Modèles linéaires - Statistical Analysis System (SAS)
Rappels de traitement statitiques en bi dimensionnel

Emmanuelle Gautherat$^{(a)}$

(a) Crieg-Regards, Université de Reims Champagne Ardenne

Second semestre - 4 ECTS

Outline

Introduction

Etude conjointe de deux variables.
–> selon la nature des variables considérées ;
–> passer en revue les indicateurs statistiques + objets graphiques.

Etude conjointe de deux variables.
–> selon la nature des variables considérées ;
–> passer en revue les indicateurs statistiques + objets graphiques.

1. Deux variables qualitatives
2. Deux variables qualitatives ordonnées
3. Deux variables quantitatives discrètes finies
4. Deux variables quantitatives discrète dénombrables
5. Deux variables quantitatives continues
6. Une variable qualitative-ordonnée ou non- et une variable quantitative -discrète ou continue-

Source

Les exemples de ce cours sont tirés de

- Référence : Chase, M. A., and Dummer, G. M. (1992), "The Role of Sports as a Social Determinant for Children", Research Quarterly for Exercise and Sport, 63, 418-424

Source

Les exemples de ce cours sont tirés de

- Référence : Chase, M. A., and Dummer, G. M. (1992), "The Role of Sports as a Social Determinant for Children", Research Quarterly for Exercise and Sport, 63, 418-424
- Les auteurs ont interrogé des enfants scolarisés dans des écoles sélectives du Michigan.
  Ils leur ont posé les questions suivantes :
  1. Qu'aimerais-tu le mieux faire à l'école : A. Avoir de bonnes notes, B. Etre bon en sport, C. Etre populaire ?
  2. Pour chacun de ces critères ils leur ont demandé de placer un chiffre marquant l'ordre d'importance que ces enfants accordent aux choix A,B,C ainsi que pour le choix "avoir beaucoup d'argent" ;
  3. Un ensemble d'informations socio-dém. ont été également recueillies.

Source

Gender	Grade	Age	Race	Urban/Rural	School	Goals	Grades	Sports	Looks	Money
girl	5	10	White	Suburban	Brentwood Elementary	Popular	3	2	1	4
girl	5	10	White	Suburban	Brentwood Elementary	Grades	1	3	2	4
girl	5	11	White	Suburban	Brentwood Elementary	Sports	1	3	2	4
girl	6	11	Other	Suburban	Brentwood Middle	Popular	2	3	1	4
girl	6	11	White	Suburban	Brentwood Middle	Popular	1	3	2	4
girl	6	11	White	Suburban	Brentwood Middle	Popular	1	3	4	2
girl	6	11	White	Suburban	Brentwood Middle	Grades	4	1	2	3
girl	6	11	White	Suburban	Brentwood Middle	Sports	1	2	3	4
girl	6	11	White	Suburban	Brentwood Middle	Grades	3	4	1	2
girl	6	11	White	Suburban	Brentwood Middle	Popular	1	2	3	4

Figure – extrait de la base de données *Chase, M. A., and Dummer, G. M. (1992), "The Role of Sports as a Social Determinant for Children"*

Deux variables qualitatives

On se restreint à deux variables qualitatives uniquement :

Gender	Goals
boy	Sports
boy	Popular
girl	Popular
girl	Popular
girl	Popular
girl	Popular
girl	Popular
girl	Grades
girl	Sports
girl	Sports
girl	Sports
girl	Grades
boy	Popular

Figure – extrait de la base de données réduite à deux variables qualitatives, *Chase, M. A., and Dummer, G. M. (1992), "The Role of Sports as a Social Determinant for Children"*

Indicateurs

tableau de contingence –> perte des données individuelles

Indicateurs

tableau de contingence –> perte des données individuelles –> vue synthétique selon les modalités des variables

		Gender		Total
		boy	girl	
Goals	Grades	117	130	247
	Popular	50	91	141
	Sports	60	30	90
Total		227	251	478

Figure – Tableau de contingence de l'extrait de la base de données réduite à deux variables qualitatives, *Chase, M. A., and Dummer, G. M. (1992), "The Role of Sports as a Social Determinant for Children"*

Indicateurs

Profils lignes –> permet de donner une "masse" égale à chaque ligne : on peut comparer les modalités des colonnes (donc les lignes).

Indicateurs

Profils lignes –> permet de donner une "masse" égale à chaque ligne : on peut comparer les modalités des colonnes (donc les lignes).

		Gender		Total
		boy	girl	
Goals	Grades	47,4%	52,6%	100,0%
	Popular	35,5%	64,5%	100,0%
	Sports	66,7%	33,3%	100,0%
Total		47,5%	52,5%	100,0%

Figure – Profils lignes du tableau de contingence, extrait de la base de données réduite à deux variables qualitatives, *Chase, M. A., and Dummer, G. M. (1992), "The Role of Sports as a Social Determinant for Children"*

Indicateurs

Profils colonnes –> permet de donner une "masse" égale à chaque colonne : on peut comparer les modalités des lignes (donc les colonnes).

Indicateurs

Profils colonnes –> permet de donner une "masse" égale à chaque colonne : on peut comparer les modalités des lignes (donc les colonnes).

		Gender		Total
		boy	girl	
Goals	Grades	51,5%	51,8%	51,7%
	Popular	22,0%	36,3%	29,5%
	Sports	26,4%	12,0%	18,8%
Total		100,0%	100,0%	100,0%

Figure – Profils colonnes du tableau de contingence, extrait de la base de données réduite à deux variables qualitatives, *Chase, M. A., and Dummer, G. M. (1992), "The Role of Sports as a Social Determinant for Children"*

Indicateurs

Test du chi 2 d'indépendance
Variables issues d'un échantillon iid + au moins 5 données dans les effectifs théoriques.

Indicateurs

Test du chi 2 d'indépendance
Variables issues d'un échantillon iid + au moins 5 données dans les effectifs théoriques.

$$H_0 : \ X \text{ indépendante de } \ Y \quad \text{contre} \quad H_1 : X \ et \ Y \text{dépendantes}$$

Le risque de rejeter H_0 à tort doit être au plus de α (risque de déclarer l'indépendance, à tort).

$$\Delta_{\chi^2} = \sum_i \sum_j \frac{\left(n_{ij} - \frac{n_{i.} n_{.j}}{n}\right)^2}{\frac{n_{i.} n_{.j}}{n}}.$$

En d'autres termes

$$\Delta_{\chi^2} = \sum_i \sum_j \frac{\left(n_{obs} - n_{theo}\right)^2}{n_{theo}}.$$

Indicateurs

Test du chi 2 d'indépendance
Variables issues d'un échantillon iid + au moins 5 données dans les effectifs théoriques.

$$H_0 : \ X \text{ indépendante de } \ Y \quad \text{contre} \quad H_1 : X \ et \ Y \text{dépendantes}$$

Le risque de rejeter H_0 à tort doit être au plus de α (risque de déclarer l'indépendance, à tort).

$$\Delta_{\chi^2} = \sum_i \sum_j \frac{\left(n_{ij} - \frac{n_{i.}n_{.j}}{n}\right)^2}{\frac{n_{i.}n_{.j}}{n}}.$$

En d'autres termes

$$\Delta_{\chi^2} = \sum_i \sum_j \frac{\left(n_{obs} - n_{theo}\right)^2}{n_{theo}}.$$

Cas de deux variables à deux modalités spécifiques (correction de Yates).

Indicateurs

	Valeur	ddl	Signification asymptotique (bilatérale)
Khi-deux de Pearson	21,455[a]	2	,000
Rapport de vraisemblance	21,769	2	,000
Nombre d'observations valides	478		

a. 0 cellules (0,0%) ont un effectif théorique inférieur à 5. L'effectif théorique minimum est de 42,74.

Figure – Chi2 et sa p-value sur la base d'un extrait de la base de données réduite à deux variables qualitatives,*Chase, M. A., and Dummer, G. M. (1992), "The Role of Sports as a Social Determinant for Children"*

Indicateurs

	Valeur	ddl	Signification asymptotique (bilatérale)
Khi-deux de Pearson	21,455[a]	2	,000
Rapport de vraisemblance	21,769	2	,000
Nombre d'observations valides	478		

a. 0 cellules (0,0%) ont un effectif théorique inférieur à 5. L'effectif théorique minimum est de 42,74.

Figure – Chi2 et sa p-value sur la base d'un extrait de la base de données réduite à deux variables qualitatives, *Chase, M. A., and Dummer, G. M. (1992), "The Role of Sports as a Social Determinant for Children"*

Rque : si n grand, tendance à rejeter l'indépendance (différence à 0 non expliquée par la fluctuation d'échantillonnage)

Indicateurs

Contribution au Test du chi 2 d'indépendance : les "petits chi 2"
Résidus standardisés$_{ij}$= (n_{ij} observé - n_{ij} théo) / $\sqrt{n_{ij}\text{théo}}$.

Petits χ^2_{ij} = (Résidus standardisés$_{ij}$)^2.

Le résidu standardisé outre le sens de la contribution (avec les signes) varie autour d'une moyenne de 0 et a un écart-type de 1 –> suit presque une loi normale (utile pour dire si grand ou pas).

Indicateurs

Contribution au Test du chi 2 d'indépendance : les "petits chi 2"
Résidus standardisés$_{ij}$= (n_{ij} observé - n_{ij} théo) / $\sqrt{n_{ij}\text{théo}}$.

Petits χ^2_{ij} = (Résidus standardisés$_{ij}$)^2.

Le résidu standardisé outre le sens de la contribution (avec les signes) varie autour d'une moyenne de 0 et a un écart-type de 1 –> suit presque une loi normale (utile pour dire si grand ou pas).

Le résidu standardisé ajusté est plus proche d'une loi normale.
Résidus standardisés ajustés$_{ij}$ = résidus standardisés$_{ij}$ / $\sqrt{\text{(1- fréquence marginale}_{i.})\ \text{(1- fréquence marginale}_{.j})}$
avec fréquence marginale$_{i.}$ = $\frac{n_{i.}}{n}$

Indicateurs

Contribution au Test du chi 2 d'indépendance : les "petits chi 2"
Résidus standardisés$_{ij}$= (n_{ij} observé - n_{ij} théo) / $\sqrt{n_{ij}\text{théo}}$.

Petits χ^2_{ij} = (Résidus standardisés$_{ij}$)^2.

Le résidu standardisé outre le sens de la contribution (avec les signes) varie autour d'une moyenne de 0 et a un écart-type de 1 –> suit presque une loi normale (utile pour dire si grand ou pas).

Le résidu standardisé ajusté est plus proche d'une loi normale.
Résidus standardisés ajustés$_{ij}$ = résidus standardisés$_{ij}$ / $\sqrt{\text{(1- fréquence marginale}_{i.})\ \text{(1- fréquence marginale}_{.j})}$
avec fréquence marginale$_{i.}$ = $\frac{n_{i.}}{n}$
–> préférable pour dire si grand ou pas en comparant aux quantiles de la loi normale

Indicateurs

		Gender	
		boy	girl
Goals	Grades	-,1	,1
	Popular	-3,4	3,4
	Sports	4,0	-4,0

Figure – "Petits χ^2" : résidus standardisés et ajustés sur la base d'un extrait de la base de données réduite à deux variables qualitatives, *Chase, M. A., and Dummer, G. M. (1992), "The Role of Sports as a Social Determinant for Children"*

Indicateurs

Test du χ^2 dit si liées ou pas -> avec quelle intensité ?
χ^2 dépend de n, de K et de L le nombre lignes et de colonnes du tableau de contingence.

Indicateurs

Test du χ^2 dit si liées ou pas -> avec quelle intensité ?
χ^2 dépend de n, de K et de L le nombre lignes et de colonnes du tableau de contingence.

- Carré moyen de contingence : $\Phi = \sqrt{\frac{\chi^2}{n}}$

Indicateurs

Test du χ^2 dit si liées ou pas -> avec quelle intensité ?
χ^2 dépend de n, de K et de L le nombre lignes et de colonnes du tableau de contingence.

- Carré moyen de contingence : $\Phi = \sqrt{\frac{\chi^2}{n}}$ –> élimine l'effet taille (n)
  rque : $\Phi_{max} = \sqrt{min(K-1; L-1)}$

Indicateurs

Test du χ^2 dit si liées ou pas -> avec quelle intensité ?
χ^2 dépend de n, de K et de L le nombre lignes et de colonnes du tableau de contingence.

- Carré moyen de contingence : $\Phi = \sqrt{\frac{\chi^2}{n}}$ –> élimine l'effet taille (n)
  rque : $\Phi_{max} = \sqrt{min(K-1; L-1)}$
- V de Cramer $V = \sqrt{\frac{\chi^2}{n\ min(L-1;K-1)}}$
  Varie entre 0 et 1.
  Ne dépend ni du nombre de lignes et de colonnes ni de la taille n de l'échantillon ;

Indicateurs

Test du χ^2 dit si liées ou pas -> avec quelle intensité ?
χ^2 dépend de n, de K et de L le nombre lignes et de colonnes du tableau de contingence.

- Carré moyen de contingence : $\Phi = \sqrt{\frac{\chi^2}{n}}$ –> élimine l'effet taille (n)
  rque : $\Phi_{max} = \sqrt{min(K-1; L-1)}$
- V de Cramer $V = \sqrt{\frac{\chi^2}{n\ min(L-1;K-1)}}$
  Varie entre 0 et 1.
  Ne dépend ni du nombre de lignes et de colonnes ni de la taille n de l'échantillon ;
- Coefficient de contingence C.

Indicateurs

Coefficient de contingence C ->

Indicateurs

Coefficient de contingence C -> issu du χ^2

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}}.$$

Indicateurs

Coefficient de contingence C -> issu du χ^2

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}}.$$

- absence d'association $C = 0$;
- concordance parfaite C devrait atteindre 1. En fait la valeur max de C dépend du nombre de colonnes et de lignes : on ne peut donc comparer que deux coefficients de contingence que pour des tableaux de même nbre de colonnes et de lignes.

Indicateurs

On obtient

- $\chi^2 = 21,455$ avec $n = 478$
- $\Phi = 0,212$
- $V_{Cramer} = 0,212$
- $C = 0,207$

Dispositifs graphiques

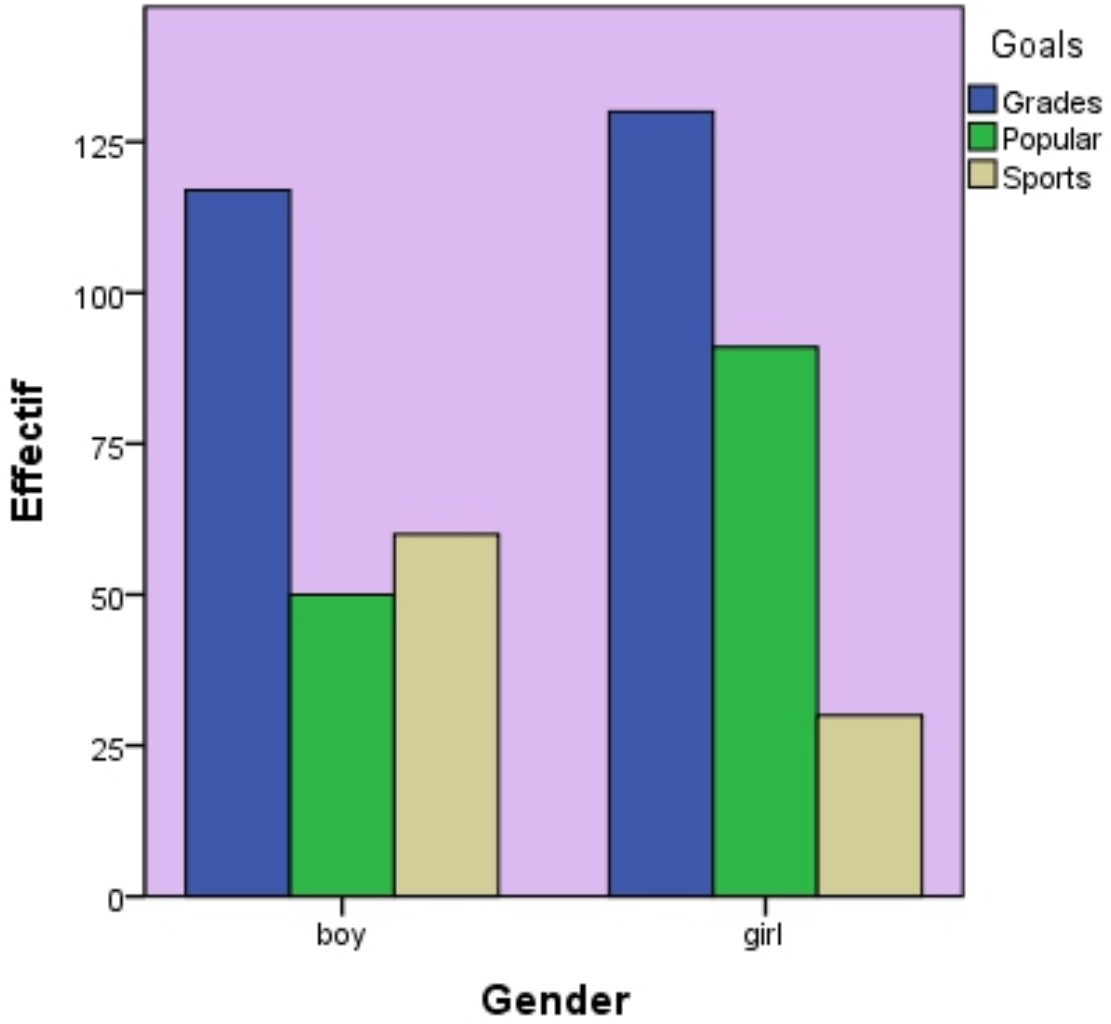


Figure – Graphe bâton superposé sur les effectifs : regroupement par sexe, *Chase, M.*

Dispositifs graphiques

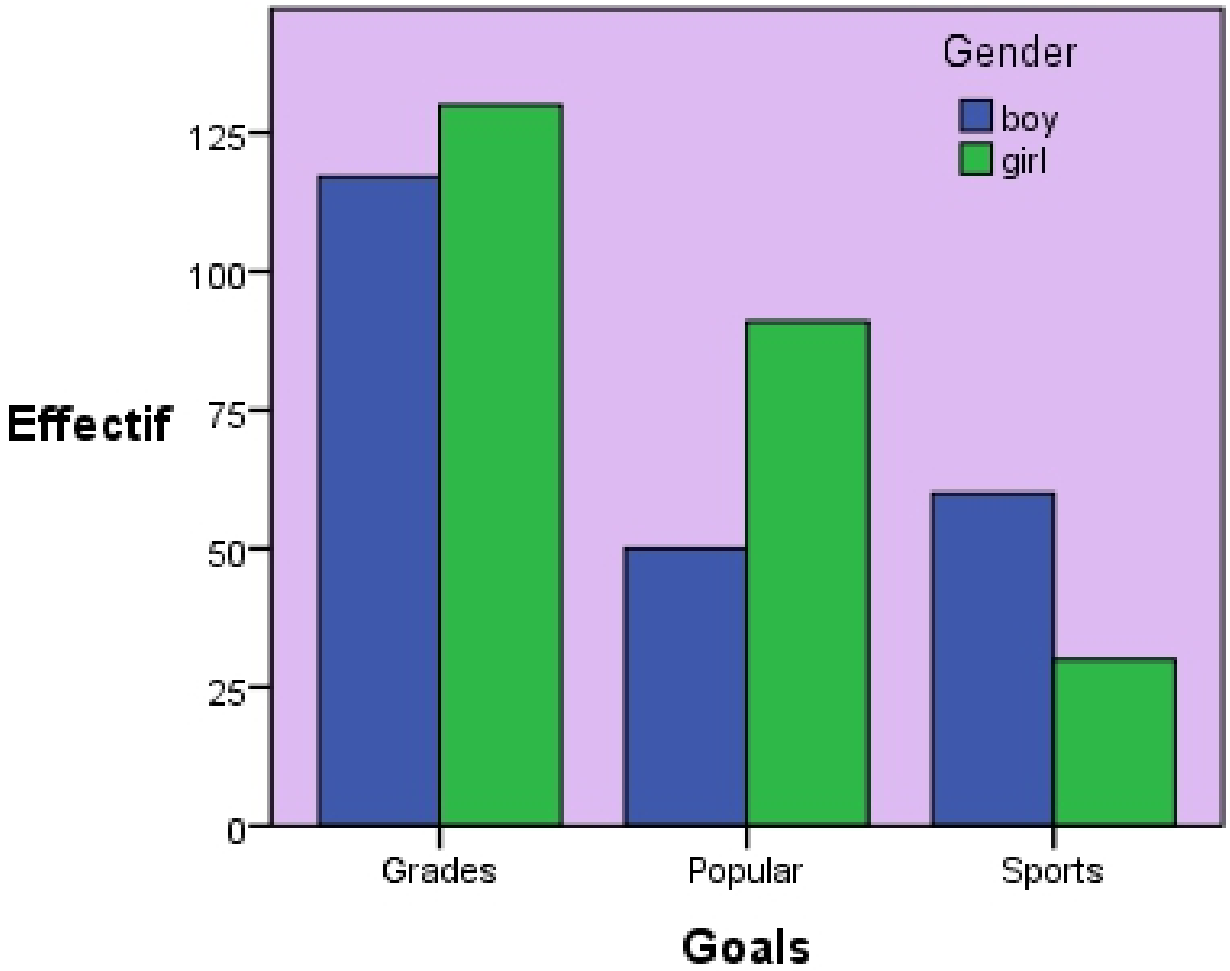


Figure – Graphe bâton superposé sur les effectifs : regroupement par choix des buts, *Chase, M. A., and Dummer, G. M. (1992), "The Role of Sports as a Social*

Dispositifs graphiques

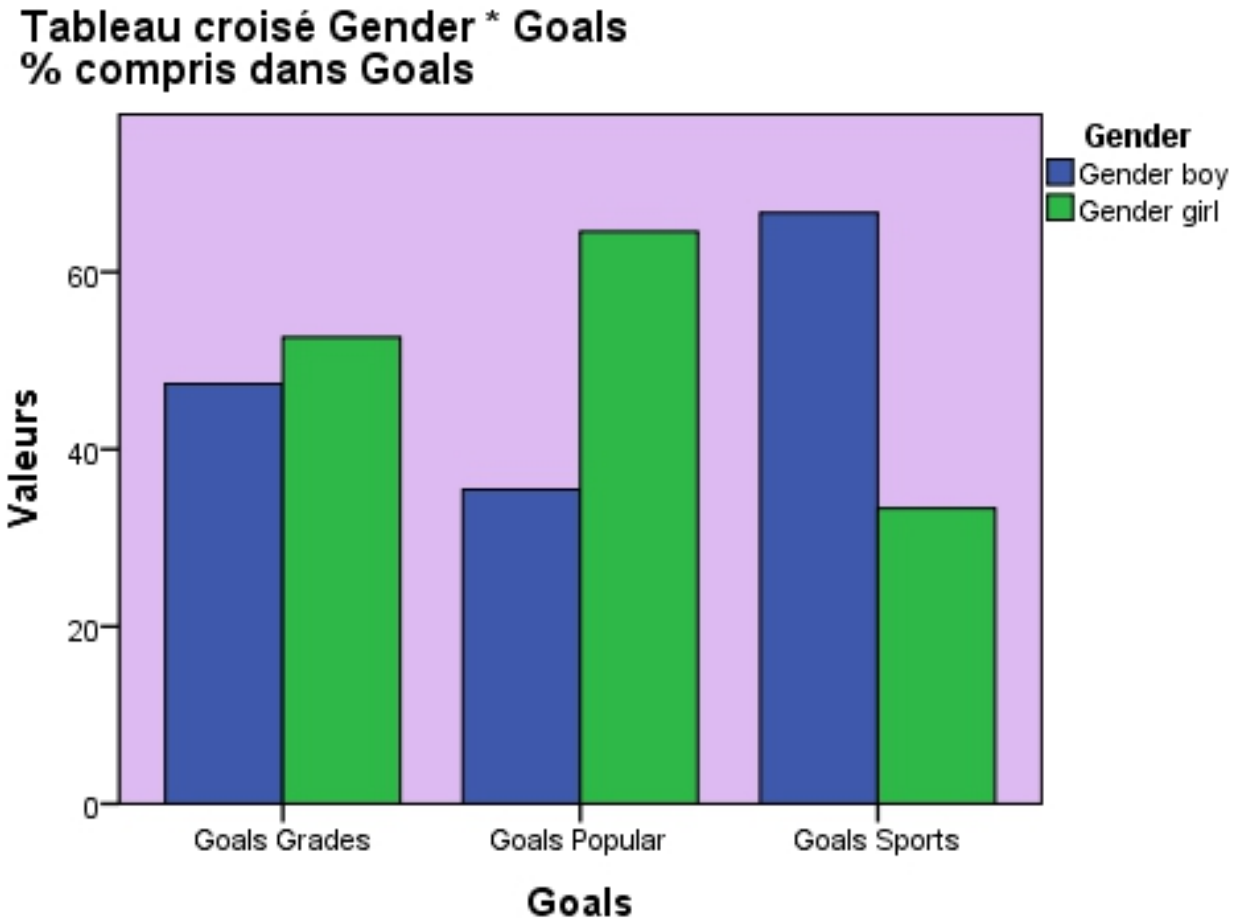


Figure – Graphe bâton superposé sur les profils lignes : regroupement par buts. Issu des données de *Chase, M. A., and Dummer, G. M. (1992), "The Role of Sports as a*

Dispositifs graphiques

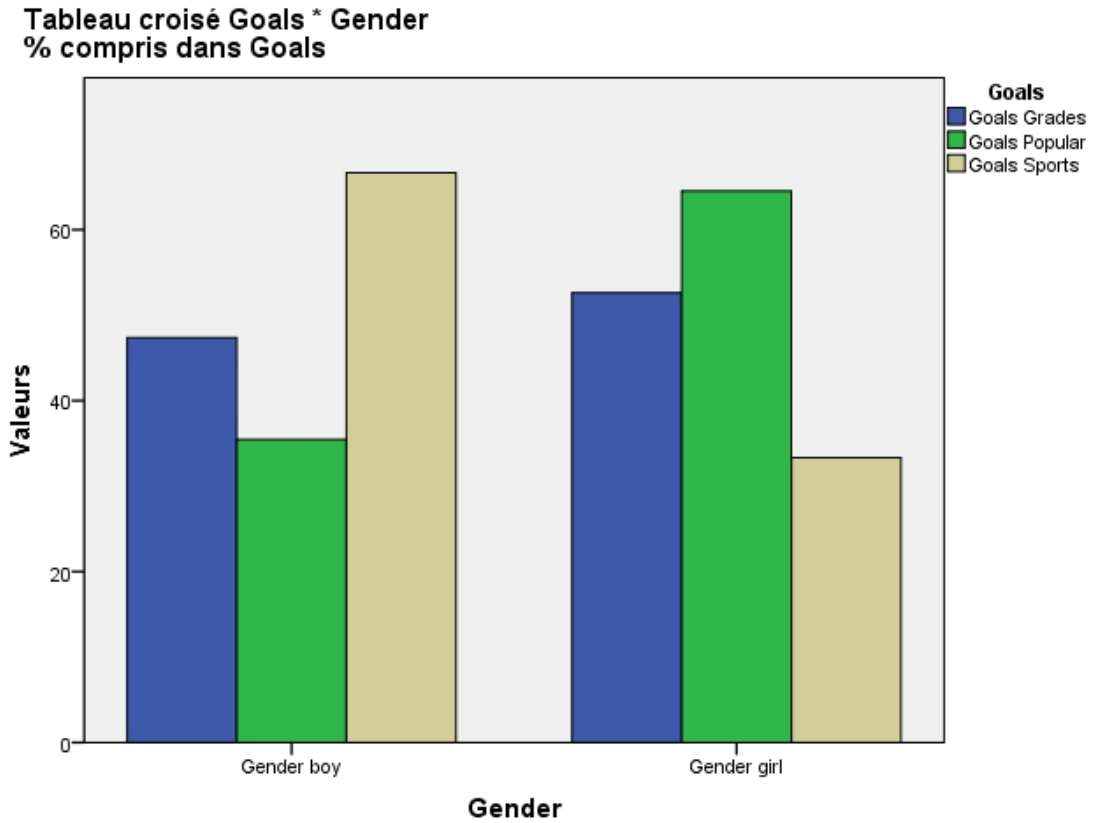


Figure – Graphe bâton superposé sur les profils lignes : regroupement par sexe. Base issue des données de *Chase, M. A., and Dummer, G. M. (1992), "The Role of Sports as a Social Determinant for Children"*

Dispositifs graphiques

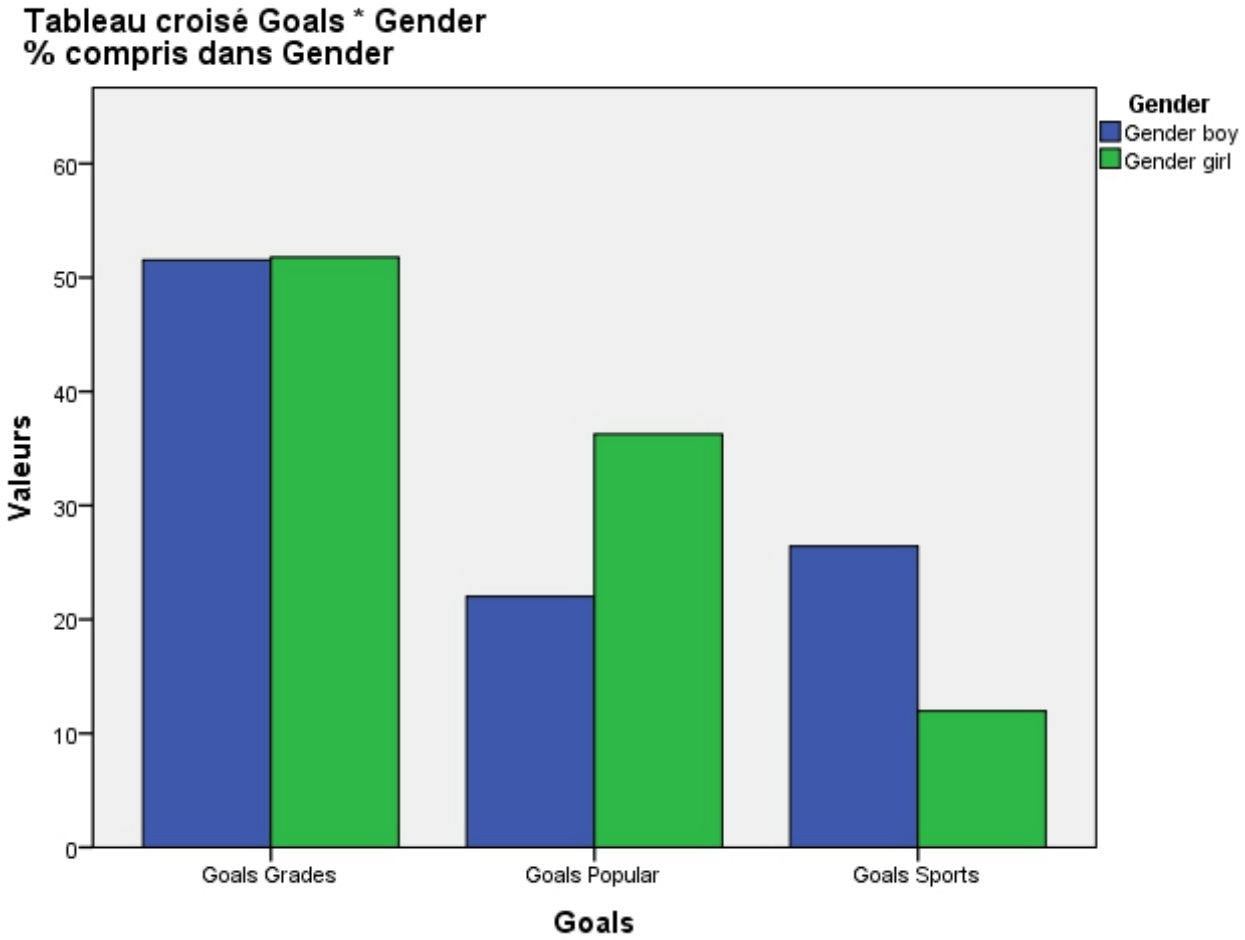


Figure – Graphe bâton superposé sur les profils colonnes : regroupement par but. Base issue des données de *Chase, M. A., and Dummer, G. M. (1992), "The Role of*

Dispositifs graphiques

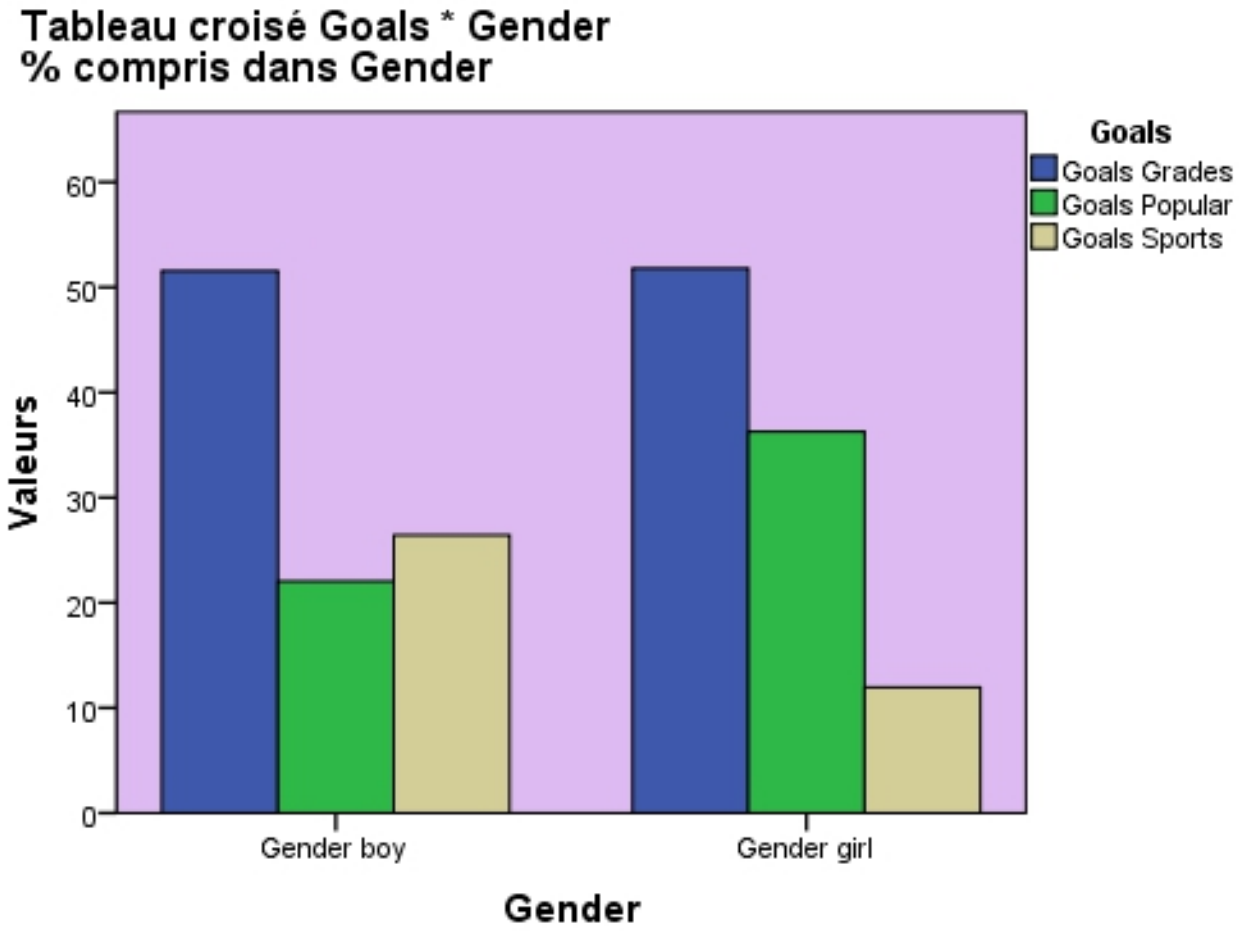


Figure – Graphe bâton superposé sur les profils colonnes : regroupement par sexe. Base issue des données de *Chase, M. A., and Dummer, G. M. (1992), "The Role of*

Dispositifs graphiques

L' analyse factorielle des correspondances propose des dispositifs graphiques mais est techniquement fondée sur l'ACP.

Deux variables qualitatives ordinales

On se restreint à deux variables ordinales uniquement :

Grades	Sports
1	2
2	1
4	3
2	3
4	2
4	2
3	4
3	4
3	2
4	3

Figure – extrait de la base de données réduite à deux variables ordinales, *Chase, M. A., and Dummer, G. M. (1992), "The Role of Sports as a Social Determinant for Children"*

-Grades : rang alloué par les écoliers à l'importance d' "avoir de bonnes notes" pour être reconnu par ses pairs
-Sports : rang alloué à l'importance d'être un "bon sportif"
Ordre : 1= le plus important, ..., 4 = le moins important

Indicateurs

On peut reprendre la totalité des indicateurs existant dans le cadre de deux variables qualitatives non ordonnées

Indicateurs

On peut reprendre la totalité des indicateurs existant dans le cadre de deux variables qualitatives non ordonnées

On ajoute des statistiques de corrélation des rangs

- Corrélation des rangs de Spearman
- Corrélation des rangs de Kendall

Corrélation de Spearman

Deux variables ordinales x et y appariées, décrites sur n individus. On peut les considérer comme des observations des variables aléatoires X et Y.

Cadre théorique : chaque observation a une place unique.
Création de 2 variables $R = rang(X)$ et $S = rang(Y)$. Représentent les rangs de chaque observation (pas deux observations de $R_i = rang(X_i)$ identiques) :

$$\min_{i=1,\dots,n}(R_i) = 1, \qquad \max_{i=1,\dots,n}(R_i) = n, \qquad \overline{R} = \frac{n+1}{2}$$

On note $\rho_{Spearman}$ le coefficient de corrélation de Spearman.

Corrélation de Spearman

On définit la corrélation de Spearman par $\rho_S = cor(R; S)$.

Par abus (clair) de notation, on confond la notation pour des variables aléatoires X et Y avec son estimation pour les observables x et y.

$$\rho_S = \frac{\sum_{i=1}^{n}(R_i - \overline{R})(S_i - \overline{S})}{\sqrt{\sum_{i=1}^{n}(R_i - \overline{R})^2}\sqrt{\sum_{i=1}^{n}(S_i - \overline{S})^2}}$$

ou encore

$$\rho_S = \frac{12\sum_{i=1}^{n} R_i S_i}{n(n^2-1)} - \frac{2(n+1)}{(n-1)}$$

et, en notant $D_i = R_i - S_i$

$$\rho_S = 1 - \frac{6\sum_{i=1}^{n} D_i^2}{(n^3 - n)}$$

Corrélation de Spearman

Interprétation :

- exprime le degré de concordance des classements entre x et y, ainsi que son sens : exprime la concordance au vu du classement global : variable x et variable y ;

Corrélation de Spearman

Interprétation :

- exprime le degré de concordance des classements entre x et y, ainsi que son sens : exprime la concordance au vu du classement global : variable x et variable y ;
- $-1 \leq \rho_S \leq 1$;

Corrélation de Spearman

Interprétation :

- exprime le degré de concordance des classements entre x et y, ainsi que son sens : exprime la concordance au vu du classement global : variable x et variable y ;
- $-1 \leq \rho_S \leq 1$;
- X et Y indépendantes $=> \rho_S = 0$;

Corrélation de Spearman

Interprétation :

- exprime le degré de concordance des classements entre x et y, ainsi que son sens : exprime la concordance au vu du classement global : variable x et variable y ;
- $-1 \leq \rho_S \leq 1$;
- X et Y indépendantes $=> \rho_S = 0$;
- Test du ρ_S valable dans un cadre non paramétrique.

Corrélation de Spearman

Interprétation :

- exprime le degré de concordance des classements entre x et y, ainsi que son sens : exprime la concordance au vu du classement global : variable x et variable y ;
- $-1 \leq \rho_S \leq 1$;
- X et Y indépendantes $=> \rho_S = 0$;
- Test du ρ_S valable dans un cadre non paramétrique. Test aussi puissant que la corrélation de Pearson, dans le cadre paramétrique Gaussien ;

Corrélation de Spearman

Interprétation :

- exprime le degré de concordance des classements entre x et y, ainsi que son sens : exprime la concordance au vu du classement global : variable x et variable y ;
- $-1 \leq \rho_S \leq 1$;
- X et Y indépendantes $=> \rho_S = 0$;
- Test du ρ_S valable dans un cadre non paramétrique. Test aussi puissant que la corrélation de Pearson, dans le cadre paramétrique Gaussien ;
- Peut traduire une situation non linéaire, mais elle doit être monotone ;

Corrélation de Spearman

Interprétation :

- exprime le degré de concordance des classements entre x et y, ainsi que son sens : exprime la concordance au vu du classement global : variable x et variable y ;
- $-1 \leq \rho_S \leq 1$;
- X et Y indépendantes $=> \rho_S = 0$;
- Test du ρ_S valable dans un cadre non paramétrique. Test aussi puissant que la corrélation de Pearson, dans le cadre paramétrique Gaussien ;
- Peut traduire une situation non linéaire, mais elle doit être monotone ;
- Robustesse par rapport aux valeurs extrêmes ;

Corrélation de Spearman

Interprétation :

- exprime le degré de concordance des classements entre x et y, ainsi que son sens : exprime la concordance au vu du classement global : variable x et variable y ;
- $-1 \leq \rho_S \leq 1$;
- X et Y indépendantes $=> \rho_S = 0$;
- Test du ρ_S valable dans un cadre non paramétrique. Test aussi puissant que la corrélation de Pearson, dans le cadre paramétrique Gaussien ;
- Peut traduire une situation non linéaire, mais elle doit être monotone ;
- Robustesse par rapport aux valeurs extrêmes ;
- Une variante permet de prendre en compte les ex-aequo : **essentiel**.

Corrélation de Spearman

Cadre avec ex-aequo.
facteur de correction -> valeurs intermédiaires.

Corrélation de Spearman

Cadre avec ex-aequo.
facteur de correction -> valeurs intermédiaires.

- Calcul des rangs moyens. Soit n_R le nombre de rangs moyens distincts.

Corrélation de Spearman

Cadre avec ex-aequo.

facteur de correction -> valeurs intermédiaires.

- Calcul des rangs moyens. Soit n_R le nombre de rangs moyens distincts. (Si $n_R = n$ pas d'ex-aequo).

Corrélation de Spearman

Cadre avec ex-aequo.

facteur de correction -> valeurs intermédiaires.

- Calcul des rangs moyens. Soit n_R le nombre de rangs moyens distincts. (Si $n_R = n$ pas d'ex-aequo).
  - on calcule les rangs ;

Corrélation de Spearman

Cadre avec ex-aequo.

facteur de correction -> valeurs intermédiaires.

- Calcul des rangs moyens. Soit n_R le nombre de rangs moyens distincts. (Si $n_R = n$ pas d'ex-aequo).
  - on calcule les rangs ;
  - pour les individus de même valeur : on alloue une valeur de rang moyenne.

Corrélation de Spearman

Cadre avec ex-aequo.

facteur de correction -> valeurs intermédiaires.

- Calcul des rangs moyens. Soit n_R le nombre de rangs moyens distincts. (Si $n_R = n$ pas d'ex-aequo).
  - on calcule les rangs ;
  - pour les individus de même valeur : on alloue une valeur de rang moyenne.
- Soit t_r le nombre d'apparition du même rang moyen r

Corrélation de Spearman

Cadre avec ex-aequo.
facteur de correction -> valeurs intermédiaires.

- Calcul des rangs moyens. Soit n_R le nombre de rangs moyens distincts. (Si $n_R = n$ pas d'ex-aequo).
  - on calcule les rangs ;
  - pour les individus de même valeur : on alloue une valeur de rang moyenne.
- Soit t_r le nombre d'apparition du même rang moyen r
- On calcule la valeur T_X le facteur de correction, fonction du nombre d'ex-aequo au sein de la variable X :

$$T_X = \sum_{a=1}^{n_R} (t_a^3 - t_a).$$

$$\rho_S = \frac{(n^3 - n) - 6\sum_{i=1}^{n} D_i^2 - (T_X + T_Y)/2}{\sqrt{(n^3 - n)^2 - (T_X + T_Y)(n^3 - n) + T_X T_Y}}$$

Corrélation de Spearman

Test du coefficient de corrélation de Spearman.
Aucune hypothèse de loi, mais un recueil iid des observations.

H_0 : X et Y indépendantes contre H_1 : X et Y sont concordantes ou anti-concordantes

Corrélation de Spearman

Test du coefficient de corrélation de Spearman.
Aucune hypothèse de loi, mais un recueil iid des observations.

H_0 : X et Y indépendantes contre H_1 : X et Y sont concordantes ou anti-concordantes

- pour n petit (inférieur à 10, mais toujours supérieur à 4), on utilise des tables exactes ;

Corrélation de Spearman

Test du coefficient de corrélation de Spearman.
Aucune hypothèse de loi, mais un recueil iid des observations.

H_0 : X et Y indépendantes contre H_1 : X et Y sont concordantes ou anti-concordantes

- pour n petit (inférieur à 10, mais toujours supérieur à 4), on utilise des tables exactes ;
- Entre 20 et 35, on utilise une approximation de Student ;

Corrélation de Spearman

Test du coefficient de corrélation de Spearman.
Aucune hypothèse de loi, mais un recueil iid des observations.

$$H_0 : X \text{ et } Y \text{ indépendantes contre } H_1 : X \text{ et } Y \text{ sont concordantes ou anti-concordantes}$$

- pour n petit (inférieur à 10, mais toujours supérieur à 4), on utilise des tables exactes ;
- Entre 20 et 35, on utilise une approximation de Student ;
- Au delà, une approximation normale.

Mais, il s'agit tours d'une corrélation de Pearson , même si elle s'exécute sur des rangs.

Corrélation de Kendall

On transforme toujouts x et y en rang R et S On dit que la paire (r_i, s_i) correspondante à l'individu i est concordante avec la paire (r_j, s_j) correspondante à l'individu j si leurs observations sur les variables R et S évoluent dans le même sens : $(r_i - r_j)(s_i - s_j) > 0$. Dans le cas où ce prodit est négatif, on parle de paires discordantes.

Corrélation de Kendall

On transforme toujouts x et y en rang R et S On dit que la paire (r_i, s_i) correspondante à l'individu i est concordante avec la paire (r_j, s_j) correspondante à l'individu j si leurs observations sur les variables R et S évoluent dans le même sens : $(r_i - r_j)(s_i - s_j) > 0$. Dans le cas où ce prodit est négatif, on parle de paires discordantes.
On note N_C le nombre de paires concordantes, et N_D le nombre de paires discordantes.

Corrélation de Kendall

On transforme toujouts x et y en rang R et S On dit que la paire (r_i, s_i) correspondante à l'individu i est concordante avec la paire (r_j, s_j) correspondante à l'individu j si leurs observations sur les variables R et S évoluent dans le même sens : $(r_i - r_j)(s_i - s_j) > 0$. Dans le cas où ce prodit est négatif, on parle de paires discordantes.
On note N_C le nombre de paires concordantes, et N_D le nombre de paires discordantes.
La comparaison de toutes les paires est longue si n est grand.

Corrélation de Kendall

On transforme toujouts x et y en rang R et S On dit que la paire (r_i, s_i) correspondante à l'individu i est concordante avec la paire (r_j, s_j) correspondante à l'individu j si leurs observations sur les variables R et S évoluent dans le même sens : $(r_i - r_j)(s_i - s_j) > 0$. Dans le cas où ce prodit est négatif, on parle de paires discordantes.
On note N_C le nombre de paires concordantes, et N_D le nombre de paires discordantes.
La comparaison de toutes les paires est longue si n est grand.

On définit ρ_K la corrélation de Kendall par

$$\rho_K = 2\frac{N_C - N_D}{n(n-1)}$$

Corrélation de Kendall

Interprétation :

Corrélation de Kendall

Interprétation :

- Exprime le degré de concordance des classements entre X et Y, ainsi que son sens : exprime la concordance individu après individu ;

Corrélation de Kendall

Interprétation :

- Exprime le degré de concordance des classements entre X et Y, ainsi que son sens : exprime la concordance individu après individu ;
- $-1 \leq \rho_K \leq 1$;

Corrélation de Kendall

Interprétation :

- Exprime le degré de concordance des classements entre X et Y, ainsi que son sens : exprime la concordance individu après individu ;
- $-1 \leq \rho_K \leq 1$;
- X et Y ont autant de chance d'être concordant que discordant (indépendance des classements) $=>$ $\rho_K = 0$;

Corrélation de Kendall

Interprétation :

- Exprime le degré de concordance des classements entre X et Y, ainsi que son sens : exprime la concordance individu après individu ;
- $-1 \leq \rho_K \leq 1$;
- X et Y ont autant de chance d'être concordant que discordant (indépendance des classements) $=>$ $\rho_K = 0$;
- Test du ρ_K valable dans un cadre non paramétrique ;

Corrélation de Kendall

Interprétation :

- Exprime le degré de concordance des classements entre X et Y, ainsi que son sens : exprime la concordance individu après individu ;
- $-1 \leq \rho_K \leq 1$;
- X et Y ont autant de chance d'être concordant que discordant (indépendance des classements) $=>$ $\rho_K = 0$;
- Test du ρ_K valable dans un cadre non paramétrique ;
- Robustesse par rapport aux valeurs extrêmes ;

Corrélation de Kendall

Interprétation :

- Exprime le degré de concordance des classements entre X et Y, ainsi que son sens : exprime la concordance individu après individu ;
- $-1 \leq \rho_K \leq 1$;
- X et Y ont autant de chance d'être concordant que discordant (indépendance des classements) $=>$ $\rho_K = 0$;
- Test du ρ_K valable dans un cadre non paramétrique ;
- Robustesse par rapport aux valeurs extrêmes ;
- Une variante permet de prendre en compte les ex-aequo : **essentiel**.

Corrélation de Kendall

Cadre avec ex-aequo.
facteur de correction -> valeurs intermédiaires.

Corrélation de Kendall

Cadre avec ex-aequo.
facteur de correction -> valeurs intermédiaires.

Corrélation de Kendall

Cadre avec ex-aequo.
facteur de correction -> valeurs intermédiaires.

- Soit n_R le nombre de rangs moyens distincts ;

Corrélation de Kendall

Cadre avec ex-aequo.
facteur de correction -> valeurs intermédiaires.

- Soit n_R le nombre de rangs moyens distincts ;
- Soit t_r le nombre d'apparition du même rang moyen r ;

Corrélation de Kendall

Cadre avec ex-aequo.
facteur de correction -> valeurs intermédiaires.

- Soit n_R le nombre de rangs moyens distincts ;
- Soit t_r le nombre d'apparition du même rang moyen r ;
- On calcule la valeur E_X le facteur de correction, fonction du nombre d'ex-aequo au sein de la variable X (resp. pour Y) :

$$E_X = \sum_{a=1}^{n_R} (t_a^2 - t_a).$$

$$\rho_K = 2 \frac{N_C - N_D}{\sqrt{((n^2 - n) - E_X)((n^2 - n) - E_Y)}}.$$

Corrélation de Kendall

Test du coefficient de corrélation de Kendall.
Aucune hypothèse de loi, mais un recueil iid des observations.

H_0 : X et Y indépendantes contre H_1 : X et Y sont concordantes ou anti-concordantes

Corrélation de Kendall

Test du coefficient de corrélation de Kendall.
Aucune hypothèse de loi, mais un recueil iid des observations.

H_0 : X et Y indépendantes contre H_1 : X et Y sont concordantes ou anti-concordantes

- pour n petit (inférieur à 10, mais toujours supérieur à 4), on utilise des tables exactes ;

Corrélation de Kendall

Test du coefficient de corrélation de Kendall.
Aucune hypothèse de loi, mais un recueil iid des observations.

H_0 : X et Y indépendantes contre H_1 : X et Y sont concordantes ou anti-concordantes

- pour n petit (inférieur à 10, mais toujours supérieur à 4), on utilise des tables exactes ;
- Pour $n > 10$, on utilise une approximation normale.

Corrélations : exemples

Relation non monotone : carrevar1=(var1 -5)^2 pour $n = 10$ observations de var1 variant de 1 en 1 entre -4 et 5

Corrélations

			var11	carrevar1
Tau-B de Kendall	var11	Coefficient de corrélation	1,000	,210
		Sig. (bilatérale)	.	,413
		N	10	10
	carrevar1	Coefficient de corrélation	,210	1,000
		Sig. (bilatérale)	,413	.
		N	10	10
Rho de Spearman	var11	Coefficient de corrélation	1,000	,276
		Sig. (bilatérale)	.	,440
		N	10	10
	carrevar1	Coefficient de corrélation	,276	1,000
		Sig. (bilatérale)	,440	.
		N	10	10

Figure – Données artificielles. Calcul des coefficients de corrélation dans le cadre d'une relation non monotone.

Corrélations : exemples

Relation monotone non linéaire : cubevar1=c(var11-5)^3 pour $n = 10$ observations de var11 variant de 1 en 1 entre 1 et 10

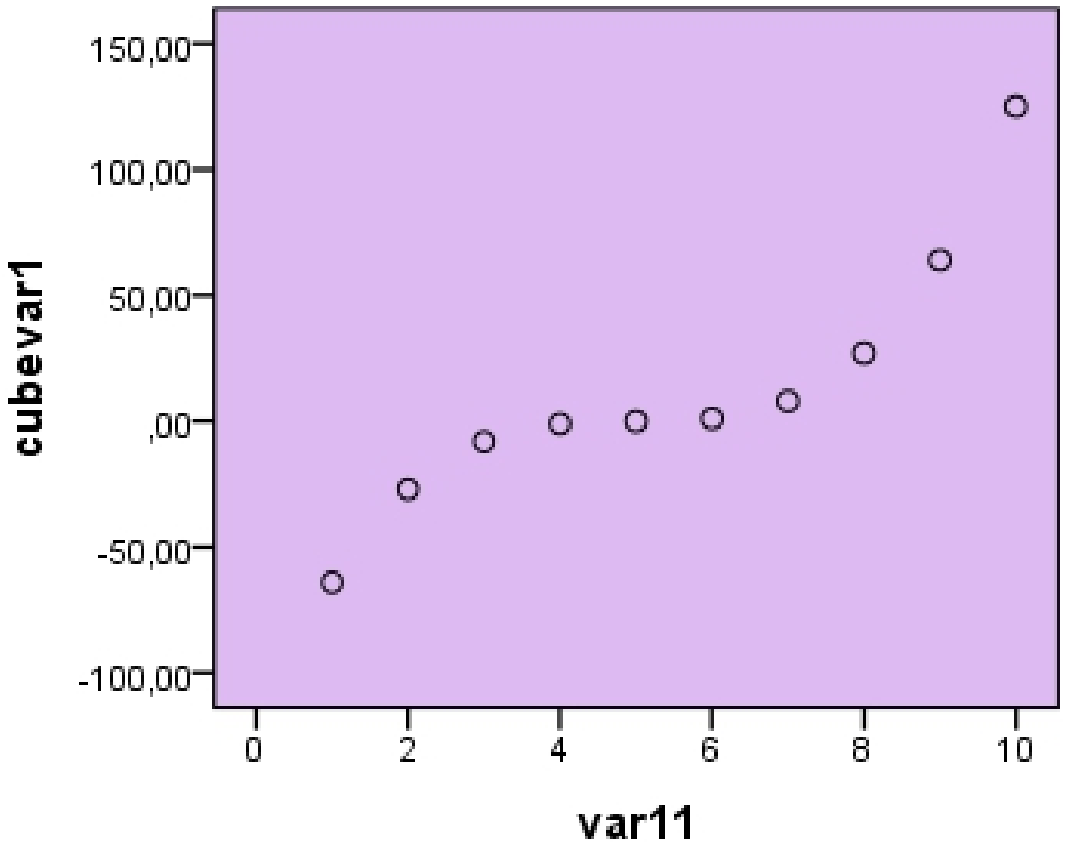


Figure – Données artificielles. Relation monotone, non linéaire.

Corrélations : exemples

Corrélations

			var11	cubevar1
Tau-B de Kendall	var11	Coefficient de corrélation	1,000	1,000**
		Sig. (bilatérale)	.	.
		N	10	10
	cubevar1	Coefficient de corrélation	1,000**	1,000
		Sig. (bilatérale)	.	.
		N	10	10
Rho de Spearman	var11	Coefficient de corrélation	1,000	1,000**
		Sig. (bilatérale)	.	.
		N	10	10
	cubevar1	Coefficient de corrélation	1,000**	1,000
		Sig. (bilatérale)	.	.
		N	10	10

**. La corrélation est significative au niveau 0,01 (bilatéral).

Figure – Données artificielles. Calcul des coefficients de corrélation dans le cadre d'une relation monotone non linéaire.

Corrélations : exemples

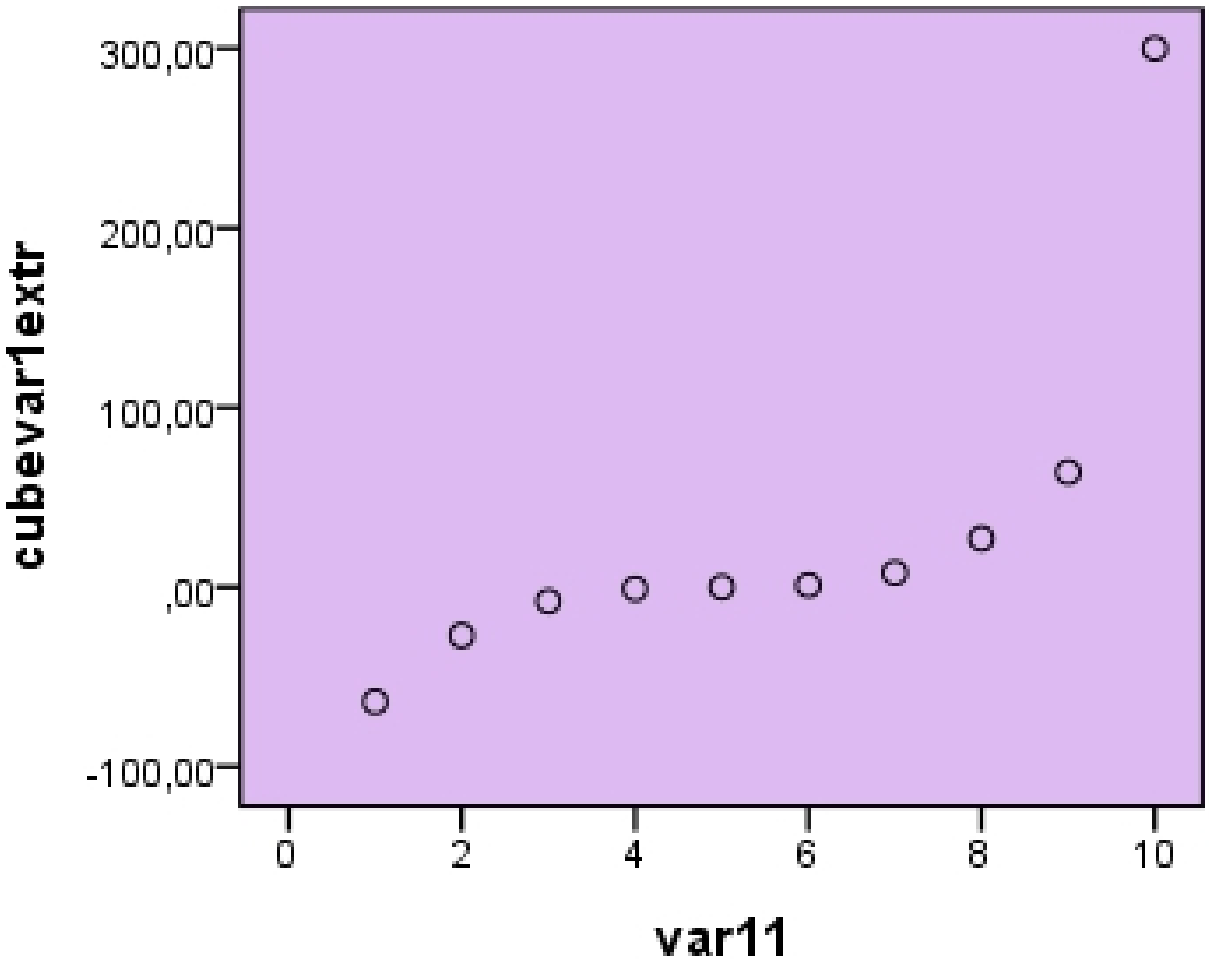


Figure – Données artificielles. Relation monotone avec données extrêmes.

Corrélations : exemples

Récapitulatif des observations[a]

var11	VAR00004	VAR00005
1	2	10
2	1	1
3	3	2
4	4	3
5	5	4
6	6	5
7	7	6
8	8	7
9	9	8
10	10	9

Figure – Données artificielles

Corrélations : exemples

Corrélations

		VAR00004
Tau-B de Kendall	var11	,956
		,000
Rho de Spearman	var11	,988
		,000

		VAR0005
Tau-B de Kendall	var11	,600
		,016
Rho de Spearman	var11	,455
		,187

Figure – Données artificielles

Corrélations : exemple sur données réelles

Récapitulatif des observations[a]

Grades	Sports
1	2
2	1
4	3
2	3
4	2
4	2
3	4
3	4
3	2
4	3

a. Limité aux 10 premières observations

Figure – Extrait des réponses de 10 écoliers aux questions "importance des notes" et "importance du sport",*Chase, M. A., and Dummer, G. M. (1992), "The Role of Sports as a Social Determinant for Children"*

Corrélations : exemple sur données réelles

Corrélations

		Sports
Tau-B de Kendall	Grades	-,111
		,004
		478
Rho de Spearman	Grades	-,149
		,001
		478

Figure – Calcul des corrélations de Spearman et de Kendall pour les réponses aux questions "importance des notes" et "importance du sport",*Chase, M. A., and Dummer, G. M. (1992), "The Role of Sports as a Social Determinant for Children"*

Corrélations : exemple sur données réelles

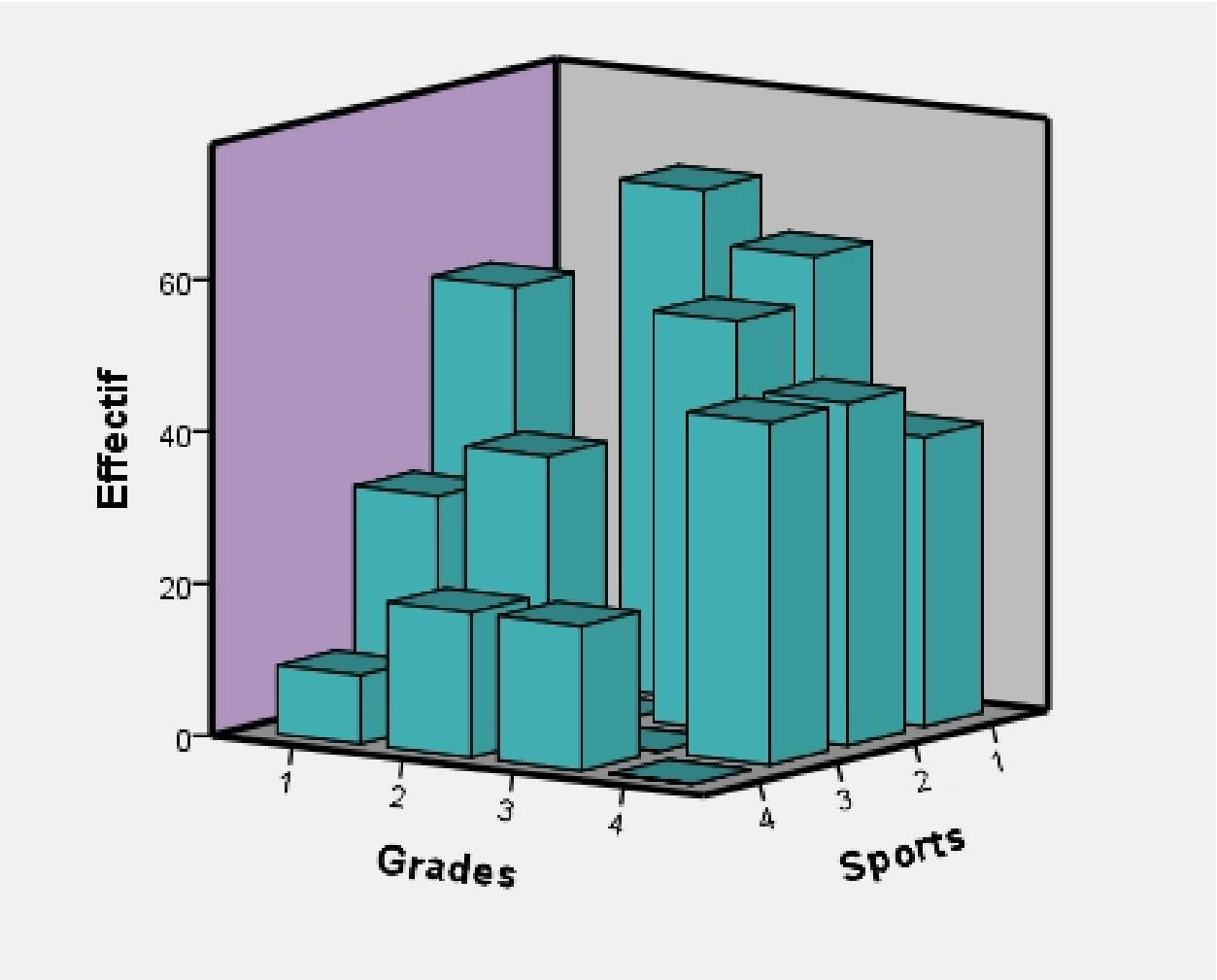


Figure – Graphe bâtons en trois dimensions correspondant aux effectifs des réponses aux questions "importance des notes" et "importance du sport", *Chase, M. A., and*

Deux variables discrètes

L'ensemble de ce qui a été présenté précédemment (variables ordinales) est valable.

Deux variables discrètes

L'ensemble de ce qui a été présenté précédemment (variables ordinales) est valable.
Indicateurs : ajout du coefficient de corrélation de Pearson (qui sera développé dans le cadre de deux variables continues)

Deux variables discrètes

L'ensemble de ce qui a été présenté précédemment (variables ordinales) est valable.
Indicateurs : ajout du coefficient de corrélation de Pearson (qui sera développé dans le cadre de deux variables continues)
Graphes : les diagrammes de dispersion, et les graphes bâtons en trois dimensions.

Deux variables continues

Le jeu de données utilisé :
Moore, David S., et George P. McCabe (1989). Introduction to the Practice of Statistics, p. 179. Original source : "Family Expenditure Survey, Department of Employment", 1981 (British official statistics)

Deux variables continues

Le jeu de données utilisé :
Moore, David S., et George P. McCabe (1989). Introduction to the Practice of Statistics, p. 179. Original source : "Family Expenditure Survey, Department of Employment", 1981 (British official statistics)

Il décrit la moyenne des dépenses hebdomadaires, exprimées en livre, des ménages de Grande Bretagne distingués en 11 régions, dans les achats de tabac et d'alcool.

Corrélation de Pearson

On se place dans le cadre d'observations issues de variables continues : aucun ex-aequo.
On note le coefficient de corrélation de Pearson par $\rho_{Pearson}$.

Corrélation de Spearman

Interprétation :

Corrélation de Spearman

Interprétation :

- exprime l'intensité d'une liaison, ainsi que son sens, dans le cadre d'une relation linéaire (sinon, n'exprime rien) ;

Corrélation de Spearman

Interprétation :

- exprime l'intensité d'une liaison, ainsi que son sens, dans le cadre d'une relation linéaire (sinon, n'exprime rien) ;
- $-1 \leq \rho_{Pearson} \leq 1$;

Corrélation de Spearman

Interprétation :

- exprime l'intensité d'une liaison, ainsi que son sens, dans le cadre d'une relation linéaire (sinon, n'exprime rien) ;
- $-1 \leq \rho_{Pearson} \leq 1$;
- X et Y indépendantes implique $\rho_{Pearson} = 0$;

Corrélation de Spearman

Interprétation :

- exprime l'intensité d'une liaison, ainsi que son sens, dans le cadre d'une relation linéaire (sinon, n'exprime rien) ;
- $-1 \leq \rho_{Pearson} \leq 1$;
- X et Y indépendantes implique $\rho_{Pearson} = 0$;
- Test du $\rho_{Pearson}$ valable dans un cadre paramétrique gaussien pour tout n. Test valable dans un cadre non paramétrique pour n grand.

Corrélation de Pearson

On définit la corrélation de Pearson par la covariance entre les variables X et Y réduites.

Corrélation de Pearson

On définit la corrélation de Pearson par la covariance entre les variables X et Y réduites.
On note $\widehat{\sigma}_X$ l'écart-type de la variable X.

Corrélation de Pearson

On définit la corrélation de Pearson par la covariance entre les variables X et Y réduites.
On note $\widehat{\sigma}_X$ l'écart-type de la variable X.

$$cov(X;Y) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})$$

$$\rho_{Pearson} = cov(\frac{X}{\widehat{\sigma}_X};\frac{Y}{\widehat{\sigma}_Y})$$

Test de la corrélation de Pearson

Dans le cas gaussien, le coefficient de corrélation peut être testé

$$H_0 : X \text{ indépendante de } Y \quad \text{contre} \quad H_1 : X \text{ et } Y \text{ dépendantes}$$

Hors du cas gaussien, seule une asymptotique dans un cadre iid permet de déterminer une statistique de test.

Deux variables continues : présentation des données

Region	Alcohol	Tobacco
North	6,47	4,03
Yorkshire	6,13	3,76
Northeast	6,19	3,77
East Midlands	4,89	3,34
West Midlands	5,63	3,47
East Anglia	4,52	2,92
Southeast	5,89	3,20
Southwest	4,79	2,71
Wales	5,27	3,53
Scotland	6,08	4,51
Northern Ireland	4,02	4,56

Figure – Données : 2 variables continues, 1 variable qualitative représentant les individus, *"Family Expenditure Survey, Department of Employment", 1981, British official statistics*

Deux variables continues : présentation des données

Region	Alcohol	Tobacco
North	6,47	4,03
Yorkshire	6,13	3,76
Northeast	6,19	3,77
East Midlands	4,89	3,34
West Midlands	5,63	3,47
East Anglia	4,52	2,92
Southeast	5,89	3,20
Southwest	4,79	2,71
Wales	5,27	3,53
Scotland	6,08	4,51
Northern	4,02	4,56

Deux variables continues : visualisation des données

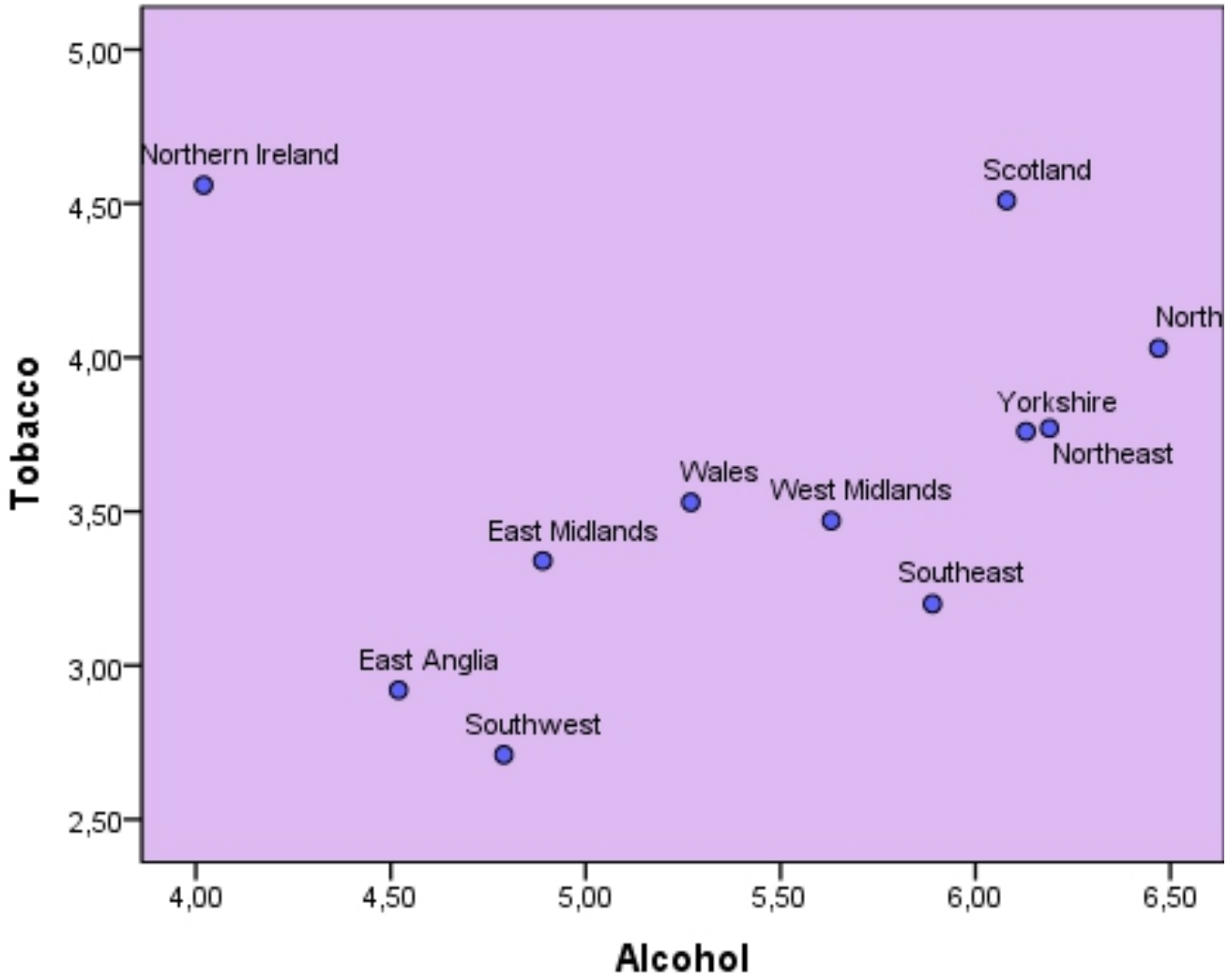


Figure – Dépenses de tabac en fonction de la dépense en alcool par un diagramme de

Deux variables continues : visualisation des données

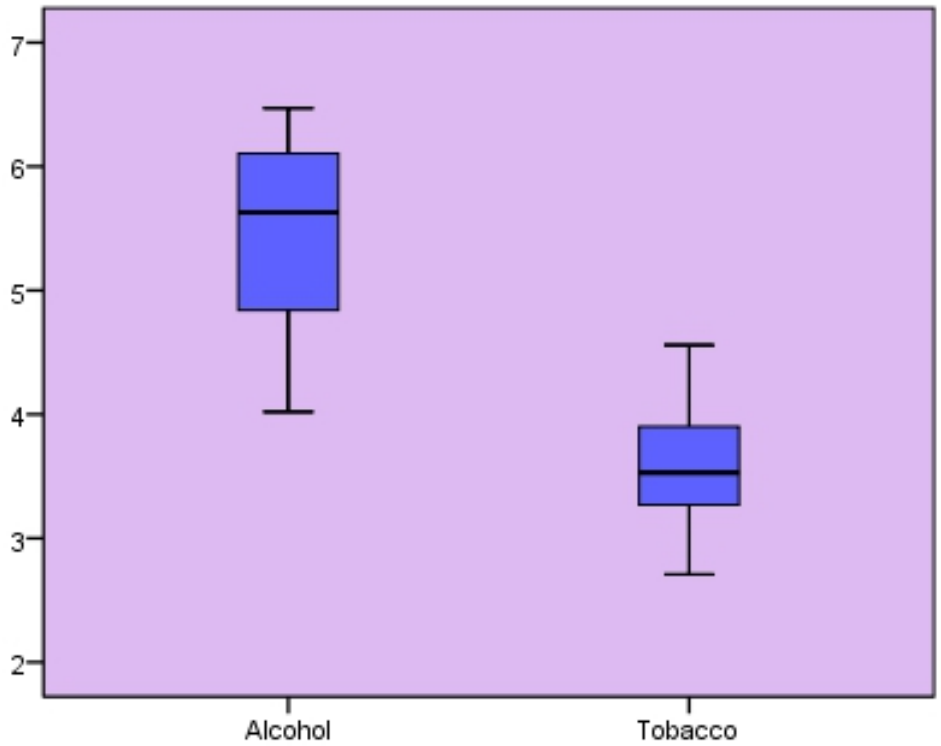


Figure – Dépenses de tabac et dépenses d'alcool par des boîtes à moustache, *"Family Expenditure Survey, Department of Employment", 1981, British official statistics*

Deux variables continues : indicateurs

- Corrélation de Pearson $= 0,224$ avec une p-value de 0,509

Deux variables continues : indicateurs

- Corrélation de Pearson $= 0,224$ avec une p-value de 0,509

Test de Kolmogorov-Smirnov à un échantillon

	Alcohol	Tobacco
Z de Kolmogorov-Smirnov	,553	,417
Signification asymptotique (bilatérale)	,920	,995

Deux variables continues : indicateurs

- Corrélation de Pearson $= 0,224$ avec une p-value de 0,509

Test de Kolmogorov-Smirnov à un échantillon

	Alcohol	Tobacco
Z de Kolmogorov-Smirnov	,553	,417
Signification asymptotique (bilatérale)	,920	,995

- Corrélation de Kendall $= 0,345$ avec une p-value de 0,139 ;
- Corrélation de Spearman $= 0,373$ avec une p-value de 0,259.

Pré-requis pour les aspects mathématiques

- Matrice de projection orthogonale ;
- Lois dérivées de la loi gaussienne ;
- Tests de Wald et tests paramétriques optimaux (UPP).

Analyse de la variance : Anova

On s'intéresse à expliquer les valeurs obtenues pour une variables observées quantitatives, selon l'appartenance à un groupe.
- taille des épis de blé, selon l'engrais placé dans le champs ;
- taille des épis de blé selon l'engrais placé dans le champs et la région considérée.

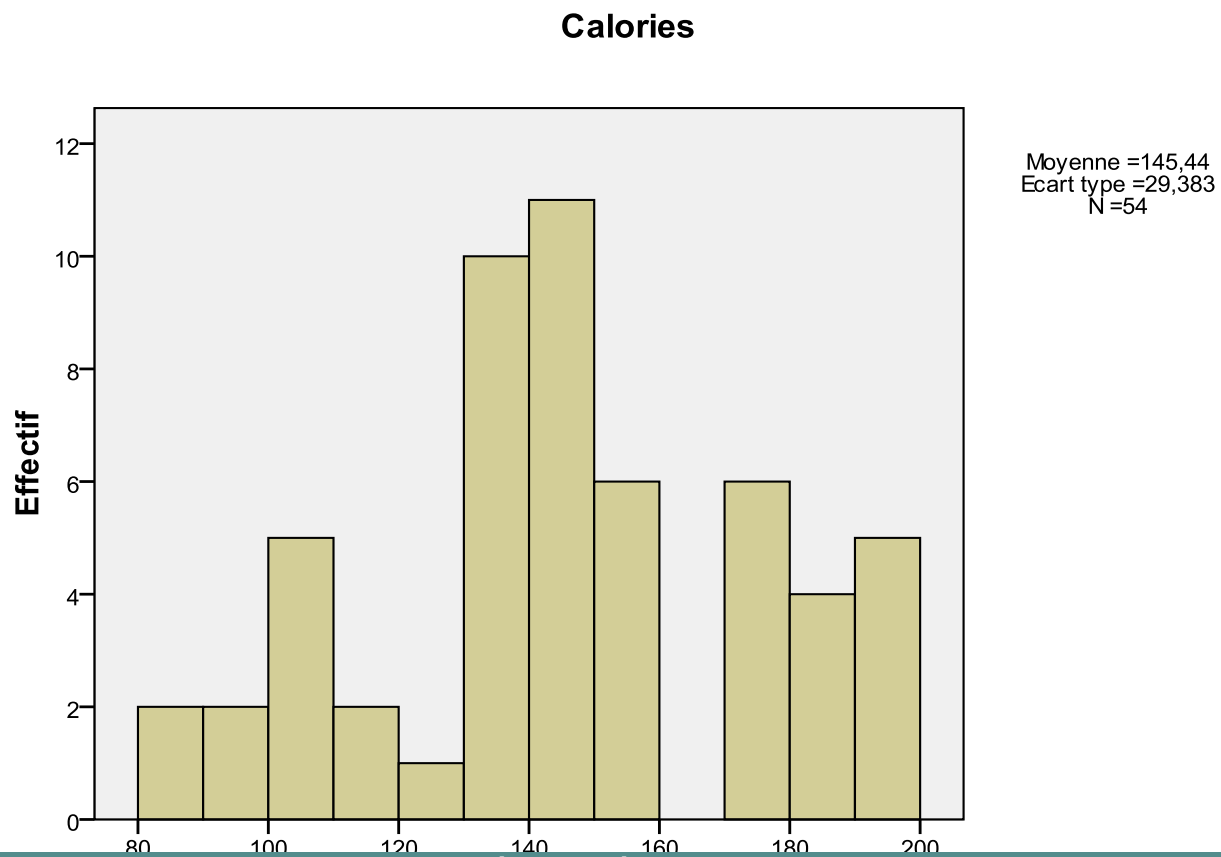
Les facteurs (appartenance à un sous-groupe) sont contrôlés, ou considéré comme tels. Situations pour lesquelles une observation n'appartient qu'à une seule classe (ou sous-groupe) déterminée par l'expérimentateur.

Relation entre une variable quantitative et une variable qualitative.

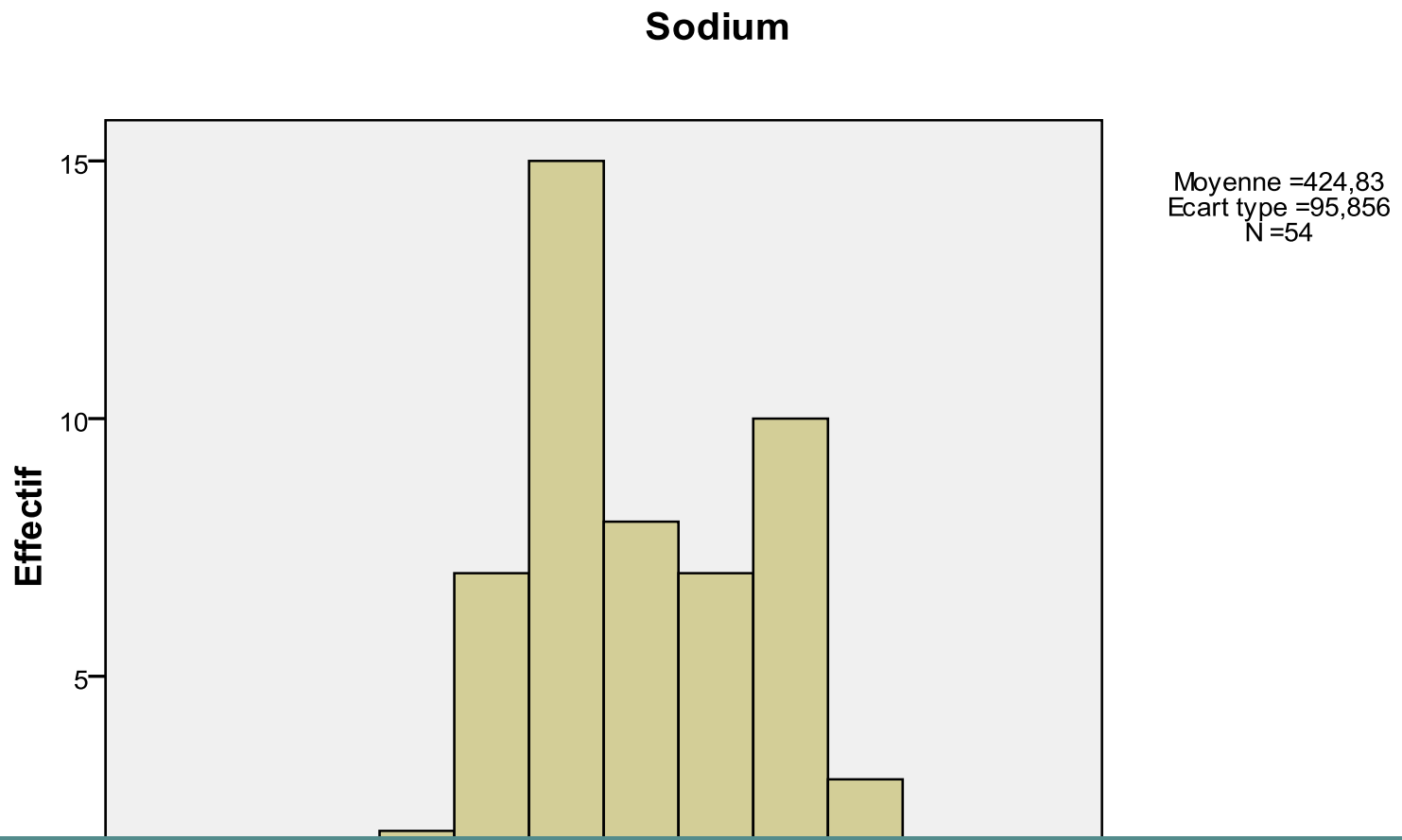
Suppose un modèle probabiliste : un aléa.

Analyse de la variance

Reprise exemple des Hot-Dog, taux de sodium et calories selon le type de viande, beuf, volaille. Référence : Consumer Reports, June 1986, pp. 366-367



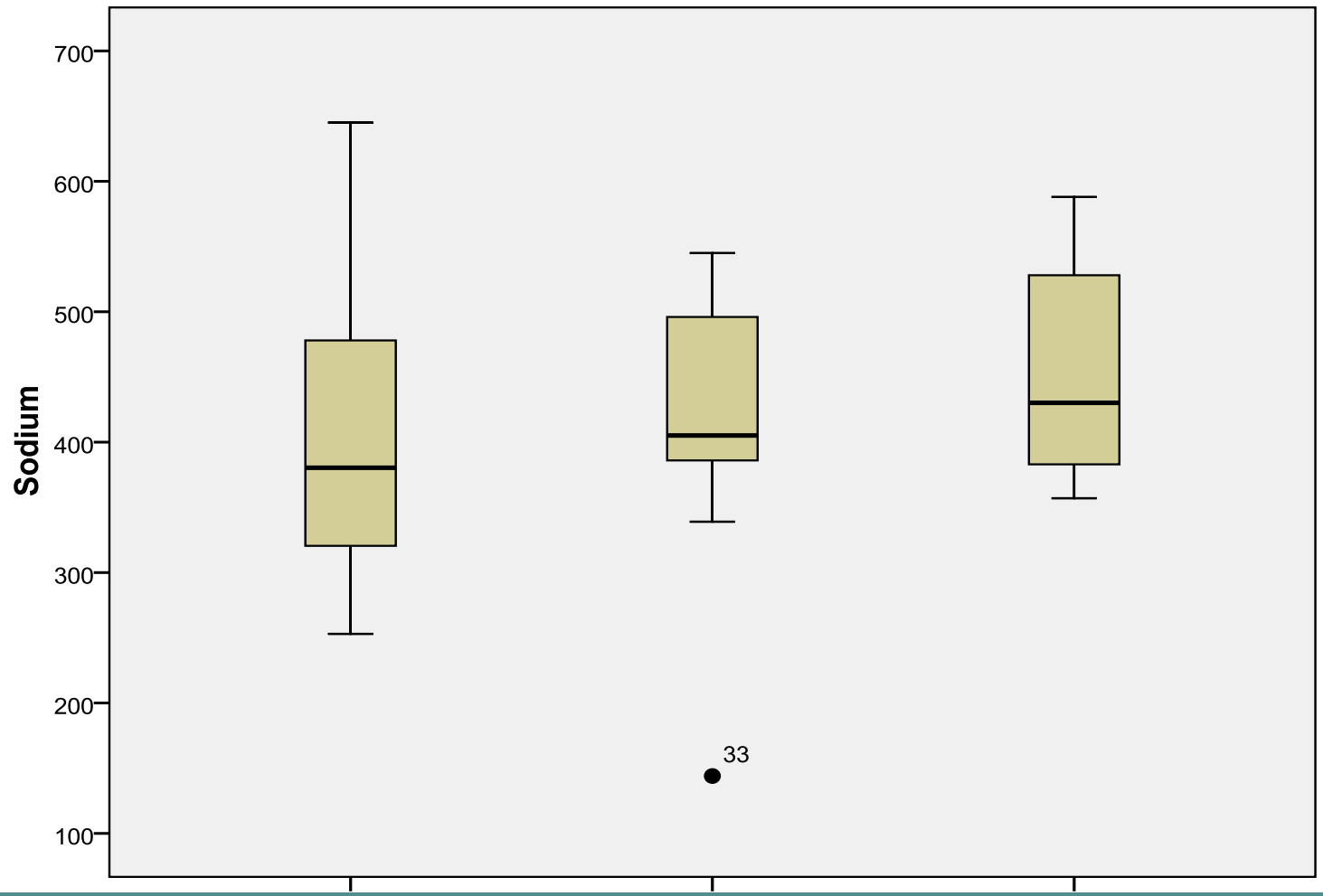
Analyse de la variance



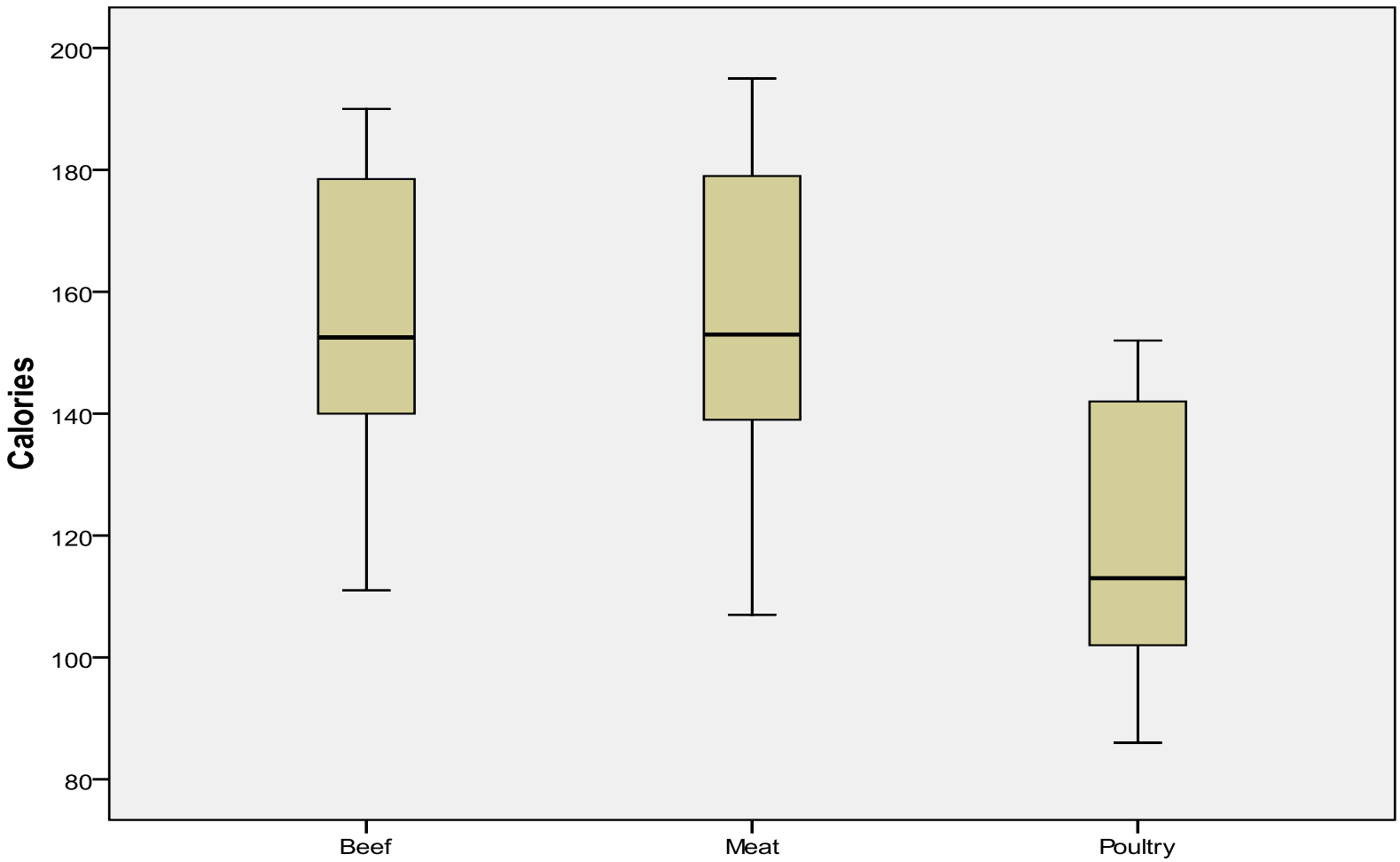
Analyse de la variance

Type_Code		Calories	Sodium
Beef	Moyenne	156,85	401,15
	N	20	20
	Ecart-type	22,642	102,435
Meat	Moyenne	158,71	418,53
	N	17	17
	Ecart-type	25,236	93,872
Poultry	Moyenne	118,76	459,00
	N	17	17
	Ecart-type	22,551	84,739
Total	Moyenne	145,44	424,83

Analyse de la variance



Analyse de la variance



Analyse de la variance à un facteur

Soient $Y = (Y_1, \cdots, Y_n)$ indépendemment distribués, de loi gaussienne (taux de calories). Pour $k = 1, \ldots, n$, on classe Y_k en p sous-groupes (3 groupes) en fonction d'une seconde variable qualitative x_k (type de viande) prenant p modalités (mélange, volaille, boeuf). On dit que $k \in Classe\ i$, lorsque $x_k \in Classe\ i$.

$$Y_k = Y_{i,j}$$

- $i = 1, \ldots, p$, numéro de groupe ;
- $n_1, \ldots, n_p$, effectif des groupes i $(\sum_{i=1}^{p} n_i = n)$;
- $j = 1, \ldots, n_i$ numéro d'individu j dans le sous-groupe i.

$$Y = \begin{pmatrix} Y_{1,1} \\ \vdots \\ Y_{1,n_1} \\ \vdots \\ Y_{p,1} \\ \vdots \\ Y_{p,n_p} \end{pmatrix}$$

Analyse de la variance à un facteur

On suppose *a priori* que Y dépend linéairement du sous-groupe auquel elle appartient.
Quantifier cette dépendance –> effet facteur.

Analyse de la variance à un facteur

On suppose *a priori* que Y dépend linéairement du sous-groupe auquel elle appartient.
Quantifier cette dépendance –> effet facteur.
Idée : décomposition de la variance (variance intra+ variance inter –> création de l'indicateur R^2)

Analyse de la variance à un facteur

Modèle probabiliste

$$\begin{cases} Y = \sum_{i=1}^{p} m_i \mathbb{I}_{Y \in Classe\ i} + \xi, \quad i = 1, \ldots, p \\ \xi \sim \mathcal{N}(0, \sigma^2) \end{cases}$$

Des modèles plus généraux cherchent à relaxer la dernière hypothèse.

Analyse de la variance à un facteur

Modèle statistique

$$\begin{cases} Y_{ij} = m_i + \xi_{ij}, \quad i = 1, \ldots, p, \quad j = 1, \ldots, n_i \\ \xi = (\xi_{ij})_{i,j} \sim \mathcal{N}_n(0, \sigma^2 Id_n) \end{cases}$$

La dernière hypothèse pourra être relaxée dans certains modèles étudiés ultérieurement.

Analyse de la variance à un facteur

Modèle statistique

$$\begin{cases} Y_{ij} = m_i + \xi_{ij}, \quad i = 1, \ldots, p, \quad j = 1, \ldots, n_i \\ \xi = (\xi_{ij})_{i,j} \sim \mathcal{N}_n(0, \sigma^2 Id_n) \end{cases}$$

La dernière hypothèse pourra être relaxée dans certains modèles étudiés ultérieurement.
Facteur pas d'influence sur Y

Analyse de la variance à un facteur

Modèle statistique

$$\begin{cases} Y_{ij} = m_i + \xi_{ij}, \quad i = 1, \ldots, p, \quad j = 1, \ldots, n_i \\ \xi = (\xi_{ij})_{i,j} \sim \mathcal{N}_n(0, \sigma^2 Id_n) \end{cases}$$

La dernière hypothèse pourra être relaxée dans certains modèles étudiés ultérieurement.
Facteur pas d'influence sur Y –> $m_1 = \cdots = m_p$.

Analyse de la variance à un facteur

$$m = \begin{pmatrix} m_1 \\ \vdots \\ m_1 \\ m_2 \\ \vdots \\ m_2 \\ \vdots \\ \vdots \\ m_p \\ \vdots \\ m_p \end{pmatrix}.$$

et ξ le vecteur du bruit.
Loi de Y ?

Analyse de la variance à un facteur

$$m = \begin{pmatrix} m_1 \\ \vdots \\ m_1 \\ m_2 \\ \vdots \\ m_2 \\ \vdots \\ \vdots \\ m_p \\ \vdots \\ m_p \end{pmatrix}.$$

et ξ le vecteur du bruit.
Loi de Y ? :

$$Y \sim \mathcal{N}_n(m; \sigma^2 Id_n).$$

Analyse de la variance à un facteur

$$m = \begin{pmatrix} m_1 \\ \vdots \\ m_1 \\ m_2 \\ \vdots \\ m_2 \\ \vdots \\ \vdots \\ m_p \\ \vdots \\ m_p \end{pmatrix}.$$

et ξ le vecteur du bruit.
Loi de Y ? :

$$Y \sim \mathcal{N}_n(m; \sigma^2 Id_n).$$

m ?
σ^2 ?

Rappels : matrice de projection orthogonale

Soient E un e.v. munit d'un produit scalaire, W un sous-e.v. de E, et B un vecteur de E.
Soit Π_W la matrice de projection orthogonale d'un élément de E sur W. Alors

$$\begin{aligned}
\Pi_W^2 &= \Pi_W \\
trace(\Pi_W) &= dim(W) \\
Id_E B - \Pi_W B &= \Pi_{W^\perp} B = B^\perp \\
\Pi_W B &= \underset{a \in W}{\text{argmin}} \|B - a\|_2 \\
\Pi_W &= ww' \text{ lorsque } W = vect(w),\ w \text{ vect. unitaire}
\end{aligned}$$

Soient W et G deux ss-ev de E,

$$\begin{aligned}
\Pi_W \Pi_G = 0 &\iff W \perp G \\
\Pi_{W+G} &= \Pi_W + \Pi_G, \text{ si } W \perp G \\
\Pi_{W+G} &= \Pi_W + \Pi_{(Id_E - \Pi_W).G}, \text{théo. de Frisch-Waugh}
\end{aligned}$$

Rappels : matrice de projection orthogonale

Soient $E = \mathbb{R}^n$, V ss-ev de E (modèle), engendré par $(1_{n_1}, \ldots, 1_{n_p})$ où pour tout i,

$$1_{n_i} = {}^t(0, \ldots, 0, 1, \ldots, 1, 0, \ldots, 0).$$

Rappels : matrice de projection orthogonale

Soient $E = \mathbb{R}^n$, V ss-ev de E (modèle), engendré par $(1_{n_1}, \ldots, 1_{n_p})$ où pour tout i,

$$1_{n_i} = {}^t(0, \ldots, 0, 1, \ldots, 1, 0, \ldots, 0).$$

Alors

$$dim(V) = p.$$

Rappels : matrice de projection orthogonale

Soient $E = \mathbb{R}^n$, V ss-ev de E (modèle), engendré par $(1_{n_1}, \ldots, 1_{n_p})$ où pour tout i,

$$1_{n_i} = {}^t(0, \ldots, 0, 1, \ldots, 1, 0, \ldots, 0).$$

Alors

$$dim(V) = p.$$

$$dim(V^{\perp}) = n - p.$$

Rappels : matrice de projection orthogonale

Soient $E = \mathbb{R}^n$, V ss-ev de E (modèle), engendré par $(1_{n_1}, \ldots, 1_{n_p})$ où pour tout i,

$$1_{n_i} = {}^t(0, \ldots, 0, 1, \ldots, 1, 0, \ldots, 0).$$

Alors

$$dim(V) = p.$$

$$dim(V^\perp) = n - p.$$

$$\mathbb{R}^n = V \overset{\perp}{\oplus} V^\perp.$$

Estimateurs -ANOVA 1 facteur

Propriété 1 : Estimateur des effets du facteur

L'EMV $\widehat{m}$ de m est défini par

$$\begin{aligned} \widehat{m}_i &= Y_{i.} \\ \widehat{m} &= \sum_{i=1}^{n} \widehat{m}_i 1_{n_i} \sim \mathcal{N}_n(m; \sigma^2 \Pi_V) \end{aligned}$$

où $Y_{i.} = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij}$ et

Preuve : **On sait que MCO=EMV pour l'espérance dans un modèle gaussien :**

$$\begin{aligned} \widehat{m}^{EMV} &= \underset{m \in V}{argmin} \| Y - m \|_2^2 \\ &= \Pi_V Y \\ &= \Pi_V m + \Pi_V \xi \\ &= m + \Pi_V \xi \end{aligned}$$

et

$$\Pi_V \xi \sim \mathcal{N}_n(0_n, \sigma^2 \Pi_V \Pi_V') = \mathcal{N}_n(O_n, \sigma^2 \Pi_V).$$

Estimateurs -ANOVA 1 facteur

Propriété B : Estimateur de la variance

ESB de σ^2 :

$$\widehat{\sigma^2} = \frac{\|Y - \Pi_V Y\|_2^2}{n-p} = \frac{1}{n-p} \sum_{i=1}^{p} \sum_{j=1}^{n_i} (Y_{ij} - Y_{i.})^2.$$

$$(n-p)\frac{\widehat{\sigma^2}}{\sigma^2} \sim \chi^2(n-p)$$

et $\widehat{\sigma^2}$ indépendant de $\widehat{m}$.

Estimateurs -ANOVA 1 facteur

Preuve : Connu dans le cadre de l'estimation de la variance dans un modèle gaussien EMV de σ^2

$$\begin{aligned}\widehat{\tilde{\sigma}^2} &= \frac{\|Y - \Pi_V Y\|_2^2}{n} \\ &= \frac{1}{n}\|\Pi_{V^\perp} Y\|_2^2,\end{aligned}$$

–>transformation linéaire d'un vecteur gaussien $\sim \mathcal{N}_n(m, \sigma^2 Id_n)$

Estimateurs -ANOVA 1 facteur

Preuve : Connu dans le cadre de l'estimation de la variance dans un modèle gaussien EMV de σ^2

$$\begin{aligned}\widehat{\tilde{\sigma}^2} &= \frac{\|Y - \Pi_V Y\|_2^2}{n} \\ &= \frac{1}{n}\|\Pi_{V^\perp} Y\|_2^2,\end{aligned}$$

–>transformation linéaire d'un vecteur gaussien $\sim \mathcal{N}_n(m, \sigma^2 Id_n)$
–> par Cochran, indépendance avec $\Pi_V Y$ et

$$\begin{aligned}\frac{1}{\sigma^2}\|\Pi_{V^\perp} Y\|_2^2 &\sim \chi^2(n-p) \\ \mathbb{E}(\widehat{\tilde{\sigma}}^2) &= \frac{n-p}{n}\sigma^2\end{aligned}$$

Test sur l'effet du facteur

Tous les tests sont réalisables avec les propriétés précédentes. Ces tests se construisent de manière similaire à l'exemple suivant.

Test sur l'effet du facteur

Tous les tests sont réalisables avec les propriétés précédentes. Ces tests se construisent de manière similaire à l'exemple suivant.
Facteur considéré : pas influence ? influence ? (dans le cadre d'une supposée relation causale *a priori*) :

Test sur l'effet du facteur

Tous les tests sont réalisables avec les propriétés précédentes. Ces tests se construisent de manière similaire à l'exemple suivant.
Facteur considéré : pas influence ? influence ? (dans le cadre d'une supposée relation causale *a priori*) : $m_1 = \cdots = m_p$.
Soit W ss-ev engendré par 1_n (pas d'influence) : $W = \{\tilde{m}1_n, \tilde{m} \in \mathbb{R}\}$.

$$H_0 \ : \ m \in W \quad \text{contre} \quad H_1 \ : m \in V \cap W^{\perp}$$

au niveau α.

Test sur l'effet du facteur

Tous les tests sont réalisables avec les propriétés précédentes. Ces tests se construisent de manière similaire à l'exemple suivant.
Facteur considéré : pas influence ? influence ? (dans le cadre d'une supposée relation causale *a priori*) : $m_1 = \cdots = m_p$.
Soit W ss-ev engendré par 1_n (pas d'influence) : $W = \{\tilde{m}1_n, \tilde{m} \in \mathbb{R}\}$.

$$H_0 \ : \ m \in W \quad \text{contre} \quad H_1 \ : m \in V \cap W^{\perp}$$

au niveau α.
–> cadre paramétrique : existence test UPP,

Test sur l'effet du facteur

Tous les tests sont réalisables avec les propriétés précédentes. Ces tests se construisent de manière similaire à l'exemple suivant.
Facteur considéré : pas influence ? influence ? (dans le cadre d'une supposée relation causale *a priori*) : $m_1 = \cdots = m_p$.
Soit W ss-ev engendré par 1_n (pas d'influence) : $W = \{\tilde{m}1_n, \tilde{m} \in \mathbb{R}\}$.

$$H_0 \ : \ m \in W \quad \text{contre} \quad H_1 \ : m \in V \cap W^{\perp}$$

au niveau α.
–> cadre paramétrique : existence test UPP,
–> test du rapport de vraisemblance :

$$D = \left\{ \frac{\sup_{H_1} V_n}{\sup_{H_0} V_n} \geq t \right\}$$

$$D = \left\{ \frac{\sup_{m \in W^{\perp} \cap V, \sigma^2} V_n(m; \sigma^2)}{\sup_{m \in W, \sigma^2} V_n(m; \sigma^2)} \geq t \right\}.$$

Test sur l'effet du facteur

Les sup sont atteints pour les projecteurs sur W et sur V. Ainsi

$$D = \left\{ \frac{\sup_{m\in V,\sigma^2} V_n(m;\sigma^2)}{\sup_{m\in W,\sigma^2} V_n(m;\sigma^2)} \geq t \right\}$$

$$D = \left\{ \frac{V_n(\Pi_V Y; \frac{1}{n}\|Y-\Pi_V Y\|_2^2)}{V_n(\Pi_W Y; \frac{1}{n}\|Y-\Pi_W Y\|_2^2)} \geq t \right\}.$$

Or

$$\begin{aligned} V_n(\Pi_U Y; \frac{1}{n}\|Y-\Pi_U Y\|_2^2) &= \left(\frac{2\pi}{n}\|Y-\Pi_U Y\|_2^2\right)^{-n/2} \exp\left(-\frac{n}{2}\frac{\|Y-\Pi_U Y\|_2^2}{\|Y-\Pi_U Y\|_2^2}\right) \\ &= C_n \|Y-\Pi_U Y\|_2^{-n} \end{aligned}$$

donc

$$D = \left\{ \frac{\|Y-\Pi_W Y\|_2^2}{\|Y-\Pi_V Y\|_2^2} \geq K \right\}.$$

Test sur l'effet du facteur

$Y - \Pi_V Y$ est orthogonal à $\Pi_V Y - \Pi_W Y$

Test sur l'effet du facteur

$Y - \Pi_V Y$ est orthogonal à $\Pi_V Y - \Pi_W Y$
–> avec Cochran : indépendance.

Test sur l'effet du facteur

$Y - \Pi_V Y$ est orthogonal à $\Pi_V Y - \Pi_W Y$
–> avec Cochran : indépendance.
–>avec Pythagore

$$\|Y - \Pi_V Y\|_2^2 + \|\Pi_V Y - \Pi_W Y\|_2^2 = \|Y - \Pi_W Y\|_2^2.$$

Test sur l'effet du facteur

$Y - \Pi_V Y$ est orthogonal à $\Pi_V Y - \Pi_W Y$
–> avec Cochran : indépendance.
–>avec Pythagore

$$\|Y - \Pi_V Y\|_2^2 + \|\Pi_V Y - \Pi_W Y\|_2^2 = \|Y - \Pi_W Y\|_2^2.$$

Ainsi

$$D = \left\{ \frac{\|\Pi_V Y - \Pi_W Y\|_2^2/(p-1)}{\|Y - \Pi_V Y\|_2^2/(n-p)} \geq k \right\}.$$

et

$$\frac{\|\Pi_V Y - \Pi_W Y\|_2^2/(p-1)}{\|Y - \Pi_V Y\|_2^2/(n-p)} \sim F(p-1; n-p).$$

Test sur l'effet du facteur

On appelle cette décomposition la somme des carrés totale

$$\|Y - \Pi_W Y\|_2^2 = \|\Pi_V Y - \Pi_W Y\|_2^2 + \|Y - \Pi_V Y\|_2^2.$$

Elle s'interprète comme

$$SCT = SCR + SCf$$

où SC signifie somme des carrés, T=Totale R= Résidu et f= facteur.
On a donc

$$\begin{aligned} \frac{SCf}{\sigma^2} &\sim \chi^2(p-1) \\ \frac{SCR}{\sigma^2} &\sim \chi^2(n-p) \\ SCR & \quad independant \;\; SCf \\ F = \frac{SCF/p-1}{SCR/n-p} &\sim F(p-1, n-p) \end{aligned}$$

Test sur l'effet du facteur

On en tire le tableau d'analyse de la variance

Source de variation	Som Car	d.d.l	Carrés moyens	F
Facteur	SCf	p-1	$CMf = \frac{SCf}{p-1}$	
Résidus	SCR	n-p	$CMR = \frac{SCR}{n-p}$	$\frac{CMf}{CMR}$

Le test de Fisher est robuste vis à vis de l'hypothèse de gaussiannité : il "résiste" si les lois sont symétriques.
En revanche il ne résiste pas à une rupture d'homoscédasticité : σ^2 ne dépend pas du groupe d'appartenance.

Deux échantillons : test de student

$$m_1 = m_2, \quad \text{contre} \quad m_1 \neq m_2, m_1 < m_2, m_1 > m_2$$

Rapport de vrais :

$$\bar{Y}_{1.} - \bar{Y}_{2.} \sim \mathcal{N}(m_1 - m_2; \sigma^2(\frac{1}{n_1} + \frac{1}{n_2})).$$

Loi tabulée

Deux échantillons : test de student

$$m_1 = m_2, \quad \text{contre} \quad m_1 \neq m_2, m_1 < m_2, m_1 > m_2$$

Rapport de vrais :

$$\bar{Y}_{1.} - \bar{Y}_{2.} \sim \mathcal{N}(m_1 - m_2; \sigma^2(\frac{1}{n_1} + \frac{1}{n_2})).$$

Loi tabulée σ^2 ?

$$\begin{aligned} \widehat{\sigma}^2 &= \frac{1}{2}(S^2_{n_1}((Y_{1j})_j) + S^2_{n_2}((Y_{2j})_j)) \\ &= \frac{1}{n-2}\left(\sum_{j=1}^{n_1}(Y_{1j} - \bar{Y}_{1.})^2 + \sum_{j=1}^{n_2}(Y_{2j} - \bar{Y}_{2.})^2\right). \end{aligned}$$

Ainsi,

$$\frac{\bar{Y}_{1.} - \bar{Y}_{2.} - (m_1 - m_2)}{\sqrt{\widehat{\sigma}^2(\frac{1}{n_1} + \frac{1}{n_2})}} \sim T(n-2).$$

Même procédure lorsque plus de 2 échantillons, en traitant 2 à 2 les modalités du facteur considéré.

Hypothèses d'applications

Gaussianité (même si symétrie convient). : test de coef de symétrie, de KS

Test d'homoscédasticité :$\forall i = 1, \ldots, p, \sigma_i^2 = \sigma_2$.

test d'homogénéité des variances

Cadre : 1 facteur, deux modalités.

$$H_0 : \sigma_1^2 = \sigma_2^2, \quad \text{contre} \quad \sigma_1^2 \neq \sigma_2^2.$$

Rapport de vraisemblance :

$$\begin{aligned} R &= \frac{\sup_{H_1} V_n}{\sup_{H_0} V_n} \\ &= \frac{\sup_{m_1,m_2,\sigma_1^2 \neq \sigma_2^2} V_n}{\sup_{m_1,m_2,\sigma_1^2 = \sigma_2^2} V_n} \\ &= \frac{N}{D}. \end{aligned}$$

test d'homogénéité des variances

$$\begin{aligned} N &= V_n(\bar{Y}_{1.}, \bar{Y}_{2.}, \frac{n_1-1}{n_1}S^2_{n_1}; \frac{n_2-1}{n_2}S^2_{n_2}) \\ &= (2\pi)^{-\frac{n}{2}} e^{-\frac{n}{2}} \left(\frac{n_1-1}{n_1}S^2_{n_1}\right)^{-\frac{n_1}{2}} \left(\frac{n_2-1}{n_2}S^2_{n_2}\right)^{-\frac{n_2}{2}}. \end{aligned}$$

$$\begin{aligned} D &= V_n(\bar{Y}_{1.}, \bar{Y}_{2.}, \frac{(n_1-1)S^2_{n_1}+(n_2-1)S^2_{n_2}}{n}) \\ &= (2\pi)^{-\frac{n}{2}} e^{-\frac{n}{2}} \left(\frac{(n_1-1)S^2_{n_1}+(n_2-1)S^2_{n_2}}{n}\right)^{-\frac{n}{2}}. \end{aligned}$$

test d'homogénéité des variances

$$\begin{aligned} R &= C(n_1, n_2)\frac{(A+B)^{\alpha+\beta}}{A^\alpha B^\beta} \\ &= \left(\frac{A}{B}\right)^{-\alpha}\left(1+\frac{A}{B}\right)^{\alpha+\beta}. \end{aligned}$$

Sous H_0 : $\{R \geq t\} = \{\frac{A}{B} < a\} \cup \{\frac{A}{B} > b\}$. avec

$$\frac{A}{B} = \frac{(n_1-1)S_{n_1}^2}{(n_2-1)S_{n_2}^2}.$$

D'ou : $\{R \geq t\} = \{F < a' \cup F > b'\}$. avec

$$F = \frac{S_{n_1}^2}{S_{n_2}^2} \sim F(n_1-1, n_2-1).$$

(rque $F \sim F(p,q) \Longrightarrow \frac{1}{F} \sim F(q,p)$)
Si plus de deux modalités : test de Levene et en cas de suspicion de non normalité test de Brown-Forsythe

test de Bartlett

Cadre : 1 facteur, p modalités.

$$H_0 : \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_p^2, \quad \text{contre} \quad H_1 : \exists i, j \sigma_i^2 \neq \sigma_j^2.$$

$$Q = (n-p) ln(\frac{SCR}{n-p}) - \sum_{i=1}^{p} (n_i - 1) ln(S_i^2)$$

pour $S_i^2 = \frac{1}{n_i - 1} \|Y - \Pi_V Y\|_2^2$ et $SCR = \sum_{i=1}^{p} (n_i - 1) S_n i^2$.
Et on a

$$\frac{Q}{C(n, (n_i)_i, p)} \sim \chi^2(p-1).$$

- Très sensible à la non normalité

test de Hartley

Si effectifs n_i égaux entre eux, ce test est plus rapide.
Stat de test :

$$\frac{S^2_{i,max}}{S^2_{i,min}} \sim Hartley(p, n-1)$$

$$H_0 : \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_p^2, \quad \text{contre} \quad H_1 : \exists i, j \sigma_i^2 \neq \sigma_j^2.$$

Remarque :
- Pour $p = 2$ Hartley équivalent à F.
- Pour $p > 2$, Hartley moins sensible que Bartlett
- Très sensible à la non normalité

Comparaison de moyennes multiples

Cadre : H_0 de l'anova est rejetée.
But : Quels groupes ont des moyennes différentes ?

On désire tester H_O : $\exists i, j, m_i = m_j$.

Comparaison de moyennes multiples

Cadre : H_0 de l'anova est rejetée.
But : Quels groupes ont des moyennes différentes ?

On désire tester H_O : $\exists i, j, m_i = m_j$.
On compare 2 à 2 les moyennes.

1. Méthode de Student (LSD)
2. Méthode de Bonferroni
3. Méthode de Scheffé
4. Méthode de Tukey

Méthode LSD

On procède comme pour le test de Student en utilisant le carré moyen des résidus (CMR) pour estimer $\sigma 2$. On rejète H_0 pour

$$|\Pi_{V_i} Y - \Pi_{V_j} Y| \geq t_{1-\frac{\alpha}{2}}^{n-p} \widehat{\sigma} \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}.$$

Si effectifs égaux :

$$|\Pi_{V_i} Y - \Pi_{V_j} Y| \geq t_{1-\frac{\alpha}{2}}^{n-p} \widehat{\sigma} \sqrt{\frac{2}{n_1}}.$$

LSD : Least Significant Difference.

Méthode de Bonferroni

On part toujours du test de Student usuel mais en tenant compte du nombre de comparaisons effectuées. Estimateur de σ^2 par CMR. Même stat de test que pour LSD, mais on choisit α de manière à contrôler le risque d'erreur global.
α^* correspond à la valeur pour laquelle au moins une égalité est rejetée à tort, donc à 1 - probabilité qu'aucune égalitée ne soit rejetée à tort.
On a $p(p-1)/2$ paires, on obtient

$$\alpha^* \leq 1 - (1-\alpha)^{\frac{p(p-1)}{2}}$$

On pose $\alpha^* \leq C$. On a alors (en utilisant, $0 \leq \alpha \leq 1$) une condition suffisante

$$\alpha \leq \frac{2c}{p(p-1)}.$$

Méthode de Scheffé

On rejète H_0 si

$$|\Pi_{V_i} Y - \Pi_{V_j} Y| \geq \sqrt{(p-1) F_{1-\alpha^*}^{(p-1,n-p)}} \widehat{\sigma} \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}.$$

avec

$$\widehat{\sigma^2} = \frac{SCR}{n-p}.$$

On peut tester tous les couples, à la fois, en calculant $K = \sqrt{(p-1) F_{1-\alpha^*}^{(p-1,n-p)}}$ pour vérifier ensuite si

$$|\Pi_{V_i} Y - \Pi_{V_j} Y| \geq K \widehat{\sigma} \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}.$$

Méthode de Scheffé

On rejète H_0 si

$$|\Pi_{V_i} Y - \Pi_{V_j} Y| \geq q_{1-\alpha^*}^{(p,n-p)} \widehat{\sigma} \sqrt{\frac{1}{2}(\frac{1}{n_i} + \frac{1}{n_j})}.$$

avec $q_{1-\alpha^*}^{(p,n-p)}$ le fractile de l'étendue studentisée (tables).

Comparaison des méthodes

- Tukey + sensible à la détection de petites différences entre couples que Scheffé.
- Tukey préférable à Bonferroni pour comparer tout
- Si que quelques comparaisons Bonferroni préférable.

Comparaison des méthodes

- Tukey + sensible à la détection de petites différences entre couples que Scheffé.
- Tukey préférable à Bonferroni pour comparer tout
- Si que quelques comparaisons Bonferroni préférable.

En non paramétrique, mais en asymptotique, Kruskall-Wallis

Traitement ANOVA des hot-dogs

Test d'homogénéité des variances

Calories

Statistique de Levene	ddl1	ddl2	Signification
,490	2	51	,616

Traitement ANOVA des hot-dogs

Test de Kolmogorov-Smirnov à un échantillon

	Calories
N	54
Paramètres normaux[a,,b]	
Moyenne	145,44
Ecart-type	29,383
Différences les plus extrêmes	
Absolue	,095
Positive	,084
Négative	-,095
Z de Kolmogorov-Smirnov	,696
Signification asymptotique (bilatérale)	,718

Traitement ANOVA des hot-dogs

ANOVA

Calories

	Somme des carrés	ddl	Moyenne des carrés	F	Signifi
Inter-groupes	17692,195	2	8846,098	16,074	,000
Intra-groupes	28067,138	51	550,336		
Total	45759,333	53			

Traitement ANOVA des hot-dogs

Comparaisons multiples

Variable dépendante:Calories

	(I) Type	(J) Type	Différence de moyennes (I-J)	Erreur standard	Signification
Scheffe	Beef	Meat	-1,856	7,739	,972
		Poultry	38,085*	7,739	,000
	Meat	Beef	1,856	7,739	,972
		Poultry	39,941*	8,046	,000
	Poultry	Beef	-38,085*	7,739	,000
		Meat	-39,941*	8,046	,000
LSD	Beef	Meat	-1,856	7,739	,811
		Poultry	38,085*	7,739	,000
	Meat	Beef	1,856	7,739	,811
		Poultry	39,941*	8,046	,000
	Poultry	Beef	-38,085*	7,739	,000