Statistiques descriptives : TP n°1

1 Conseils généraux

1.1 Forme des données

En statistique on veillera toujours à ce que les données aient la forme suivante : les variables en colonne et les individus en ligne. La première ligne devra comporter le nom des variables. Si les données n'ont pas initialement cette forme on veillera à la leur donner. Ceci simplifiera grandement les traitements ultérieurs dans Excel ou tout autre logiciel statistique. Si initialement on a les individus en colonnes et les variables en ligne on pourra par exemple utiliser un collage spécial avec l'option "Transposée" dans Excel.

1.2 Importation des données

Une dernière possibilité pour importer les données consiste en un copier-coller par exemple à partir d'un document pdf ou d'une page web. Dans ce cas après collage vous devez voir apparaître dans Excel des options de mise en forme en bas à droite de la zone collée. Ces options vous permettent entre autres d'accéder à l'assistant d'importation pour la zone collée. Il est possible d'accéder directement à cet assistant par un collage spécial. Cette méthode n'est pas la plus recommandée mais elle permet dans de nombreux cas de gagner un temps précieux par rapport à une saisie manuelle ...

1.3 Mise en garde sur les graphiques

Attention à la 3D : même si les histogrammes et diagrammes circulaires en 3D peuvent être attrayants, mis à part leur effet esthétique, ils n'apporte pas d'information supplémentaire par rapport aux représentations basiques 2D. Bien au contraire ils ont tendance à fausser nos perceptions.

Attention à la lisibilité des graphiques : prendre des polices assez grandes et s'assurer que la figure est assez grande.

Attention à l'origine des axes : toujours veiller pour la représentation de proportions à ce que la ligue de base se trouve bien à 0.

Attention aux couleurs : toujours se poser la question de l'utilisation finale du document produit ; ce dernier sera-t-il imprimé en noir et blanc ou en couleur ? Si il est imprimé en noir et blanc adapter les couleurs pour que le document reste lisible à l'impression.

1.4 Les tableaux croisés dynamiques (TCD)

Dans Excel l'insertion de tableaux croisés dynamiques ainsi que les graphiques associés permet de répondre à nombre de questions en statistiques descriptives 1D et 2D.

Principe général : sélectionner la plage de données puis choisir dans le menu "Insertion", "TblCroiséDynamique" puis "Graphique croisé dynamique" (dans le cas ou on souhaite à la fois le tableau croisé dynamique et le graphique). Ensuite il ne reste plus qu'à faire glisser les différents champs dans les catégories souhaitées.

1.4.1 Diagramme en bâtons

Ici il suffit de faire glisser la variable à la fois dans le "Champs Axe (Abscisses)" et le champs "Valeurs". On obtient ensuite directement le tableau de fréquences et le digramme en bâtons. Il est aussi possible de modifier le graphique en fonction des besoins.

Attention pour les variables quantitatives discontinues, bien vérifier que les valeurs sont rangées par ordre croissant.

1.4.2 Histogramme

Pour faire un histogramme à partir des donnée brutes en utilisant un TCD :

1. Sélectionner les données
2. Choisir "Insertion", "TblCroiséDynamique" puis "Graphique croisé dynamique"
3. Sélectionner la variable considérée dans les champs à inclure dans le rapport, elle apparaît alors dans "Valeurs"
4. Faire glisser cette même variable dans la zone "Champs Axe (Abscisse)"
5. Dans la feuille de calcul, sélectionner une cellule, puis dans le menu "Outils de tableau croisé dynamique", "Options" choisir "Grouper les champs", modifier les valeurs de début, fin et pas à votre convenance (on pourra s'appuyer sur la règle de Sturges). Attention quand même à ce que la valeur de début soit inférieure au minimum de la série et la valeur fin supérieure au maximum de la série.
6. Remplacer dans le champs "Valeurs" le choix de "Somme" par "Nombre" dans "Paramètres de champs de valeurs"
7. Sélectionner la figure puis dans le menu "Outils de graphique croisé dynamique" puis "Création" et dans la zone "Disposition du graphique" choisir la disposition telle qu'il n'y ait pas d'espace entre les barres (Mise en forme 8).
8. Enfin dans les "Outils de graphique croisé dynamique" peaufiner la figure : "Titre", "Axes", ...

1.5 Lister toutes les modalités d'une variable

Pour lister toutes les modalités d'une variables qualitative, on peut copier-coller la plage de données dans un autre colonne puis dans la plage de données ainsi copiée supprimer les doublons à l'aide du menu "Données" puis supprimer les doublons. On obtient ainsi l'ensemble des modalités possibles.

1.6 Tableau de fréquences à partir de la fonction "nb.si"

A partir de la liste de l'ensemble des modalités d'une variable on peut utiliser la fonction "nb.si" pour compter dans la plage de données initiale le nombre d'occurrences de chacune des modalités, on pensera à utiliser à bon escient les $ pour le glissement de la formule.

1.7 Découper les données en classe avec la fonction "nb.si.ens"

Une fois les limites des classes définies on peut facilement déterminer le nombre de données appartenant à chacune d'entre-elles grâce à la fonction "nb.si.ens".

1.8 Calcul des fréquences cumulées croissantes et décroissantes

Calcul des fréquences cumulées croissantes et décroissantes. La solution la plus simple est de repartir de la définition des fréquences cumulées croissantes, c'est-à-dire d'utiliser la fonction "somme" en plaçant judicieusement les $ de sorte à faire glisser la formule.

Remarque : la valeur de la cellule A2 est égale à "somme(A2 :A2)".

1.9 Courbes cumulatives croissantes et décroissantes

1.9.1 Données quantitatives discontinues

Ici on opte pour une représentation graphique légèrement différente de celle vue en cours car la forme de fonction en escalier n'est pas très adaptée dans Excel. On va simplement réaliser une une nuage de points avec en abscisse la valeur de la variable et en ordonnée la fréquence cumulée correspondante. On pourra superposer sur une même graphique les fréquences cumulées croissantes et décroissantes.

1.9.2 Données quantitatives continues

Courbes des fréquences cumulées croissantes et décroissantes. Attention une attention toutes particulière doit être portée à la construction de cette courbe. Noter avant toute chose que la courbe cumulée croissante vaut 0 au niveau du minium et 1 au niveau du maximum. Inversement la courbe cumulée décroissante vaut 1 au niveau du minimum et 0 au niveau du maximum. Le points d'intersection des deux courbes a pour ordonnée $0,5$ et on retrouve en abscisse la valeur de la médiane. Sur machine on peut soit construire cette courbe à partir des données brutes, soit à partir de données regroupées en classe lors de la construction de l'histogramme (ces classes étant généralement obtenues en utilisant la règle de Sturges). L'intérêt de la seconde version est qu'elle permet d'obtenir une version "lissée" de la courbe cumulative, tandis que la première produit une version plus "bruitée".

À partir des données brutes On commence par réaliser une TCD sur les données afin de regrouper les valeurs identiques on prendra soin dans le champ "Valeurs" du TCD de bien choisir "Nombre". Trier dans ce tableau les valeurs par ordre croissant puis à partir du tableau ainsi obtenu calculer les fréquences cumulées croissantes pour chacune des valeurs. Enfin faire le tracé de la courbe.

À partir des données groupées en classes Supposons qu'on ait le tableau de fréquences suivant :

Classe	$[x_1; x_2[$	$[x_2; x_3[$	$[x_3; x_4[$
Fréquence	$0,3$	$0,5$	$0,2$
Fréquence cumulée croissante	$0,3$	$0,8$	1
Fréquence cumulée décroissante	1	$0,7$	$0,2$

On construit la courbe cumulative croissante en reliant les point suivants :

x	x_1	x_2	x_3	x_4
y	0	$0,3$	$0,8$	1

Il ne reste plus qu'à créer le tableau précédent dans Excel puis de réaliser la courbe à partir de ce graphique.

De même, on construit la courbe cumulative décroissante en reliant les point suivant :

x	x_1	x_2	x_3	x_4
y	1	0,7	0,2	0

1.10 L'utilitaire d'analyse statistique

Enfin un outil particulièrement utile pour réaliser des statistiques à partir d'Excel est l'utilitaire d'analyse statistique ; il permet de réaliser des histogrammes (en lui précisant les limites de classes), de calculer l'ensemble des statistiques descriptives, ainsi que de nombreuses autres choses ...

Avant toutes choses, il convient d'installer le module "Utilitaire d'analyse" : "Fichier", "Options", "Compléments", sélectionner "Analysis ToolPack", cliquer sur "Atteindre...", cocher la case "Analysis ToolPack" puis cliquer sur "OK". L'utilitaire d'analyse statistique peut ensuite être utilisé en se rendant dans "Données", "Utilitaire d'analyse".

1.11 Réalisation d'une boite à moustache dans Excel

La réalisation d'une boîte à moustache dans Excel n'est pas automatique. Pour la réaliser, il faut suivre rigoureusement un certain nombre d'étapes, le lien http ://www.youtube.com/watch ?v=s8ZW4PVar permet d'obtenir un très bon rendu pour cette dernière.

1.12 Moyenne, variance, écart-type, médiane, quartiles, centiles

- La fonction MOYENNE permet de calculer la moyenne.
- La fonction VAR.P.N permet de calculer la variance (attention les autres variantes produisent des valeurs légèrement différentes).
- La fonction ECARTYPE.PEARSON permet de calculer l'écart-type (attention la fonction ECARTYPE donne des résultats légèrement différents).
- La fonction MEDIANE permet de calculer la médiane.
- La fonction QUARTILE.INCLURE permet des calculer les quartiles, elle réclame en second argument le numéro du quartile à calculer 0,1,2,3, ou 4.
- La fonction CENTILE permet de calculer les centiles, elle réclame en deuxième argument le centile à calculer, par exemple 0,1 si l'on souhaite calculer le premier décile.

Tous ces indicateurs peuvent aussi être calculés simplement à partir de formules. Il n'existe en général pas dans Excel de fonction bien adaptée au traitement des données par classe. Dans ce cas, il faut repartir du cours et adapter les formules.

Ces indicateurs peuvent aussi être facilement obtenus à l'aide de l'utilitaire d'analyse statistique (Données, puis Utilitaire d'analyse, puis Statistiques descriptives), attention de l'avoir installé au préalable.

1.13 Analyse de la concentration

Il n'y a pas de fonction Excel disponible pour réaliser l'analyse de la concentration. Ici on décrit l'approche à partir de données brutes, l'approche pour des données par classes ne posant normalement pas de difficulté supplémentaire. Pour réaliser cette analyse on suivra les étapes suivantes :

1. Calcul de la médiane
2. Réalisation de la courbe des fréquences cumulées croissantes en (%).
3. Calculer D_1 et D_9 ainsi que le rapport inter-déciles.
4. Calculer la somme cumulée croissante des salaires et la remettre en (%).

5. Trouver la médiale (valeur qui découpe la masse salariale en 2), on pourra utiliser une interpolation linéaire.
6. Calculer la concentration.
7. Tracer la courbe de Lorenz.
8. Calculer l'indice de Gini.

2 Exercices sur la réalisation de tableaux et de graphiques

Exercice 1 Faire un diagramme circulaire 3D représentant la somme reversées aux travailleurs et aux actionnaires. Disons 1/4 pour les travailleurs et 3/4 pour les actionnaires. En jouant sur la rotation 3D, produire une version de ce diagramme pour un présentation aux actionnaires, et une version de ce diagramme pour un présentation aux salariés. Attention cet exercice vise seulement à vous faire prendre conscience des risques de manipulation à partir de tels graphiques et non pas à vous apprendre à les pratiquer ! Dans la pratique nous recommandons bien sur d'opter pour le diagramme circulaire 2D, voire mieux pour diagramme en bâtons 2D.

Exercice 2 Les diagrammes en bâtons peuvent aussi être manipulés, attention à la ligue de base ! Supposons maintenant que les bénéfices se partagent à 45% pour les salariés et 55% pour les actionnaires. Produire un digramme en bâtons donnant l'impression que la part reversée aux actionnaires est beaucoup plus grande que celle reversée aux salariés, ceci à l'aide d'un changement de la ligne de basse. Dans la pratique on veillera que pour la représentation de proportions la ligue de base se trouve bien à 0.

Exercice 3

1. Ouvrir le fichier Exercice 7 feuille TD (dans la partie jeux de données de moodle).
2. Faire les questions de l'exercice 7 de la feuille de TD (utiliser l'approche par TCD et aussi celle faisant appel à la fonction **nb.si**)
3. Calculer les fréquences cumulées décroissantes en % et faire la courbe cumulative associée

Exercice 4

1. Refaire l'exercice 12 de la feuille de TD
2. Faire l'histogramme associé
3. Construire la courbe des fréquences cumulées croissantes et décroissantes

Exercice 5 Le fichier Excel temps10km.xls contient les temps de 793 coureurs d'un 10 kilométres exprimés en minutes.

1. Réaliser un histogramme de cette distribution.
2. Tracer les courbes cumulatives croissantes et décroissantes.
3. Calculer moyenne, variance, écart-type, coefficient de variation. Faire ces calculs à partir des fonctions Excel et à partir de formules.
4. Médiane, quartiles, déciles, écart-interquatiles (absolus et relatif), écart-interdéciles (absolus et relatif). Faire ces calculs à partir des fonctions Excel et à partir de formules.
5. Réaliser la boîte à moustaches.

Exercice 6 Ouvrir le ficher ozone.txt dans Excel en prenant garde à bien définir les paramètres d'importation. Puis enregistrer le fichier au format xls ("ozone.xls") pour une réutilisation plus facile. Mettre en forme ce fichier pour que chaque colonne corresponde à une variable (si tel n'est pas le cas). Le jeu de données traite de variables climatiques et d'une variable pollution à l'ozone mesurées durant l'été 2011 à Rennes. Pour l'étude qui va suivre, ne retenir que les variables `max03` (maximum d'ozone journalier), `T12` (température à midi), `vent` (direction), `pluie` et `Vx12` (projection du vecteur vitesse du vent sur l'axe Est-Ouest).

1. Décrire la la population, l'unité statistique, les variables étudiées ainsi que leurs natures
2. Réaliser l'étude de chacune des variables

Exercice 7 Le fichier donnees_philippine.csv contient les salaires de 632 foyers Philippins ainsi que d'autre variables qui ne seront pas analysée ici. Il s'agit du jeu de données Ilocos décrit dans le fichier http ://cran.r-project.org/web/packages/ineq/ineq.pdf.

1. Étudier la distribution des salaires.
2. Réaliser une étude de la concentration des salaires.

Exercice 8 Afin d'accélérer leur traitement statistique, les données de l'exercice 31 de la fiche de TD ont été exportées dans le fichier exercice31.xls. Répondre aux questions de l'exercice rappelée ci-dessous.

On veut réorganiser pour gagner du temps l'activité d'un service qui exécute diverses tâches répétitives. Certaines de ces tâches sont rapides, d'autres demandent un grand nombre d'heures ouvrées.

Tracer la courbe de Lorenz mettant en relation le nombre de postes et le nombre d'heures de travail. Utiliser cette courbe pour mettre en évidence trois types de postes :

1. les postes A peu nombreux et représentant beaucoup d'heures,
2. les postes B peu nombreux et utilisant peu d'heures,
3. les postes C nombreux mais utilisant peu d'heures,
4. proposez une politique de réorganisation.

3 Quelques liens utiles

- http ://www.google.com/publicdata/ : ce site vous permet de visualiser en direct un grand nombre de statistiques publiques d'intérêt de manière interactive.
- http ://partenaires.onisep.fr/wp-content/uploads/2011/08/2011_ZOOM_STATISTIQUE_WEB_150dp brochure ONISEP sur les métiers de la statistique.
- http ://www.decideo.fr : ce site vous donnera un aperçu de l'informatique décisionnel. La section "étude de cas" permet de visualiser son intérêt sur des exemples concrets. L'exemple http ://www.decideo.fr/Etudes-de-cas_r12.html est particulièrement parlant.
- 10 outils du web qui rendent la vie plus facile http ://www.slideshare.net/lucos/10-usages-et-outils-qui-rendent-la-vie-plus-facile