SAS®
Initiation et manipulation de données

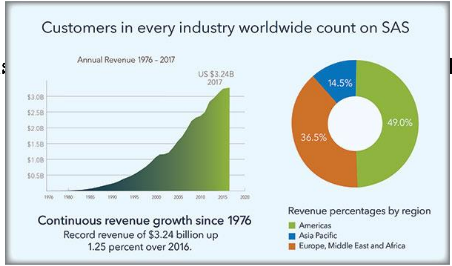
Nicolas Poulin

poulin@math.unistra.fr

Quelques mots sur SAS®.

- SAS Institute Inc. : société fondée en 1976 à l'Université de Caroline du Nord.

Source :

http://www.sashtml#history

poulin@math.unistra.fr

Quelques mots sur SAS®.

- SAS® :
  - Statistical Analysis System.
  - Langage de programmation (depuis 1976).
  - Version actuelle : SAS 9.4
  - SAS Entreprise Guide

poulin@math.unistra.fr

Quelques mots sur SAS®.

Un script SAS® se décompose généralement en 2 étapes :

- Etape DATA :
  - voir UE : PIP : certification SAS®.
  - Importation des données.
  - Gestion des variables.
  - Création de variables.
  - ...
- PROCédures : traitement des données...
  - commence par PROC
  - fini par RUN

poulin@math.unistra.fr

Préparation des données.

- télécharger le jeu de données *mais.csv* sur `http://www-irma.u-strasbg.fr/~gardes/teaching.html`
- remplacer les NA par des espaces vides dans excel.
- sauvegarder le fichier au format *.xlsx*, dans un répertoire nommé UElogiciels.

poulin@math.unistra.fr

Import d'un fichier *.xlsx* en SAS®.

1. On donne un nom au fichier :

```
FILENAME mais "/folders/myfolders/Documents/UElogiciels/mais.xlsx";
```


poulin@math.unistra.fr

Import d'un fichier *.xlsx* en SAS®.

1. On donne un nom au fichier :

```
FILENAME mais "/folders/myfolders/Documents/UElogiciels/mais.xlsx";
```

2. On importe les données :

```
PROC IMPORT DATAFILE= mais
    DMBS= xlsx
    OUT= work.mais1;
    GETNAMES=YES;
RUN;
```


poulin@math.unistra.fr

Exploration des données

1. Affichage des attributs des variables :

poulin@math.unistra.fr

Exploration des données

1. Affichage des attributs des variables :

```
PROC CONTENTS DATA= work.mais1; RUN;
```


poulin@math.unistra.fr

Exploration des données

1. Affichage des attributs des variables :

```
PROC CONTENTS DATA= work.mais1; RUN;
```

2. Impression des données :

Exploration des données

1. Affichage des attributs des variables :

```
PROC CONTENTS DATA= work.mais1; RUN;
```

2. Impression des données :

```
PROC PRINT DATA= work.mais1; RUN;
```


poulin@math.unistra.fr

Statistiques descriptives pour des variables quantitatives

1. Sur l'ensemble du jeu de données :

```
PROC MEANS DATA= work.mais1;
    var Hauteur Masse;
RUN;
```

2. Statistiques descriptives dans les niveaux d'une variable qualitative :

```
PROC MEANS DATA= work.mais1;
    var Hauteur Masse;
    class Parcelle;
RUN;
```

3. Statistiques descriptives dans les niveaux de plusieurs variables qualitatives nichées :

```
PROC MEANS DATA= work.mais1;
    var Hauteur Masse;
    class Parcelle Couleur;
RUN;
```

Remarque

L'ordre dans lequel les variables qualitatives sont écrites est important. Ici il s'agit de Couleur dans Parcelle.

poulin@math.unistra.fr

PROC FREQ de SAS®.

```
PROC FREQ <options> ;
    BY variables ;
    EXACT statistic-options </ computation-options> ;
    OUTPUT <OUT=SAS-data-set> options ;
    TABLES requests </ options> ;
    TEST options ;
    WEIGHT variable </ option> ;
RUN;
```

PROC FREQ de SAS®.

- BY : pour des analyses par groupe
- EXACT : pour faire des tests exacts
- OUTPUT : création d'un dataset en sortie
- TABLES : spécification des tableaux et analyses
- TEST : tests de mesure d'association
- WEIGHT : définition d'une variable de pondération.

poulin@math.unistra.fr

PROC FREQ de SAS®.

- BY : pour des analyses par groupe
- EXACT : pour faire des tests exacts
- OUTPUT : création d'un dataset en sortie
- TABLES : spécification des tableaux et analyses
- TEST : tests de mesure d'association
- WEIGHT : définition d'une variable de pondération.

Remarque

Pour pouvoir utiliser le paramètre BY il faut trier le tableau de données en fonction des niveaux de la variable qualitative.

Statistiques descriptives pour des variables qualitatives

1. variables qualitatives :

```
PROC FREQ DATA= work.mais1 nlevels;
    TABLES Parcelle Couleur;
RUN;
```

Remarque

L'option nlevels permet d'afficher le tableau des niveaux du facteur des variables spécifiées dans le paramètre TABLES.

poulin@math.unistra.fr

Statistiques descriptives pour des variables qualitatives

2 variables qualitatives emboîtées :

- Il faut tout d'abord trier le jeu de données.
- Si on veut les fréquences des parcelles pour les différentes couleurs :

```
PROC SORT DATA= work.mais1;
    BY Couleur Parcelle;
RUN;
```

- On peut ensuite utiliser la PROC FREQ

```
PROC FREQ DATA= work.mais1 nlevels;
    TABLES Parcelle;
    BY Couleur;
RUN;
```


poulin@math.unistra.fr

La loi du χ^2.

En probabilité et statistique des variables aléatoires sont distribuées selon certaines lois de probabilité. Parmi elles, la loi du χ^2 est une des plus connues.

poulin@math.unistra.fr

La loi du χ^2.

En probabilité et statistique des variables aléatoires sont distribuées selon certaines lois de probabilité. Parmi elles, la loi du χ^2 est une des plus connues.

Le paramètre ν d'une loi du χ^2 est appelé nombre de dégrés de liberté.

poulin@math.unistra.fr

La loi du χ^2.

En probabilité et statistique des variables aléatoires sont distribuées selon certaines lois de probabilité. Parmi elles, la loi du χ^2 est une des plus connues.

Le paramètre ν d'une loi du χ^2 est appelé nombre de dégrés de liberté.

La loi du χ^2 à ν degrés de liberté correspond à la somme de ν carrés de lois normales $\mathcal{N}(0,1)$.

poulin@math.unistra.fr

Tests du χ^2.

La loi du χ^2 est souvent utilisée pour les tests d'hypothèses.

poulin@math.unistra.fr

Tests du χ^2.

La loi du χ^2 est souvent utilisée pour les tests d'hypothèses.

Le nom de tests du χ^2 correspond aux tests du χ^2 de Pearson.

poulin@math.unistra.fr

Tests du χ^2.

La loi du χ^2 est souvent utilisée pour les tests d'hypothèses.

Le nom de tests du χ^2 correspond aux tests du χ^2 de Pearson.

Il existe 2 types de tests du χ^2 de Pearson :

- le test d'ajustement à une loi théorique : teste si un échantillon a pu être tiré selon une loi spécifique.

poulin@math.unistra.fr

Tests du χ^2.

La loi du χ^2 est souvent utilisée pour les tests d'hypothèses.

Le nom de tests du χ^2 correspond aux tests du χ^2 de Pearson.

Il existe 2 types de tests du χ^2 de Pearson :

- le test d'ajustement à une loi théorique : teste si un échantillon a pu être tiré selon une loi spécifique.
- le test d'indépendance : teste si 2 variables qualitatives observées sur 1 échantillon sont indépendantes.

poulin@math.unistra.fr

Test du χ^2 d'ajustement à une loi.

But : tester si une variable qualitative observée sur 1 échantillon est issue d'une loi de probabilité $\mathscr{L}$ spécifiée.

poulin@math.unistra.fr

Test du χ^2 d'ajustement à une loi.

But : tester si une variable qualitative observée sur 1 échantillon est issue d'une loi de probabilité $\mathscr{L}$ spécifiée.

Les hypothèses testées sont :

- H_0 : la variable suit la loi $\mathscr{L}$.
- H_1 : la variable ne suit pas la loi $\mathscr{L}$.

poulin@math.unistra.fr

Test du χ^2 d'ajustement à une loi.

But : tester si une variable qualitative observée sur 1 échantillon est issue d'une loi de probabilité $\mathscr{L}$ spécifiée.

Les hypothèses testées sont :

- H_0 : la variable suit la loi $\mathscr{L}$.
- H_1 : la variable ne suit pas la loi $\mathscr{L}$.

Le test est basé sur le tableau des effectifs (resp. de fréquences) qui présentent les occurences (resp. les fréquences) de la variable qualitative sur l'échantillon.

⚠ Anglais : la frequency table est le tableau des effectifs.

poulin@math.unistra.fr

Test du χ^2 pour l'ajustement à une loi.

- On dispose d'un dé à 6 faces dont on ne sait pas s'il est équilibré.
- On suppose que ce dé a été lancé $N = 100$ fois et que l'on a obtenu :

Face 1	**Face 2**	**Face 3**	**Face 4**	**Face 5**	**Face 6**
17	14	15	18	16	20

poulin@math.unistra.fr

Test du χ^2 pour l'ajustement à une loi.

- On dispose d'un dé à 6 faces dont on ne sait pas s'il est équilibré.
- On suppose que ce dé a été lancé $N = 100$ fois et que l'on a obtenu :

Face 1	**Face 2**	**Face 3**	**Face 4**	**Face 5**	**Face 6**
17	14	15	18	16	20

- Est-ce que ce dé est équilibré?

poulin@math.unistra.fr

Test du χ^2 pour l'ajustement à une loi.

Les hypothèses testées sont :

- H_0 : le dé est équilibré.
- H_1 : le dé n'est pas équilibré.

poulin@math.unistra.fr

Test du χ^2 pour l'ajustement à une loi.

Les hypothèses testées sont :

- H_0 : le dé est équilibré.
- H_1 : le dé n'est pas équilibré.

C'est-à-dire :

- H_0 : le résultat du lancé de dé suit une loi uniforme.
- H_1 : le résultat du lancé de dé ne suit pas une loi uniforme.

poulin@math.unistra.fr

Conditions d'application du test du χ^2.

- les individus composant l'échantillon ont été choisis aléatoirement (i.e. indépendance des observations).
- les classes des variables sont exclusives.
- Règle de Cochran : au moins 80% des effectifs théoriques sont au moins égaux à 5.
- La taille de l'échantillon doit être assez grande.

poulin@math.unistra.fr

Conditions d'application du test du χ^2.

- les individus composant l'échantillon ont été choisis aléatoirement (i.e. indépendance des observations).
- les classes des variables sont exclusives.
- Règle de Cochran : au moins 80% des effectifs théoriques sont au moins égaux à 5.
- La taille de l'échantillon doit être assez grande.

Sous H_0, chaque face à une probabilité de réalisation de

poulin@math.unistra.fr

Conditions d'application du test du χ^2.

- les individus composant l'échantillon ont été choisis aléatoirement (i.e. indépendance des observations).
- les classes des variables sont exclusives.
- Règle de Cochran : au moins 80% des effectifs théoriques sont au moins égaux à 5.
- La taille de l'échantillon doit être assez grande.

Sous H_0, chaque face à une probabilité de réalisation de $\frac{1}{6}$.

Remarque

Le nombre suffisant d'observations est assez subjectif : pour certains 25 observations sont suffisantes, pour d'autres cela peut être 30 ou 50.

poulin@math.unistra.fr

Tableau de contingence théorique.

Si H_0 est vraie, sur 100 essais nous nous attendons à obtenir le tableau :

Face 1	Face 2	Face 3	Face 4	Face 5	Face 6
16.667	16.667	16.667	16.667	16.667	16.667

poulin@math.unistra.fr

Statistique de décision du test du χ^2.

La statistique de décision est une mesure de la distance entre le tableau observé et le tableau théorique :

$$\chi^2_{obs} = \sum_{i=1}^{k} \frac{(n_i - t_i)^2}{t_i}$$

où

- k est le nombre de classes.
- n_i sont les effectifs observés de la classe i.
- t_i sont les effectifs théoriques de la classe i.

Sous H_0 la statistique de décision suit une loi du χ^2 à $\nu = k - 1$ où k est le nombre de classes.

poulin@math.unistra.fr

Test du χ^2 d'ajustement avec SAS®.

1. Grâce à l'étape DATA, créer le jeu de données.

poulin@math.unistra.fr

Test du χ^2 d'ajustement avec SAS®.

1. Grâce à l'étape DATA, créer le jeu de données.

```
DATA dice;
```


poulin@math.unistra.fr

Test du χ^2 d'ajustement avec SAS®.

1. Grâce à l'étape DATA, créer le jeu de données.

```
DATA dice;
    INPUT face $ effectif;
```


poulin@math.unistra.fr

Test du χ^2 d'ajustement avec SAS®.

1. Grâce à l'étape DATA, créer le jeu de données.

```
DATA dice;
    INPUT face $ effectif;
    CARDS;
```


poulin@math.unistra.fr

Test du χ^2 d'ajustement avec SAS®.

1. Grâce à l'étape DATA, créer le jeu de données.

```
DATA dice;
    INPUT face $ effectif;
    CARDS;
    Face1 17
    Face2 14
    Face3 15
    Face4 18
    Face5 16
    Face6 20
;
```


poulin@math.unistra.fr

Test du χ^2 d'ajustement avec SAS®.

2. Grâce à PROC FREQ, procéder au test du χ^2 d'ajustement.

poulin@math.unistra.fr

Test du χ^2 d'ajustement avec SAS®.

2. Grâce à PROC FREQ, procéder au test du χ^2 d'ajustement.

Options de TABLES :

- mises après le symbole /
- CHISQ : indique que l'on veut un test du χ^2
- TESTP : proportions théoriques testées
- TESTF : effectifs théoriques testés
- NOCUM : effectifs cumulés non affichés dans le tableau

poulin@math.unistra.fr

Test du χ^2 d'ajustement avec SAS®.

2. Grâce à PROC FREQ, procéder au test du χ^2 d'ajustement.

Options de TABLES :

- mises après le symbole /
- CHISQ : indique que l'on veut un test du χ^2
- TESTP : proportions théoriques testées
- TESTF : effectifs théoriques testés
- NOCUM : effectifs cumulés non affichés dans le tableau

Paramètre de WEIGHT : donne le nom de la variable contenant les effectifs.

poulin@math.unistra.fr

Test du χ^2 d'ajustement avec SAS®.

```
PROC FREQ DATA =work.dice ORDER=DATA;
    TABLES face / CHISQ NOCUM
    TESTF=(16.667 16.667 16.667 16.667 16.667 16.667);
    WEIGHT effectif;
RUN;
```


poulin@math.unistra.fr

Test du χ^2 d'ajustement avec SAS®.

```
PROC FREQ DATA =work.dice ORDER=DATA;
    TABLES face / CHISQ NOCUM
    TESTF=(16.667 16.667 16.667 16.667 16.667 16.667);
    WEIGHT effectif;
RUN;
```

Remarque

L'option ORDER=DATA permet que les niveaux des variables qualitatives soient ordonnés comme dans le jeu de données.

poulin@math.unistra.fr

Test du χ^2 d'ajustement avec SAS®.

- BUT : tester si la répartition des couleurs dans le jeu de données *mais* est uniforme.

- Il faut d'abord trier le jeu de données :

```
PROC SORT DATA= work.mais1;
      BY Couleur;
RUN;
```


poulin@math.unistra.fr

Test du χ^2 d'ajustement avec SAS®.

- On peut ensuite utiliser la PROC FREQ :

```
PROC FREQ DATA= work.mais1 ORDER=DATA;
      TABLES Couleur / CHISQ TESTP=(0.334 0.334 0.334);
RUN;
```


poulin@math.unistra.fr

Exercice.

- On suppose que 30% des parcelles sont orientées au nord, 20% au sud, 17% à l'est et 33% à l'ouest.
- Tester si la supposition ci-dessus peut être considérée comme vrai sur l'ensemble de la population.

poulin@math.unistra.fr

Test du χ^2 d'indépendance.

But : tester si 2 variables qualitatives observées sur 1 échantillon sont indépendantes.

poulin@math.unistra.fr

Test du χ^2 d'indépendance.

But : tester si 2 variables qualitatives observées sur 1 échantillon sont indépendantes.

Les hypothèses testées sont :

- H_0 : les variables sont indépendantes.
- H_1 : les variables ne sont pas indépendantes.

poulin@math.unistra.fr

Test du χ^2 d'indépendance.

But : tester si 2 variables qualitatives observées sur 1 échantillon sont indépendantes.

Les hypothèses testées sont :

- H_0 : les variables sont indépendantes.
- H_1 : les variables ne sont pas indépendantes.

Le test est basé sur les tableaux de contingence qui présentent les occurences croisées de 2 variables qualitatives.

poulin@math.unistra.fr

Test du χ^2 d'indépendance.

But : tester si 2 variables qualitatives observées sur 1 échantillon sont indépendantes.

Les hypothèses testées sont :

- H_0 : les variables sont indépendantes.
- H_1 : les variables ne sont pas indépendantes.

Le test est basé sur les tableaux de contingence qui présentent les occurences croisées de 2 variables qualitatives.

Exemple de tableau de contingence :

	BLOND	**CHATAIN**	**BRUN**	**ROUX**
BLEU	25	9	3	7
GRIS ou VERT	13	17	10	7
MARRON	7	13	8	5

poulin@math.unistra.fr

Conditions d'applications.

- les individus composant l'échantillon ont été choisis aléatoirement (i.e. indépendance des observations).

Conditions d'applications.

- les individus composant l'échantillon ont été choisis aléatoirement (i.e. indépendance des observations).
- les classes des variables sont exclusives.

poulin@math.unistra.fr

Conditions d'applications.

- les individus composant l'échantillon ont été choisis aléatoirement (i.e. indépendance des observations).
- les classes des variables sont exclusives.
- Règle de Cochran : au moins 80% des effectifs théoriques sont au moins égaux à 5.

poulin@math.unistra.fr

Conditions d'applications.

- les individus composant l'échantillon ont été choisis aléatoirement (i.e. indépendance des observations).
- les classes des variables sont exclusives.
- Règle de Cochran : au moins 80% des effectifs théoriques sont au moins égaux à 5.
- La taille de l'échantillon doit être assez grande.

Remarque

Le nombre suffisant d'observations est assez subjectif : pour certains 25 observations sont suffisantes, pour d'autres cela peut être 30 ou 50.

poulin@math.unistra.fr

Tableau des effectifs théoriques.

- Le test consiste à mesurer si la différence entre ce qu'on observe et ce qui devrait idéalement se passer en cas d'indépendance est statistiquement significatif.

Tableau des effectifs théoriques.

- Le test consiste à mesurer si la différence entre ce qu'on observe et ce qui devrait idéalement se passer en cas d'indépendance est statistiquement significatif.
- Il faut construire un tableau d'effectifs théoriques basé sur les marges du tableau de contingence observé :

$$t_{ij} = \frac{n_{i.} \times n_{.j}}{n}$$

où

  - t_{ij} est l'effectif théorique pour la ligne i et la colonne j.
  - $n_{i.}$ la somme des effectifs observés pour la ligne i.
  - $n_{.j}$ la somme des effectifs observés pour la colonne j.

Remarque

Le tableau des effectifs théoriques peut comporter des effectifs qui ne sont pas des entiers naturels.

poulin@math.unistra.fr

Tableau des effectifs théoriques.

- Tableau de contingence observé $(n_{i,j})$:

	BLOND	**CHATAIN**	**BRUN**	**ROUX**
BLEU	25	9	3	7
GRIS ou VERT	13	17	10	7
MARRON	7	13	8	5

- Tableau des effectifs théoriques $(t_{i,j})$:

	BLOND	**CHATAIN**	**BRUN**	**ROUX**
BLEU	15.96774	13.83871	7.451613	6.741935
GRIS ou VERT	17.05645	14.78226	7.959677	7.201613
MARRON	11.97581	10.37903	5.588710	5.056452

poulin@math.unistra.fr

Prise de décision.

- Statistique de décision : $\chi^2_{obs} = \sum_{i=1}^{k} \sum_{j=1}^{c} \frac{(n_{ij} - t_{ij})^2}{t_{ij}}$.

- Loi sous H_0 : loi du χ^2 à $\nu = (k-1)(c-1)$ ddl où k et c sont les nombres de classes des deux variables.

poulin@math.unistra.fr

Prise de décision.

- Statistique de décision : $\chi^2_{obs} = \sum_{i=1}^{k} \sum_{j=1}^{c} \frac{(n_{ij} - t_{ij})^2}{t_{ij}}$.

- Loi sous H_0 : loi du χ^2 à $\nu = (k-1)(c-1)$ ddl où k et c sont les nombres de classes des deux variables.

- La décision est prise grâce à la p-value.

poulin@math.unistra.fr

Test du χ^2 d'indépendance avec SAS®.

1. Grâce à l'étape DATA, créer le jeu de données.

poulin@math.unistra.fr

Test du χ^2 d'indépendance avec SAS®.

1. Grâce à l'étape DATA, créer le jeu de données.

```
DATA COULEUR;
```


poulin@math.unistra.fr

Test du χ^2 d'indépendance avec SAS®.

1. Grâce à l'étape DATA, créer le jeu de données.

```
DATA COULEUR;
    INPUT YEUX $ CHEVEUX $ EFFECTIF;
```


poulin@math.unistra.fr

Test du χ^2 d'indépendance avec SAS®.

1. Grâce à l'étape DATA, créer le jeu de données.

```
DATA COULEUR;
      INPUT YEUX $ CHEVEUX $ EFFECTIF;
      CARDS;
```

Test du χ^2 d'indépendance avec SAS®.

1. Grâce à l'étape DATA, créer le jeu de données.

```
DATA COULEUR;
        INPUT YEUX $ CHEVEUX $ EFFECTIF;
        CARDS;
        BLEUS BLONDS 25
        BLEUS BRUNS 9
        BLEUS NOIRS 3
        BLEUS ROUX 7
        VERTS BLONDS 13
        VERTS BRUNS 17
        VERTS NOIRS 10
        VERTS ROUX 7
        MARRONS BLONDS 7
        MARRONS BRUNS 13
        MARRONS NOIRS 8
        MARRONS ROUX 5

;
```


poulin@math.unistra.fr

Test du χ^2 d'indépendance avec SAS®.

2. Grâce à PROC FREQ , créer le tableau de contingence observé.

poulin@math.unistra.fr

Test du χ^2 d'indépendance avec SAS®.

2. Grâce à PROC FREQ , créer le tableau de contingence observé.

```
PROC FREQ DATA=COULEUR ORDER=DATA;
```

Test du χ^2 d'indépendance avec SAS®.

2. Grâce à PROC FREQ , créer le tableau de contingence observé.

```
PROC FREQ DATA=COULEUR ORDER=DATA;
    TABLES YEUX*CHEVEUX ;
```


poulin@math.unistra.fr

Test du χ^2 d'indépendance avec SAS®.

2. Grâce à PROC FREQ , créer le tableau de contingence observé.

```
PROC FREQ DATA=COULEUR ORDER=DATA;
    TABLES YEUX*CHEVEUX ;
    WEIGHT EFFECTIF ;
```


poulin@math.unistra.fr

Test du χ^2 d'indépendance avec SAS®.

2. Grâce à PROC FREQ , créer le tableau de contingence observé.

```
PROC FREQ DATA=COULEUR ORDER=DATA;
    TABLES YEUX*CHEVEUX ;
    WEIGHT EFFECTIF ;
    TITLE1 "Tableau de contingence observé";
    TITLE2"- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -";
RUN;
```


poulin@math.unistra.fr

Test du χ^2 d'indépendance avec SAS®.

Le quadrant supérieur gauche du tableau indique (en anglais) le contenu de chaque case (i, j), à savoir :

- l'effectif $n_{i,j}$ (Fréquence)
- le Pourcentage correspondant à $f_{i,j} = n_{i,j}/N$
- le Pourcentage en ligne correspondant à $n_{i,j}/n_{i,.}$
- le Pourcentage en colonne correspondant à $n_{i,j}/n_{.,j}$

poulin@math.unistra.fr

Test du χ^2 d'indépendance avec SAS®.

Sur la ligne Total on peut lire :

- les effectifs $n_{.,j}$ des modalités de la variable colonne,
- les pourcentages ligne correspondant aux proportions $f_{.,j} = n_{.,j}/N$

C'est la ligne marginale donnant la distribution (le tri-à-plat) de la variable CHEVEUX sans distinction de la couleur des yeux.

Sur la colonne Total, colonne marginale, on lit de même la distribution de la variable YEUX dans l'ensemble de la population (effectifs $n_{i,.}$ et pourcentages colonne correspondant à $f_{i,.} = n_{i,.}/N$).

poulin@math.unistra.fr

Test du χ^2 d'indépendance avec SAS®.

3. Grâce à PROC FREQ , créer le tableau de contingence théorique.

poulin@math.unistra.fr

Test du χ^2 d'indépendance avec SAS®.

3. Grâce à PROC FREQ , créer le tableau de contingence théorique.

```
PROC FREQ DATA=COULEUR ORDER=DATA;
```


poulin@math.unistra.fr

Test du χ^2 d'indépendance avec SAS®.

3. Grâce à PROC FREQ , créer le tableau de contingence théorique.

```
PROC FREQ DATA=COULEUR ORDER=DATA;
    TABLES YEUX*CHEVEUX / EXPECTED;
    WEIGHT EFFECTIF ;
RUN;
```


poulin@math.unistra.fr

Test du χ^2 d'indépendance avec SAS®.

4. Grâce à PROC FREQ , tester si les variables YEUX et CHEVEUX sont indépendantes.

Test du χ^2 d'indépendance avec SAS®.

4. Grâce à PROC FREQ , tester si les variables YEUX et CHEVEUX sont indépendantes.

```
PROC FREQ DATA=COULEUR ORDER=DATA;
```


poulin@math.unistra.fr

Test du χ^2 d'indépendance avec SAS®.

4. Grâce à PROC FREQ , tester si les variables YEUX et CHEVEUX sont indépendantes.

```
PROC FREQ DATA=COULEUR ORDER=DATA;
    TABLES YEUX*CHEVEUX / CHISQ;
    WEIGHT EFFECTIF ;
RUN;
```


poulin@math.unistra.fr

Exercice.

Vérifier, sur l'exemple *mais* si la parcelle est la couleur sont des variables aléatoires indépendantes.

poulin@math.unistra.fr

Exercice.

- Il faut tout d'abord trier le jeu de données :

poulin@math.unistra.fr

Exercice.

- Il faut tout d'abord trier le jeu de données :

```
PROC SORT DATA= work.mais1;
    BY Parcelle Couleur;
RUN;
```


poulin@math.unistra.fr

Exercice.

- Il faut tout d'abord trier le jeu de données :

```
PROC SORT DATA= work.mais1;
    BY Parcelle Couleur;
RUN;
```

- On peut ensuite obtenir le tableau observé :

Exercice.

- Il faut tout d'abord trier le jeu de données :

```
PROC SORT DATA= work.mais1;
    BY Parcelle Couleur;
RUN;
```

- On peut ensuite obtenir le tableau observé :

```
PROC FREQ DATA= work.mais1 ORDER= DATA;
    TABLES Couleur*Parcelle;
RUN;
```

Remarque

Pas besoin de paramètre WEIGHT : les poids sont calculés automatiquement car il s'agit du tableau complet.

poulin@math.unistra.fr

Exercice.

- On peut ensuite obtenir le tableau théorique :

poulin@math.unistra.fr

Exercice.

- On peut ensuite obtenir le tableau théorique :

```
PROC FREQ DATA= work.mais1 ORDER= DATA;
    TABLES Couleur*Parcelle / EXPECTED;
RUN;
```


poulin@math.unistra.fr

Exercice.

- On peut ensuite obtenir le tableau théorique :

```
PROC FREQ DATA= work.mais1 ORDER= DATA;
    TABLES Couleur*Parcelle / EXPECTED;
RUN;
```

Remarque

Les conditions d'application du test du χ^2 d'indépendance ne sont pas vérifiées (trop d'effectifs théoriques inférieurs à 5). Il faut donc soit :

- *fusionner les niveaux Jaune.rouge et Rouge pourla variable Couleur*

poulin@math.unistra.fr

Exercice.

- On peut ensuite obtenir le tableau théorique :

```
PROC FREQ DATA= work.mais1 ORDER= DATA;
    TABLES Couleur*Parcelle / EXPECTED;
RUN;
```

Remarque

Les conditions d'application du test du χ^2 d'indépendance ne sont pas vérifiées (trop d'effectifs théoriques inférieurs à 5). Il faut donc soit :

- *fusionner les niveaux Jaune.rouge et Rouge pourla variable Couleur*
- *utiliser un test exact de Fisher*

poulin@math.unistra.fr

Exercice.

- Si on voulait tout de même faire le test du χ^2 :

Exercice.

- Si on voulait tout de même faire le test du χ^2 :

```
PROC FREQ DATA= work.mais1 ORDER= DATA;
     TABLES Couleur*Parcelle / CHISQ EXPECTED;
RUN;
```


poulin@math.unistra.fr

Alternatives : test exact de Fisher

Il peut être utilisé pour tester l'indépendance entre deux variables qualitatives quand la condition de Cochran n'est pas vérifiée ou pour de petits échantillons.

poulin@math.unistra.fr

Alternatives : test exact de Fisher

Il peut être utilisé pour tester l'indépendance entre deux variables qualitatives quand la condition de Cochran n'est pas vérifiée ou pour de petits échantillons.

Hypothèses :

- H_0: les variables sont indépendantes.
- H_1: les variables ne sont pas indépendantes.

poulin@math.unistra.fr

Alternatives : test exact de Fisher

Il peut être utilisé pour tester l'indépendance entre deux variables qualitatives quand la condition de Cochran n'est pas vérifiée ou pour de petits échantillons.

Hypothèses :

- H_0: les variables sont indépendantes.
- H_1: les variables ne sont pas indépendantes.

Remarque

Le test est dit exact car il calcule en fait la probabilité qu'un tableau de contingence observé le soit sous une hypothèse d'indépendance (par des techniques de dénombrement).

poulin@math.unistra.fr

Test exact de Fisher

Conditions d'application :

- les individus composant l'échantillon ont été choisis aléatoirement (i.e. indépendance des observations).

poulin@math.unistra.fr

Test exact de Fisher

Conditions d'application :

- les individus composant l'échantillon ont été choisis aléatoirement (i.e. indépendance des observations).
- les classes des variables sont exclusives.

poulin@math.unistra.fr

Test exact de Fisher

Conditions d'application :

- les individus composant l'échantillon ont été choisis aléatoirement (i.e. indépendance des observations).
- les classes des variables sont exclusives.

Test exact de Fisher avec SAS®.

```
PROC FREQ DATA= work.mais1 ORDER= DATA;
    TABLES Couleur*Parcelle / EXPECTED;
    EXACT Fisher;
RUN;
```


poulin@math.unistra.fr

Alternatives : G-test (Likelihood ratio Chi-Square)

- Le test du χ^2 de Pearson est basé sur une approximation d'un ration de log-vraisemblance.

poulin@math.unistra.fr

Alternatives : G-test (Likelihood ratio Chi-Square)

- Le test du χ^2 de Pearson est basé sur une approximation d'un ration de log-vraisemblance.
- L'utilisation du G-test est recommandé dans le livre de Robert R. Sokal et F. James Rohlf (1981), *Biometry : the principles and practive of statistics in biological research*, New-York: Freeman.

poulin@math.unistra.fr

Alternatives : G-test (Likelihood ratio Chi-Square)

- Le test du χ^2 de Pearson est basé sur une approximation d'un ration de log-vraisemblance.
- L'utilisation du G-test est recommandé dans le livre de Robert R. Sokal et F. James Rohlf (1981), *Biometry : the principles and practive of statistics in biological research*, New-York: Freeman.
- Le G-test n'utilise pas l'approximation mais calcule le vrai rapport de log-vraisemblance ce qui permet d'obtenir des résultats plus fiables.
- Les hypothèses et conditions d'application du G-test sont les mêmes que celles du test du chi^2 mais sans la limite de la taille d'échantillon.

G-test avec SAS®

Paramétres :

poulin@math.unistra.fr

G-test avec SAS®

Paramétres :

- Loi sous H_0 : loi du χ^2 à $\nu = (k-1)(c-1)$ ddl où k et c sont les nombres de classes des deux variables.

- Résultat donné dans le tableau des tests lorsqu'on utilise l'option CHISQ du paramètre TABLES

poulin@math.unistra.fr

Correction de continuité pour les tableaux 2×2

Lorsque l'on traite un tableau de contingence pour 2 variables qualitatives comportant chacune 2 niveaux, il est nécessaire d'appliquer la correction de continuité de Yates.

poulin@math.unistra.fr

Correction de continuité pour les tableaux 2×2

Lorsque l'on traite un tableau de contingence pour 2 variables qualitatives comportant chacune 2 niveaux, il est nécessaire d'appliquer la correction de continuité de Yates.

En effet, le test du χ^2 suppose que des probabilités discrétes (loi binomiale) peut être approchée par un distribution du χ^2 qui est continue.

poulin@math.unistra.fr

Correction de continuité pour les tableaux 2×2

Lorsque l'on traite un tableau de contingence pour 2 variables qualitatives comportant chacune 2 niveaux, il est nécessaire d'appliquer la correction de continuité de Yates.

En effet, le test du χ^2 suppose que des probabilités discrétes (loi binomiale) peut être approchée par un distribution du χ^2 qui est continue.

poulin@math.unistra.fr

Correction de continuité pour les tableaux 2×2

Cette approximation n'est plus valalble dans le cas de tableaux de contingences 2×2. Yates a proposé la correction :

- Statistique de décision : $\chi^2_{\text{Yates}} = \sum_{i=1}^{k} \sum_{j=1}^{c} \frac{\left(\left|n_{ij} - t_{ij}\right| - 0.5\right)^2}{t_{ij}}$.
- Loi sous H_0 : loi du χ^2 à $\nu = (k-1)(c-1)$ ddl où k et c sont les nombres de classes des deux variables.

poulin@math.unistra.fr

Correction de continuité pour les tableaux 2×2

Cette approximation n'est plus valalble dans le cas de tableaux de contingences 2×2. Yates a proposé la correction :

- Statistique de décision : $\chi^2_{\text{Yates}} = \sum_{i=1}^{k} \sum_{j=1}^{c} \frac{\left(\left|n_{ij} - t_{ij}\right| - 0.5\right)^2}{t_{ij}}$.
- Loi sous H_0 : loi du χ^2 à $\nu = (k-1)(c-1)$ ddl où k et c sont les nombres de classes des deux variables.
- La décision est prise grâce à la p-value.
- Résultat donné dans le tableau des tests lorsqu'on utilise l'option CHISQ du paramètre TABLES

poulin@math.unistra.fr

Correction de continuité pour les tableaux 2×2

SAS® donne aussi automatiquement le résultat du test exact de Fisher pour les tableaux 2×2 car certains statisticiens considèrent que l'approximation de Yates est obsoléte.

poulin@math.unistra.fr

Correction de continuité pour les tableaux 2×2

SAS® donne aussi automatiquement le résultat du test exact de Fisher pour les tableaux 2×2 car certains statisticiens considèrent que l'approximation de Yates est obsoléte.

Réaliser le test du χ^2 avec correction de Yates sur le tableau de contingence des variables *Attaque* et *Verse_Traitement*.

poulin@math.unistra.fr