Scénario 7

TABLEAUX DE CONTINGENCE

1 Présentation générale

1.1 Objectifs :

- Etudier un tableau de contingence
- Interpréter les marges d'un tableau
- Résumer graphiquement un tableau de contingence

1.2 Les étapes :

1. Rappel
   - Définition des termes : variable qualitative et modalités
   - Représentations graphiques de la distribution d'une variable qualitative
2. Indépendance entre 2 variables qualitatives
   - Reconnaissance des différents types de variables (qualitatives ou quantitatives)
   - Détermination du nombre de modalités
   - Construction du tableau croisé dit tableau de contingence
   - Construction des tableaux des : profils en lignes, profils en colonnes et des fréquences relatives
   - Résumé graphique associé à chacun des tableaux précédents
   - Calcul des effectifs ajustés
   - Ecarts entre les effectifs observés et ajustés
   - Calcul du χ^2 et schématisation des contributions au χ^2
3. Evaluation de l'étape 2
   - Application des notions acquises en 2 à un nouvel exemple
   - Présentation d'autres exemples
4. Homogénéité de 2 échantillons
   - Différence entre une étude d'homogénéité et une étude d'indépendance
   - Calcul au χ^2 pour mesurer la non homogénéité de 2 échantillons
   - Présentation de différents exemples
5. Evaluation globale

1.3 matériel

Stagiaire : calculette, papier millimétré

Formateur : projecteur de transparents, fichier INRA, 3 feuilles de transparent quadrillé

1.4 La forme

Tirage d'une sous-population INRA qu'on étudie pour les critères âge et grade

Travail à partir de tableaux à compléter par les stagiaires, puis sur le transparent projeté (étape 2)

Présentation de plusieurs exemples sur transparents (étapes 3 et 4)

1.5 Evaluation

Au début de l'étape 3 un autre exemple d'utilisation des notions de l'étape 2. QCM pour la distinction indépendance/homogénéité (étape 4)

1.6 Prérequis

Exécuter les 4 opérations avec une calculette, connaître la différence entre proportion et pourcentage, les calculer.

1.7 Durée

Une séance de 3 heures.

2 Les Etapes

2.1 Etape 1

But : Rappeler ou introduire : la définition d'une variable qualitative : lorsque les observations ne sont pas mesurées mais classées selon un certain critère on dit que la variable est de nature qualitative.

La définition des modalités : les valeurs que peut prendre la variable. Par exemple la variable sexe a 2 modalités (Homme ou Femme)

Montrer les différentes représentations graphiques associées; les plus courantes sont : le camembert, l'histogramme, le diagramme en barres empilées ...

Scénario : Présentation d'un exemple sur transparent

Cet exemple concerne la production automobile pour l'année 1980 d'une entreprise (Transparent 1)

Type de véhicule	effectif (en milliers de véhicules)	pourcentage
cabriolet 2 places	10.2	40.8
4 places	8.3	33.2
5 places	5.5	22.0
modèle luxe	1.0	**4.0**
TOTAL	**25.0**	**100.0**

Les chiffres en gras n'apparaissent pas sur le transparent; le formateur complètera avec les stagiaires.

On peut poser les questions suivantes :

- Quelle est la variable étudiée? La variable est le type de véhicule.

- Quel est le nombre de modalités? Il y a 4 modalités : "cabriolet", "4 places", "5 places", "modèle luxe". Elles caractérisent le type de véhicule. Cette variable est différente d'une variable quantitative, elle n'a pas de valeurs numériques. On a donc une variable qualitative à 4 modalités. Pour chaque modalité on a le nombre de voitures qui correspond à cette catégorie (=effectif).

  On complète les cases vides du transparent en expliquant :

  1. calcul du nombre total de véhicules :
     $10.2 + 8.3 + 5.5 + 1.0 = 25.0$
  2. calcul de la fréquence relative ou proportion :
     $10.2 / 25 = 0.408$
  3. calcul de la fréquence en
     $100 \times 10.2 / 25 = 40.8$

Question : Avez-vous une idée de la représentation graphique permettant de visualiser le tableau?

Réponse : On présente 4 sortes de représentations graphiques associées (fig.1, 2) que l'on commente.

1. les camemberts : Transparent 2

   C'est une représentation que l'on rencontre fréquemment en économie. Elle n'est pas toujours simple à interpréter principalement quand la variable qualitative a de nombreuses modalités. Dans ce type de graphique, les effectifs ou fréquences des diverses modalités sont représentés par les secteurs d'un cercle dont l'angle (ou la surface du secteur) est proportionnel à l'effectif ou au pourcentage.

2. les histogrammes : Transparent 2

   Dans cette représentation chaque type de véhicule est représenté par un rectangle dont la hauteur est proportionnelle au pourcentage ou à l'effectif (à condition de construire des rectangles de même largeur).

3. les diagrammes en barres empilées : Transparent 3

   Le principe est le même que pour le cercle mais ici le cercle est remplacé par un seul rectangle partagé en bandes de hauteur proportionnelle à l'effectif de la modalité.

4. le tableau graphique : Transparent 3

   On visualise la valeur des pourcentages par des carrés dont l'intensité de gris est proportionnelle à la valeur du pourcentage. On calcule l'amplitude des pourcentages (pourcentage maxi - pourcentage mini = 40.8 - 4.0 = 36.8). On en déduit les limites des classes d'intensité de gris.

   de blanc à noir : (4%, 13.2%) (13.2%,22.4%)(22.4%,31.6%)(31.6%,40.8%)
   pour les types : luxe, 5 places, 4 places, cabriolet

2.2 Etape 2 : Indépendance entre 2 variables qualitatives

But :

- Reconnaitre les différents types de variables et les modalités
- Construction du tableau croisé contenant les effectifs de chaque combinaison de modalités
- Extraire le maximum d'informations en calculant les divers pourcentages
- Faire des résumés graphiques

Scénario :

Présentation du fichier I.N.R.A. : pour chaque agent on possède 4 variables (centre, grade, sexe et âge).

Questions :

- parmi les 4 variables quelles sont les variables qui sont de nature qualitative et celles qui sont de nature quantitative? (âge = variable quantitative, les autres sont qualitatives).
- pour celles qui sont qualitatives, quelles sont les différentes modalités? Combien de niveaux par variable?
- Comment procéder pour vérifier s'il existe une liaison entre le grade et le sexe

Réponse attendue : effectuer un échantillonnage, ou bien tout compter

Remarque : le recensement peut être effectué sur la totalité du fichier mais trop long manuellement, on se contentera d'un échantillon de taille n=100, extrait au hasard du fichier.

Procédure d'échantillonnage :

- Compte tenu du nombre de modalités de la variable grade on procède à une réduction en regroupant les grades dans 4 catégories uniquement :

  ADMINISTRATIF ..... A
  INGENIEUR I
  SCIENTIFIQUE ..... S
  TECHNICIEN ..... T

  La correspondance entre grade et catégorie se trouve dans le tableau 2 que l'on distribuera (feuille stagiaire 1).
  Après ce recodage on se trouve en présence d'une variable sexe à deux niveaux (H et F) et de la variable catégorie à 4 niveaux (A,I,S,T).

- Avant d'effectuer le dénombrement il faut établir la liste des couples possibles.

Question : combien de couples?

Il y a 8 possibilités : (F,A), (F,I), (F,S), (F,T) puis (H,A), (H,I), (H,S), (H,T) (transparent 4). On va tirer au hasard n_i individus par stagiaire, de telle sorte que $n1 + n2 + \ldots + nk = 100$ (distribution de la table de nombres au hasard (feuille stagiaire 2 et 3).

- Pour l'échantillonnage : prendre soit des dés à 10 faces soit des nombres au hasard de 4 chiffres (il y a 8321 agents); chaque chiffre correspond à l'individu ayant le même numéro dans le fichier (colonne qui précède la variable centre). Quand le tirage est terminé on réunit les dénombrements afin de n'avoir qu'un seul échantillon de 100 individus et une répartition commune plus simple analyser. On reporte les effectifs des 8 couples dans le tableau du transparent 5 et on fait la somme des effectifs pour l'ensemble des stagiaires.
- A ce stade il faut montrer que ce tableau n'est guère lisible ni pratique à interpréter. Il faut amener les stagiaires à construire un tableau semblable au tableau transparent 5 (distribuer au fur et à mesure les feuilles stagiaires).

CONSTRUCTION DES TABLEAUX DE CONTINGENCE

- Construire le tableau des effectifs (feuille 3, transparent 5) : on transfère les données du tableau 5 dans ce tableau croisé (dit tableau de contingence). Calculer ensuite le nombre total d'individus par sexe et par catégorie (sommes marginales).

  Comment représenter de façon plus parlante ces résultats?

  Une solution consiste à calculer des pourcentages. Mais quels pourcentages? Faut-il diviser les effectifs par le total des lignes, des colonnes ou par l'effectif total de l'échantillon?

  Toutes les solutions sont possibles, chacun donne une information différente.

- Calcul des profils en lignes : On y porte le pourcentage de chaque catégorie pour un sexe donné.

catégorie sexe	A	I	S	T	TOTAL
F	f				100
H					100

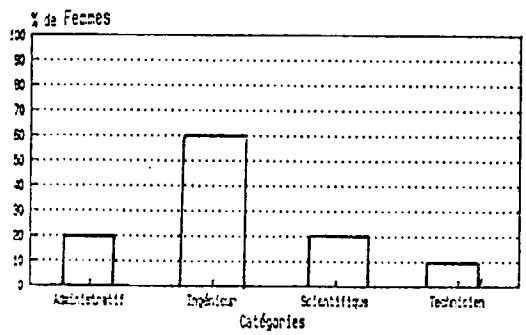
f = (effectif pour le couple FA divisé par le nombre total de Femmes) multiplié par 100.

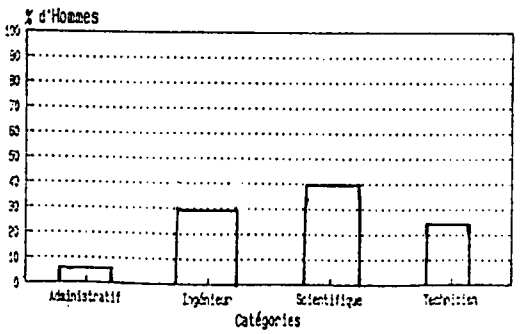
On calcule le pourcentage pour une ou deux cases et le stagiaire calcule la suite seul.

Commentaire : grâce à ce tableau, on peut comparer les valeurs d'une même catégorie et les profils par sexe.

Effectuer la représentation graphique (transparent 7 feuille stagiaire4)

C'est une forme visuelle plus facilement interprétable qu'un tableau de chiffres.



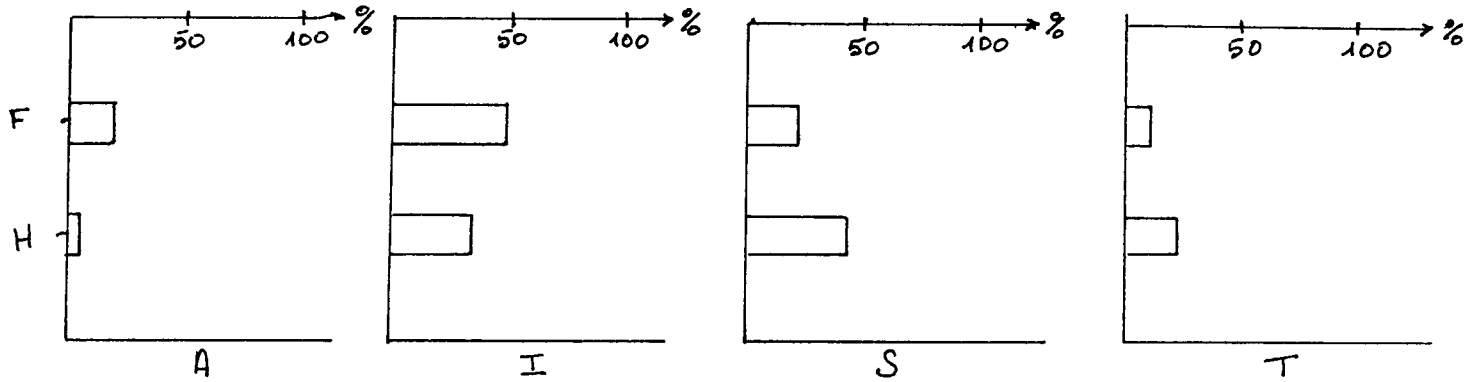


On complète les fiches graphiques en représentant par exemple l'importance de la catégorie des administratifs parmi les femmes par un rectangle dont la hauteur est proportionnelle au pourcentage. Prendre la précaution de construire des rectangles de même largeur. On note qu'il y a peu d'hommes dans cette catégorie.

Question : A partir de ce graphique que peut on dire sur l'indépendance des deux caractères sexe et catégorie? Si les deux profils sont identifiques cela signifie qu'il y a la même répartition entre les catégories quelque soit le sexe : on dit qu'il y a indépendance des deux caractères.

- Tableau des profils en colonnes (transparent 8, feuille 5)

  Il donne les pourcentages d'agents d'un sexe donné par catégorie. Effectuer cette nouvelle représentation graphique (transparent 9, feuille 6).



Même commentaires et questions que pour les profils en lignes.

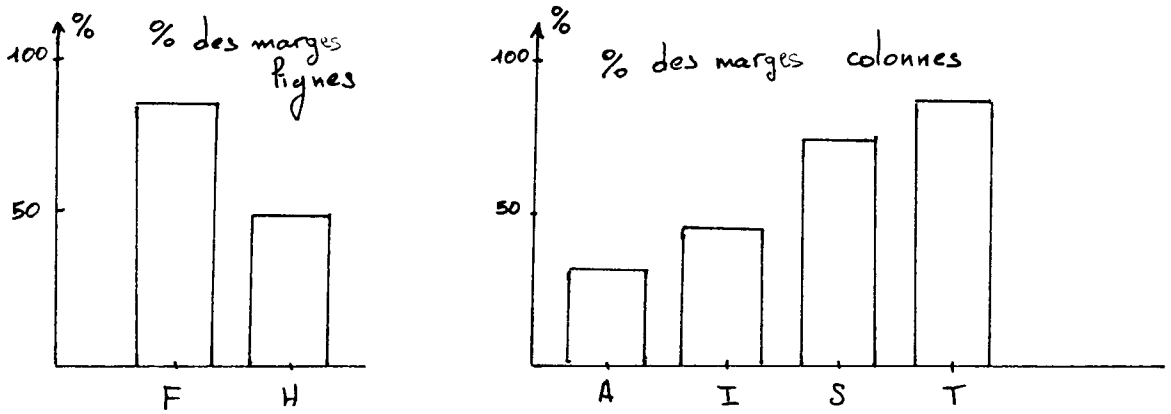
- Tableau des fréquences relatives (transparent 10, feuille 6)

  Chaque effectif du tableau est divisé par le nombre total d'individus dans l'échantillon. On porte sur le transparent 10 les fréquences relatives multipliées par 100. Comme pour 2 ou 3 on calcule 1 ou 2 cases et les stagiaires complètent.

  Effectuer la représentation grahique correspondant à ce tableau.

Plusieurs représentations sont possibles :

1. représentation des fréquences relatives marginales (sur papier millimétré)



2. Représentation des fréquences relatives pour chaque couple de variables (sur papier millimétré)

3. tableau graphique (transparent 11, feuille 6)

On prend le pourcentage minimum et maximum et on calcule l'écart Δ entre ces 2 pourcentages extrêmes : $\Delta = \%max - \%min$. On fait 4 classes d'amplitude égale et à chaque classe on associe un niveau de gris du blanc au noir pour les 4 classes $(min, min+\Delta/4), (min+\Delta, min+\Delta/2), (min+\Delta/2, min+3\Delta/4), (min+3\Delta/4, max)$.

On complète un tableau vide par ces différents hachurages en fonction des fréquences relatives obtenues.

- Calcul des nombres ajustés et comparaison avec les nombres observés (transparents 5, 12, feuille 7)

  Un des buts est de rechercher une éventuelle relation entre les 2 variables. Voir ce que l'on peut déduire des différents graphiques que l'on a construit. Les graphiques sont-ils suffisants? Pour chiffrer l'indépendance ou la non indépendance entre 2 variables il faut aborder le calcul des fréquences ajustées sous l'hypothèse d'indépendance entre les 2 critères sexe. Pour le calcul des nombres ajustés prendre le tableau 8. Parmi les n individus on a T_f femmes soit une proportion de T_f/n. S'il y a indépendance entre le sexe et la catégorie on doit retrouver cette même proportion parmi les femmes scientifiques la quantité $T_s * T_F/n$ représente le nombre ajusté de femmes que l'on devrait trouver dans la catégorie scientifique sous l'hypothèse d'indépendance. On aurait trouvé le même résultat si on avait commencé par le calcul de la proportion de scientifiques parmi les n individus. On fait calculer et remplir le tableau des nombres ajustés par les stagiaires.

  Que faire des nombres ajustés? Les comparer aux nombres observés.

  Différence entre les observés et ajustés (transparent 13). Montrer que la somme des différences est nulle.

- Calcul des contributions au χ^2 : (transparent 14, 15 et feuille stagiaire 8).

Pour chaque combinaison de variables (F,S ou H,I) ...) on calcule la différence entre le nombre observé et sa valeur ajustée, on porte cette différence au carré (donc la somme ne sera plus nulle) puis on la divise par le nombre ajusté pour se ramener à une même unité). Les valeurs maxi correspondent aux couples qui présentent l'écart maxi à l'indépendance. La somme de toutes les valeurs du tableau donne un critère que l'on appelle χ^2.

Dire que l'on calcule ce χ^2 parce qu'il existe des tables toutes faites qui fournissent la valeur du χ^2 à ne pas dépasser quand deux variables ne sont pas liées.

Si χ^2 est supérieur on dira que sexe et catégorie sont des variables liées ou dépendantes.

On peut ainsi avoir une mesure de cette "dépendance".

Souvent ce qu'on appelle la contribution au χ^2 est la valeur de chaque carré écart divisé par le χ^2 total.

2.3 Etape 3

But :

- faire un test d'évaluation de l'étape 2 (ex. 1 de Snedecor),
- présenter d'autres exemples de tableau de contingence.

Scénario :

Distribution aux stagiaires de l'exemple "recherche de l'existence d'une liaison entre l'usage du tabac et la mortalité" feuille 9. Leur laisser 5 à 10 minutes puis analyser cet exemple ensemble (transparents 16,17,18, 19).

Autres exemples de dépendance et d'indépendance (à choisir)

- Couleur des yeux - couleur des cheveux. Transparents 20 à 23. Feuille formateur 20. Le test rejette l'indépendance. Comme on n'introduit pas ici la notion de test, on dit aux stagiaires que l'écart à l'indépendance est trop important pour qu'on puisse remplacer les valeurs observées par les valeurs ajustées au modèle d'indépendance; qu'une table donne les valeurs du χ^2 à dépasser pour rejeter l'indépendance. Ces valeurs dépendent du nombre de lignes et de colonnes. Ce sont les tables du χ^2.

- Sexe et groupe sanguin. Transparents 24 et 28. On peut accepter l'indépendance. Feuille formateur 24.

- Espèce d'arbre - résistance à un débroussaillant. Transparents 29 à 32. Feuille formateur 29.

  Pour l'expérimentation du débroussaillement sélectif (dégagements) des plantations résineuses, une application aérienne de 2.4.5-T a été pratiquée sur une parcelle de taillis enrésiné mettant en œuvre 120 l/ha d'une émulsion à 3 % d'un produit commercial à 710 g/l d'acide 2.4.5-T. Pour apprécier les résultats de ce traitement -et notamment la réduction du couvert- on a procédé sur les interbandes à un sondage systématique au 1/10. Pour chaque brin de taillis d'une placette de sondage, on a ainsi relevé l'essence et une note de résistance (échelle conventionnelle de 0 à 6).

- profession des parents / études poursuivies par les enfants. Transparents 33 à 36. Feuilles formateur 33 à 35.

  Y a-t-il une relation entre le type d'études supérieures poursuivies et l'origine sociale des étudiants?

  Un échantillon de 10 000 étudiants en 1975-1976 se ventile selon le tableau de contingence suivant (d'après : Données Sociales, 3e édition, INSEE, 1978).

  Résultats : on rejette l'hypothèse d'indépendance : l'origine sociale joue sur le choix de la filière universitaire.

Remarque : Les origines sociales sont :

- exploitant agricole
- salarié agricole
- patron
- profession libérale + cadre supérieur
- cadre moyen
- employé
- ouvrier
- personnel de service
- autres

Les études poursuivies sont :

- droit
- sciences économiques
- lettres
- médecine et dentaire

- pharmacie
- pluridisciplinaire
- IUT

2.4 Etape 4

But :

- Montrer ce qui différencie l'étude de l'homogénéité de l'étude de l'indépendance entre deux variables :

  $\longrightarrow$ le mode d'échantillonnage
  et
  $\longrightarrow$ les hypothèses de départ

- présentation de divers exemples pour mettre en relief ces différences
- test d'évaluation sous forme QCM

Scénario :

On part de l'exemple INRA :

Au lieu de tirer un échantillon de 100 individus dans la population totale, on tire :

- un échantillon de 50 femmes, parmi toutes les femmes de cette population, et
- un échantillon de 50 hommes, parmi tous les hommes de cette population.

On considère toujours les caractères catégorie et sexe de l'agent.
On récupère le tableau des effectifs correspondant à ce tirage. (voir feuille stagiaire 10). On peut calculer le tableau des fréquences relatives et le tableau des profils en lignes. Transparents 37 à 39.

1) **Question :** différences avec le problème traité dans l'étape 2?

Dans l'étape 2, les % des marges lignes estimaient les proportions d'homme et de femme dans la population totale INRA. Or ici, on a imposé ces proportions (50 individus de chaque sexe); on ne peut donc pas les utiliser. Les deux caractères : catégorie et sexe, n'ont plus ce rôle symétrique, puisqu'on a imposé des effectifs a priori, pour les hommes et pour les femmes. La principale différence d'avec le problème d'indépendance est donc du essentiellement au mode d'échantillonnage.

On peut donc comparer que les profils lignes entre eux (soient les distributions des catégories pour les hommes et pour les femmes); ou avec les % des marges colonnes.

Cela revient donc à comparer l'homogénéité des distributions des catégories (A,I,S,T), selon le sexe, et non plus l'indépendance entre les deux caractères.

2) On montre ensuite que les calculs sont identiques :

- effectifs ajustés? Faire faire le calcul à partir des marges colonnes.
- Valeur de χ^2. Les 2 distributions ne sont pas comparables (pas d'homogénéité).

Autres exemples

- tension artérielle. Résultat d'un traitement.

  On veut savoir si la tension artérielle (basse ou élevée) peut modifer la réussite d'un traitement. 125 individus tirés dans chacun des 2 groupes de tension artérielle sont traités.

  Résultat : on ne rejette pas l'hypothèse d'homogénéité : il y a autant de succès du traitement pour une tension basse qu'une tension haute.

  transparents 40 à 43
  feuille formateur 40

- milieu de culture/développement des œufs.

  Recherche d'un milieu de culture adéquat pour la culture d'œufs de bovins fécondés in vivo.

  On veut savoir si l'addition d'un facteur de croissance (TGF-β) dans un milieu de culture C peut entrainer le développement des œufs jusqu'au stade éclosion. On doit donc comparer deux milieux C et T. Le milieu C (contrôle) doit mimer la situation existant in vivo, la culture est faite pendant 3 jours sur tapis de cellules tubaires de vaches puis, pendant 5 jours, sur tapis de cellules utérines. Le milieu à tester T est identique au milieu C mais avec addition quotidienne de TGF-β au 3ème jour de culture.

  Après 8 jours de culture on observe le stade atteint par les œufs. Les observations effectuées sur 41 œufs se trouvent dans le tableau du transparent 44

  éclosion : oui = +, non = -

- type de maladie/groupe sanguin

  Dans une étude sur la liaison entre le groupe sanguin et la maladie, un grand échantillon composé de malades ayant un ulcère peptique, de malades ayant un cancer de l'estomac et de personnes indemnes de ces maladies comme témoin, ont été classées selon leur groupe sanguin (O,A,B,AB). Dans cet exemple on n'a pas fait figurer le nombre relativement faible de patients AB pour simplifier.

  On veut savoir si la distribution des groupes sanguins est la même pour les trois échantillons.

  Remarque : les types de maladies sont :

  1. l'ulcère peptique ("ulcère pep.")
  2. le cancer de l'estomac ("cancer est.") et
  3. l'absence de maladie ("témoins")

  transparents 49 à 52, feuille formateur 49.

  On rejette l'homogénéité : la distribution des groupes sanguins n'est pas identique selon le type de maladie. C'est la distribution pour l'ulcère peptique qui se différencie des autres.

3) Evaluation. Distribuer le questionnaire QCM (feuilles 11, 12, 13).

2.5 Points stratégiques

- Les marges sont informatives ou non : c'est la différence entre indépendance de deux variables qualitatives (ou caractéristiques) et identité (homogénéité) des distributions (lois) d'une même caractéristique sur 2 (ou plusieurs) populations. La distinction n'est pas toujours évidente : pour certaines expériences les deux problèmes ont un sens. De plus on fait les mêmes calculs.

- L'existence d'une population de référence dont le tableau résume un échantillon est ici sous-jacent. Mais on ne traite pas ici de ce problème qui sera abordé dans les scénarios 9 et 10.

  Cela rend l'utilisation éventuelle d'une table un peu troublant : si toute la population est répertoriée il y a indépendance seulement si les fréquences relatives sont égales aux produits des fréquences marginales, dépendance dès qu'il y a non égalité.

  Alors qu'on leur dit : on peut accepter le modèle d'indépendance si le χ^2 n'est pas trop grand, sans leur expliquer que la statistique du χ^2 calculée sur un échantillon est aléatoire.

2.6 Liste du vocabulaire

Ajustement ou calcul d'une valeur ajustée (fitting, smoothing)
Diagramme en bâtons (bar diagram)
Distribution (distribution)
Fréquence (frequency)
Fréquence relative ou proportion (relative frequency)
Histogramme ou histogramme des fréquences (histogram)
Homogénéité de deux distributions
Indépendance de deux variables qualitatives
Marge colonne
Marge ligne
Modalité ou classe (classe)
Profil en colonne
Profil en ligne
Tableau de contingence ou tableau croisé (contingency table)
Variable qualitative (qualitative variable)
Variable quantitative (quantitative variable)

2.7 Références bibliographiques

FOURASTIE J. , SAHLER B. (1978). *Probabilités et statistique.* Série J. Quinet Dunod.

LERON A.J., LAZAR P. (1966). *Exercices programmés de statistique.* Flammarion Médecine Sciences.

VESSEREAU A. (1956). *La statistique.* P.U.F.

NOETHER G.E. (1971). *Introduction to statistics.* Houghton Mifflin Company.

SNEDECOR W.G.C. (1957). *Méthodes statistiques.* (6ème édition - traduction 1971). Association de Coordination Technique Agricole Paris.

JAMBU M. (1989). *Exploration et statistique.* Dunod informatique.

3 Annexes

3.1 Documents stagiaires

Présentation du scénario

Feuille 1 : codage de la caractéristique grade
Feuille 2 : Nombres au hasard
Feuille 3 à 8 : Tableaux à remplir (version réduite des transparents 1 à 15, étapes 1 et 2).
Feuille 9 : Tabac et mortalité : évaluation.
Feuille 10 : 50 hommes et 50 femmes agents de l'INRA.

Test d'évaluation indépendance/homogénéité (feuilles 11, 12, 13)

Vocabulaire du scénario 7.

3.2 Documents formateurs

Transparents 1 à 15. Etapes 1 et 2. A compléter avec les stagiaires
Transparents 16 à 19. Tabac et mortalité : résultats de l'évaluation
Transparents 20 à 36, feuilles formateur 20, 24, 29, 33 à 35. 4 exemples de dépendance ou d'indépendance (au choix).
Transparents 37 à 52, feuilles formateur 40, 44, 49. 3 exemples d'homogénéité
Feuille formateur 50 : solution au test d'évaluation de l'étape 4.

3.3 Les transparents