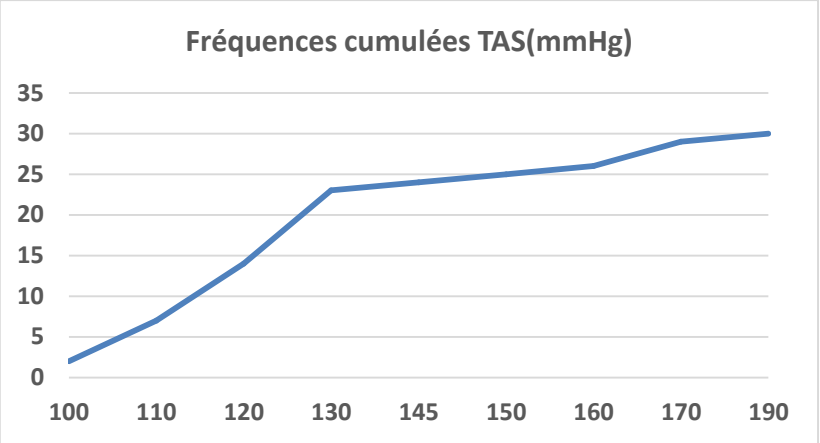
4.1 Dans l'image ci-dessous on peut voir l'histogramme des valeurs TAS. Pour l'histogramme on choisit le graphique colonne (2D). Puisque les valeurs pour TAS sont des valeurs discrètes, les colonnes sont séparées par l'espace (gap). Si les valeurs représentées sont continues, ces espaces ne doivent pas apparaitre (gap=0). Pour contrôler le"gap" on utilise l'option Format data seriee (clik droite).



4.2. Pour la représentation du graphique des fréquences relatives on utilise le graphique PIE.

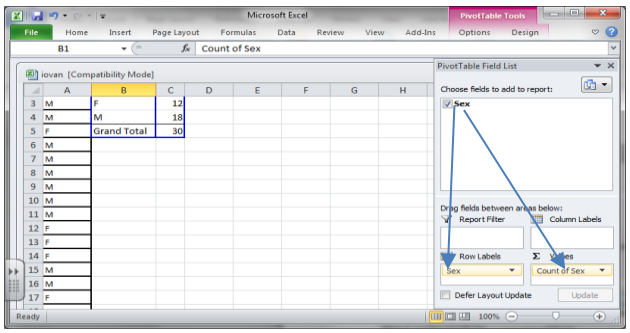


4.3. Pour la représentation graphique des fréquences cumulées on utilise le graphique ligne.

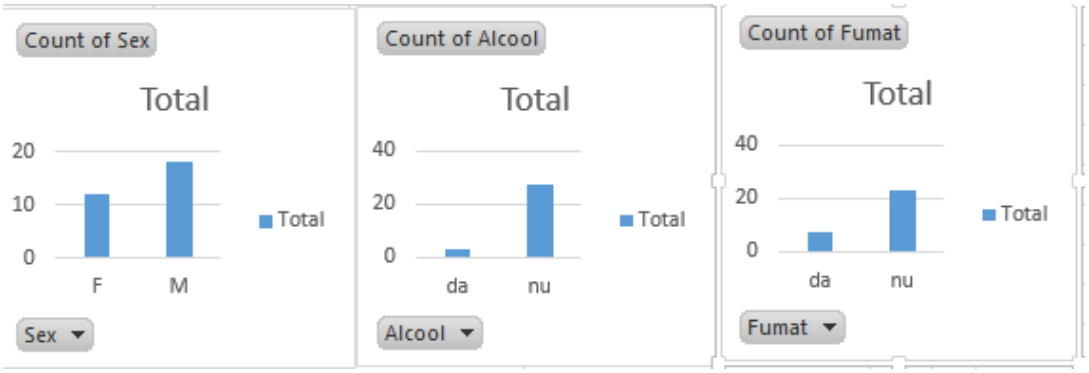


5. La réalisation de la table de fréquence pour les variables qualitatives.

Dans la table de ce travail les variables qualitatives sont: Sexe, Alcool et habitude de fumer. Ces variables prennent 2 valeurs (oui ou Non, respectivement F ou M). La table de fréquence consiste à découvrir le nombre d'observations du total (30) qui ont la valeur Oui et le nombre d'observations qui ont la valeur Non. Pour réaliser la table de fréquence pour ce type de variables, le plus simple est d'utiliser le tableau croiuse dynamique.



- On copie la colonne respective dans une feuille de calcul
- On sélectionne la colonne, y compris la cellule étiquette
- On sélectionne du menu Insert l'option Insert Pivot Table
- Dans la fenêtre Pivot Table Field List, par Drag &Drop on place dans Row Labels et Σ Values (Count Of ...) la variable en cause. Voir l'image ci-contre.
- Pour obtenir l'histogramme pour ces variables (qualitatives), au moment où on obtient la table de fréquence en utilisant Pivot table, dans le menu spécifique de cette option, apparait le menu spécifique ANALYZE d'où l'on choisit Pivot Chart. On obtient les graphiques ci-dessous.


Count of Sex
Total
20
10
0
F
M
Total
Sex
Count of Alcool
Total
40
20
0
da
nu
Total
Alcool
Count of Fumat
Total
40
20
0
da
nu
Total
Fumat