Master 2 MI Analyses statistiques multivariées.
Cours de P.Beau.

Document n°3. Inférer. Tests statistiques.

1 Test statistique.....1
2 Tests sur les moyennes – "tests t" – 2
2.1 Test de comparaison d'une moyenne à une valeur fixe, ou test t sur 1 échantillon.....2
2.2 Test de comparaison des moyennes de 2 populations.....2
3 Test d'indépendance du Chi2 de Pearson.....3

L'inférence considère que les données sont un échantillon d'un ensemble plus vaste et essaie d'estimer des propriétés de cet ensemble à l'aide de l'échantillon observé. En particulier nous allons voir comment **tester des hypothèses** sur l'ensemble d'une population, en se fondant sur les données d'un échantillon de cette population.

1. Test statistique.

On dispose d'un échantillon issu d'un ensemble plus vaste. Dans l'exemple du point de vente, on n'a pas observé toute la clientèle, uniquement certains clients. On suppose que **cet échantillon a été tiré au hasard**, à cette condition on pourra évaluer des probabilités d'erreurs, et des marges d'incertitude. Il arrive qu'on ne tire pas l'échantillon au hasard, par exemple en réalisant un sondage selon la méthode des quotas. Dans ce cas on ne **peut pas** calculer ces caractéristiques on ne peut que supposer que les résultats s'appliquant au tirage aléatoire en donnent une approximation.

Un **test statistique** est utilisé pour **choisir entre 2 hypothèses**, à l'aide d'observations d'un échantillon choisi au hasard. De ce fait le résultat est aléatoire : il est toujours possible qu'un test se trompe. Simplement, on essaie dans la mesure du possible d'évaluer et de **maîtriser le risque d'erreur**.

Un test statistique est fondé sur :

- 2 hypothèses, parmi lesquelles le test en choisira une. Ces 2 hypothèses n'ont pas le même rôle.
  - L'une d'elle est appelée **Hypothèse nulle**, notée **H0**, elle peut être adoptée au bénéfice du doute.
  - L'autre est appelée **Hypothèse adverse**, notée **H1**, elle n'est adoptée que si les données la désignent de façon **flagrante**.
- Une statistique : une valeur numérique fonction des données. Les individus ayant été choisis au hasard, cette statistique est une variable aléatoire. Selon les valeurs de cette statistique on adopte l'hypothèse nulle ou l'hypothèse adverse.
- Une règle de décision : la donnée des valeurs de la statistique (en général, un ou des intervalles) pour lesquelles on choisit H0, et des valeurs pour lesquelles on choisit H1.
- Un **risque d'erreur de première espèce**. C'est la **probabilité que la règle de décision désigne H1 alors que H0 est vraie**. Ce risque est **choisi arbitrairement** avant de construire le test. Il est nécessaire pour construire la règle de décision. Il est très courant qu'on choisisse 0,05, mais rien n'y oblige.
- Un risque d'erreur de 2nde espèce : c'est la probabilité que la règle de décision désigne H0 alors que H1 est vraie. Alors qu'on choisit le risque d'erreur de 1ère espèce, on ne peut en général pas choisir le risque d'erreur de 2nde espèce, on ne peut que l'évaluer, dans certains cas.
- La p-valeur ou **seuil de significativité** : c'est une valeur calculée par les logiciels de statistiques, **la probabilité que la valeur observée soit atteinte ou dépassée sous l'hypothèse H0**. Plus cette p-valeur

est grande, plus H0 est vraisemblable. Lorsque ce seuil est fourni la décision est simple, c'est la même pour tous les tests :

Si le seuil de significativité , ou p valeur, est inférieur au risque d'erreur de 1ère espèce choisi, on rejette H0. Sinon on accepte H0.

2. Tests sur les moyennes d'une variable quantitative.

Il s'agit de

- comparer une moyenne à une valeur fixe,
- ou de comparer les moyennes de deux populations.

Ce test suppose que la variable étudiée suit une loi Gaussienne et, dans le cas de comparaison de deux moyennes, que leurs variances sont égales.

2.1. Test de comparaison d'une moyenne à une valeur fixe, ou test t sur 1 échantillon.

On dispose d'une variable X observée sur un échantillon de n individus – indépendants ! – d'une population. On suppose que

- X suit une loi Gaussienne de paramètres μ et σ^2 – la même loi de probabilité, donc, pour tous les individus – .
- ou bien, si la taille de l'échantillon est supérieure à 50, simplement que X suit une même loi pour tous les individus, en particulier de même espérance et de même variance.

On veut savoir si l'on peut raisonnablement estimer que **la moyenne de cette variable**, sur l'ensemble de la population, est égale à **une valeur m_0 fixée o**u en est significativement différente.

1. Les hypothèses testées :
   Hypothèse nulle H0: L'espérance commune aux variables X est m_0.
   Hypothèse adverse H1: cette espérance est différente de m_0.
   Il existe deux variantes de ce test, où l'hypothèse adverse est que cette espérance est supérieure à m_0 ou encore que cette espérance soit supérieure à m_0. On par dans ce cas de test unilatéral.
2. La statistique du test est $t = \frac{\overline{X_n} - m_0}{\sqrt{\frac{n}{n-1}\left(\left(\overline{X_n^2}\right) - \left(\overline{X_n}\right)^2\right)}}$. Si les hypothèses sur la loi de la variable X sont satisfaites, alors t suit – au moins approximativement – une loi de Student de paramètre (n-1).
3. On doit choisir un risque d'erreur de 1ère espèce : $P_{H0vraie}\left(Rejeter\ H0\right)$
4. La règle de décision est la suivante :
   On rejette H0 si $|t| > t_{max}$, t_{max} étant déterminé par le risque d'erreur de 1ère espèce choisi, et lu dans une table statistique de la loi de Student.
   Mais Jamovi fournit le seuil de significativité. On n'a pas besoin de déterminer t_{max}, il suffit de comparer le seuil de significarivité au risque d'erreur de 1ère espèce choisi :
   - **Si le seuil de significativité est inférieur au risque d'erreur de 1ère espèce choisi , on rejette H0 , on accepte que la moyenne est significativement différente de m_0 .**
   - **Sinon on accepte H0**, selon laquelle la moyenne n'est pas significativement différente de m_0.

On réalise ce test par : Analyse → **Tests t comparaison des moyennes → Test t pour un échantillon.**
Une fenêtre s'ouvre. On vous propose d'indiquer une variable quantitative dans « Variable dépendante » . Un peu plus bas, dans la section Hypothèses , on vous demande d'indiquer la valeur m_0 de la moyenne à tester. Un

bouton vous permet d'indiquer si l'hypothèse adverse que vous adoptez est que l'espérance soit différente de cette valeur, uniquement inférieure , ou uniquement supérieure.

Dans la fenêtre de résultats, la p-valeur ou seuil de significativité est notée ***p***.

Q 1 Voici le résultat du test d'hypothèse H0 : la variable prix_chaussure suit une loi de moyenne 55 euros.

Test t pour un échantillon

		Statistique	ddl	p
dépense	t de Student	−1.02	67.0	0.310

Note. H_a $\mu \neq 55$

En adoptant un risque d'erreur de 1ère espèce de 0,05, peut-on juger que la moyenne est significativement différente de 55 Euros ?

Q 2 Utiliser un test pour juger si la moyenne de la variable dépense est significativement différente de 35 Euros Euros , puis de 40 Euros.

2.2. Test de comparaison d'une moyenne entre deux populations, ou test sur deux échantillons indépendants.

On considère deux populations indépendantes l'une de l'autre, on suppose que l'on ne dispose que d'un échantillon de chacune de ces 2 populations, et on émet les hypothèses suivantes :

1. Les hypothèses testées :
   Hypothèse nulle H0 : La moyenne est la même dans les deux populations entières
   Hypothèse adverse H1 : les moyennes dans les populations entières sont différentes
2. La statistique du test : On construit $t = \frac{\overline{X_1} - \overline{X_2}}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$ où $s = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$,alors *t* suit une loi de Student de paramètre $n_1 + n_2 - 2$
3. La règle de décision : on rejette H0 si t atteint des valeurs trop peu probables sous HO. Une table de la loi de Student fournit , en fonction du risque d'erreur de 1ère espèce et de la taille $n_1 + n_2$, la valeur de t à partir de laquelle on rejette HO. Mais à nouveau , Jamovi comme tous les logiciels de statistique fournit la p-valeur , ou seuil de significativé , du test: la probabilité que la valeur observée arrive, ou soit dépassée sous l'hypothèse H0 . On se fixe simplement à partir de quel seuil on rejette raisonnablement l'hypothèse nulle.

Les options de menus à utiliser sont
Analyse → tests t → tests t pour échantillons indépendants.

Jamovi demande

- de choisir la variable quantitative dont on compare les moyennes – variable dépendante -- . On peut indiquer plusieurs variables , Jamovi construira un test pour chacune d'elles ;
- puis de choisir une variable qualitative qui prend toujours une valeur parmi deux possible. Elle définit donc deux groupes dans la population, un pour chacune de ces deux valeurs.

Ce test suppose que la variable quantitative a même variance dans chacun des deux groupes. Jamovi propose de tester cet hypothèse par un « test d'homogéneité des variances », d'hypothèse nulle : les variances des deux groupes sont égales et d'hypothèse adverse : ces variances sont différentes.

Q 3 pour un risque d'erreur de 1ère espèce de 0,05, chercher si la dépense moyenne peut être considérée comme différente entre les hommes et les femmes, puis entre les sondés fréquentant le point de vente et ceux qui ne le fréquentent pas

3. Tests d'indépendance. Test du Chi2 de Pearson ;

Il porte sur 2 variables, chacune étant qualitative (c'est-à-dire, en termes jamovi, nominale) ou quantitative discrète (en termes Jamovi, ordinale)

1. **Les 2 hypothèses testées** sont
   (H0) les 2 variables sont indépendantes (hypothèse nulle)
   (H1) les 2 variables ne sont pas indépendantes (hypothèse adverse)
2. **La statistique utilisée** est $D^2 = \sum_{classes\ i,j} \frac{(n_{i,j}^{obs} - n_{i,j}^{theo})^2}{n_{i,j}^{theo}}$. Sous l'hypothèse H0, elle suit une loi de Chi2 de paramètre (m-1)(p-1).
   $n_{i,j}^{obs}$ Est l'effectif – c'est-à-dire le nombre d'observations – dans la cellule de ligne i et colonne j
   $n_{i,j}^{theo}$ est l'effectif attendu, ou théorique, en cas d'indépendance:
   $n_{i,j}^{theo} = \frac{(effectif\ total\ de\ la\ ligne\ i) \times (effectif\ total\ de\ la\ ligne\ j)}{effectif\ total}$

Il est couramment admis que le test peut s'appliquer si

- (nombre de lignes -1) * (nombre de colonnes – 1) ≥ 2
- les effectifs attendus $n_{i,j}^{theo}$ doivent être supérieurs ou égaux à 5 dans au moins 80% des cases du tableau

Dans le cas où ces deux conditions ne seraient pas satisfaites, Jamovi fournit un test qui reste applicable : « test exact de Fisher »

La règle de décision est la suivante : on rejette H0 si D^2 est supérieure à une valeur d^2max
d^2_{max} est fixée en choisissant un risque d'erreur de 1ère espèce :

$$\alpha = P_{/H0\,vraie}(rejeter\ H0) = P(D^2\ de\ loi\ \chi 2\,(m-1)(p-1) > d^2_{max}).$$

Une valeur couramment utilisée pour le risque α est 0,05 , on rencontre aussi 0,01.

Jamovi fournit le seuil de significativité – ou p-valeur – , c'est-à-dire la probabilité de rejeter H0 à tort avec les données présentes dans le tableau. Donc

- **si le seuil de significativité est supérieur à** α **choisi** , on accepte H0 et on accepte que les deux variables ne sont **pas significativement liées**.
- **si le seuil de significativité est inférieur à** α **choisi** , on rejette H0 , on accepte que les variables **sont** significativement liées.

Sous Jamovi le test du Chi2 s'effectue à l'occasion de la construction d'un tableau contingence ou tableau croisé. Il est obtenu par : **Analyses→ Fréquences → tableaux de contingence → Echantillons indépendants. .** Une fenêtre s'ouvre pour demander quelles sont les 2 variables à placer en ligne et en colonne.
Une cellule « Quantité » permet d'indiquer une variable quantitative . Dans ce cas la case du tableau de ligne i et de colonne j , au lieu de contenir le nombre d'individus concernés par la ième valeur de la variable définissant la ligne et de la j ème valeur de la variable définissant la colonne, contient la somme des valeurs de cette variable quantitative pour ces individus – par exemple, leur consommation totale d'un produit --.
La cellule nommée « couches » permet d'indiquer des variables nominales ou discrètes. Un tableau de contingence est élaboré pour chacune de leurs valeurs.

Dans la section « Cellules » de la fenêtre, on vous propose de cocher une case pour afficher dans chaque cellule du tableau l'effectif attendu $n_{i,j}^{theo}$,

Un tableau "Tests du Khi carré" apparaît dans la fenêtre sortie. Il fournit la valeur de la statistique du Chi2, et un seuil de significativité.

Q 4 Effectuer un test du Chi2 pour décider si le fait d'écouter régulièrement la radio est lié au statut marital. Donner les hypothèses testées, vérifier si les conditions de validité du test sont satisfaites, et conclure.

Q 5 On cherche à juger si les variables Radio et Genre sont liées. Construire un tableau croisé portant sur ces variables. Les conditions de validité du test du Chi2 sont-elles réunies ? Que proposez-vous ?

Q 6 On voudrait juger si le revenu du foyer est lié au statut marital. Peut-on appliquer un test du Chi2 d'indépendance à ces deux variables ?

Discrétisation de variables.

A partir d'une variable numérique continue – ici par exemple le revenu du foyer, ou la dépense – **on construit une variable ne pouvant prendre que quelques valeurs**. Cela permet de regrouper les individus en quelques classes de valeurs. .

Ici à partir de la variable revenu_foyer, on construit une variable classe_revenu_3 qui ne peut prendre que 3 valeurs : *modéré* pour les revenus t inférieurs à 30 , *moyen* pour les revenus de 30 inclus à 50 exclus , *élevé* pour les revenus supérieurs à 50 inclus. Voici comment :

- ➔ Si le revenu du foyer est inférieur à 30, alors la classe de revenu sera *modéré*
- ➔ Sinon :
  - Si le revenu est strictement inférieur à 50 – il est donc entre 30 inclus et 50 exclus –, alors la classe est *moyen*
  - Sinon – le revenu est donc supérieur à 50 – , alors la classe de revenu est *élevé*.

L'instruction IF de Jamovi permet cette construction.

Q 7 Dans l'onglet **Variables** ou l'onglet Données, cliquer sur l'icône « Calcul »

..Jamovi crée **une nouvelle variable, calculée à partir des variables déjà présentes**, son nom est une lettre, vous pouvez le changer pour classe_revenu_3 , ou n'imprte quel terme que vous jugez explicite. Une cellule « description » permet d'indiquer en toutes lettres qu'il s'agit d'une classe de revenus parmi 3.

Une cellule Formule permet de calculer cette nouvelle variable .

Cliquer sur le bouton d'assistance à la création de formules, . Des fonctions et les noms de variables s'affichent. Choisir la fonction IF, parmi les fonctions de logique. Une aide nous indique qu 'une instruction IF se construit ainsi *IF*(*expression*, *value, else*).
Dans *expression* doit se trouver l'expression dont jamovi décidera si en fonction des données, elle est vraie ou fausse. A la place de *value*, se trouve la valeur à prendre si *expression* est vraie , à la place de *else* , se trouve la valeur à prendre si *expression* est fausse.

Vous pouvez entrer à la main, la formule suivante, ou vous aider de Jamovi en cliquant successivement sur IF, puis sur les noms des variables qui apparaissent à droite.

```
IF(revenu_foyer < 30,"modéré", IF(revenu_foyer <50, "moyen", "élevé"))
```

La variable classe_revenu_3 est créée, elle apparaît dans la fenêtre données.

Q 8 Test du Chi2 d'indépendance portant sur la classe de revenu du foyer et le statut marital.

Vous pouvez maintenant exécuter le test du Chi2 d'indépendance entre la classe de revenu et le statut marital. Conclusion ?