Analyses bivariées - Introduction

Jusqu'à présent : études portant sur une seule variable.

Etude simultanée de deux variables nominales :

Analyse croisée de deux variables (par ex. questionnaire d'enquête)

– Loisir préféré et sexe

– Opinion sur l'immigration et sensibilité politique

Question posée : les deux variables sont-elles *indépendantes* ou *dépendantes* ?

Outil : analyse d'un tableau de contingence à l'aide d'un test du χ^2.

Etude de la liaison entre deux variables numériques

– Population d'étudiants. Variables : note de février et note de juin. Lien éventuel ?

– Population de sujets : lien entre taille et poids

Question posée : Y a-t-il un lien, une *corrélation* entre les deux variables ?

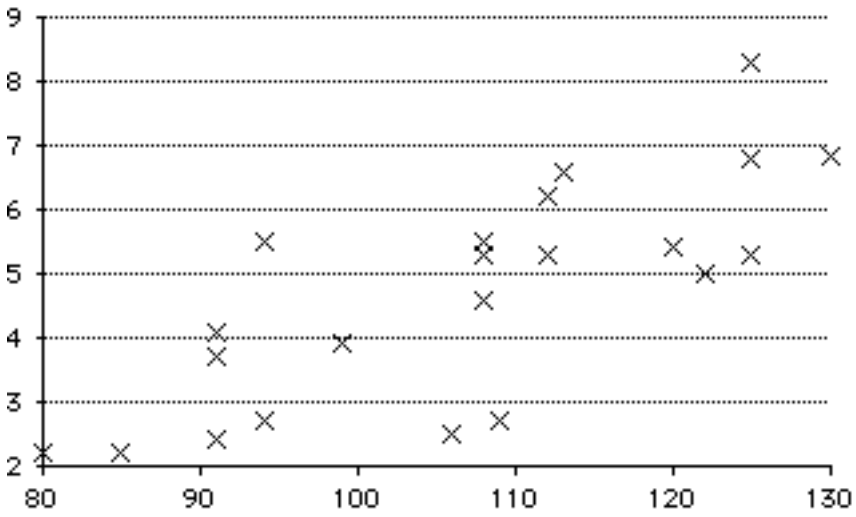
Outil : étude de la corrélation linéaire entre les deux variables

Corrélation linéaire

Données : deux variables numériques définies sur la même population

	X	Y
s_1	x_1	y_1
s_2	x_2	y_2
...	...	...

Nuage de points : points (x_i, y_i)



Covariance des variables X et Y

$$Cov(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$$

ou

$$Cov(X,Y) = \left(\frac{1}{n}\sum_{i=1}^{n} x_i y_i\right) - \overline{x}\ \overline{y}$$

Coefficient de corrélation de Bravais Pearson

$$r = \frac{Cov(X,Y)}{\sigma(X)\sigma(Y)}$$

Remarques

– Formules analogues avec corrélation, écarts types, . . .corrigés

– Il existe des relations non linéaires

– Corrélation n'est pas causalité

Régression linéaire

Rôle "explicatif" de l'une des variables par rapport à l'autre. Les variations de Y peuvent-elles (au moins en partie) être expliquées par celles de X ? Peuvent-elles être prédites par celles de X ?

Modèle permettant d'estimer Y connaissant X

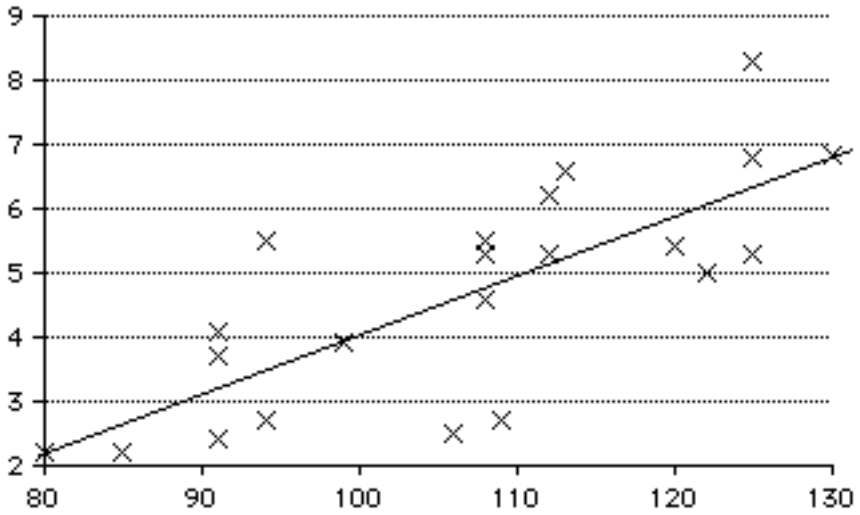
Droite de régression de Y par rapport à X :

La droite de régression de Y par rapport à X a pour équation :

$$y = ax + b$$

avec :

$$a = \frac{Cov(X,Y)}{\sigma^2(X)} \quad ; \quad b = \overline{y} - a\overline{x}$$



Comparaison des valeurs observées et des valeurs estimées

Valeurs estimées : $\widehat{y}_i = ax_i + b$
Erreur (ou résidu) : $e_i = y_i - \widehat{y}_i$

On montre que :

$$\sigma^2(Y) = \sigma^2(\widehat{Y}) + \sigma^2(E)$$

avec :

$$\frac{\sigma^2(E)}{\sigma^2(Y)} = 1 - r^2 \quad ; \quad \frac{\sigma^2(\widehat{Y})}{\sigma^2(Y)} = r^2$$

$\sigma^2(\widehat{Y})$: variance *expliquée* (par la variation de X, par le modèle)

$\sigma^2(E)$: variance *perdue* ou *résiduelle*

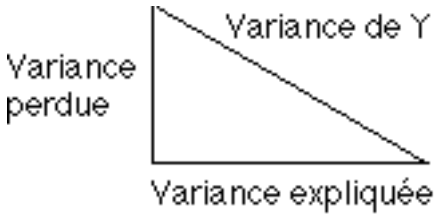
r^2 : part de la variance de Y qui est expliquée par la variance de X. *Coefficient de détermination*

Exemple : $r = 0.86$

$r^2 = 0.75$; $1 - r^2 = 0.25$; $\sqrt{1 - r^2} = 0.5$.

• La part de la variance de Y expliquée par la variation de X est de 75%.

• L'écart type des résidus est la moitié de l'écart type de Y.



Etude de la liaison entre deux caractères qualitatifs

Exemple : préférences des publics masculin et féminin.
Effectifs observés

	H	F	Total
Comédie	90	75	165
Drame	50	45	95
Variétés	160	80	240
Total	300	200	500

Goûts dépendants du sexe ?

Exemple : $\frac{240}{500} = 48\%$. 48% des personnes interrogées préfèrent les variétés. Si les deux variables étaient parfaitement indépendantes, on devrait retrouver :
– 48% des hommes, c'est-à-dire 144 hommes préférant les variétés
– 48% des femmes, c'est-à-dire 96 femmes préférant les variétés.

Effectifs attendus (ou théoriques) si indépendance :

$$\text{Dans chaque case : effectif} = \frac{\text{total ligne} \times \text{total colonne}}{\text{total général}}$$

	H	F
Comédie	99	66
Drame	57	38
Variétés	144	96

Problème : évaluer la distance entre les deux tableaux ?

Calcul de la "distance" du χ^2 définie par :

$$\chi^2_{obs} = \sum_{i,j} \frac{(n_{ij} - t_{ij})^2}{t_{ij}}$$

Calcul pratique :

Mod.	n_{ij}	t_{ij}	$\frac{(n_{ij} - t_{ij})^2}{t_{ij}}$
H.C.	90	99	0.82
H.D	...		0.86
H.V.			1.78
F.C.			1.23
F.D.			1.29
F.V.			2.67
			8.64

Quelle interpétation peut-on donner de χ^2_{obs} ?

– Coefficient de contingence de Pearson :

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

– Coefficient Phi de Cramér :

$$\Phi_c = \sqrt{\frac{\chi^2}{N(k-1)}}$$

où N est l'effectif total et k le minimum des deux valeurs nombre de lignes, nombre de colonnes.

– En fait, la véritable interprétation de χ^2 est faite en statistiques inférentielles, à l'aide d'un test du χ^2.

Test du χ^2

– 500 personnes : échantillon

– 2 sources de variation : effet du sexe, hasard

– *Si seul le hasard est en cause*, la distance suit une *loi du χ^2 à 2 ddl.*

– On se fixe un seuil de 5% (par exemple)

– *Si seul le hasard est en cause*, on a seulement 5% de chances d'observer un χ^2 supérieur à la *valeur critique* $\chi^2_c = 5.991$.

– Or on a observé : $\chi^2_{obs} = 8.64$.

– Conclusion : différence de goûts selon le sexe.

Résumé

Tableau de contingence : effectifs observés n_{ij}
Totaux par ligne : $n_{i.}$ par colonne : $n_{.j}$
Total général : N ou $n_{..}$
l lignes et c colonnes

Effectifs théoriques : tableau (t_{ij}) avec :

$$t_{ij} = \frac{n_{i.}\, n_{.j}}{n_{..}} = \frac{total\ \ ligne \times total\ \ colonne}{total\ \ général}$$

Distance du χ^2 :

$$\chi^2_{obs} = \sum_{i,j} \frac{(n_{ij} - t_{ij})^2}{t_{ij}}$$

Test proprement dit :

– Hypothèses :
H_0 : Les variables sont indépendantes.
H_1 : Les variables sont dépendantes.

– On fixe un seuil α (5%, 1%, . . .)

– Lecture de la table : valeur critique χ^2_{crit} pour le seuil α et $(l-1)(c-1)$ ddl

– Intervalles d'acceptation et de rejet

– Comparaison de χ^2_{obs} et de χ^2_{crit}

– Conclusion :
  – Si $\chi^2_{obs} < \chi^2_{crit}$, indépendance acceptée
  – Si $\chi^2_{obs} > \chi^2_{crit}$, indépendance rejetée ;
  les variables dépendent l'une de l'autre.

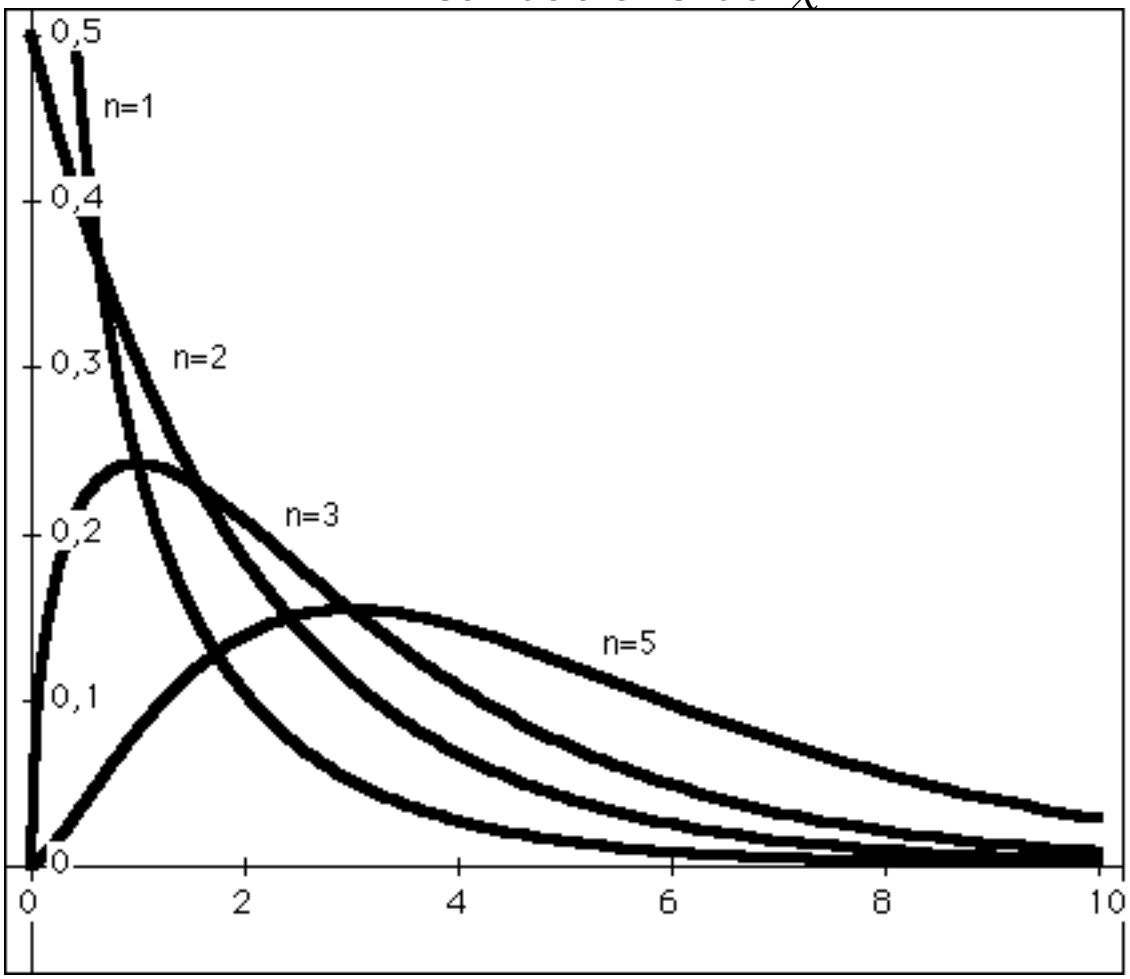
Remarques

– Condition sur les effectifs théoriques minimaux ; regroupement éventuel des modalités

– Correction de Yates (tableau 2×2)

$$\chi^2_{corr} = \sum_{i,j} \frac{(|n_{ij} - t_{ij}| - 0.5)^2}{t_{ij}}$$

Distributions du χ^2



Une autre utilisation du χ^2 : adéquation à une loi théorique

Confronter un tableau d'effectifs "théoriques" à un tableau d'effectifs observés.

Exemple : Etude de Durkheim sur le suicide selon la saison (entre 1835 et 1843).

Pour un échantillon de 1000 suicides :

Print.	Eté	Autom.	Hiver	Total
283	306	210	201	1000

Fréquences et effectifs théoriques :

Print.	Eté	Autom.	Hiver	Total
25%	25%	25 %	25 %	100%
250	250	250	250	1000

Contributions au χ^2 :

	Print.	Eté	Autom.	Hiver	χ^2
Obs.	283	306	210	201	
Théo.	250	250	250	250	
Ctri	4.356	12.544	6.40	9.604	32.904

Test proprement dit :

H_0 : Le tableau de fréquences théoriques correspond aux fréquences dans la population
H_1 : Les fréquences dans la population sont différentes de celles du tableau de fréquences théoriques

Statistique de test : χ^2 à 3 ddl.
Pourquoi 3 ? : 4 observations. Total théorique (1000) calculé à partir des observations. $4 - 1 = 3$.

Seuil : 1%. Valeur critique : $\chi^2_{crit} = 11.35$

Règle de décision :
– Si $\chi^2_{obs} \leq 11.345$, on accepte H_0.
– Si $\chi^2_{obs} > 11.345$, on refuse H_0 et on accepte H_1.

Conclusion :
Ici, $\chi^2_{obs} = 32.9$. On accepte donc H_1. Autrement dit, le taux de suicides varie avec la saison.