Cours de statistiques Descriptives et Inferentielles (Inductive)

Philip Wood 2005

Cours de Statistique Descriptive:

1. Introduction:

Defn. : Science du groupement méthodique des faits qui se présent a une évaluation numérique.

3 éléments essentiels

1. Groupement (anémie, VIH ….), 2. numérique (mathématique, nombre avec une maladie, nombre sans maladie, nombre des cabinets…) (sexe 1 & 2, niveau d'études), 3. évaluation – essentielle surtout pour inferentielle – analyse, comparaison…

Science jumeau = épidémiologie : Etude des différents facteurs de l'apparition et évolution des phénomènes de la santé.

Simple description des résultats = Statistiques Descriptive

Analyse à profondeur = Statistiques Inferentielle.

Exemple 2: Etude anémies : Enfants admis à l'hôpital d'Oicha pendant 3 mois en 2003.

Age	Effectif	Avec	Sans	%
0-1	209	51	158	24
1-2	54	26	28	48
2-3	17	11	6	65
3-4	15	9	6	60
4-5	5	5	0	100

Total 102

Les enfants de quel âge souffrent d'avantage de l'anémie ?

Réponse 1. Enfants de l'age 0-1 = 51 enfants parmi 102 avec anémie

2. Enfants de 4-5 – 100% étaient anémique.

Mais il y a seulement 5 enfants de l'age de 4-5ans. Est-ce que ce résultat est par hasard ? Les statistiques inferentielle peut nous dire.

Ici on a fait une simple description de ce que nous avons trouvé.

IMPORTANT. C'est toujours nécessaire de comparer un échantillon avec la (une) population duquel l'échantillon fait part.

Exemple 2 : Est-ce le tabac contribue la bronchite? (remplissez les boites vides)

FUMEURS	Quotidienne	Peu ou jamais	TOTAL	%
Avec bronchite	49	111	160	
Saines	270	1230	1500	

Hypothèse- Fumer le tabac n'a aucun effet néfaste.

Exemple 3 : Etude: Malades hépatiques qui boivent de l'alcool (remplissez les boites vides)

ALCOOL	Quotidienne	Peu ou jamais	TOTAL	
Malades	15	35	50	
Saines	311	1417	1728	

Hypothèse- Boire de l'alcool n'a aucun effet néfaste.

Exemple 4 : Est-ce que les moustiquaires diminue la paludisme ? Enfants admis avec paludisme nov 05

Paludisme	Avec moustiquaire	Sans moustiquaire	TOTAL	
Malades	21	29	50	
Autres maladies	115	235	350	

2. Dénombrement, Dépouillement

Données numériques = dénombrement. Defn : Trouver le nombre de quelque chose.

La valeur de chaque caractère est à découvrir

Quatre descriptions des études (8 possibilités) (Types d'études) :

1. Direct ou 2. indirect: Données collectionnées directement sur terrain ou 2 indirectement par les registres etc. Par documents (indirects) ou par observation des sujets = direct.

3. Etudes prospectives (Etude commence à partir d'aujourd'hui), 4. rétrospectives (Ancien cas retrouvés d'un registre) On trouve que dans le registre il y a les éléments que les gens ont oublie à noter. L'étude rétrospective peut être moins exacte mais plus facile que cela soit grande.

5.. Dénombrement instantané = étude transversale. (Par exemple nombre de cas de telle maladie aujourd'hui.) 6. Continue = étude longitudinale (Par exemple évolution de traitement d'un group des malades.)

7.. Etude de toute une population (souvent trop difficile) ou 8. une étude d'un échantillon de la population.

Besoin de précision, honnêteté, organisation, patience, pratique et une certaine connaissance du sujet sans préjugé, puis une analyse avec soin pour éviter les erreurs.

Besoin souvent d'une étude préliminaire. Est-ce qu'il y aura assez de cette maladie (événement) de rendre l'étude utile ?

Dépouillement Defn : Analyse minutieux du dénombrement. Par :

a) pointage – un questionnaire est entré dans un registre et puis on addition des colonnes

b) fiches perforées (Carte perforée) (Carte trouée et une aiguille pour sélectionner certaines caractéristiques qui sont représenté par un trou dans la carte)

c) ou ordinateur {avec programme accès ou Excel) Exemple les résultats de nos examens)

N.B. Vous voulez étudier un sujet précise (par ex une maladie) mais vous devrez toujours recueilli l'information sur la population dans laquelle se trouve cette maladie.

Exemple 1. : Etude porte à porte pour vérifier la cicatrice de BCG chez les gens des ages différentes dans un quartier :

= étude directe, prospectif, transversale, d'un échantillon

A étudier le nombre avec cicatrice par rapport au nombre total vu.

Exemple 2 : Analyse de tous les gens opéré pour une hernie a Beni en 2005 à partir du registre a la salle d'opération.

= étude indirecte, rétrospectif, étude transversale, d'un échantillon.

A étudier le nombre des hernies par rapport au nombre total des opérations.

Exemple 3 : Etude des cas de paludisme chez 2 groups des enfants 1. Qui a eu une vaccine expérimentale, 2. Un groupe semblable sans vaccin

= étude directe, prospectif, longitudinale, d'un échantillon.

A étudier le nombre des cas de paludisme dans un group par rapport auautre.

Exemple 4 : Suivie de 2 groupes des diabètes un qui reçoive insuline l'autre diabinase pour voir leur longueur de vie.

= étude directe, prospectif, longitudinale d'un échantillon

A étudier les deux groupes.

Soit consciemment ou inconscieusement on propose une hypothèse, à accepter ou rejeter, puis on établir une étude pour l'épreuve de cette hypothèse. Dans les exemples en haut l'hypothèse puisse être 1. C'est bon d'avoir le BCG 2. Les hernies sont plus fréquentes chez les hommes. 3. Un vaccin contre le paludisme est efficace. 4. Les gens vivent plus longtemps avec insuline.

T.P.

1. Faites une étude des WC dans un quartier ou vous habitez. Voir 10 maisons, combien ont un WC?

2. Comment faire une étude rétrospective longitudinale du paludisme ?

3. Comment faire une étude des cancers du sein dans notre coin ?

3. Taux et ratio :

Ratio : Partie de la population avec un caractère par rapport aux autres dans la population avec un autre caractère. Exemple : Nombre de la population qui sont hypertendus par rapport avec les normotensives.

Rapport: nombre avec un caractère par rapport au total. Exemple : Nombre des WCs dans un quartier par rapport au nombre de maisons dans ce quartier.

N.B. Tout le monde qui faites cette étude n'aura pas le même résultat – on appelle ceci la **variance**.

La variance dépende souvent de la taille de l'échantillon. Votre échantillon est une fraction (%) de la population totale qui est trop difficile à étudier en totale. La plus grande votre échantillon le moins la variance. Il y a moins de variance si votre échantillon est pris par hasard et qu'il n'y ait pas trop d'un ou autre caractère spécifique dans la totale.

Aleantoire : La choix de la population, qu'on va étudier, doit être normalement par hasard. Il y a une gamme des règles pour choisir un échantillon dans une manière complètement au hasard (voir chapitre 8). On appelle une telle échantillon sans biasis un échantillon aleantoire.

Dans vos conclusions sans doute vous tirez les conclusions qui puissent être appliquer à une population plus grande que votre échantillon. Par exemple vous pouvez tirer la conclusion que tous le monde doit … dormir sous une moustiquaire. Mais faites attention. Est-ce que votre échantillon est représentatif de la population mondiale, ou de RDC, de Nyankunde, d'Oicha, de l'Hôpital d'Oicha, ou de salle 10 de l'hôpital d'Oicha ? C'est inutile de dormir sous moustiquaire la ou il n'y a pas de paludisme. Tirez vos conclusions avec soin.

Prévalence et incidence.

A. **La prévalence**

La prévalence est la mesure du nombre de cas d'une maladie donnée, à un moment donné dans une population.

On l'obtient par le recensement des individus malades de la collectivité. C'est donc un paramètre qui nous renseigne sur l'importance d'une maladie ou d'une infection dans une population à un moment déterminé. C'est pourquoi on l'appelle **un indice statique**. L'indice prévalence = nombre total des cas à un moment donné pour chaque 1000 personnes dans la population totale

Exemple : Donc La prévalence de tuberculose au Congo (en 2004) est vers $20\%_0$

Quand on multiplie par 1000, on exprime l'indice en %o. Il arrive qu'on multiplie par 100 000: on l'exprime alors en "pour 100 000".

On peut distinguer deux indices de prévalence:

a) **la prévalence instantanée**, celle dont nous venons de parler, qui concerne le nombre de malades recensés à un moment donné

b**) la prévalence de période** qui compte tous les cas ayant existé pendant la période étudiée. Tous deux se calculent par rapport à la même population de référence.

T.P.

Exemple : 1. Dans un village de 3 450 habitants, il y a 79 cas de tuberculose; calculer le taux de prévalence par mille.

2. Dans un village de 3 450 habitants le centre de santé reçoive 152 cas de paludisme sévère en novembre et 74 en février. Calculer la prevelance de période de paludisme pour novembre et février.

B. **L'incidence**

L'incidence est une mesure dynamique, de mouvement. On l'obtient

En dénombrant **les cas nouveaux de la maladie étudiée, dans la population, survenus pendant une période donnée.** La période est souvent une année.

Taux d'incidence = Nombre de nouveau cas/ population totale x1000

Il y a les autres taux qui sont tout à fait semblable a l'incidence = la mortalité (par an), la natalité, la mortalité maternelle etc..

T.P. : 1. Dans un village parmi 256 examens de la peau on trouve 217 positive pour onchocercose. Le village compte 2147 habitants. Quelle est la prévalence?

2. En 1969 dans 7 pays africains avec une population de 38,141,000 on a compte 131,581 cas de rougeole. Calculez le taux d'incidence.

3. Au Congo en 1984 il y avait 1,125,000 naissances parmi une population de 25 millions et 598,000 décès. Calculez le taux de natalité et mortalité.

4. Le nord Kivu on compte 2 785 632 habitants. Pendant 2002 on a compté 54 876 cas de TBC dont 9 471 nouveau cas. Calculer le taux d'incidence et prévalence.

4. Fractions, décimales, pourcentage, pour mille… (Révision de mathématique simple)

Définition de la statistique : Science du groupement méthodique des faits qui se présent a une évaluation numérique.

Nombres, mensuration

1. Pour-cent 60/100 % Pour mille 600/1000 Pour dix milles 6000 par 10.000

2. Fraction 6/10 = 3/5

3. Décimale – la plus facile a comprendre – partie d'une unité 0.6

4. Pourcentile - ou centile – place d'un individu parmi 100 individus. Le pourcentile en poids est marqué comme le « chemin de la santé » dans la carte graphique. Tous les enfants normales doivent se trouver entre ces 2 lignes (le 3eme et 97eme pourcentiles des enfants en bonne santé).

T.P:1. Exprimez en % - 1. 0.5, 2. 0,035, 3. ¼, 4. 1/20, 5. 19/20, 6. ¾, 7. 75pourmille 8. 1.5 9. 1/3 , 10. 0.4

2. Exprimez en termes décimales: 11. 80% , 12 110% , 13 ¾ , 14 2/5 , 15 1/3 , 16 26.5% ,17 75pourmille, 18 25% , 19 1/3 , 20 100pourmille

3. Exprimez comme une fraction 21 80%, 22 26.5%, 23 0.65 , 24 0.5, 25 66% , 26 0.3333, 27 1.75, 28 200pourmille, 29 0.25, 30 80%

Apprendre a formulez votre résultat en mots. 800/1000 = 80% = 8/10 = 0.8 = Huit cent pour mille ou quatre vingt pour cent.

T.P. 4. Exprimer « six cent pour mille » dans les plus grands nombres de moyens possibles.

5. Exprimer en décimale: 61%,100%,1/4, 1/6, 300 $^{0}/_{00,}$ 78%, ½, 5/8, 9/15 , 75pourmille

5. Variables qualitatives et quantitatives:

On recherche un phénomène, un trait, une propriété, caractère et ceci sont les **variables**. (A ne pas confondre avec la variance – en haut). Un autre nom pour les variables est les données.

La valeur de chaque caractère est à découvrir.

Une valeur aléatoire est une valeur qu'on ne peut pas prédire avec certitude. (Aléatoire veut dire au hasard)

Les caractères quantitatifs sont ceux avec une valeur numérique. Les qualités ne sont pas numériques mais on peut donner un numéro a une qualité pour qu'on puisse l'analyser par les moyens mathématiques.

Il y a 2 Caractères quantitatifs – **continue** (taille, poids, vitesse…..) = nombre exact avec virgule ou fraction. Pour les valeurs continues il se peut que ce soit plus claire, plus facile à étudier, si on groupe ces résultats (donnes ou variables) en classes ou tranches (par exemple hommes entre 40ans et 49ans).

- **- discontinue** (ou discrètes ou nominales)(no d'enfants, no lits…….). Discontinue veut dire un nombre exact. Ceci sont les valeurs quantitatives nominales (= discrètes ou discontinues) (par ex, nombre d'années d'études, nombre de bouteilles de coke, nombre des lits). Donc ce sont toujours les nombres arrondis.

Les valeurs qualitatives – Les études peuvent mesurer les variables qualitatives. Les valeurs qualitatives ont besoin d'être codifiées, c'est à dire exprime par un numéro. Le sexe puisse être exprime 1=homme 2=femme. (N.B. Le sexe est une valeur dichotomique (di- veut dire 2)– parce qu'il n'y a que 2 possibilités.)

Qualitatives – codifiées – 1 célibat, 2 marié, 3 veuf etc…. Années d'écolage.

La différence entre qualitative discontinue et qualitative est parfois difficile a comprendre. Tous les deux sont un numéro exact (arrondi).

Les variables doivent être cohérentes pour faire une comparaison. C'est à dire tous les données appartiennent

au même groupement pour permettre les chiffres décrire ce groupement. Impossible a comparer les oranges et les pommes – ils sont différents ! Les valeurs sont **homogènes** – quand il traite le même sujet

Les valeurs doivent être précises et établies d'une façon systématique.

T.P. : 1. On veut savoir le besoin en lits a la maternité.

Dans un hôpital on a eu 905 accouchements dans l'année 2000. Ceci représente combien par mois ? Par semaine ? Par jour ? Si tous les mamans restent à l'hôpital pour un moyen de 4 jours on a besoin de combien de lits ?

2. Les suivants sont quels caractères (Quantitative continue, discontinue ou qualitative) ? 1. Sexe d'un malade, 2. Age d'un enfant 3. Nombre de lits dans un hôpital 4. Profession d'un consultant 5. Région d'origan d'un malade 6. Tp d'un malade 7. Durée de vie d'un cancerese 8. Salaire d'un ouvrier 9. Population d'un pays 10 Etat civil d'un malade.

6. La Moyenne La Médiane Le Mode

a) **Moyenne = somme des valeurs divise par nombre des cas**

T.P. A calculer : L'age moyen de 10 élèves en G1 : 20 23 23 29 21 24 30 23 22 19

Moyen = μ (Grec m minuscule)

μ = x1 + x2 +……/ N x1 = l'age du premier étudiant (20 ans), x2= l'age de la prochaine (23) etc.. N= le nombre total des étudiants = 10 dans cet exemple.

On exprime la moyenne en termes universels comme $\mu = \Sigma x / N$ $\sum$ = S majuscule en grec et veut dire « la somme de ». x = chacun des 10 ages. Divisé par le nombre d'étudiants total = N = 10 (dans cet exemple)

.

Avantage de la moyenne

C'est un paramètre parfaitement compréhensible pour le public et son calcul est simple.

Inconvénient de la moyenne

Avec peu des données (distributions à faible effectif), le calcul de la moyenne est très influencé par les valeurs les plus grandes et les plus petites. C'est donc un paramètre insuffisant, qui devra être complété par d'autres.

Souvent on estime la moyenne par l'étude d'un échantillon. Exemple : Pour estimer l'age moyenne d'une classe on doit savoir l'age de tous les étudiants pour trouver μ mais on peut calculer pour 10 étudiants pris au hasard pour dire que leur moyenne estimée (m) sera semblable au classe complète.

b. Le mode: Le mode est la valeur de la variable à laquelle correspond l'effectif le plus grand (= la fréquence la plus élevée). Dans le cas d'une série groupée en classes, on parle de classe modale, celle qui correspond à la plus grande fréquence (cas d'une variable continue).

Le mode des âges de 10 élèves en G1 : 20 23 23 29 21 24 30 23 22 19 = 23 (23 apparaisse 3 fois)

Avantages du mode :

Sa détermination est immédiate aussi bien sur le graphique que sur le tableau statistique. Sa signification est évidente, car il est intéressant de connaître la valeur de la variable qui revient le plus souvent au cours des observations.

Inconvénients du mode :

Le mode n'a de signification que si l'effectif correspondant est nettement supérieur aux autres effectifs. En outre, une série statistique peut posséder plus qu'un ou plusieurs modes. Le mode perd alors beaucoup de sa signification. Il se peut aussi que le mode n'existe pas.

Exemple d'une série avec 2 modes : 20 23 23 21 19 25 23 25 22 25

c. La médiane et les quantiles

La médiane est la valeur de la variable qui se trouve au milieu d'une distribution quand les données sont rangées par ordre croissant ou décroissant. Autrement dit, la médiane est la valeur de la variable telle que la somme du nombre de toutes les valeurs qui se trouve en dessous est égale à la somme de toutes les valeurs qui se trouvent au-dessus. Pour la trouver c'est le donnée tout au milieu d'un groupe impair ou entre 2 résultats d'un groupe en nombre pair.

La médiane des âges de 10 élèves en G1 : 20 23 23 29 21 24 30 23 22 19 – arrangé en ordre 19 20 21 22 23 23 23 24 29 30 - nombre pair donc entre le 5eme et 6eme =23

Avantage de la médiane

Elle est indépendante des valeurs les plus grandes et les plus petites: elle prend donc un grand intérêt dans la description des séries très asymétriques.

Inconvénient de la médiane

Elle n’est pas aussi familière au public que la moyenne.

On peut également marquer la médiane sur les graphiques

Les quantiles sont les divisions d’une série de données pour que chaque division ait le même nombre des données.

La médiane est parfois appelée quantile d'ordre 2, parce qu'elle divise la série en deux parties égales.

Exemple: le cas de l’age des enfants à charge d’un personnel. Classons les 83 données (enfants) dans l'ordre croissant:

1 1 1 1 1 11 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 4 4 4 4 4 4 4 4 4 4 4 4 5 5 5 5 6

Il y a 83 enfants (données).

La médiane est donc la valeur se trouvant au milieu, soit, ici, la 42eme. Il y en a 41 au-dessous et 41 au-dessus. La 42ème donnée est 2. La médiane est donc 2ans.

Il existe d'autres quantiles. Ce sont, principalement, les quartiles, les déciles et les centiles. Nous avons déjà dit que la médiane divise une série en deux parties égales. De même, les quartiles sont les 3 valeurs de la variable qui divisent la série en 4 parties contenant chacune le même nombre d’observations. Les déciles sont les 9 valeurs qui divisent la série en 10 parties contenant chacune le même nombre d'observations. Les centiles sont les 99 valeurs qui divisent la série en 100 parties contenant chacune le même nombre d'observations.

T.P. Dans l’exemple en haut on veut diviser les enfants en 4 groupes – donner les 3 quartiles d’age qui décrivent ces 4 divisions.

On utilise les centiles dans la carte graphique des poids des enfants normales. Ces 99 valeurs décrivent les poids de 100 enfants typiques normaux. On l’appelle « la route de la santé ».

TP 1. Dans la maternité de Nyankunde en 2000 on a eu les accouchements suivants :

J81, F 68 M73 A78 M80 J78 J82 A75 S73 O73 N72 D73

Trouvez la moyenne, la médiane et les modes des accouchements.

2. Dans un examen de pharmacologie les étudiants ont gagné les points sur 20 suivants :

10.1 12.7 10.9 10.5 11.3 12.3 12.8 11.6 10.9 13.8

Trouvez la moyenne, la médiane et les modes de leur points.

7. Graphiques:

Une graphique doit toujours avoir 1. un titre, 2. Les coordonnées (axes, y et x, avec échelles), 3. Une trace (des résultats).

Il y a plusieurs types des graphiques : 1. Graphique a courbes. 2. Graphique en colonne = Histogramme (en colonnes (par groupe). Une colonne dans un histogramme montre les résultats des groupes qui sont semblables. Par exemple une colonne puisse être les gens de 20 a 29ans. Donc la prochaine colonne doit être les 30 a 39 etc.. On appelle ceci « l'intervalle de classe ».

Donc il y a une limite supérieure, une limite inférieur et un centre de chaque classe.

Exemple : 1. Résultats des examens : nombre des étudiants avec 45-49 points, 50-54, 55-59 etc….

2. Histogramme de la pluie journalière pendant octobre.

3. On peut exprimer le résultat comme un diagramme en cercle.

4. Graphique a courbes de la mortalité à cause de TBC en France entre 1900 et 2000. A partir de ce graphique on peut formuler une hypothèse pour expliquer la courbe.

5. Carte graphique des enfants (Enf et Santé p108). La courbe de poids

Comment être sûr que l'enfant grandit normalement et est en bonne santé? Déjà l'aspect général de l'enfant est révélateur de son état de nutrition et de santé: il est robuste, fort, sa peau est souple, ses muscles fermes; il joue, il ne pleure pas à tout moment, et il a bon appétit.

Mais un moyen certain pour affirmer le bon état de santé et de nutrition est la pesée mensuelle lors de la consultation préscolaire. Si le poids de l'enfant se trouve dans le "bon chemin » et augmente régulièrement, la nutrition et santé de l'enfant sont bonnes.

Courbe de poids et santé de l'enfant – utilité :

L'enregistrement du poids de l'enfant peut servir à 1.surveiller le rythme de croissance

2. détecter les premiers symptômes révélateurs d'une carence en calories ou en protéines

3. évaluer les effets d'un traitement destiné à corriger ces carences 4. juger du succès ou de l'échec des programmes d'éducation nutritionnelle.

Comme un médecin se base sur certains signes cliniques extérieurs de la maladie, tels que l'évolution de la température pour l'évolution de la maladie le médecin utilise le poids pour évaluer le rythme de croissance. Le corps, comme tout organisme vivant, grandit depuis sa naissance jusqu'à la taille adulte. Sa croissance est particulièrement intense durant les cinq premières années de la vie. Ses besoins nutritionnels sont exigeants. Chez l'enfant de moins de cinq ans, si le poids n'augmente pas normalement, c'est que quelque chose ne va pas. Le ralentissement, et par conséquent le retard de croissance, constitue le premier signe d'un danger imminent de malnutrition latente. Durant cette période, croissance et santé sont étroitement liées. Un facteur infectieux ou une carence alimentaire se manifestent aussitôt par un retard ou un arrêt de croissance. Les symptômes de malnutrition franche surviennent donc le plus souvent après une période plus ou moins longue pendant laquelle le poids de l'enfant est demeuré stationnaire. C'est cette période de ralentissement de la croissance qui permet d'avoir l'attention attirée sur le problème de la nutrition et qui facilite la prévention des formes graves.

En résumé, nous retenons qu'un enfant bien nourri est en bonne santé, ce qui se traduit par une courbe de poids excellente: "bonne alimentation = bonne santé = bonne croissance = bon chemin sur la courbe de poids".

Présentation de la courbe de poids (système proposé par Jelliffe)

Voici comment se présente la courbe de poids se trouvant sur la fiche de consultation préscolaire (v. plus loin, schéma I).

a. Cinq rectangles se suivent, divisés chacun en 12 parties dans le sens de la largeur et 18 parties dans le sens de la hauteur. Ces divisons horizontales et verticales leur donnent l'aspect de grilles.'

Chaque grille représente une année de vie

- la première grille à gauche = de 0 à 1 an
- la deuxième grille = de 1 à 2 ans
- la troisième grille = de 2 à 3 ans
- la quatrième grille = de 3 à 4 ans
- la cinquième grille = de 4 à 5 ans.

b. Les cases inférieures de chaque grille servent à indiquer les douze mois de l'année. La colonne de droite de chaque grille est numérotée de 2 à 19 = le poids de l'enfant, en kilos.

C. Chaque grille est traversée par deux lignes grasses obliques. Elles partent de l'extrémité inférieure gauche et suivent une direction ascendante. Elles sont presque parallèles au départ, mais s'écartent l'une

de l'autre en progressant vers l'extrémité de droite.

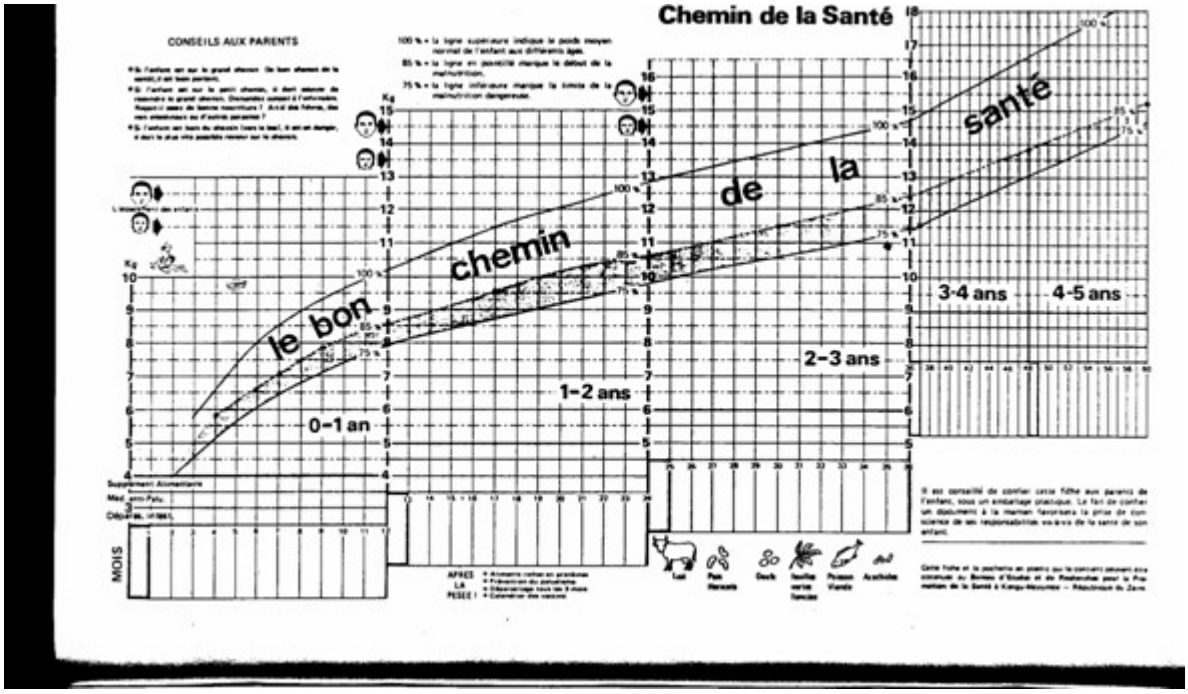
d. Entre ces lignes grasses, trois lignes obliques, plus minces, suivent la même direction, allant de gauche à droite.

e. Au centre de la troisième grille se trouvent cinq chiffres superposés les centiles (voir en haut):. Chaque chiffre désigne une ligne de poids. Ordinairement la ligne maximale est 95% des poids moyens jugé normaux et pris de référence, donc 95% des enfants normaux. Le ligne inférieur est 5% des enfants normales de cet age – ceux qui sont petits mais tout à fait normaux. On trouvera ci-dessous la signification des espaces situés entre ces lignes, espaces appelés, lés "routes" ou "chemins". En pesant l enfant au moins une fois par mois, on voit en effet que les points qui représentent son poids forment une ligne propre à l'enfant. Cette ligne de poids semble suivre l'un des "chemins" qui se trouvent devant l'enfant (à sa droite sur le graphique).

f. A l'extrémité supérieure gauche se trouve un "tableau alimentaire". Ce tableau attire l'attention de la maman et de l'éducateur sur l'importance du régime alimentaire de l'enfant et sur le rôle fondamental que joue l'alimentation dans la croissance de l'enfant.

g. Dans le coin inférieur droit, on trouve une petite grille: «Prévention du paludisme". Elle présente cinq divisions horizontales = les cinq premières années de la vie de l'enfant, et douze divisions verticales = les douze mois de ces cinq premières années. Ces petites cases sont destinées à noter chaque mois si l'enfant reçoit un médicament préventif de la malaria. Actuellement, on abandonne la chimioprophylaxie systématique du paludisme chez tous les enfants de 0 à 5 ans. On la réserve à des enfants à risque: anémie falciforme, enfant fragile... Pour tous les autres, on propose plutôt un traitement présomptif de toutes les fièvres par la chloroquine. Ceci pour retarder l'apparition d'une résistance à la chloroquine.

h. Au verso voir la composition familiale, les vaccinations, et du développement etc.



T .P. ;1. Trouvez un graphique de l’évolution de l’infection de VIH en Afrique.

2. Chercher les graphiques pour montrer les défis du millenium en Afrique.

8. Echantillons

Choix de l'échantillon

Toutes les personnes pouvant être retenues dans une enquête constituent la population de référence, mais il est rare d'en étudier la totalité. Il est plus courant d'en sélectionner un échantillon également appelé population étudiée, de telle sorte que chaque personne appartenant à la population de référence ait une chance égale d'être incluse dans l'étude. De cette façon, la population étudiée sera probablement représentative de la population de référence. Un échantillonnage incorrect ou insuffisant est une erreur fréquente dans les enquêtes.

L'étude de la population entière peut demander trop de temps, de personnel et d'argent. De plus, les dimensions d'une telle étude pourraient être source d'erreurs supplémentaires. Dans certaines circonstances, l'examen de la population entière est cependant inévitable, par exemple, lorsqu'on veut recenser tous les cas survenant au cours d'une épidémie ou lorsque la sélection d'un groupe de personnes serait ressentie comme une discrimination.

Il existe deux méthodes principales pour tirer un échantillon d'une population de référence :

1. **Echantillon aléatoire** (tout à fait par hasard) et

2. **Echantillon systématique**. Exemple s’il s’agit d’une étude maison à maison pour étudier par

exemple la possession d'une moustiquaire ; il existe les tables des nombres aleatoire qui vous indique que vous devrez (par exemple) aller à maison 45 puis 32 ….. Mais c'est plus facile d'aller systématiquement – par exemple chaque 3eme maison.

Donc pour des raisons statistiques, l'échantillonnage aléatoire a plus de chances d'être représentatif, mais l'échantillonnage systématique peut être plus facile à réaliser en pratique.

Donc on doit décider quelle sera l'unité d'échantillonnage. Ce peut être des personnes, des maisons ou des villages, selon les cas. Puis on doit sélectionner le point de départ en utilisant une méthode aléatoire garantissant que toute unité a une chance égale d'être retenue. Cela peut être fait soit par tirage au sort, soit en utilisant une table de nombre aléatoire. Puis on continue en utilisant les nombres aléatoires ou dans une manière systématique (une admission hospitalière sur 3, une maison sur dix dans une rue etc.)

3**. Echantillonnage en grappe:**

Il est souvent impossible d'obtenir une base d'échantillonnage pour des individus. Une solution consiste à utiliser des villages tirés au sort ou des foyers plutôt que des individus. On recommande un tel échantillonnage en grappe, par exemple, la « technique des 30 grappes de 7 unités ». Trente villages - ou grappes de maisons - sont tirés au sort, dans chaque grappe 7 maisons sont alors choisies aléatoirement. Cette technique a été mise au point à l'origine pour estimer la couverture vaccinale, mais elle est maintenant largement utilisée pour toutes sortes d'enquêtes descriptives.

Inconveniences : Cette méthode d'échantillonnage ne donne pas une estimation suffisamment précise pour les maladies rares. Elle ne convient pas non plus pour mesurer des changements intervenant dans l'état de santé d'une population.

Avantages : Les échantillons en grappe ont plusieurs avantages :

1. Il ne nécessite qu'une, base d'échantillonnage simple, par exemple liste des villages.

2. . L'enquête est plus facile et plus rapide car les gens sont regroupés.

3.. C'est une façon de faire qui est souvent mieux acceptée par la population.

Taille de l'échantillon

En général, plus l'échantillon est grand, plus l'estimation de la fréquence sera fiable. En revanche, lorsqu'une précision plus importante est requise ou que la prévalence est faible, un échantillon beaucoup plus grand sera nécessaire. Pour les études mathématiques des résultats (statistiques inferentiels) on a besoin d'un minimum de 30 résultats qui puisse être comparés.

T.P. 1. Pour estimer la couverture vaccinale de BCG dans une aire de santé on visite 210 maisons (dans 2 villages) pour examiner les cicatrices sur le bras des enfants. Dans un village on trouve 95% de 108 enfants avec une cicatrice dans, dans un autre 34% de 183 enfants. Donnez la prevelance d'immunisation dans les villages, et dans l'aire. Donnez quelques explication de ces résultats.

2. Dressez une table de récolte des données pour une enquête sur la relation entre le taux d'hémoglobine et

l'ankylostomiase en fonction de l'âge et du sexe de sujets dans un village.

9. Erreurs et biasis:

1. **Erreurs d'inscription et d'observation**.

Les mesures peuvent être facilement inexactes. C'est en général la faute de l'enquêteur qui mesure mal et non celle des instruments ou des sujets, on parle alors **d'erreur due à l'observateur**. Il peut cependant exister des erreurs dues aux instruments si ceux-ci ne sont pas contrôlés régulièrement, par exemple le zéro ajusté sur les balances. Une autre source courante d'erreur est la mauvaise transcription de l'information sur les formulaires ou les questionnaires d'enquête.

Donc les erreurs puissent être :

1. De l'observateur – expérience, vision, ouïe, personnalité….

2. D'inscription – oublie, vraies erreurs, des vrais semblables

3. Des instruments – manquer de la précision (balance), colorant de laboratoire trop âgé, bandelettes périmés.

4. Faute de calcule, confondre les classes

5. Fautes de réponse de la population.

Les inexactitudes peuvent être diminuées par:

1. La formation soigneuse du personnel et le contrôle fréquent du respect des méthodes.

2. L'observation de directives écrites standardisées et reconnues indiquant, par exemple, comment peser un enfant ou comment poser les questions d'un questionnaire.

3. L'utilisation d'observations dites « à l'aveugle» lorsque c'est possible. Cela signifie que le sujet et/ou l'observateur ne connaissent pas les éléments d'information importants (par exemple, le but précis de l'observation ou de la mesure ou encore s'il est possible qu'un enfant soit ou non sous-alimenté) qui pourraient les amener à biaiser leurs réponses ou leurs techniques par des à priori.

4. L'obligation pour chaque enquêteur d'apposer son nom sur le compte-rendu de chaque interrogatoire, examen clinique, mesure ou test biologique afin qu'on sache clairement qui l'a fait. Ceci incite à un travail plus précis et facilite le contrôle des dossiers.

5 .La vérification des instruments de mesure au moins une fois par jour à l'aide d'une unité connue, par exemple les balances pour nourrissons devraient être contrôlées avec un poids, toujours le même, de 10 kg.

2. **Erreurs de taux de réponse**

Même si les échantillons sont bien choisis, le résultat des enquêtes peut être faux si une proportion importante des foyers ou des individus n'est pas contactée ou ne répond pas aux questions. C'est ce qu'on appelle les **non-réponses**. Un biais peut être introduit par la sélection de ceux qui sont vus et l'oubli des absents. Par exemple, une enquête dans un village d'une région rurale, si elle est effectuée dans la journée, peut ne pas tenir compte des jeunes hommes ou femmes travaillant aux champs. Dans les enquêtes sur la lèpre, les patients atteints peuvent être délibérément évasifs ou ne pas se présenter du tout, on trouvera alors une prévalence faible. Inversement, les gens ne peuvent se présenter que s'ils pensent en retirer quelque chose comme dans les enquêtes nutritionnelles donnant lieu à une distribution gratuite d'aliments. Ceux qui ne sont pas vus peuvent avoir autant d'importance que ceux qui sont vus. Dans les enquêtes portant sur des maladies fréquentes, l'importance des non-réponses peut être moins critique que pour les maladies rares. Mais les problèmes du mauvais échantillonnage et d'un taux de réponse insuffisant s'appliquent à toutes les enquêtes.

Dans toute enquête, il est donc nécessaire de 1.Voir au moins 80 % de l'échantillon original. 2. Repérer tous les non-répondeurs au moins une fois.

Exemple : Etude tous les VIH a Oicha

On va compter :	**On va manquer :**
Symptômes	Sans symptômes
Plaintes	Pas des plaintes – (Timide stoïque…)
Proche au Service de Santé	Trop loin de S de S
Maladie soupçonnée	Erreur du médecin
Diagnostic établi	Diagnostic manque (trop tôt, trop tard – mort ..)
Cas rapporté	Cas non rapporté
Inclus dans l'étude	Par faute – exclue

3. Erreurs de reproductibilité

La reproductibilité d'une mesure est l'aptitude à reproduire régulièrement la même information lors d'examens répétés dans les mêmes conditions et dans la même population.

Même les mesures les plus simples sont sujettes à erreur, parfois à un degré étonnant. Les erreurs relatives à l'exécution des examens, déterminent la reproductibilité de la mesure, tandis que celles qui sont inhérentes à la méthode elle-même déterminent sa validité.

Plus la méthode est fiable, plus les données seront reproductibles. Si la variabilité d'une méthode conduit à des fluctuations aléatoires, on pourra méconnaître une relation existante, mais on ne pourra pas conclure faussement à une relation inexistante. D'un autre côté, s'il existe une sur- ou sous-estimation régulière de la valeur réelle, appelée **biais**, des conclusions erronées sont probables ; cela est possible lorsque les mesures

sont régulièrement plus basses ou plus élevées que ce qu'elles devraient être.

La reproductibilité d'une mesure peut être affectée par :

1. La variation liée à l'observateur. Cela peut se produire, que les observations soient faites par une même personne (**variation intra-observateur**) ou par des personnes différentes (**variation inter observateur**). Un exemple en est la variation bien connue dans l'aptitude des techniciens à déterminer la présence de parasites du paludisme sur une même lame.

2.La variation liée au sujet. La réponse à une question peut être affectée, par exemple, par les motivations et croyances du sujet et par le lieu de l'entrevue.

3.La variation liée aux instruments et aux méthodes. Certains sont de toute évidence plus fiables que d'autres.

Autrement dit les biasis possibles sont:

1. d'échantillon (– plus de femmes qu'hommes…… a l'hôpital (hommes stoïque??)) Déplacés traites gratuitement ou

2.. d'estimateur Trop d'enthousiasme pour une maladie…

10. Les questionnaires

Les questionnaires peuvent paraître simples mais en fait ils sont étonnamment difficiles à concevoir. Ils sont utilisés, habituellement par un enquêteur, pour recueillir des informations, par exemple, sur ce que les gens ont fait récemment, ce qu'ils mangent, les maladies qu'ils ont eues, les décès qui sont survenus et où ils vont se faire soigner. Ces informations seraient impossibles à obtenir d'une autre manière. Il est plus facile, par exemple, de demander à quelqu'un où il s'approvisionne en eau que de l'observer pour le découvrir. Il faut se rappeler cependant, que ces informations correspondent à ce que les gens prétendent, ce qui peut être très différent de ce qu'ils font en réalité.

Il y a les **questionnaire auto-administré** ou formulaire d'enquête.

Les questionnaires posent fréquemment les problèmes suivants

1. Mauvaises questions, peu claires, mal formulé et comportant en réalité plus d'une question. Chaque question doit être simple, claire et ne pas susciter la méfiance.

2. Questions orientées pouvant influencer la réponse. Les questions ne devraient pas suggérer de réponse.

3. Questions délicates ou personnelles favorisant des réponses évasives. Commencer par des questions générales, passer ensuite aux questions plus délicates.

Mesure des variables :

Lorsque les variables ont été choisies, l'étape suivante consiste à prévoir comment elles seront mesurées sur le terrain.

Chaque variable doit répondre à deux exigences:

-Une bonne définition.

-Une bonne méthode de mesure.

La maladie a une signification différente selon les personnes. Par exemple, ce qu'une personne appelle « rhume banal »peut être interprété par une autre comme une « grippe». Ces différences de perception peuvent conduire à des situations où l'appréciation des variables diffère selon les personnes, c'est-à-dire que les résultats ne sont pas reproductibles.

Il est donc nécessaire de définir toutes les variables clairement et au moyen de critères qui en permettent une mesure objective. Le paludisme, par exemple, pourrait être défini comme 1. la présence de Plasmodium dans le sang circulant du patient, 2. identifié sur un étalement sanguin, ou 3. comme une splénomégalie chez l'enfant, ou 4. comme une fièvre avec frissons, ou 5. une combinaison de ces éléments. On doit établir dans les explications (méthodes) de l'étude « **la définition opérationnelle**» que le rechercher à utilisé. Lorsqu'on formule la définition opérationnelle des variables, on devrait toujours être conscient que seules des techniques simples et standardisées sont applicables à grande échelle. Les techniques d'examen sophistiqué comme celles qui sont utilisées dans les hôpitaux sont souvent peu pratiques. On doit admettre que des techniques simplifiées peuvent omettre un petit pourcentage de cas ou indure des non-cas, mais il est tout aussi important de s'assurer que les

résultats sont reproductibles.

Dans le choix des méthodes de mesures, deux aspects doivent être considérés. Ce sont :

1. la précision ou reproductibilité (voir en haut)

2. la validité de la mesure.

La validité fait référence à la capacité d'un test à diagnostiquer correctement la présence ou l'absence de la maladie envisagée.

Une définition stricte du cas, de la maladie ou de l'événement étudié est d'une importance extrême pour

obtenir une validité élevée parce que les mots peuvent avoir différentes significations selon les personnes. Un diagnostic exact est aussi important pour l'épidémiologiste qu'il l'est pour le clinicien. Mais la tâche du clinicien est de répondre à la question : « quelle affection ce patient présente-t-il ?» . Il est libre de pratiquer des examens complémentaires jusqu'à ce que le diagnostic soit certain. A l'opposé, l'épidémiologiste devra présélectionner des critères diagnostiques pour répondre à la question: « Est-ce que cet individu, appartenant à mon échantillon de population présente ou non» de l'affection que j'étudie.

Les critères diagnostiques qu'utilisent l'épidémiologiste peuvent faire appel à un questionnaire standardisé, à un examen clinique, aussi bien qu'à des examens tels que la radiographie (tuberculose), l'électrocardiographie (maladie de Chagas), l'ophtalmoscopie (onchocercose),

l'échographie (hépato - et splénomégalies du paludisme et de la schistosomiase), et l'anatomopathologie (lèpre). Dans la sélection des critères diagnostiques, l'épidémiologiste devra avant tout prendre en considération l'exactitude et la validité des différentes méthodes.

La validité qui présente deux propriétés importantes, d'un examen ou test diagnostic, sont **la sensibilité et la spécificité**. On dit, par exemple, qu'un test a une sensibilité de 90 % s'il détecte 90 % des personnes qui ont réellement la maladie. Par ailleurs, on dit qu'un test a une spécificité de 90 % s'il est négatif chez 90 % des personnes n'ayant pas la maladie.

La valeur prédictive d'un test, qui dépend de la prévalence de la maladie aussi bien que de sa sensibilité et de sa spécificité, est la mesure la plus importante permettant de déterminer l'utilité du test sur le terrain. La valeur prédictive positive mesure la probabilité qu'une personne ayant un test positif soit réellement atteint par la maladie.

11 Statistiques de la population

Les populations changent (Quelques statistiques pour 2003):

Mouvement naturel = bilan des naissances et décès - accroissement naturel (DRC 2.9%/an ; France 0.5%/an)

Mouvement social = immigrations et émigrations - accroissement migratoire

La natalité = Nombre naissances vivantes/ effectif population x 1000 DRC = 45%o (1984)

Taux de fécondité = Naissances vivantes / femmes âgées 15 a 49 (France = 1.89 ; DRC = 6.70)

La mortalité = No décès/ pop x 1000 = 20%o (recensement Congo 1984)

Mortalité infantile no décès 0 – 365j / naissances vivantes x 1000 (DRC= 119.6 ; France = 5.0)

Mortalité neo-natale = no décès 0 – 28j / naiss viv x 1000

Mortalité par cause de décès – OMS classification internationale des maladies.

Létalité : Pourcentage des gens avec une maladie quelconque qui vont mourir.

Mortalité maternelle = Femme décédées / naissance vivantes x 1000

L'espérance de la vie = moyen âge de décès - variabilité mondiale – une indication des conditions sociales sanitaires. (En 2002 Norvège = 78.9, Brésil = 68.0 Soudan 55.5 D.R.Congo = 41.4ans)

Pourcentage de la population ayant au moins 15ans France = 18.6% DRC = 46.9%

Pourcentage de la population 65 et plus. France = 16.1% DRC = 2.6%

Migrations – temporaires, définitives. Population flottante. Etc..

Taux d'alphabétisation. Taux de scolarisation. Parité de pouvoir d'achat par habitant (PPA)

T.P. Au Zaïre en 1975 on a pu relever les données suivantes:

Population: 22 582 230 Naissances vivantes: 981 638 Décès 486 192 Décès entre 0 a 365j 115 672

Calculez le taux brut de natalité, le taux brut de mortalité et le taux de la mortalité infantile.

12. Statistiques sanitaires

Pour l'état de santé d'une population on analyse plusieurs valeurs:

1. Une valeur idéale; 2. une valeur objective (ce qu'on cherche à atteindre); 3. Une valeur d'alarme (qui nécessite une action par le service de santé); 4. une valeur de mesure (ce qu'on mesure sur terrain).

La morbidité = quantité de maladies dont souffre une population. Il y a une déclaration obligatoire des maladies transmissibles; statistiques des zones de santé; qui provoquent les enquêtes sur terrain.

La Prévalence, L'incidence (voir en haut).

L'invalidité (ne peut pas travailler) ou incapacité (de vivre « normalement »). Tous ces deux puisse être total ou partielle, permanente ou partielle.

L'incapacité puisse être définie comme : No personnes avec une incapacité de longue durée/ personnes examinées x 100

Coefficient de fréquentation hospitalière = malades hospitalisées/ pop x 1000

L'indice lits/pop = lits/pop x 1000

L’indice des agents de santé = agents de santé x1000

Journées d’hospitalisation. Coefficient d’occupation des lits moyenne par mois. (Nombre de lits occupe/lits vides).

Le séjour moyen a l’hôpital.

Couverture de l’accessibilité = pop ayant bénéficie des soins / pop ayant besoin x 100

Personnel = no personnel de santé / pop x 10 000

T.P. 1. Pendant l'année 1984, la zone de santé de Nyankunde comprenait 100 000 habitants. 22 centres de santé y fonctionnaient. Le taux de natalité était de 43,62 0/00. Dans 16 centres de santé, fonctionnaient des consultations préscolaires et prénatales. On comptait 17 000 enfants âgés de 0 à - de 5 ans et 13 650 femmes âgées de'15 à - de 50 ans. 9 688 enfants ont été inscrits à la CPS et 3 203 femmes ont été inscrites à la CPN. Il y eu 35 589 consultations préscolaires et 8 199 consultations prénatales. Calculez

1. le % de centres de santé organisant des CPN et des CPS

2. la couverture en CPN et CPS

3. le nombre moyen de consultations de CPN et CPS.

2. Faites à partir des données de l’année passe, un graphique a courbe en mettant les étudiants de l’ISTM en ordre – le plus bas au plus haut. Puis 2. un histogramme des résultats d’examen ISTM Nyankunde avec le nombre d’étudiants dans les groupes de 5% (40 a 45%, 46 a 50% etc..) .

Introduction à la statistique inductive (ou inférentielle)

1. Introduction à la probabilité

Les dés sont les petits cubes dont chaque face est marquée avec de un à six points : Donc 1 à 6 sur chaque surface. On peut jeter (tirer) les dés pour voir quel numéro sort sur la surface supérieure, dans une manière qui est tout à fait par hasard (aléatoire).

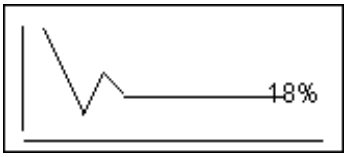
Parmi le nombre de tires du dé, combien de fois un 4 est sur la surface supérieures ?

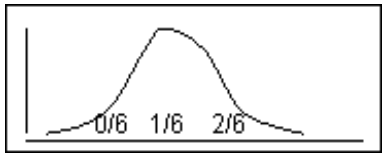
En théorie parmi 6 tirages on pense qu’il y aura 1 quatre ; c’est a dire 1 quatre sur 6 tires. Ceci (1 sur 6) c’est **la probabilité**. Mais avec une distribution au hasard (disons normale) on note que quelques fois il y aura 2 quatres parmi 6 tires ou d’autres fois 0 quatres après 6 tires.

On peut exprimer nos résultats de tire de dé sur un graphique. Le 4 puisse arriver le deuxième tire (1 sur 2) mais de plus en plus qu’on tire au hasard ; les inégalités vont disparaître et on trouvera que par moyen il y a un quatre chaque 6 tires = 1 sur 6 = 1/6 = 0.18 = 18% des fois qu’on tire (jette) le dé.

Ce graphique est un **Graphique de régression** à 18% (0.18)

On peut exprimer le même résultat dans un autre graphique - **Graphique de la densité de la probabilité.** On pense qu'en théorie on doit avoir un 4 pour 6 tires. Donc calculer combien de fois il y a un 4 pour 6 tires, combien de fois 0, combien de fois 2. Le nombre des tires est sur l'axe y et les trois possibilités (0 fois, 1 fois 2 fois) sont sur l'axe x. Le nombre maximal sera 1 sur 6 avec moins avec soit 0 sur 6 ou 2 sur 6. Donc le courbe obtenu c'est un peu comme le contour d'une cloche. Donc on l'appelle « courbe en cloche ». Voyant comme il y a beaucoup de distribution comme ceci on l'appelle le courbe normale. Le premier mathématicien d'étudier cette courbe était Carl Guass, donc on l'appel la courbe Gaussienne. On peut décrire les caractéristiques d'une telle distribution qu'on appelle la loi normale La courbe décrit la densité de la probabilité de l'arrivé des 4s quand on tire un dé.





Régression Densité de la probabilité

T.P. 1 :Pendant qu'un ami jette un dé, 2 autres amis analyse les résultats. A. Un ami note chaque 4 et compte si ceci arrive apres combien de tires (jets). Exprimez le résultat comme un graphique de régression. B. Un autre ami compte chaque le nombre des 4s pendant chaque 6 jets du de. Exprimez le résultat comme un graphique de la densité de probabilité.

2. Une maladie pulmonaire chez les gens qui habite proche à une industrie d'acier qui dégage beaucoup de fumée à travers une cheminée très haute est prévalant dans 15% des gens qui habite 3,5- 3,9km de l'usine : 12% de ceux qui habitent 0 – 0,4km de l'usine ; 14% de ceux qui habitent 4 – 4,4km de l'usine : 14% pour 0,5-0,9km : 14% pour 1- 1,4km : 18% pour 2.5 – 2.9km : 14% pour 1,5 – 1,9 : 16% pour 3-3,4km : et 15% pour 2 – 2,4km . Faites un graphique (y vertical = % ; x horizontale = distance de l'usine) de la distribution (densité) de la probabilité que quelqu'un souffre d'une maladie pulmonaire proche à cette usine.

3. On trouve la température de 10 enfants normales : 36.0 – 36.4 : 1enf : 36.5-36.9 , 2enf : 37.0-37.4 :

4enf : 37.5-37.9 : 2enf : et 38.0-38.4 : 1enf

Température de 10 enfants à l'hôpital : 36.5-36.9 – 1enfant : 37.0 – 37.4 , 2 : 37.5-37.9 ,1 : 38.5-38.9, 1 : 39.0-39, 4 : 39.5-39.9 : 1enfant.

Faites 2 graphiques de la densité de la probabilité qu'un enfant possède une température quelconque. Comment expliquer ces graphiques ?

4. On mesure la quantité exacte de coke (cola) dans 70 grandes bouteilles et on trouve 750ml exacte en seulement 10. On trouve 753ml en 1; 749 en 8; 751,5 en 6; 747.5 en 2; 752 en 4; 749.5 en 9; 751 en 8; 747en 1; 748,5 en 6; 750,5 en 9; 752,5 en 2; 748 en 4. Faites le graphique.

2. La loi normale, Courbe en cloche, courbe Gaussienne, courbe normale:

Les lois mathématiques qui décrivent les caractéristiques d'une distribution normale se voit sur un graphique comme une courbe en cloche (comme la courbe de la densité de la probabilité en haut)..

Courbe normale : Courbe en forme de cloche est symétrique par rapport à la moyenne qui normalement est tout au centre de la courbe. Une courbe en cloche représentant une population qui se distribue normalement. (On l'appelle courbe en cloche parce qu'elle est semblable a la configuration d'une cloche qui sonne.)

Il se peut que dans une étude vous n'etes pas en train d'étudier un seul variable. Il se peut qu'il y ait 2 influences qui peuvent donner naissance à 2 populations différentes. Dans notre exemple des dés de chapitre 1 vous pouvez étudier uniquement le 4 sur le dé ou le 4 et le 5. On peut calculer la possibilité qu'il y ait 2 distributions surimposées. NB que dans le TP 3 du chapitre 1 on a bien séparé les 2 populations des enfants – un à la maison, l'autre à l'hôpital. Chaque population a les caractéristiques différentes en ce qu'on étudie (la température).

Dans la statistique ordinairement on étudie chaque population séparément – mais il y a les moyens mathématiques savoir si ce qu'on étudie est un seule ou deux populations différentes de comparer les 2 populations.

Moyenne, variance et écart type :

Voir vos résultats d'estimations de quantité de coke (TP 1.4). Votre graphique est en forme d'une cloche. Ceci est la « Courbe de Gauss » ou la « Courbe Gausienne » ou la « courbe normale » ou la « courbe en cloche. ». Les règles mathématiques qui gouverne une telle courbe s'appelle la « loi normale »,ou « loi de Gauss » ou « loi Laplace Gauss ». Le graphe de cette fonction est une courbe en cloche. Quelques exemples d'une telle distribution sont : 1. Les résultats d'un examen. 2. Le poids des bébés a la naissance, 3. Le nombre de cas pendant une épidémie d'Ebola. 4. Le nombre d'étudiants qui sont soit en avance ou en retard pour une session d'enseignement. Donc la majorité arrive à l'heur mais quelques-uns uns, peu, sont en avance et un nombre semblable en retard. etc. etc.

Ceci est la forme d'une distribution des données la plus fréquente. Par exemple dans les examens on trouve la grande plupart des étudiants sont médiocres (représentés par la hauteur centrale de la courbe) , quelques-uns uns très intelligent (représente par le peu au fin de la courbe) et un petit nombre sont en train d'échouer (représente par le peu au commencement de la courbe).

Ce petit nombre qui sont **sur les bords d'une distribution** est une estimation de la répartition de la population (ou la distribution) et on appelle cette répartition la **variance**.

La variance est toujours en bas et en haut de la moyenne c'est à dire au tour de la moyenne. La **quantité** d'une distribution est montré par l'hauteur de la courbe tandis que la **qualité** de la distribution se montre par la largeur de la courbe. On peut appelé cette variance « l'écart » de la moyenne (L'écart d'un résultat = le distance qui le sépare de la moyenne.) (On donne le symbole σ « s » en grec pour representer un ecart). Par convention on decrit un ecart standard (normale ou classique) de 34,13% de la moyenne, qu'on appelle **l'écart type**.

Par une autre convention les mathématiciens ont décidé que la mesure mathématique de **la variance <u>= la moyenne des carrés des écartes entre les valeurs d'x et leur moyenne.</u>** C'est à dire :

Un écart = le nombre – la moyenne. Par exemple si vous avez eu 75% en examen de mathématiques et le moyenne pour toute votre classe est 55% votre écart est 75-55 = 20. = ($x - \mu$) (NB μ grec m est le symbole universel de la moyenne) Maintenant faire le carré de l'écart = 20x20 = 400. (= $x - \mu)^2$ Maintenant faire le même calcule pour tous les étudiants dans la classe, puis additionner les écartes carrés. (Le symbole Σ en grec est « s » majuscule et veut dire la somme de). Maintenant vous devrez trouver la moyenne de tous ces écartes carres, donc diviser par le nombre de résultats que vous avez analyse (n= le nombre d'étudiants en classe) donc: la moyenne des écartes carrés = la variance = $\Sigma(x-\mu)^2/n$

On donne le lettre grec σ (s minuscule) pour les écarts – donc la variance est σ^2 (Lire sigma 2 ou sigma carre) = l'écart carre. Pour trouver le vrai écart on prend la racine carré de ce numéro.

L'écart type = $\sqrt{\Sigma(x-\mu)^2/n}$

Le racine carré de la variance est un chiffre important dans les statistiques – il s'appelle **l'écart type**. Ce chiffre est très intéressant parce qu'on trouve que ceci comprend toujours 34.13% des résultats a chaque cote de la moyenne.

Exemple: Si la moyenne dans un examen est 50% et l'écart type 45%, ceci veut dire que 64,26% des étudiants ont reçu entre 45 et 55%.

Estimations :

Supposons qu'on veut savoir l'age moyen de la population d'une ville. Evidement c'est impossible de contacter *chaque* personne dans la ville donc on prend un échantillon dans l'espoir que cet échantillon est représentatif de la ville entière. Donc nous faisons notre estimation de l'age moyen. Pour différencier cette estimation de la vrai moyen on utilise les symboles suivants :

1. μ c'est le vrai moyenne d'une distribution.

2. m est notre estimation de la moyenne

3. σ est le vrai écart type

4. s est notre estimation de l'écart type

Révision :

Moyenne = $\mu = \frac{\Sigma x}{N}$

Ordinairement on ne peut pas compter tous les malades dans un pays, donc on compte un échantillon (soit représentative ou pris tout à fait par hasard (aléatoire)) on fait la moyenne de cet échantillon =

$m = \frac{\Sigma x}{N}$

Variance = la quantité par lequel les (ou un) résultats varient de la moyenne

variance (dispersion) = σ^2 par définition = $\Sigma(x - \mu)^2$

On fait l'estimation de σ^2 dans un échantillon = $s^2 = \Sigma(x-m)^2$

La racine carrée de la variance s'appelle l'écart type.

On trouve que σ = Ecart Type = 34% des résultats = l'écart de la moyenne (Donc de chaque coté de la moyenne = 2 fois σ = 68% des résultats)

TP. Faites 100 prise de la TA diastolique. Faites un graphique qui montre vos résultats. Probablement le plus grand no des gens sont au milieu au tour d'une tension diastolique de 80mmHg. Donc on l'appelle ceci la densité de la probabilité parce que c'est le plus probable que vous serez au milieu, vers la moyenne de 80mmHg. Il y aura les variances que nous pouvons appelle hypotension ou hypertension. Si on accepte que 68% des résultats soient normales les hypo et hyper tension seront plus grande ou moins grande que l'écarte type.

3. Ecart Type (Révision et exemples)

Dans la courbe en cloche on voit la moyenne, et la variance (la dispersion)

Par exemple l'arrivé en classe des étudiants le 2 avril 03 après la pluie était à 13.55 : 1etudiant, 14.03 1, 14.04 1, 14.05 3, 14.06 4, 14.07 3, 14.10 2, 14.11 2, 14.16 1, 14.20 1. Faites une courbe en cloche et calculer l'écart type le 34% des étudiants qui arrivent à chaque cote de la moyenne.

La variance = σ^2 Racine carrée du variance = écart type $\sqrt{\sigma^2} = \sigma$

L'écart type = 68% des données. Il est exprimes dans les même unités que les données.

L'écart type = est 34% a chaque cote de la moyenne.

Les mesures de dispersion les plus courantes s'appuient sur la mesure des écarts entre chaque donnée et la moyenne, $(x-\mu)$.

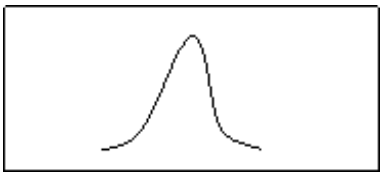
En effet, lorsqu'on connaît la valeur de cette distance moyenne, on peut conclure que plus celle-ci est grande, plus les données sont dispersées et plus l'échantillon est donc hétérogène. En revanche, on peut être assuré que plus cette distance est courte, plus les données sont donc concentrées autour de la tendance centrale et plus on a affaire à un échantillon homogène.

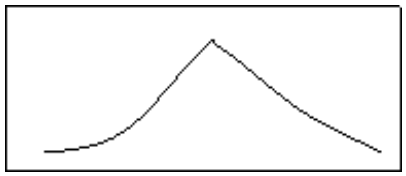
Hétérogène

(du grec hétéro = autre et-genos = origine). Se dit d'un échantillon dont les données sont largement dispersées sur l'échelle de mesure de la distribution, ce qui se traduit par un écart type important et indique que les données sont différentes les unes des autres. Il se peut qu'on soit en train d'étudier 2 populations, pas une seule. Donc dans notre exemple en haut il se peut que les étudiants continue à arriver jusqu'à 15hrs. Mais a 15hrs on est en train de compter les étudiants qui arrivent pour la classe a 15hrs et non PAS pour la classe de 14hrs – donc une autre population.

Homogène

(du grec homo = semblable et-genos = origine). Se dit d'un échantillon dont les données sont concentrées autour de la moyenne ou de la médiane, ce qui se traduit par un petit écart type et indique que les données différant peu les unes des autres.





Population homogène Population hétérogène

Parmi 6 de nos étudiants, ils ont arrives comme suit: (14).03, 05 06 09 11 14, (après 14 heures) on calcule la moyenne de cet échantillon, soit:

$\frac{3 + 5 + 6 + 9 + 11 + 14}{6} = \frac{48}{6} = 8$ minutes = 14h08 est la moyenne heure d'arrivée

puis on calcule la distance entre chaque donnée et la moyenne et on fait la somme des différences

-5 -3 -2 +1 +3 +6

(3-8)+(5-8)+(6-8)+(9-8)+(ll -8)+(14-8)

La variance est la somme des carrées des ces écarts

$= (-5)^2 + (-3)^2 + (-2)^2 + (+1)^2 + (+3)^2 + (+6)^2 = 25+9+4+1+9+36 = \frac{84}{6} = 14$

Donc la variance est 14 minutes (le résultat obtenu de cette façon qui correspond à la formule : $\Sigma(x-\mu)2/n = \sigma^2$)

Donc l'écart type est la racine carre de ce numéro = $\sqrt{\sigma^2} = \sigma = \sqrt{14} = 3{,}74$

$= \sqrt{\Sigma(x-\mu)2/n} = \sigma$

D'après les données de notre exemple, l'écart type = 3,74. C'est à dire 68 % des étudiants sont arrives entre 14h.08 $\pm$ 3.74 minutes. (Parmi 6 étudiants 68% sont arrivés 3,74 minutes avant ou après 14h.08)

Il faut cependant encore ajouter qu'afin d'obtenir une meilleure estimation de l'écart type pour des petits échantillons, C'est-à-dire dont le nombre de données est inférieur à 30, on divise par n - 1 plutôt que par n. Ainsi, la formule de l'écart type d'un échantillon s'écrit : $\sqrt{\Sigma(x-\mu)2/n-1} = \sigma$

La variance constitue un indice de dispersion utilisé dans certains tests statistiques.

Le symbole de l'écart type d'une population est représenté par la lettre grecque sigma σ alors que dans le cas d'un échantillon, on le représente par la lettre s. Il en va de même pour la variance, qui correspond au carré de l'écart type et qui est donc représentée par le symbole σ^2, dans le cas d'une population, et par s^2 pour un échantillon.

Quelques définitions à comprendre :

Variance

Indice de dispersion des données représenté par la moyenne des carrés des écarts de chacune des données par rapport à la moyenne de la distribution. La variance constitue le carré de l'écart type.

Indice de dispersion

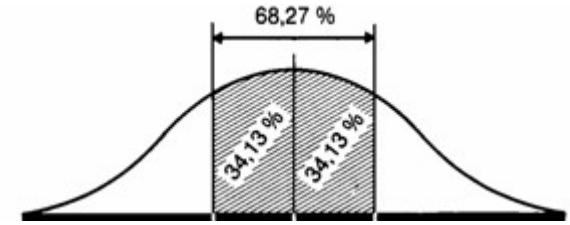
Grandeur mesurable qui traduit la manière dont les données s'éparpillent, se dispersent sur l'échelle de mesure de la distribution. La variance et l'écart type sont les indices de dispersion les plus utilisés.

Écart type

Indice de dispersion le plus utilisé. Il représente la racine carrée de la variance et est symbolisée par la lettre grecque σ (sigma) lorsqu'il se rapporte à une population et par la lettre s dans le cas d'un échantillon.

Σ (sigma majuscule)

Dix-huitième lettre de l'alphabet grec symbolisant le processus de sommation dans les formules mathématiques lorsqu'elle est majuscule, et l'écart type d'une population lorsqu'elle est minuscule (σ).



Par convention, dans une distribution on pense que la limite de normal est 95% des résultats d'une distribution.

T.P. 1.: On fait le poids de tous les enfants dans une classe de l'école primaire et on trouve les poids (kg) suivants : 25, 21 26 23 28 24 23 25 28 25 21 24 21 22 25

Calculez 1.la moyenne m =Σx/N 2.la variance = $s^2 = \Sigma(x-m)^2$ et 3. l'écart type. $s = \sqrt{\Sigma(x-m)^2}$ 4. 68% des enfants pèsent entre combien de kilos ?

2. Dans le laboratoire on trouve les résultats suivants parmi 20 personnes :

10 Avec œufs d'ankylostomes hémoglobine de 10.3g/dl 8.4 8.4 10 10.9 8.8 10.9 10.9 9 9.8

10 Sans ankylostomes hémoglobine de 12.7g/dl 9.4 8.3 11.5 8.3 9.7 8.9 11.6 9.2 10.9

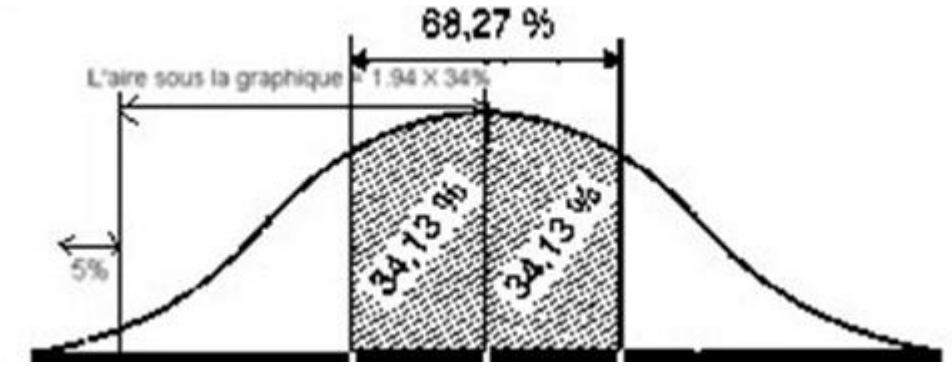
Calculez les moyennes et les écarts types pour voir la différence de ces 2 groupes.

3. On doit faire une comparaison de 2 traitements pour paludisme. Inventer une épreuve de l'efficacité de chloroquine et fansidar donné alternativement aux malades qui arrivent à l'hôpital avec malaria. Estimer la vitesse de chute de la température après ces deux traitements. Comment est-ce que vous pouvez mesurer les 2 écarts types pour comparer les 2 traitements ?

La statistique inferentielle (inductive) vise à indiquer s'il est probable ou non que ces deux échantillons proviennent de la même population.

4. Exemple : On veut savoir si les déplacés de guerre sont malnouris ou non. C'est à dire quand on les étudie avec les autochtones est-ce qu'ils semble d'être dans la même population par moyen mathématiques que les gens locales? A utiliser la circonférence de bras infantile entre l'age de 1-5 ans : (norme plus que 16cm)

Ecart type, 68% On va dire que d'être normale on doit être dans la 95% de la population. 95% = 1.96 fois l'écart type.



Résultats Gp témoin (école primaire) et gp expérimentale (camp de déplacés) Circonférence bras :

15.4 16.2 15.8 15.6 16 16.4 15.2 15.7 15.9 15.8 cm Déplacées

15.8 16.2 16 15.9 16.1 16 15.7 16.3 15.9 16.1cm Ecole Prim

Graphique densité de la probabilité

Calculer l'écart type

La même population? 1.Plus le moyen est diffèrent plus qu'on va penser qu'il y a une vraie différence entre la nutrition des enfants sain et malade. 14 16?

2. Plus grande les échantillons plus de possibilité qu'ils qu'on va penser qu'il y a une vraie différence entre la nutrition des enfants sain et malade.

Aire sous le graphique = la meilleure combinaison de moyenne et variance.

T.P. On administre la même teste d'anatomie à 2 ISTM différentes. Resultats :

ISTM 1 : 4 6 7 6 7 5 6 7 8 5

ISTM 2 : 7 8 9 6 6 4 3 7 4 8

Calculer le moyen et l'écart type de ces deux institutions.

T.P. 15 jeunes sont testés dans une automobile stationnaire pour leur temps de réaction sans ou après la consomption de deux bouteilles de bière. En milliseconds on trouve une réaction:

Avant : 15 11 16 13 18 14 13 15 18 15 11 14 11 12 15 milliseconds

Apres alcool : 17 13 20 18 21 22 19 20 17 19 14 12 18 21 17 milliseconds

Calculez l'écart type en milliseconds.

TP :Le poids de naissance des nouveaux nés moyenne (parmi 200 naissances à Oicha) = 3.3kg. Ecart type = 0.5kg Qu'est-ce que ceci veut dire ?

Il y a 68% des enfants qui sont + ou mois un demi-kilo de 3.3kg.

Question 1 : Calculer la probabilité qu'un nouveau-né ait un poids de moins que 2.8kg ?

Nous savons que 68% sont un écart de 3,3kg = 34% sont + que 3.3 par un écart de 0,5kg (3.3 à 3.7) et que 34% sont moins que 3.3kg (3.3 à 2.8),) donc 16% sont à chaque cote = 0.16 = 16% sont moins que 2.8kg

Question 2 : 2,5 kg ?

Pour la calculer on décide combien d'écart types de le moyenne se ramène à une loi centrée réduite

Z= 2,5 – 3.3/ 0,5 = -1.6 fois l'écart type (Moins 1,6) = par table 1 de la loi centrée réduite (B p 308)

= 0.945 = 94.5% plus que 2,5kg = 0,055 moins = 5.5% moins que 2.5kg

On peut le calculer, mais on a besoin de « calculus » parce que les intervalles sont logarithmiques au lieu d'être simple et égaux.: 1 écart type = 16% 1.96 écart type = 2.5% 1.6 fois écart type = combien ?

X = 16, 1.96x = 2.5 .96 = 13.5 .6 = 13.5/9,6X6 = 8.4%.

T.P. Une étude trouve l'âge de mort des Congolais est en moyen 55ans avec un écart type de 10 ans.

Quel est la probabilité que vous aller mourir avant 45 ans ?

= 16% : 45-55/10 = - 1 = selon les tables, 840 = 16%

Quel est la probabilité de votre mort à 30 ans ?

T.P. Sur un échantillon de 41 sujets on trouve un poids moyen de 58kg. L'écart type est de 12 kg Quels sont la possibilités que quelqu'un a un poids plus que 70Kg ?

4.. Fluctuations d'échantillonnage – estimation par intervalle (intervalle de confiance)

Vous pouvez exprimer vos résultats d'un calcule statistical dans les manieres differentes :

1. Estimation ponctuelle – le seul chiffre (exacte) que vous avez obtenu

2. Estimation par intervalle – L'intervalle dans lequel on peut avoir la confiance que le vrai chiffre réside.

Les chiffres statistiques d'OMS sont toujours exprimer en intervalle. Par exemple le nombre de cas de choléra dans le monde ce mois c'est entre ……. Et……..

Intervalle et variance sont de la même famille.

Une variance standard = un écart type = 68% des résultats dans une distribution normale

Normalement 2 écarts types (plus exactement 1.96 écarts types) = 95% de tous les résultats.

Si on compte tous la population il n'y a pas d'intervalle de confiance. Mais ordinairement c'est pas possible de compter CHAQUE cas de choléra et on doit faire une estimation a partir d'un échantillon qu'on a pu compter. Le plus grand l'échantillon le plus petit l'intervalle, et vice versa.

Intervalle de confiance (certitude à 95%) de la prévalence déterminée par l'enquête en fonction de la prévalence attendue et de la taille de l'échantillon :

Prévalence attendue %	Nombre de personnes dans l'échantillon				
	50	100	200	500	1000
	Intervalle de confiance des estimations de prévalence				
1	-	0- 5	0.1 - 4	0.3- 3	0.5- 2
5	-	2-11	2- 9	3- 8	4- 7
10	3-22	5-18	6- 15	7-13	8-12
20	10-34	13-29	15-26	16-24	18-23
30	18-45	21 -40	24-37	26-35	27-33
40	26-55	30-50	33-47	35-45	37-43
50	36-64	40-60	43-57	45-55	47-53
60	45-74	50-70	53-67	55-65	57-63
70	55-82	60-79	63-76	65-74	67-73
80	66-90	71 -87	74-85	76-84	77-82
90	78-97	82-95	85-94	87-93	88-92

T.P. Selon vous, lesquels des résultats suivants sont significatifs :

Parmi 1000 GE on trouve 48% des hommes positives et 52% des femmes positive.

Parmi 202 sucs dermiques on trouve 43% des hommes positive et 57% des femmes positive.

Parmi 50 bronchitiques on trouve 65% qui fume et 35% qui ne fume pas.

Parmi 450 malnourris on trouve 35% avec tuberculose.

Parmi 2 villages de 500 population chacun on trouve 50% avec schistosomiase dans un village et 37% dans

l’autre.

Calcule de l’Intervalle de confiance

L’estimation ponctuelle consiste à attribuer une valeur au paramètre étudie à partir des observations faites sur l’échantillon. Mais ordinairement on fait une estimation d’intervalle parce qu’on ne peut jamais dire qu’on a compter tous.

Paramètre	Valeur théorique (absolu, réel)	Estimation en pratique
Pourcentage	P	P_o
Moyenne	μ	M
Variance	$\sigma 2$	S2
Ecart type	σ	S
Coefficient de corrélation	ρ	R

Estimation d’un pourcentage, moyenne, variance

A cause des fluctuations d’échantillonnage l’estimation ponctuelle change. Donc on a plus de confiance dans un intervalle des valeurs plutôt qu’une seule. On appelle ceci l’intervalle de confiance et par convention c’est l’intervalle qui doit inclure 95% des résultats.

Si vous voulez un intervalle de 68% = vous aurez 1 écart type d’intervalle.

L’intervalle habituelle est de 95% = 1,96 fois l’écart type = Zα/2 Z décrit l’aire sous la courbe en cloche, la courbe normale, on estime une variance de 0,025 (2.5%) à chaque coté de la courbe en cloche :

$$\text{L'intervalle} = Po + \text{ou} - Z\alpha/2 \sqrt{\frac{po \times qo}{N}}$$

Po = pourcentage observé Za/2 = 1,96 pour 95% qo = reste du pourcentage observé

N = nombre de l'échantillon

Dans un échantillon de 60 sujets on trouve 18 paludiques. Quel est la prévelance de paludisme ?

18 par 60 = .3 = 30% = votre estimation ponctuelle

Intervalle de confiance à 95% = 0.30 + ou moins 1.96 √ 0.30 X 0.70 / 60

= 0.30 + ou moins 1.96 X√ 0.0035 = 1.96 x 0.06 = 0.12

= 0.30 + ou moins 0.12 = .18 - .42 = 18 – 42% Donc 95% des résultats de paludisme doivent être entre un prévelance de 18% à 42%. On suggère que, même si on teste le monde entier il y n'y aura pas moins que 18% ou plus que 42% de paludiques.

NB dans le table en haut pour un échantillon de 50 avec un prevelace de 30% on liste 18 a 45 comme les intervalles de confiance à 95%.

T :P : Dans un échantillon de 85 enfants on trouve 34 avec ascaris. Quel est l'estimation ponctuelle et d'intervalle de la prévelance d'ascaridose?

T :P : Voir statistiques pour VIH de l'OMS en Afrique

Il y a les tables d'intervalle de confiance (certitude 95%) des prévalences en fonction de la taille de l'échantillon. (V&Mp 78)

T.P. Parmi 100 utilisateurs des moustiquaires on trouve 45 hommes et 55 femmes. Quel est la fréquence d'utilisation des moustiquaires chez les femmes ? Donnez une estimation ponctuelle et par intervalle.

TP Dans des séries de 7 matches pour la coupe d'Afrique Congo gagne 0.6 (60%) des fois contre Cameroun. Qui va gagner le séries ? Dans combien de matches ? Si vous avez un billet pour le 7eme match quelle est la possibilité que vous pouvez l'utiliser ?

5.Hypothèse nul et alternatif

Hypothèse : Supposition qu'on fait pour expliquer une chose mais qui reste à vérifier.

Information pour donner un lien de causalité.

Echantillons dépendent ou indépendant

La première étape d'un test consiste à spécifier une hypothèse.

Hypothèse = une explication qui selon vos connaissances semble expliquer les faits.

Hypothèse : 1.TBC est à cause des bacilles de Koch. Koch a posé cette hypothèse et la prouver par ses postulats :

2. SIDA c'est à cause de VIH. Cueillir les anticorps contre VIH pour voir quel % des SIDA en ont.

OU 3. Doxycycline ne guérisse pas paludisme. Prenez 2 populations, avec paludisme – donnez rien (ou chloroquine) à un ; puis doxycycline à l'autre et suivre leurs courbes de temperatures.

On vérifie cette hypothèse relative à la façon dont se distribuent les données recueillies.

Ordinairement il y a un choix entre 2 hypothèses :

0. Hypothese nulle. Répond à la question oui ou non. =Ho On suppose que les différences vues ne sont pas significatives. Hypothèse selon laquelle les différences sont le fait du hasard et n'ont aucune signification. Si on trouve que la différence est significative on rejet l'hypothèse. Donc elle est habituellement formulée dans le but de la voir éventuellement rejetée au profit de l'hypothèse alternatif.

Exemple : Est-ce qu'il y a moins d'infection si on lave les mains ?

1. H. Alternative. Répond à la question est-ce que ces deux séries sont différentes dans une façon significative? = H1. Il y a 2 distributions différentes on veut voir si c'est de façon significative.

Exemple : Les praticiens lavent leurs mains 7 fois sur 10, les médecins 3 sur dix – est ce que cette différence est significative ?

Ordinairement ce qui rejet Ho est au profit de H1 qui peut être accepté.

On général on n'accepte ou rejet sauf s'il n'y a que 5 chances sur 100 de se tromper. On dit que cette différence est significative. Il y a un seuil de probabilité (p) de 0.05 (=5%)

Il y a d'autres niveaux de signification. (0.1 ou 0.01 etc). (Il y a n'import quel niveau de signification mais on juge que le 5% est le seuil le plus important.)

Etude: Malades hépatiques qui boivent de l'alcool

ALCOOL	Quotidienne	Peu ou jamais	TOTAL	

Malades	15	35	50	
Saines	311	1417	1728	

Hypothèse- Boire de l'alcool n'a aucun effet néfaste. H0

NB nécessité de comparer un échantillon avec la population en général.

Formuler une hypothèse doit être faite AVANT la récolte des données.

Il y a plusieurs testes qu'on peut utiliser selon le type de comparaison ou la taille de l'échantillon.

Exemples des hypothèses : = Ho ou H1 ?

1. Le nombre des accès de paludisme sont moins chez les élevés qui dort sous moustiquaire.
2. Les femmes qui ont besoin de césarienne sont moins que 150cm de taille
3. Risques des accouchements a domicile sont plus qu'à la maternité.
4. Paludisme au N .E . Congo est résistant a chloroquine.
5. La pénicilline est efficace pour la prise en charge des infections respiratoires.
6. Le bactrim est plus efficace que pénicilline pour les affections respiratoires.
7. La mortalité pour les perforations typhique est identique pour un traitement médical ou chirurgical.

Echantillons dépendent et indépendant :

Dépendant – on utilise la même population 2 fois. Exemple – Pour vérifier l'utilité d'une moustiquaire on demande celui qu'on étudie d'utiliser pour une semaine, puis de dormir SANS moustiquaire pour une semaine.

Indépendant – on compare deux populations, une avec la moustiquaire l'autre sans.

T.P. 20 infirmiers devant les examens. (2 populations indépendantes.). 10 reçoivent une leçon de révision, 10 n'en a pas. Problème: Est-ce que la révision a diminué leur anxiété avant les examens ? H0 Anx = anx (le même dans les deux groupes) H1: Prev < Po (La révision groupe a moins d'anxiété que la groupe sans révision.)

Taux d'anxiété :

Avec rev : 5 5 4 4.5 4.5 4 5.5 3.5 4.5 4.5

Sans rev : 5 6 4.5 5.5 5.5 4.5 6.5 4.5 5.5 5

Calculer le moyen et l'écart type et faites votre jugement selon le résultat :

T.P. Suggérer une hypothèse pour vérifier :

1. L'utilité des latrines.
2. La valeur d'un filtre d'eau de fabrication locale.
3. L'haut taux de césariennes a Oicha par rapport au taux nationale.
4. L'importance de savon
5. L'importance de bonne aération d'une maison

6. Principes des tests statistiques :

Formuler une hypothèse

Trouver les données

En déduire ce que devraient les observations si l'hypothèse est vraie.

Vérifier si les observations faites sont contradictoires ou conformes à ce qu'on attende.

Accepte ou rejet l'hypothèse.

Par convention en sciences humaine on considère que l'hypothèse peut être "significative" s'il n'y a pas plus que 5 chances sur 100 de se tromper en affirmant que la différence est significative (seuil de 5%). Au-dessus d'un tell niveau de signification (ou niveau de confiance) on considère qu'il est plus probable que la différence soit le fait du hasard et ce fait l'hypothèse n'a pas de signification (ou une hypothèse nulle ne peut être rejetée).

On doit utiliser la teste approprié.

1. Données cardinales (quantitatives) et la distribution est normale (en cloche) on utilise les moyens + écart types - = test t. = testes paramétriques. Donc dépendent des coefficients de certaines équations :

Syn. : Teste de Student, « T » teste

Ex : Parmi 6 de nos étudiants de G3 ils ont arrivés comme suit: (14h).03, 05 06 09 11 14, (après 14 heures) on calcule la moyenne de cet échantillon, soit:

$$\frac{3 + 5 + 6 + 9 + 11 + 14}{6} = \frac{48}{6} = 8 \text{ minutes} = 14h08 \text{ est le moyen heure d'arrivée}$$

puis on calcule la distance entre chaque donnée et la moyenne et on fait la somme des différences

-5 -3 -2 +1 +3 +6

(3-8)+(5-8)+(6-8)+(9-8)+(ll -8)+(14-8)

(On utilise une autre mesure de temps en temps – le moyen de ces écarts tous en positive (voyant que comme telle leur moyen et toujours 0). Ici l'écart moyen est 5+3+2+1+3+6 = 20 /6 = 3.33)

La variance est la somme des carrées des ces écarts

$$= (-5)^2 + (-3)^2 + (-2)^2 + (+1)^2 + (+3)^2 + (+6)^2 = 25+9+4+1+9+36 = \frac{84}{6} = 14$$

On appelle variance 14 le résultat obtenu de cette façon qui correspond à la formule : $\Sigma(x-\mu)2/n = \sigma 2$

Donc σ (l'écart type = racine carre de 14 = 3.74 = 68% des étudiants sont arrives entre 14h08 + ou – 3.74 secondes. Ou 95% des étudiants sont arrives entre 14h08 + ou – 1.96 X 3.74 = + ou – 7.3 seconds = entre 14h00,7 et 14h15,3

2. Donnés non quantitatives (qualitatives) ou échantillons trop petits pour savoir s'ils sont de distribution normale on utilise le teste X^2 (khi carré) = Teste non paramétrique.

Donc il y a plusieurs testes et c'est difficile à savoir lequel à utiliser dans quelles circonstances.

Toujours ultiliser l'écart type et 1,96 X écart type si possible.

Dans ces testes il y a les calcules ordinairement de l'aire sous la courbe de distribution ; un calcule qui nécessite le 'calculus' et qui est difficile a faire. Donc on utilise les tables ou les calcules qui sont déjà faites et qui se trouve a la fin de ce cours ou n'import quel livre des statistiques.

7. Testes paramétriques pour comparer les moyens.

1. Pour comparer une moyenne et une valeur théorique :

$$Z = \frac{m - \mu Ho}{\sqrt{\frac{S2}{n}}}$$

Z = valeur de l'aire sous la courbe normale. m = la moyenne observé. µHo = valeur théorique

S2 = la variance (L'écart type carrée) n = taille de l'échantillon

Exemple: Une firme de produits pharmacologiques veut savoir si le procédé de fabrication qu'elle utilise fournit effectivement des flacons de désinfectant de 250ml. Le volume de 200 flacons est mesuré; On trouve en moyenne m = 249.8ml la variance des volumes étant de 3.5. Doit-on considérer que la moyenne observée m (est différent de la valeur exacte en 95% des cas)(en termes mathématiques on dit) s'écarte de la valeur 250 ?

$$Z = \frac{m - \mu Ho}{\sqrt{\frac{S2}{n}}} = \frac{249,8 - 250}{\sqrt{\frac{3,5}{20}}} = 1.51$$ = moins que 1.96 donc différence insignificatif

Pour étudier le pouvoir irritant de deux substances on a badigeonné deux parcelles de peau de 15 souris, l'une avec iode l'autre avec goudron. Pour chaque souris on mesure la différence X de surface irritée. La moyenne est 2,2mm^2 et sa variance s^2 = 9,1mm^2. La moyenne observée diffère-t-elle de 0, valeur correspondant à l'absence de différence entre iode et goudron.

N = 15, m = 2,2 s^2 = 9,1

$$Z = \frac{2,2 - 0}{\sqrt{\frac{9,1}{15}}} = 2,82$$ = plus que 1.96 donc la différence est significative.

Pour être sure que la différence est due aux produits chimiques on tire au sort pour savoir quel souris reçoit quel produit et l'observateur sont aveugle, c'est à dire il ne sait pas quel produit le souris a reçu.

T.P. On essaye l'efficacité de chloroquine dans 25 malades. Apres un traitement on trouve que le goutte épaisse est devenue négative dans un moyen de 3,8 jours. La variance était de 11. Doit –on considère que chloroquine est moins efficace que artemesat qui rendre le GE négatif en 2,5j.

(On considère que l'artemesat donne le meilleur résultat qui est théoriquement possible !)

8. Test t de student : Utilisé pour des échantillons indépendants, pour comparer les moyennes

Revision : faire une definition de : Echantillion, Indépendant (Dépendant) Moyennes (μ ou m m1 m2 etc..) Ecart type σ ou s Variance σ^2 ou s^2

$$T = \frac{m1 - m2}{\sqrt{(s1carré/n1 + s2carré/n2)}}$$

Hypothèse (H0) : la tension artérielle systolique est la même chez les fumeurs et chez les non-fumeurs. On tire au sort 32 sujets (17 fumeurs et 15 non-fumeurs) chez qui on a mesuré la tension artérielle (en mm hg) Les résultats sont les suivants

Fumeur	TA	Moyenne	Ecart à la moy.	Carrée écart	Non F	TA	M	Ecart	Carrée écart
	147	148,8				134	139,8		
	130	148,8				121			
	163	148,8				129			
	161	148,8				147			
	145	148,8				151			
	181					137			
	167					134			

	139					141			
	141					143			
	135					136			
	146					131			
	143					119			
	151					160			
	150					151			
	133					163			
	127								
	171								
Moy.	148,8		Somme	231,4		139,8		Somme	166,46

Donc on compare ces deux moyennes

T = 148,8 – 139,8

√ 231,4/17 + 166,46/15 = 9 / √ 13.6 + 11.1 24.7 4.9 = 1,8.

Tables de t (B p 310) Pour 30 (31) *ddl (degrés de liberté)* seuil = 2.042 – donc cette différence n'est pas significative.

*Defn: DDL = Nombre de termes d'un échantillon dont la valeur peut être assignée librement.

Degrés de liberté :

Supposons qu'on a un résultat total de 8 et qu'il y a 3 résultats qui ont donne ce résultat. Si la première donné est 3 le deuxième 1 donc le troisième est connu ne peut être que 4. Dans un tell série on dit qu'il y a donc 2 dégrées de liberté. Plus généralement dans une série il y a toujours n – 1 dégrées de liberté (Le nombre total des observations moins 1)..

La plus grande l'échantillon le moins important les dégrées de liberté.

Plus le nombre est grand plus le teste de t approche la loi normale (voir tableau) B310

TP Un échantillon de 112 malades atteintes de cancer du colon a été compare avec un échantillon de 185 témoins non malades quant à leur consommation de caféine. Pour les malades elle est égale a m1=147,2mg per jour (écart type 101,8mg/j) et pour la population m2 = 132.9mg j. (écart type 115,7) Ces deux moyennes sont elles différentes ?

(C'est à dire est ce que caféine puisse être une cause de cancer du colon)

$$Z = \frac{147,2 - 132,9}{\sqrt{\frac{101,8^2}{112} + \frac{115,7^2}{185}}} = 1,11$$ = pour plus que 100 dégréé de liberté moins que 1,96 Bp310

T.P. Pour comparer 2 somnifères Diazepam et nitrazepam on les a donné à 2 groupes de 50 étudiants tirés au sort. Ceux qui ont reçu Nitrazepam a dormi en moyen 5,6 heures et ceux qui ont reçu Diazepam 4,9 heures. 17 étudiants avec Nitrazepam ont dormi entre 5,6 et 6,7 heures et 17 étudiants avec Diazepam ont dormi entre 4,9 et 5,8 heures.

Les deux somnifères ont-ils les effets différents ?

9.Comparison de deux pourcentages ou effectifs :.

Test du X2 = Khi carré ou chi carré = lettre de l'alphabet grec (ch) (pour savoir si Ho est vrai :

Teste de khi carré - plus facile parce qu'on ne doit pas calculer l'écart type.)

$X2 = \text{Somme } \frac{(\text{Observés} - \text{Calculés})2}{\text{Tous les calculés}}$ $X2 = \Sigma \frac{(Oij - Cij)2}{ij\ Cij}$ $X2 = \text{Somme tous } (O - C)^2 / C$

Plus facile a comprendre avec un exemple concrète :

Exemple : 120 patients atteints de Ulcère de Buruli reçoivent soit Rifampicin ou INH

Hypothèse 1 : Rif = plus efficace que INH pour Buruli

	Traitement avec :		
	INH	Rif	Tot
Guéri	22 = 31%	25 = 50%	47
Non Guéri	48	25	73
	70	50	120

= effectifs observés (O)

Il semble que Rif est plus efficace mais est ce que c'est vrai ?

Effectifs théoriques doivent être calculés :

On pense que 47 parmi 120 de 70 doit être guéri d'INH = 47/120X70 = 27.4

Ou 47 parmi 120 de 50 avec Rif = 19.6

Ou 73 parmi 120 de 70 seront non guéri avec INH = 42.6

Ou 73 parmi 120 de 50 seront non guéri avec Rif = 30.4

= effectifs calculés (C)

Donc ensemble :	22 (27.4)	25 (19.6)
	48 (42.6)	25 (30.4)

Xo2 = Somme tous (O – C)2 / C

$$Xo2 = \frac{(22 - 22{,}4)^2}{27.4} + \frac{(25 - 19{,}6)^2}{19.6} + \frac{(48 - 42.6)^2}{42.6} + \frac{(25 - 30{,}4)^2}{30.4} = 4.2$$

2 traitements donc 1 ddl

Chercher dans les tables de Chi carré. Le resultat est plus haut que 3,84 à 0,05% moins que 5,02 a 0,025% donc Rif n'est pas significativement plus efficace que INH.

TP

Dans une étude de l'anémie on trouve:

	Anémie	Non anémie
Enfants 0 – 2	62	104
Enfants 2 – 4	24	35

Est- ce qu'il y a un nombre significative plus d'anémie chez les enfants de 2 à 4 ?

T.P. 160 malades atteints de cancer de la vessie et 160 malades pris comme témoins ont été interroges sur leur passe professionnel. 51 sujets parmi les malades (soit 32%) et 37 parmi les non malades (soit 23%) ont indiqué avoir exposés aux solvants chimiques. Les pourcentages d'exposition aux solvants sont-ils diffèrent chez les malades et chez les témoins ?

Effectifs observés

Exp	oui	non	Tot
Mal	51	109	160
Non Mal	**37**	**123**	**160**
	88	232	320

Effectifs théoriques :

160/320 de 88 = malade exp = 44

Xo2 = Somme tous (O – C)2 / C

(44) (116)

(44) (116)

51-44)2 …. = 3,07

1ddl tables X2 = moins que 3,84 = non significatif

T.P. Lors d'une enquête réalisée sur un échantillon de taille 500, représentatif des décès enregistres au Nord Kivu on a observe que 190 décès (soit 38%) étaient dus à une maladie infectieuse. On se demande si ce pourcentage diffère de la valeur de référence pour le Congo ou 40% est la mortalité nationale des infections.

10. Mesure du risque : rapport des cotes (odds ratio) = risque relatif

Il est possible dans les enquêtes de calculer le risque relatif de développer la maladie chez les sujets exposés par rapport avec aux sujets non exposés. Dans les enquêtes cas/ témoins le rapport des cotes (odds ratio en anglais) est une mesure approximative du risque relatif.

La mesure de ce risque et de son intervalle de confiance permet d'une part de réaliser une teste d'association et d'autre part de mesurer l'intensité de la liaison entre les variables étudiées.

Si l'intervalle de confiance du risque relatif passe par 1, cela signifie qu'il est des circonstances ou ce risque est de 1 et donc qu'il n'est pas supérieur chez les exposes par rapport aux non-exposes. On peut alors rejeter l'hypothèse H1 et par contre l'accepter au risque alpha choisi s'il ne passe pas par 1

RR = AD/BC

Test d'homogénéité (= comparaison de risque parmi les exposes par rapport au non exposés

$$X2 = \frac{((AD - BC) - N/2)2 \quad N}{N1 \;\; N2 \quad M1 \; M2}$$

IC = RR $^{1+ \text{ ou } - 1.96/ x}$

Ou IC = Log OR + ou – 1,96 x racine carrée 1/A + 1/B + 1/C + 1/D

Si contient 1 l'association est non significative. Ne contient 1 significatif

Exemple : Etude de l'urticaire chez les gens prenant allopurinol

	Avec urticaire	Sans	Total
Avec allopurinol	15 (A)	52 (B)	67 M1 (22,4%)
Sans allopurinol	94 (C)	1163 (D)	1257 M2 (7,4%)
	109 N1	1215 N2	1324

RR = 15 X 1163 = 17445

52 X 94 4888 = 3.5 3,5 fois plus grande chance d'avoir l'urticaire avec allopurinol que sans.

X2 = 15X1163 - 52X94 - 1324/2)2 1324

109 X 1215 X 67 X 1257 = 18749164/1114072 = 16,82 X = 4,1

IC = 3.5 1 + ou – 1.96/4,1

= 1,77 5.05

Exemple 2 : Chez un groupe de femmes hospitalisées Miettinen a recherche une association entre la prise de contraceptifs oraux et le risque de thrombose veineuse :

	Avec thrombose	Sans	Total
Avec contraceptifs	12 A	53 B	65 M1
Sans contraceptifs	30 C	347 D	377 M2
	42 N1	400 N2	442

RR = 12 x 347 / 30 X 53 = 2,62 fois plus que thrombose chez les gens qui prend contraceptif.

X2 12 x347 - 30X51 - 442/2)2

42 X 400 X 65 X 377 = 5.94

X = 2.44

IC = 2,62 1 + ou – 1,96/2,44

= 1,21 5,68

T.P. : Dans la maternité d'Oicha parmi 136 femmes de 15 à 19 on trouve 12 positive pour HIV. Parmi 207 de 20 a 24 on trouve 9 positive. Est-ce que cette différence est significative ?

11. Le calcul des corrélations

L'étude des corrélations cherche à établir s'il existe une relation entre deux mesures effectuées sur le même échantillon (comme cela pourrait être le cas pour la taille et le poids des enfants, par exemple, ou encore pour le niveau de Q.I. et celui de la réussite scolaire) ou de mesures obtenues auprès de deux échantillons distincts (lors d'une comparaison entre couples de jumeaux, par exemple), et, si une telle relation existe, elle vise à vérifier si l'augmentation des valeurs d'une des deux mesures correspond à l'augmentation (corrélation positive) ou à la diminution (corrélation négative) de l'autre mesure.

En d'autres termes, le calcul de corrélation permet de savoir si la connaissance des valeurs d'une mesure permet de prédire celle de l'autre.

Jusqu'à présent, dans l'analyse des résultats de l'expérience qui porte sur l'effet de la marijuana, nous avons volontairement négligé les temps de réaction des sujets. Or, il serait intéressant de vérifier s'il n'existe pas une relation entre la performance proprement dite et la vitesse à laquelle les réponses des sujets sont émises, de façon qu'on puisse éventuellement prédire que plus un sujet est lent, plus il risque d'être précis et de fournir de meilleures performances, ou l'inverse.

On peut utiliser deux types de tests pour y arriver: le coefficient de Bravais-Pearson, ou test r, qui est un test paramétrique, et le coefficient de corrélation de rang de Spearman, ou test r,, qui s'applique à des données ordinales et qui est, de ce fait, un test non paramétrique. Mais avant d'aborder l'étude de ces tests, voyons tout d'abord ce qu'on entend par coefficient de corrélation.

Coefficient de corrélation

Le coefficient de corrélation est une valeur toujours comprise entre + 1 et - 1. Lorsque la corrélation est parfaite et positive, ce coefficient est de + 1 ; lorsqu'elle est parfaite et négative, il est de - 1. Ceci se traduit sur un graphique par une ligne droite déterminée par les points de rencontre des valeurs de chacune des paires.

Defn : Corrélation :

Relation entre deux variables qui peut être parfaite, de telle façon qu'en connaissant les valeurs de l'une on connaît les valeurs de l'autre, ou imparfaite, indiquant simplement un lien

plus ou moins systématique entre elles, ou encore nulle s'il n'existe aucun lien; d'autre part,

la corrélation peut être positive lorsque les variations de chacune des variables se produisent dans le même sens, ou négative lorsque celles-ci se produisent dans des sens opposés.

Corrélation parfaite positive (r+1)

Corrélation parfaite négative (r = -1)

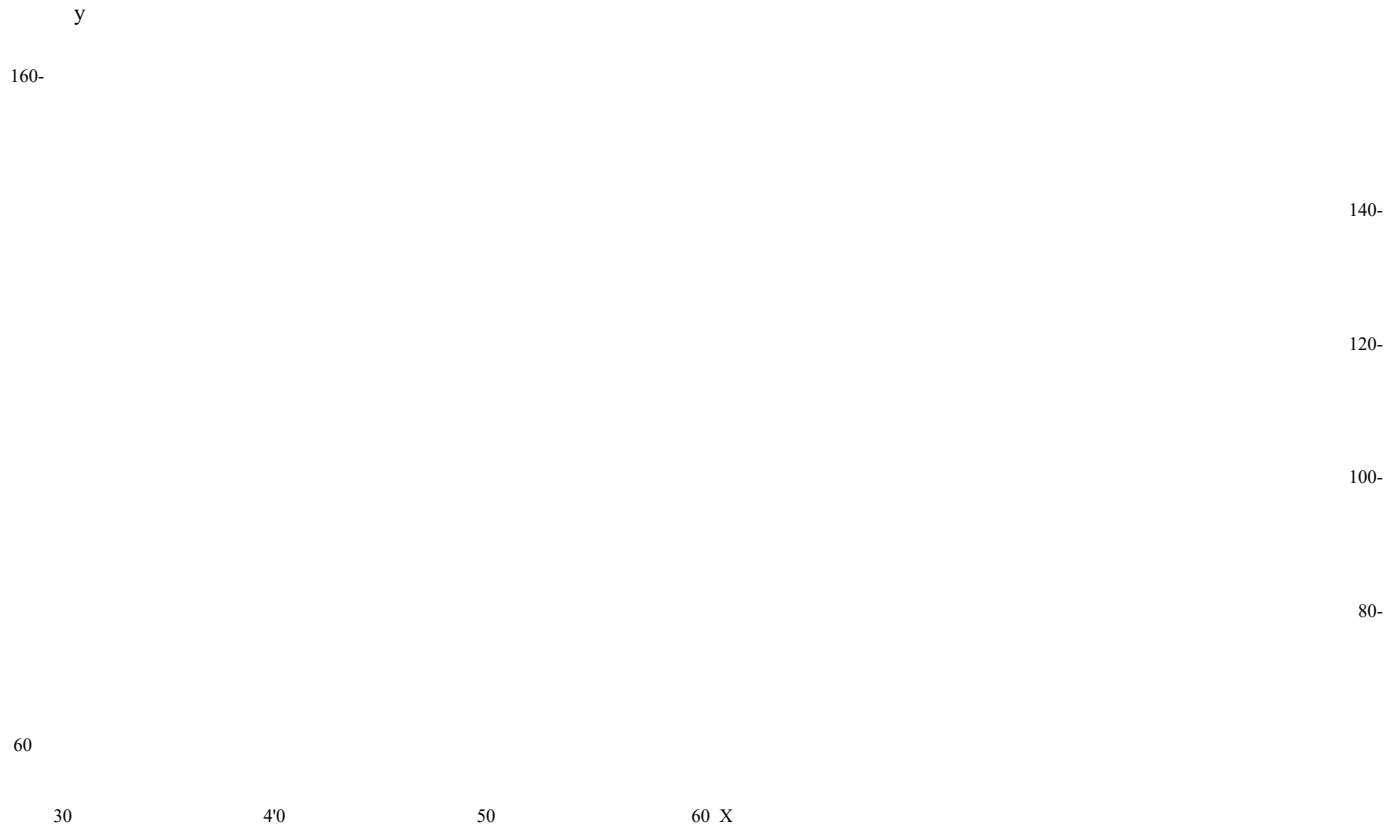
u 1 : Age X et valeur d'un dosage biologique Y

Age	Y	Age	Y	Age	Y	Age	Y
40	62	44	82	48	102	52	122
40	77	44	97	48	117	52	137
41	67	45	87	49	107	53	127
41	82	45	102	49	122	53	142
42	72	46	92	50	112	54	132
42	87	46	107	50	127	54	147
43	77	47	97	51	117	55	137
43	92	47	112	51	132	55	152

Le droite de régression estimée à partir de l'ensemble de ces données est représentée sur la figure 1.

Elle a pour équation: y 130,5 + 5,00 x. Le coefficient de corrélation est égal à r 0,904.

1 : Droite de régression de Y sur l'âge X (données du tableau 1)



Lorsque les points ne forment plus une ligne droite mais un « nuage », le coefficient de corrélation va admettre des valeurs d'autant plus proches de zéro que le nuage se rapproche de la forme d'un cercle.

Le fait que le coefficient soit égal à zéro indique que les deux variables sont totalement indépendantes l'une de l'autre".

En sciences humaines, on considère qu'une corrélation est élevée lorsque le coefficient est supérieur à 0,60; ce n'est cependant qu'au-dessus de 0,9019 qu'on considère la corrélation comme étant très élevée.

Tout dépend cependant de la grandeur de l'échantillon : plus celui-ci est important et plus la valeur du coefficient obtenu est significative.

Il existe à cet égard des tables indiquant les valeurs critiques que les coefficients de corrélation de Bravais-Pearson ou de Spearman doivent atteindre, compte tenu du nombre de degrés de liberté égal au nombre de paires moins 2 (n - 2), pour être considérés comme significatifs.

'Les tests de corrélation de Bravais-Pearson et de Spearman, que l'on utilise habituellement, servent à évaluer les relations en ligne droite. Il peut donc arriver que r soit déclaré égal à 0 alors que les points peuvent suivre le dessin d'une courbe, indiquant par là une corrélation qui peut parfois être parfaite (voir notamment le cas de la loi de Yerkes-Dodson, figure 4.1). Une telle corrélation, dès le moment où elle a été repérée graphiquement, peut être mesurée à l'aide du rapport de corrélation il (êta) effectué entre les deux

parties de la courbe. Il n'en sera pas question ici.

Defn : coefficient de corrélation

valeur située entre - 1 et + 1 qui mesure le degré de corrélation existant entre deux variables. r est le coefficient de corrélation utilisé pour les données cardinales et r. celui mesurant la corrélation entre des données ordinales.

T.P. Dans plusieurs pays on calcule le taux de malnutrition des enfants 1-5, et on trouve les chiffres suivantes : Congo 34%, Afghanistan 58%, Zimbabwe 39%, Mali 62%. Le pourcentage des femmes dans ces pays qui peuvent lire est Mali 3%, Afghanistan 7%, Congo 30%, Zimbabwe 43%.

Est-ce qu'il y a une corrélation entre le fait de lire et qu'il y a les enfants mal nourris ?

12. Révision

La statistique comprend trois secteurs principaux : la statistique descriptive, la statistique inductive et la mesure des corrélations.

1. La statistique descriptive

1. La statistique descriptive a pour but de classer les données, d'en distribuer les fréquences, de découvrir les tendances centrales de cette distribution et la façon dont les données se dispersent autour d'elles.

2. Le classement des données s'effectue tout d'abord en plaçant celles-ci par ordre croissant en une suite ordonnée. Elles sont alors regroupées, selon leur fréquence, en classes dont les intervalles sont déterminés par le chercheur en fonction de ce qu'il veut mettre en évidence dans la distribution.

3. Parmi les paramètres les plus utilisés pour décrire une distribution, on distingue, d'une part, les mesures de *tendance centrale* telles que le mode, la médiane ou la moyenne et, d'autre part, des indices de dispersion tels que la variance ou l'écart type.

4. Le mode correspond à la valeur de la donnée apparaissant le plus souvent parmi toutes les autres, ou au milieu de la classe dont la fréquence est la plus élevée.

La médiane correspond à la valeur de la donnée centrale, une fois que toutes les données ont été classées par ordre croissant.

La moyenne se calcule en divisant la somme des valeurs de toutes les données par leur nombre.

Une distribution est considérée comme normale lorsqu'elle se présente sous la forme d'une courbe en cloche dont les mesures de tendance centrale se superposent et indiquent par là sa symétrie.

5. *L'étendue d'une* distribution est constituée par la différence existant entre le plus grand résultat et le plus petit.

6. L'écart moyen constitue un indice de dispersion plus précis que l'étendue. Il consiste à calculer la *distance moyenne des* différentes données par rapport à la moyenne de la distribution. Soit, de façon simplifiée.

$$\mu = \Sigma x / n$$

7. La variance est une autre mesure de dispersion, découlant de la précédente, qui correspond à la moyenne des carrés des différences entre chaque donnée et la moyenne, soit

$$\sigma 2 = \Sigma (x - \mu)2 / n$$

8. L'écart type est l'indice de dispersion le plus utilisé. Il est obtenu en extrayant la racine carrée de la variance. Il représente donc la racine carrée de la somme des carrés de chaque écart par rapport à la moyenne de la distribution. Sa formule est la suivante.

$$\sigma = \sqrt{\sigma 2}$$

9. La propriété essentielle de l'écart type réside dans le fait que, quelle que soit sa valeur, il détermine toujours, dans une distribution normale, un pourcentage semblable de résultats se situant de part et d'autre de la moyenne. Ainsi:

68 % des résultats se situent à plus ou moins 1 écart type de la moyenne;

95 % des résultats se situent à plus ou moins deux fois l'écart type de la moyenne;

99,7 % des résultats se situent à plus ou moins trois fois l'écart type de la moyenne.

10. C'est grâce à ces mesures de tendance centrale et aux indices de dispersion que vont pouvoir être évaluées les différences existant entre deux ou plusieurs distributions, afin de vérifier jusqu'à quel point ces différences peuvent être extrapolées à la population dont les échantillons sont issus. C'est le rôle de la statistique *inductive.*

2. La statistique inductive

1. La statistique *inductive cherche* à cerner les conséquences des différences qui peuvent apparaître entre deux distributions afin d'induire éventuellement une loi s'appliquant à la population dont les échantillons sont issus.

2. Afin de vérifier si les différences sont significatives, il s'agit de poser une *hypothèse qu'on* va alors tester à l'aide d'une épreuve statistique.

On appelle hypothèse nulle l'hypothèse qui avance que la différence n'est pas significative et hypothèse alternative celle qui avance le contraire.

3. La vérification de l'hypothèse s'effectue à l'aide d'un test paramétrique pour peu qu'on possède suffisamment de données, exprimées de façon quantitative, et que ces données se distribuent selon une courbe normale. Si, par contre, les données sont en nombre restreint ou encore qu'elles sont ordinales ou nominales (voir l'encadré B. 1), on utilise alors un test non paramétrique.

4. Parmi les tests paramétriques, le plus courant et le plus efficace est le test t de Student qui consiste à comparer les moyennes et les écarts types de chacune des deux distributions. Lorsque celles-ci appartiennent à des échantillons indépendants, on utilise la formule

$$Z = \frac{m - \mu Ho}{\sqrt{\frac{S2}{n}}}$$

alors que pour deux échantillons reliés, la formule est la suivante

$$T = \frac{m1 - m2}{\sqrt{(s1carré/n1 + s2carré/n2}}$$

5. L'analyse de variance est un autre test paramétrique utilisé lorsqu'il s'agit de comparer plus de deux distributions. À l'aide du test de Scheffé, il est alors possible, à la suite de l'analyse de variance, de déterminer les paires dont la différence est ou non significative.

6. Le test du X2 (« khi deux ») est un test non paramétrique qui cherche à vérifier si deux variables sont indépendantes ou non l'une de l'autre. Ce test vise à comparer la façon dont les fréquences observées en cours d'expérience se distribuent en fonction des critères de chacune des variables par rapport à la manière dont elles se distribueraient théoriquement si les variables étaient indépendantes. À partir d'un tableau de contingences dans lequel sont reportées les différentes fréquences, on calcule le X2 en comparant, pour chaque case, la fréquence observée (0) à la fréquence théorique (E) correspondante, puis en faisant la somme de ces comparaisons, soit:

$$X^2 = \Sigma \ (O - C)^2$$

7. Le test du signe (ou test binomial) est un autre test non paramétrique qui permet de vérifier facilement si l'introduction de la variable indépendante a modifié de façon suffisamment importante les données obtenues lors de l'établissement du niveau de base. Il suffit pour cela de compter le nombre de détériorations (-) ou le nombre d'améliorations (+), puis de comparer la valeur d'un de ces deux nombres avec celle que le hasard aurait permis d'obtenir (1 chance sur 2 ou n en appliquant la formule

$$Z = \frac{(X \pm 0{,}5) - \frac{n}{2}}{\frac{\sqrt{n}}{2}}$$

8. Il existe bien d'autres tests non paramétriques dont l'utilisation est requise, dans la vérification d'hypothèses, lorsqu'on ne peut employer un test paramétrique. C'est le cas notamment du test de *séquences* chargé de vérifier si l'ordre d'apparition des événements d'une série s'effectue ou non selon les lois du hasard. C'est également celui du test de U ou du test de T nécessaires dans les cas de variables ordinales et selon qu'il s'agit respectivement d'échantillons indépendants ou dépendants.

9. Dans tous les cas, il suffit de comparer le résultat obtenu à l'aide du test avec celui figurant dans la table correspondante, au niveau de signification de 0,05 et en tenant compte du nombre de degrés de liberté. Si le résultat obtenu est supérieur à celui figurant dans la table, on peut rejeter l'hypothèse nulle et affirmer que la différence est significative.

3. Le calcul de corrélation

1. Le calcul de corrélation vise à établir la relation existant éventuellement entre deux mesures effectuées sur le même échantillon ou sur deux échantillons distincts afin de vérifier si l'augmentation des valeurs correspond à l'augmentation ou la diminution de l'autre.

2. Les valeurs du coefficient de corrélation se situent toujours entre + 1, qui représente une corrélation parfaite positive, et - 1, qui représente une corrélation parfaite négative. Un coefficient de 0 signifie qu'il n'existe aucune corrélation entre les deux séries de données.

3. Le coefficient de corrélation de *Bravais-Pearson (r)* est un test paramétrique qui s'appuie sur la comparaison des moyennes et des écarts types des résultats provenant des deux mesures. Sa formule est la suivante :

$$r = \frac{(\Sigma XY) - nXY}{}$$

(n - 1) sx sy

4. Quant au coefficient de corrélation de rang de Spearman (rs), il s'agit d'un test non paramétrique qui tente d'établir une relation entre le rang occupé par les valeurs dans chacune des deux séries de mesures.

5. Un coefficient de corrélation ne peut cependant revêtir une quelconque signification que si le nombre de paires est suffisant, ce qui peut être vérifié à partir d'une table des valeurs significatives (critiques) de r ou de r, pour un seuil de signification de 0,05.