STATISTIQUES

I. Séries statistiques simples

A. Définitions

1. Population

Ensemble des éléments auxquels se rapportent les données étudiées.
Dans une population, chaque éléments est appelé individu ou unité statistique.
Le nombre d'individus est la taille.
Les observations constituent un ensemble de données.

2. Caractère statistique

⇨ caractère qualitatif : Ex: sexe => 2 **Modalités** = Homme ou Femme

⇨ caractère quantitatif discret = valeurs observées non continues, isolées les unes des autres. Ex: Nombre d'enfants (0, 1, 2, 3, 4,…). Mais toujours petit nombre de valeurs possibles.

⇨ caractère quantitatif continu: = toutes les valeurs intermédiaires sont possibles entre la plus petite et la plus grande. Ex: Taille, poids…

B. Séries classées / représentations graphiques.

1. Séries classées

Classement par tableau des données pour chaque modalité => Série classée, d'où on déduit des nombres n, pour chaque modalité, et des fréquences f.

- $f_1 = n_1/n * 100$
- $n_1 = n/100 * f_1$

2. Représentations graphiques

⇨ tuyaux d'orgues ou camemberts : pour caractère qualitatif, représentation selon la taille (tuyau d'orgue) ou la fréquence (camembert)

⇨ diagramme en bâtons : hauteur proportionnelle à la valeur ou à la fréquence de l'effectif.

⇨ histogramme : transformation de la série en série classée à classe de même amplitude = juxtaposition de rectangles dont la surface est proportionnelle à la taille ou aux fréquences des effectifs de la classe.

C. Résumé numérique

Pour les caractères quantitatifs seulement.

Appréhender l'unité ou la diversité de la population = indicateurs de tendance centrale ou indicateurs de dispersion.

1. Indicateurs de tendance centrale

a) Mode

C'est la valeur la plus fréquente d'un caractère discret (bâton le plus élevé dans un diagramme en bâtons).

Centre de la classe la plus fréquente pour un caractère continu.

Classe modale = classe la plus haute sur le graphique

b) Moyenne

Somme des données divisée par nombre de données:

Moyenne = $\overline{x} = 1/n(x_1 + x_2 + \ldots.. + x_n)$

Moyenne = $\overline{x} = 1/n(n_1c_1 + n_2c_2 + \ldots. + n_kc_k)$ => n = effectif de classe et c = centre de classe

Sous Excel: somme.prod. (Centre de classe;fréquence)

c) Médiane

Valeur telle que 50% des effectifs sont supérieurs et 50% sont inférieurs

Cf. Ecart interquartiles

2. Caractéristiques de dispersion

a) Etendue

C'est la différence entre la valeur Max. observée et la valeur Min.

b) Ecart-type / Variance

Mesure de l'étalement par rapport à la moyenne

$\sigma = \sqrt{\text{variance}} = \sqrt{\text{moyenne des carrés – carré de la moyenne}}$

$= \sqrt{\sum \overline{x}_i^2/n - \overline{x}^2}$

$= \sqrt{1/n \sum[(\text{centre de classe})^2 * \text{effectif de classe}] - \overline{x}^2}$

Sous Excel : somme.prod. (Centre de classe; centre de classe; fréquence)

c) Ecart interquartiles

- ⇨ Q1 = 25% des individus ont une valeur < à Q1
- ⇨ Q2 = 50% des individus ont une valeur comprise entre Q2 et Q1 = 50% des individus < Q2 = médiane
- ⇨ Q3 = 75% des individus ont une valeur comprise entre Q3 et Q2 = 75% des individus<Q3

Exemple:

Âge	Fréquences	Fréquences cumulées
25	10%	10% de la pop < 25ans
Q1		25% de la pop < Q1
35	22%	32% de la pop < 35 ans
45	12%	44% de la pop < 45 ans
Q2		50% de la pop < Q2
55	12%	56% de la pop < 55 ans
…/…		

Q1 (25%) est compris entre 32% et 10%, donc entre les classes d'âge 1 et 2, d'où:

$$\frac{Q1 - 25}{35 - 25} = \frac{25\% - 10\%}{32\% - 10\%}$$

Q1 = 25 + 10*15/22 = 31,8 ans

Q2 (médiane = 50%) est compris entre 56% et 44%, donc entre classes d'âge 3 et 4, d'où:

$$\frac{Q2 - 45}{55 - 45} = \frac{50\% - 44\%}{56\% - 44\%}$$

Q2 = 45 + 10*6/12 = 50 ans

II. Séries statistiques doubles: régression linéaire

A. Corrélation entre deux caractères quantitatifs

1. Nuage de points

Dans un tableau à données doubles, deux caractères quantitatifs x et y. On recherche une relation entre les deux. Si nuage de point en ballon de rugby, alors il existe une relation linéaire entre les deux. Si y augmente quand x augmente => corrélation positive; si y diminue quand x augmente => corrélation négative.

C'est une relation de type: **y = ax + b**

⇨ a est la pente (sous Excel : pente) = coefficient directeur => Si indépendantes, alors, y = 0x +b, la droite est horizontale et y = b.
⇨ b est l'ordonnée à l'origine (sous Excel: ordonnee.origine)

2. Coefficient de corrélation linéaire

La covariance :
cov(x,y) = Moyenne des produits – produit des moyennes
$= \sum \overline{x_i}\overline{y_i}/n - \overline{xy}$

Le coefficient de corrélation: r:
La somme des carrées des distances des points à la droite doit être minimale = **r^2**

$\sqrt{r^2} = r$ = coefficient de corrélation
$= cov(x,y)/ \sigma_x * \sigma_y$

Nombre sans dimension
Toujours entre -1 et 1
Toujours du même signe que a

Si r est proche de –1 ou 1 => les quantités sont corrélées linéairement
Si r proche de 0 => les deux quantités sont totalement indépendantes

Modèle bon si r entre 0,8 et 1

B. Droite de régression

Si r proche de -1 ou 1 => corrélation linéaire => on cherche à déterminer l'équation linéaire y = ax + b

Calcul de a et b:

⇨ $a = r * \sigma_x/\sigma_y$
⇨ $b = y – ax$

III. Séries doubles: Liaisons Indépendance

A. Tableau de contingence

Résumé de données nombreuses de 2 variables qualitatives ou quantitatives.
Tableau à double entrée:
⇨ verticale (nombre de pièce)
⇨ horizontale (loyer)

Les lignes "Total" donnent les répartitions selon une des 2 valeurs.

Si il y a corrélation, les lignes sont concentrées sur une diagonale descendante.

Les loyers moyens = Centres de classe * Effectifs de classe / Total ligne
D'où:
⇨ loyer moyen = y = ∑centre de classe * effectif/ total
⇨ C'est une **moyenne Conditionnelle**

La classe modale suit la diagonale descendante principale (= liaison entre les deux variables)

B. Répartition conditionnelle

1. Profil ligne

Chaque ligne (nombre de pièce) est une sous population répartie selon la deuxième variable (loyer). Sachant que le nombre de pièces est fixé = C'est une **Répartition conditionnelle**.

Pour comparer les lignes entre elles (effectifs différents) on fait une répartition en fréquence = **Profil ligne**.

⇨ Si les profils se ressemblent => Caractères indépendants
⇨ Si les profils diffèrent => caractères dépendants

2. Courbe de régression de y en x:

Etablie par positionnement des points moyens (y) en fonction du nombre de pièces
Elle est exclusivement descriptive => on ne peut pas extrapoler.
Si on fait la courbe de régression inverse, (nombre de pièces en fonction du loyer) et que les deux courbes sont proches => les 2 variables sont fortement liées.

C. Répartition sous l'hypothèse d'indépendance

Si les deux caractères sont indépendants, toutes les lignes sont réparties de la même façon.

Sous l'hypothèse (H0) d'indépendance entre les deux variables => **tableau de répartition théorique**:

	Fumeurs	Non fumeurs	
L	12 (calculé)	18	30
ES	20 (calculé)	30	50
S	8	12	20
	40%	60%	100%

On calcule la distance entre le tableau de répartition théorique et le tableau de répartition observée. C'est le test du **Khi-2 (X2).** On regarde l'écart classe par classe:

Différence² classes croisées / effectif théorique

⇨ $\sum$(effectif observé – effectif théorique)² /effectif théorique.
(pour toutes les classes)

C'est la distance critique (dc)
La distance théorique est données par la loi du Khi-2 (en fonction de la taille du tableau).

⇨ Si do (distance observée) < dc => non rejet de l'hypothèse d'indépendance
⇨ Si do > dc = rejet de l'hypothèse d'indépendance

Sous Excel, la distance est donnée par la fonction KHI-2.INVERSE (5%; degré de liberté). Le degré de liberté = (nombre de ligne –1) * (nombre de colonnes –1).