***COURS ET TRAVAUX DIRIGES***

***L1 Sciences sociales ORGANISATION ET TRAVAIL***

***M. LAOUDI***

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  | | --- | |  | | **Cours : Corrélation entre deux variables**    Jusqu'à présent, nous nous sommes intéressés à des questions du type:   * quelle est la taille moyenne des garçons belge âgés d'une vingtaine d'années ? * quelle est la probabilité pour qu'un médicament soit efficace ? * quel pourcentage de voix un parti politique recueillera-t-il aux prochaines élections ? * quelle fraction des barres métalliques produites par une usine sera-t-elle rejetée par le client ? * le poids moyen des pains produits dans une boulangerie est-il supérieur à 800 grammes ?   Dans toutes ces questions, nous étudions le comportement statistique d'une seule variable: taille, efficacité du médicament, pourcentage de voix, longueur des barres, poids des pains.  Il existe cependant toute une gamme de problèmes statistiques où l'on s'intéresse à la relation entre plusieurs variables.  Exemples:   * les individus les plus grands sont-ils les plus lourds ? * le revenu d'une famille a-t-il une influence sur les résultats scolaires des enfants ? * y a-t-il une relation entre le tabagisme et les cancers du poumon ? * le rendement en céréales dépend-il de la quantité d'engrais utilisée ? * la productivité d'une entreprise est-elle liée au salaire des ouvriers ou employés ?   Dans ces questions, nous désirons savoir si le comportement d'une variable est influencé par la valeur d'une autre variable:   |  |  |  | | --- | --- | --- | | taille | http://www.astro.ulg.ac.be/cours/magain/STAT/Stat_Main_Fr/images/flfl.gif | Poids | | tabagisme | http://www.astro.ulg.ac.be/cours/magain/STAT/Stat_Main_Fr/images/flfl.gif | cancer | | revenu | http://www.astro.ulg.ac.be/cours/magain/STAT/Stat_Main_Fr/images/flfl.gif | résultats | | rendement | http://www.astro.ulg.ac.be/cours/magain/STAT/Stat_Main_Fr/images/flfl.gif | engrais |     La relation peut être causale ou non.  Pour étudier les relations ou corrélations entre deux variables statistiques, on peut les porter sur un graphique.  Exemple:  Relation entre la taille et le poids des individus.  Pour chaque individu de l'échantillon, on porte sur un graphique:   * sa taille en [abscisse](javascript:Def6()) * son poids en [ordonnée](javascript:Def7())   chaque individu est donc, dans ce graphique, représenté par un point (point représentatif).  Soit un individu mesurant 172 cm et pesant 66 kg:  http://www.astro.ulg.ac.be/cours/magain/STAT/Stat_Main_Fr/images/chap7i1.gif    Dans le graphe, il y aura donc autant de points qu'il y a d'individus dans l'échantillon.  http://www.astro.ulg.ac.be/cours/magain/STAT/Stat_Main_Fr/images/chap7i2.gif  *Relation entre le poids et la taille dans un échantillon de 30 individus.*  On peut (par la pensée ou réellement) tracer une droite qui passe au mieux par ces points (au milieu du "nuage" de points).  Si cette droite "monte", on dira qu'il y a corrélation positive entre les deux variables.  Si elle "descend", c'est une corrélation négative.  Si elle est "horizontale", ou si on ne peut pas décider, c'est qu'il y a absence de corrélation.  Corrélation positive:  [http://www.astro.ulg.ac.be/cours/magain/STAT/Stat_Main_Fr/images/chap7i3.gif](http://www.astro.ulg.ac.be/cours/magain/STAT/Stat_Main_Fr/Chapitre7.html#dia)    Corrélation négative:  [http://www.astro.ulg.ac.be/cours/magain/STAT/Stat_Main_Fr/images/chap7i4.gif](http://www.astro.ulg.ac.be/cours/magain/STAT/Stat_Main_Fr/Chapitre7.html#dia)    Absence de corrélation:  [http://www.astro.ulg.ac.be/cours/magain/STAT/Stat_Main_Fr/images/chap7i5.gif](http://www.astro.ulg.ac.be/cours/magain/STAT/Stat_Main_Fr/Chapitre7.html#dia)      La qualité de la corrélation entre deux variables peut se mesurer par la dispersion des points autour de la relation moyenne.  Corrélation parfaite:  [http://www.astro.ulg.ac.be/cours/magain/STAT/Stat_Main_Fr/images/chap7i9.gif](http://www.astro.ulg.ac.be/cours/magain/STAT/Stat_Main_Fr/Chapitre7.html#dia2)    Bonne corrélation (corrélation forte):  [http://www.astro.ulg.ac.be/cours/magain/STAT/Stat_Main_Fr/images/chap7i10.gif](http://www.astro.ulg.ac.be/cours/magain/STAT/Stat_Main_Fr/Chapitre7.html#dia2)    Mauvaise corrélation (corrélation faible):  [http://www.astro.ulg.ac.be/cours/magain/STAT/Stat_Main_Fr/images/chap7i11.gif](http://www.astro.ulg.ac.be/cours/magain/STAT/Stat_Main_Fr/Chapitre7.html#dia2)    Exemples:   * Corrélation entre le poids et la taille pour les garçons   http://www.astro.ulg.ac.be/cours/magain/STAT/Stat_Main_Fr/images/chap7i15.gif    On constate une augmentation du poids avec la taille (corrélation positive): les garçons les plus grands sont généralement les plus lourds.  Mais la dispersion des points est assez grande: la corrélation est assez faible.   * Corrélation entre le poids et la taille pour les filles   http://www.astro.ulg.ac.be/cours/magain/STAT/Stat_Main_Fr/images/chap7i16.gif    On ne constate pas de relation entre le poids et la taille (absence de corrélation): le poids des filles est pratiquement indépendant de leur taille.  (les filles les plus grandes sont donc les plus minces)  ***Exercice***  Le tableau suivant nous donne la répartition de l’âge et le poids de six individus   |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | | X  Poids | 70 | 75 | 55 | 80 | 65 | 80 | | Y  L’âge | 25 | 40 | 15 | 30 | 20 | 35 |     1)- calculer le coefficient de corrélation (X, Y).  2)- y-a-t’il une liaison entre le poids et l’âge ?  CORRIGE  ͞Calculons le coefficient de corrélation linaire r (X, Y) =  Pour la variable X, calculons :   * La moyenne   x͞ = = [70+75+55+80+65+80] = = 70.83   * La variance   V(x)= -  = [ + + + + + ] -  = [4900+5625+3025+6400+4225+6400]- 5016.88  = 78.95   * L’écart type = = = 8.88   Calculons pour la variable Y   * La moyenne   Y͞ = = [25+40+15+30+20+35] =   * La variance   V(Y)= -  = [ + + + + + ] -  = 72.91   * L’écart type = = = 8.53   Calculons la covariance entre X et Y :  Cov(X, Y)= Y͞  [(70x25)+(75x40)+(55x15)+(80x30)+(65x20)+(80x35) ]-(27.5x70.83)  = 64.68  r (X, Y) = = = 0.85  conclusion  le coefficient de corrélation r (x, y) = 0.85 ≈ +1, donc il existe une corrélation linéaire positive forte entre le poids (x) et l’âge (y).    Le coefficient de corrélation est compris entre -1 et +1.  Plus il s'éloigne de zéro, meilleure est la corrélation.   |  |  | | --- | --- | | *r* = +1 | corrélation positive parfaite | | *r* = -1 | corrélation négative parfaite | | *r* = 0 | absence totale de corrélation |      |  |  | | --- | --- | |  |  |   ***Exercice***  Une entreprise veut mener une étude sur la liaison entre les dépenses mensuelles en publicité et le volume des ventes qu’elle réalise.  Nous avons obtenu au cours des cinq derniers mois les données suivantes :   |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | | X  Dépenses publicitaires  En milliers de DA | 10 | 15 | 20 | 30 | 35 | | Y  Volume des ventes  En milliers de DA | 400 | 700 | 800 | 900 | 950 | | |

***Les questions***

1)- Calculer le coefficient de corrélation.

2)- Interpréter le coefficient de corrélation.

3)- estimer le volume des ventes pour un mois ou les dépenses publicitaires sont de 40.000 DA.

Calculons le coefficient de corrélation linaire r (x, y) =

Pour la variable X, calculons :

La moyenne

x͞ = = [10+15+ 20+30+35] =

La variance= [ + + + +] -

=

L’écart type = = = 9.27

Calculons pour la variable Y

* La moyenne

Y͞ = = [400+700+800+900+950] =

* La variance

V(Y)= -

= [ + + + + ] -

= 38000

* L’écart type = = = 194.93

Calculons la covariance entre X et Y :

Cov(X, Y)= Y͞

[ (10x400)+(15x700)+(20x800)+(30x900)+(35x950)]-(22x750)

= 1650

r (X, Y) = = = 0.91

Conclusion

Le coefficient de corrélation r (x, y) = 0.91 ≈ +1, Donc il existe une forte corrélation linéaire positive entre les dépenses publicitaires et le volume de ventes.

2)- Le coefficient de corrélation

r = 0.91, donc il y a une forte corrélation linéaire positive entre X et Y.

**Cours : Le Khi Deux**

Le khi deux est un test, on l’utilisera comme un outil de comparaison entre une distribution théorique et une distribution observée.

On cherche à savoir si la distribution expérimentale étudiée peut s’apparenter à une distribution théorique ; le test permet de juger de la qualité de l’ajustement.

Après avoir tiré un échantillon aléatoire, on vérifie si la distribution de l’échantillon s’apparente à une distribution théorique.

Tableau N°1 : corrélation entre situation matrimoniale et l’origine géographique.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Situation  matrimoniale  Origine  géographique | Célibataire | Marié | Divorcé | Total |
| Rurale | 4 | 42 | 3 | 49 |
| Urbaine | 12 | 23 | 0 | 35 |
| Total | 16 | 65 | 3 | 84 |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| tx | tʹx | D  ǀ tx- tʹx ǀ | CORRECTION DE YATES  D˂5 |  | |  | |
| 4  42  3  12  23  00 | 9.33  37.91  1.75  6.66  27.08  1.25 | 5.33  4.09  1.25  5.34  4.08  1.25 | 5.33  3.58  0.75  5.34  3.58  0.75 | 28.40  12.81  0.56  28.51  12.81  0.56 | | 3.04  0.33  0.32  4.28  0.47  0.44 | |
|  | | | | | 8.88 | |

tx les effectifs dans les cases

Respectez l’ordre

t1  = 4 ; t2 = 42 ; t3 = 3 ; t4 = 12 ; t5 = 23 ; t6 = 0.

tʹx =

tʹ1= = 9.33

tʹ2= = 37.91

tʹ3= = 1.75

tʹ4= = 6.66

tʹ5= = 27.08

tʹ6 =

La correction de Yates l’en effectue si D5.

Donc : 4.08 5 Avec la correction de Yates 4.08-0.5 = 3.58.

1.25 5 Avec la correction de Yates 1.25-0.5 =0.75.

4.08 Avec la correction de Yates 4.08-0.5 =3.58.

1.25 Avec la correction de Yates 1.25-0.5 = 0.75.

***La formule mathématique pour calculerle khi deux*** :

= Σ = 8.88

8.88 C’est le Khi deux calculer (voir le tableau).

On doit repérer le khi deux théorique a partir de quelques informations statistiques.

D’abord le ***seuil de signification***

|  |
| --- |
| Le niveau de signification (ou niveau α) est un seuil qui détermine si le résultat d'une étude peut être considéré comme statistiquement significatif après que les tests statistiques prévus ont été réalisés. Le niveau de signification est le plus souvent défini sur 5 % (ou 0,05). |

Donc on a α = 5% (ou 0.05)

***Degré de liberté***

D. L = (nombre de ligne -1) x (nombre de colonnes -1)

= (2-1) x (3-1) =2

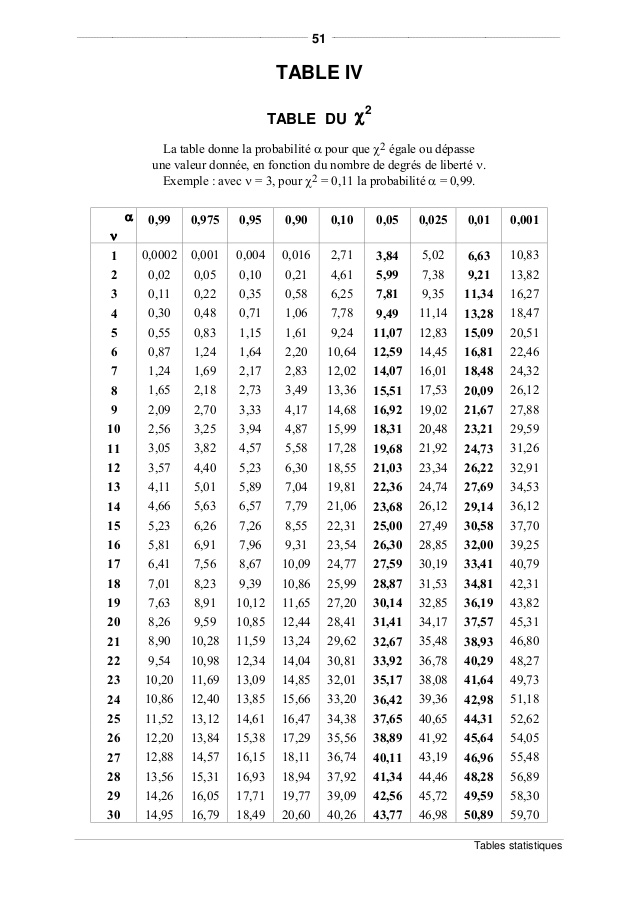
Ici, on parle des lignes et des colonnes du tableau n°1 relatif au tableau initial de cet exercice.

Une remarque très importante, on ne prend pas les lignes et les colonnes liées aux totaux en considération.

C’est pour cela on à seulement deux lignes et trois colonnes.

|  |
| --- |
| Remarque 1 : Dans un tableau à double entrée, le degré de liberté correspond au nombre de cases dont il faut connaître le contenu pour pouvoir reconstituer l'ensemble du tableau à l'aide des totaux de colonnes et de lignes. Il constitue un reflet de la taille du tableau et, dans les tables statistiques, il permet de repérer la valeur correspondant à la taille du tableau.  Remarque 2 : Dans un tableau à double entrée, on détermine le degré de liberté correspondant en calculant le produit du nombre de colonnes moins 1 et du nombre de lignes moins 1. Par exemple, dans un tableau à 4 cases, ce calcul donne : (2 - 1) \* (2-1) = 1. |

On peut repérer le KHI deux théorique, a partir de la table de Khi 2 suivante :



Pour le degré de liberté on a trouvé *v*=2. (Ici *v*) qui correspond a la deuxième ligne dans la table de Khi deux.

Concernant le seuil de signification : On a α = 5% (ou 0.05) qui correspond à la sixième colonne.

En conséquence le khi deux théorique = 5.99.

On termine l’exercice par l’interprétation du résultat final.

Donc, le Khi deux calculé est supérieur au Khi deux théorique

˃

8.88 5.99

La différence entre la distribution des données dans l’échantillon étudié et la population dont il provient est significative et due à une cause systématique à 95 % de chance. On ne peut pas parler de représentativité. Donc, on ne peut pas aborder la question de corrélation entre la variable x et la variable y. (Origine géographique et état matrimonial).

Si : ˂

C’est un autre cas différent (par rapport à notre exercice), automatiquement l’interprétation sera différente.

Donc on répond comme suit :

La différence entre la distribution des données dans l’échantillon étudié et la population dont il provient, n’est pas significative (on parle de la différence).

La différence n’est pas due à une cause systématique à 95 % de chance. Raisonnablement, il y a une représentativité. Donc on peut aborder la question de corrélation entre la variable x et la variable y.