Analyse de données quantitatives (LPOLS1221)
Séance 7 : Statistiques descriptives et corrélations

Rappel de certaines notions de base en statistique

Effectif : nombre d'occurrences

Moyenne (arithmétique) : $\bar{X} = \frac{\sum x_i}{n}$

Mode : modalité la plus fréquente

Médiane/quartile/décile : modalité qui répartit l'échantillon en 2/4/10 parts strictement égales (50%/25%/10%)

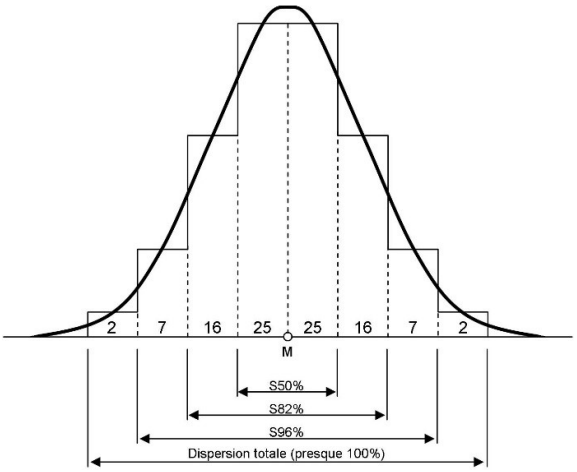
Variance (mesure de dispersion) : $s^2 = \frac{\sum(x_i - \bar{x})}{n-1}$ (si mesurée sur un échantillon)

Écart-type (mesure de dispersion) : $s = \sqrt{s^2} = \sqrt{\frac{\sum(x_i - \bar{x})}{n-1}}$ (si mesuré sur un échantillon)

Étendue : écart entre valeur minimale et valeur maximale

Ecart interquartile (mesure de dispersion/d'asymétrie) : $EI = Q_1 - Q_3$

Le menu Explore : tests de normalité et statistiques descriptives

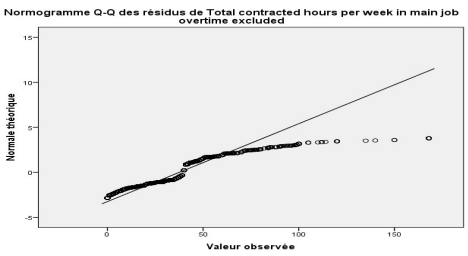


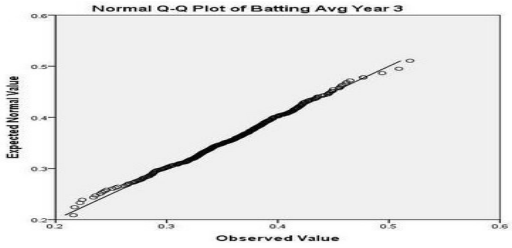
Beaucoup de statistiques demandent que les données soient distribuées d'une façon normale, avec une courbe qui suit la loi gaussienne (voir image). Les tests de normalité (Shapiro-Wilk pour des échantillons de 2000 observations ou moins et Kolmogorov-Smirnov pour des échantillons de plus de 2000 observations) mesurent la distribution des observations. Si la significativité (Sig.) du test est en dessous de 0,05, les données sont considérées comme non-normalement distribuées. Dans le cas contraire, on peut les considérer comme normalement distribuées.

Il y a aussi des graphiques pour tester l'hypothèse de normalité : les Q-Q plots et les P-P plots.

*Le tracé quartile-quartile (**QQplot**) en est le plus utilisé et compare les données existantes avec leur distribution hypothétique, si elles étaient normales. Si les données suivent la diagonale du graphique, les données sont (presque) normales. Dans le cas contraire, ce n'est pas le cas.*

Exemples de variables : non-normalement distribuée (gauche) et normalement distribuée (droite)





Explore (Explorer) : Permet de décrire des variables **quantitatives.** Elle fournit d'emblée des mesures de la tendance centrale et de dispersion, ainsi que deux indices de normalité des données : la symétrie (skewness) et l'aplatissement (kurtosis). Via l'option **Plots** (Diagrammes) -> **Normality plots with tests** (graphe de répartition gaussien avec test), vous pouvez également demander un **test de normalité** et un tracé **quartile-quartile (QQ plot),** ainsi qu'un histogramme. Par défaut, elle génère un **box-plot** (boîte à moustache). Les variables quantitatives à explorer sont à rentrer dans l'encadré **Dependent list**. L'encadré **Factor list** permet de les explorer par modalités d'une variable qualitative.

Toutes ces manipulations peuvent être effectuées par sous-groupes d'observations. L'option ***Split File****, du menu* ***Data****, permet de scinder l'échantillon en plusieurs sous-échantillons qui seront traités séparément, en sélectionnant l'option « Organize output by groups » (organiser le résultat par groupes). Ainsi, on obtient un résultat pour chacun des groupes considérés.*

Les graphiques dans SPSS

Comme déjà dit, la fabrication des graphiques peut se faire via certaines procédures d'analyse dans SPSS telles que les procédures Frequencies ou Explore. Cependant, le menu **Graphs** -> **Legacy dialogs** (Boîte de dialogue) permet de générer une panoplie de types de graphiques. De manière générale, nous vous conseillons d'utiliser les graphiques de la façon suivante.

- Pour décrire des variables nominales/ordinales : **Bar Chart/simple** (diagrammes en barres simple)
- Pour décrire des variables quantitatives : **Histogram** (histogramme)
- Pour décrire la relation entre 2 variables nominales/ordinales : **Bar Chart/Clustered** (Diagramme en barres groupées)
- Pour décrire la relation entre une variable quantitative et une variable nominale/ordinale : **Box-plot** (Boîte à moustache)
- Pour décrire la relation entre 2 variables quantitatives : **Scatter dot** (Nuage de points)

De préférence, évitez les graphiques qui rendent la lecture et l'interprétation difficiles comme : les camemberts, diagrammes en barres pilés, etc.

L'option récurrente ***Panel by*** *vous permet de produire plusieurs graphiques en une seule fois selon des modalités précises (par exemple, la distribution des niveaux d'instruction selon le sexe : un graphique pour les hommes et un graphique pour les femmes).*

La corrélation

Le test de corrélation est un test inférentiel bivarié qui permet d'étudier l'association entre **deux variables quantitatives**. Son coefficient correspond à la pente d'une droite et sert ainsi à mesurer la variation de la variable Y, si la variable X est modifiée.

Le coefficient de corrélation (r) est obtenu en calculant le rapport entre la covariance observée entre deux variables x et y ($COV(x,y)$) et le produit des écarts-types observés pour ces deux mêmes variables (S_xS_y). La covariance est une statistique permettant de connaître les écarts conjoints d'une observation vis-à-vis de deux variables ou plus. Son calcul est donc sensiblement différent à celui de la variance.

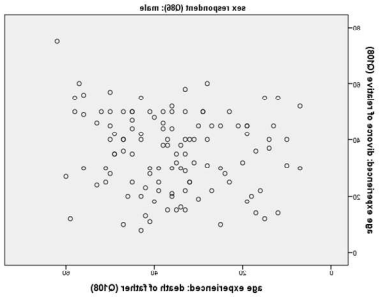
$$r = \frac{COV(x,y)}{S_xS_y} \quad \text{où } COV(x,y) = \frac{1}{n-1}\sum(x_i - \bar{X})(y_i - \bar{Y})$$

Le coefficient de corrélation est compris entre **-1** (corrélation négative parfaite) et **+1** (corrélation positive parfaite). Un score proche de 0 renverra à une absence de relation. En sciences sociales, on considère généralement les coefficients de corrélation ainsi :

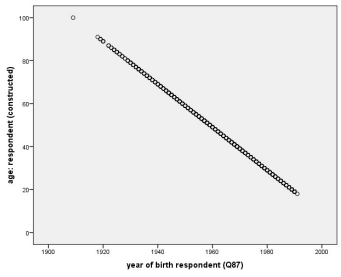
- Autour de (-) 0.1 : corrélation faible

- Autour de (-) 0.3 : corrélation moyenne
- Autour de (-) 0.5 : corrélation forte

On peut inspecter visuellement la corrélation entre deux variables quantitatives, via le menu **Graphs** -> **Legacy Dialogs** -> **Scatter/Dot**. Des points qui se regroupent autour d'une droite imaginaire suggèrent l'existance d'une corrélation entre les variables considérées, tandis que des points semblent distribués aléatoirement suggèrent l'absence d'une corrélation significative.



Pas de correlation significative



Corrélation parfaite (négative)

Comme les autres tests d'association, la corrélation part d'une hypothèse nulle qui postule l'absence de relation entre les deux variables. Si la p-valeur associée au test de corrélation est inférieure au seuil choisi (exemple 5%), on peut rejeter H_0 et conclure donc qu'il y a une relation entre les deux variables. Dans le cas inverse, on ne pourra pas conclure que les variables sont associées.

Mémo pour interpréter le test de corrélation :
- Analyse de la p-valeur, par rapport au seuil choisi, pour décider s'il y a une association
- Analyse de la valeur absolue du coefficient R, qui indique la force de l'association
- Analyse du signe du coefficient, qui indiqe le sens de la relation : corrélation négative (y diminue si x augmente) ou positive (y augmente si x augmente).

Manipulation dans SPSS

Analyze (Analyse) ➔ Correlate (Corrélation) ➔ Bivariate (Bivariée)

Applications

Utilisez la base de données de l'Eurobaromètre 2017 pour réaliser ces exercices.

Application 1.

Sortez le graphique le plus pertinent pour afficher la relation entre l'âge (**d11**) et l'âge quand le répondant a arrêté ses études à temps complet (**d8**)* dans l'échantillon belge. Faites-le séparément pour les hommes et les femmes (**d10**). Testez ensuite l'association avec un test statistique, toujours différencié par genre. Commentez.

* Enlevez de l'analyse toutes les valeurs manquantes de la variable d8, en créant une nouvelle variable.

Application 2.

On voudrait étudier les facteurs associés à la satisfaction générale de vie chez les membres de l'Union Européenne (**eu28**) (en excluant la Grande Bretagne et l'Irlande du Nord (**country**)). Pour cela, on va entreprendre une série de procédures.

1. Recodez la variable portant sur la satisfaction générale (**d70**), pour en créer une nouvelle variable (**satisfaction**) avec deux catégories :
   "très satisfait(e) et "plutôt satisfait(e)" -> "satisfait"
   "plutôt pas satisfait(e)" et "pas du tout satisfait(e)" -> "pas satisfait"
   Réalisez un graphique approprié de cette nouvelle variable. Puis, réalisez un deuxieme graphique, par sexe, utilisant des pourcentages. Commentez.

2. Testez si la satisfaction de vie des individus dépend de leur niveau de richesse. Pour cela :
   a. Créez une nouvelle variable, qui exprimera un index sommatif de richesse, sur base des variables suivantes: Une télévision (**d46.1**), Un lecteur DVD (**d46.2**), Un lecteur CD audio (**d46.3**), Un ordinateur de bureau (**d46.4**), Un ordinateur portable (**d46.5**), Une tablette tactile (**d46.6**), Un smartphone (**d46.7**), Une connexion Internet à la maison (**d46.8**), Une voiture (**d46.9**), Un appartement ou une maison que vous avez fini de payer (**d46.10**), Un appartement ou une maison que vous êtes en train de payer (**d46.11**). Attribuez-lui une description et un niveau correct de mesure. Ensuite, sans effectuer aucune manipulation: quels sont le minimum et le maximum possibles pour cette variable? Analysez la variable par la suite, est-ce que sa distribution suit une loi normale ?
   b. Choisissez un graphique approprié pour examiner la relation entre cet index de richesse et la satisfaction de vie (**satisfaction**).
3. On suspecte qu'il y a aussi une relation entre la classe sociale et la satisfaction. Pour cela :
   a. Recodez la variable portant sur la classe sociale auto-perçue du répondant (**d63**), pour en créer la variable **classe_soc**, avec trois modalités :
      "ouvrière" et "moyenne inférieure" -> "ouvrière"
      "moyenne" -> "moyenne"
      "moyenne suppérieure et "la plus élevée" -> "supérieure"
   b. Choisissez le graphique le plus approprié pour représenter la relation entre les deux variables et commentez
4. A l'aide d'un graphique, estimez s'il y a une différence d'âges (**d11**) en termes de satisfaction générale de vie (**satisfaction**)
5. Sans vous aider par une représentation graphique, testez s'il y a une relation entre la richesse (**richesse**) et l'âge des individus (**d11**).

Que peut-on conclure ? Quels sont les facteurs qui semblent être liés à la satisfaction de vie des individus ?