Statistique descriptive bivariée

Anicet E. T. Ebou, ediman.ebou@inphb.ci

01

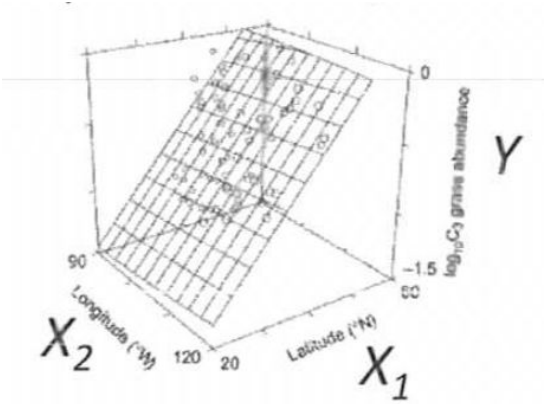
Que veut dire “linéaire” ?

Que veut dire “linéaire”?

- Ce chapitre traite de corrélation et régression **linéaires**
- La définition de linéaire dans ce contexte n’est pas seulement “une relation en forme de droite entre 2 variables”
- C’est plus général: un modèle linéaire exprime que la relation est décrite par une combinaison linéaire de paramètres; aucun paramètre n’apparaît en exposant, ni multiplié ou divisé par un autre paramètre

Que veut dire “linéaire”?

- Exemple: $Y = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2$
- Ce modèle n’est pas une droite, mais un plan:



- Mais c’est un modèle linéaire car aucun des 3 paramètres β_i ne se trouve en exposant, multiplié ou divisé par un autre

02

Série statistique bivariée

Série statistique bivariée

On s'intéresse à deux variables *x* et *y*. Ces deux variables sont mesurées sur les n unités d'observation. Pour chaque unité, on obtient donc deux mesures.

La série statistique est alors une suite de n couples des valeurs prises par les deux variables sur chaque individu: (x_1, y_1), …, (x_i, y_i), …(x_n, y_n). Chacune des deux variables peut être, soit quantitative, soit qualitative.

03

Deux variables quantitatives

3.1

Introduction

Relation et dépendance

Soit deux caractères quantitatifs *X* et *Y*, décrivant le même ensemble d'unités.

On dit qu'il existe une relation entre *X* et *Y* si l'attribution des modalités de *X* et de *Y* ne se fait pas au hasard, c'est-à-dire si les valeurs de *X* **dépendent** des valeurs de *Y* ou si les valeurs de *Y* **dépendent** des valeurs de *X*.

Relation et dépendance

Dire que *Y* **dépend** de *X* signifie que la connaissance des valeurs de *X* permet de prédire, dans une certaine mesure, les valeurs de *Y*. En d'autres termes, si *Y* dépend de *X*, on peut trouver une fonction ***f*** telle que ***Y = f(X)***.

Représentation graphique de deux variables

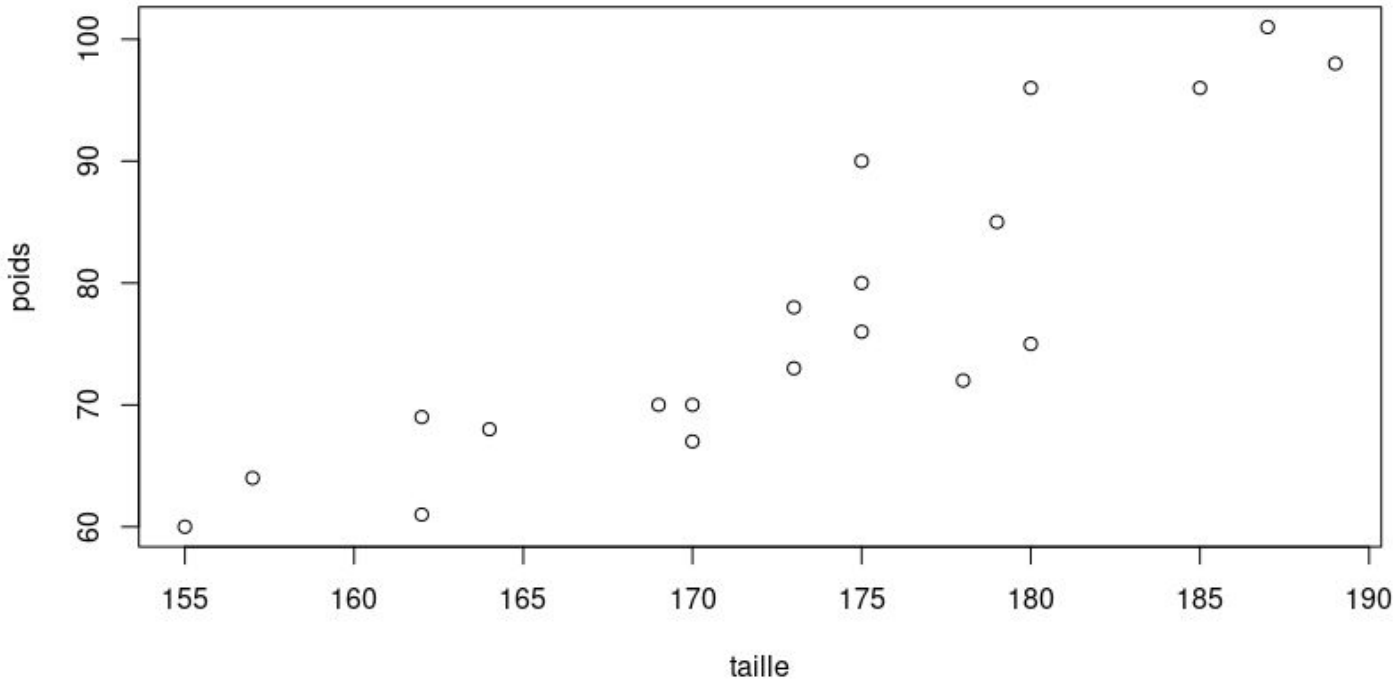
Dans ce cas, chaque couple est composé de deux valeurs numériques. Un couple de nombres (entiers ou réels) peut toujours être représenté comme un point dans un plan.

Représentation graphique de deux variables

Exercice: On mesure le poids *Y* et la taille *X* de 20 individus. Représenter le nuage de points de ces deux variables.

y_i	x_i	y_i	x_i
60	155	75	180
61	162	76	175
64	157	78	173
67	170	80	175
68	164	85	179
69	162	90	175
70	169	96	180
70	170	96	185
72	178	98	189
73	173	101	187

Représentation graphique de deux variables



Nuage de points en langage R

```
> poids <- c(60,61,64,67,68,69,70,70,72,73,75,76,78,80,85,90,
96,96,98,101)
> taille <- c(155,162,157,170,164,162,169,170,178,173,180,175,
173,175,179,175,180,185,189,187)
> plot(taille, poids)
```

Analyse des variables

Les variables *x* et *y* peuvent être analysées séparément. On peut calculer tous les paramètres dont les moyennes et les variances :

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad s_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2,$$

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i, \qquad s_y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2.$$

Ces paramètres sont appelés *paramètres marginaux*: *variances marginales*, *moyennes marginales*, *écarts-types marginaux*, *quantiles marginaux*, etc..

3.2

Covariance

Covariance

La covariance est une mesure de la variabilité conjointe de deux variables aléatoires.

Elle est définie par la formule suivante: $s_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}).$

En développant l'expression précédente, on obtient une seconde forme de la formule précédente: $\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y}.$

Covariance

La covariance peut prendre des valeurs positives, négatives ou nulles.

Si les valeurs supérieures d'une variable correspondent principalement aux valeurs supérieures de l'autre variable, et qu'il en va de même pour les valeurs inférieures (c'est-à-dire que les variables ont tendance à présenter un comportement similaire), la covariance est positive.

Covariance

Dans le cas contraire, lorsque les valeurs supérieures d'une variable correspondent principalement aux valeurs inférieures de l'autre, (c'est-à-dire que les variables tendent à avoir un comportement opposé), la covariance est négative.

Quand $x_i = y_i$, pour tout $i = 1,..., n$, la covariance est égale à la variance.

3.3

Corrélation de Pearson

Corrélation de Pearson

Le coefficient de corrélation (ou coefficient de corrélation de Pearson) est la covariance normalisée par les deux écart-types marginaux:

$$r_{xy} = \frac{s_{xy}}{s_x s_y}.$$

Le coefficient de corrélation permet d'analyser les relations linéaires.

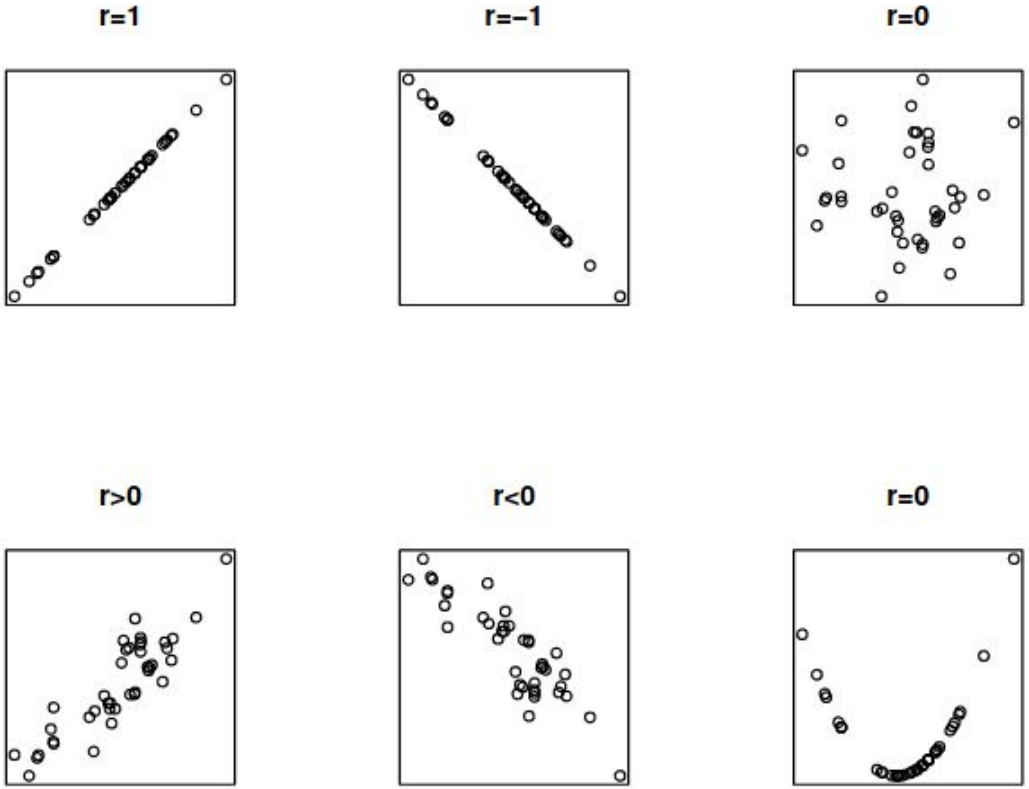
Corrélation de Pearson

- Le coefficient de corrélation mesure **la dépendance linéaire** entre deux variables;
- Le coefficient de corrélation varie entre -1 et 1;
- Si le coefficient de corrélation est positif, les points sont alignés le long d'une droite croissante;

Corrélation de Pearson

- Si le coefficient de corrélation est négatif, les points sont alignés le long d'une droite décroissante;
- Si le coefficient de corrélation est nul ou proche de zéro, il n'y a pas de dépendance linéaire. On peut cependant avoir une dépendance non-linéaire avec un coefficient de corrélation nul.

Corrélation de Pearson



Conditions d’application

- Deux variables continues;
- Les observations dans une même variable doivent être indépendantes;
- La distribution des données suit une loi normale.

Exercice

Calculez le coefficient de corrélation de Pearson et interprétez le résultat.

Individus	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Poids	68	55.4	69	61.6	65.7	68.4	75	53	85.5	93.6	72.3	68	57	58	53	94	71.5	72	56	73
Glycémie	1.08	0.84	0.72	0.62	0.75	0.77	0.72	0.57	1.02	1.09	0.99	0.78	0.81	0.91	1.06	1.05	0.94	1.15	1.03	1.07

3.4

Corrélation de Spearman

Corrélation de Spearman

Le coefficient de corrélation de Spearman mesure les relations monotones quelque soit la forme. Il est utilisé lorsque les conditions d'application du coefficient de corrélation de Pearson ne sont pas remplies.

Corrélation de Spearman

Il examine s'il existe une relation entre le rang des observations pour deux caractères *X* et *Y*, ce qui permet de détecter l'existence de relations monotones (croissante ou décroissante), quelle que soit leur forme précise (linéaire, exponentiel, puissance, ...).

Corrélation de Spearman

Le coefficient de corrélation est la covariance de variables de rang sur les écarts-types des variables de rang.

$$r_s = \frac{\operatorname{cov}(\mathrm{rg}_X, \mathrm{rg}_Y)}{\sigma_{\mathrm{rg}_X} \sigma_{rg_Y}}$$

Un rang est un nombre consécutif affecté à une observation spécifique dans un échantillon d'observations triées par leurs valeurs, qui ainsi reflète la relation ordinale de l'observation.

Corrélation de Spearman

Par exemple, la suite de données suivante : 10, 20, 13, 45, 0, 12 est d'abord classé par ordre croissant 0, 10, 12, 13, 20, 45 et sera ensuite remplacée par : 2, 5, 4, 6, 1, 3

Exercice

Calculez le coefficient de corrélation de Spearman et interprétez le résultat.

Individus	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Poids	68	55.4	69	61.6	65.7	68.4	75	53	85.5	93.6	72.3	68	57	58	53	94	71.5	72	56	73
Glycémie	1.08	0.84	0.72	0.62	0.75	0.77	0.72	0.57	1.02	1.09	0.99	0.78	0.81	0.91	1.06	1.05	0.94	1.15	1.03	1.07

3.5

Coefficient de détermination

Coefficient de détermination

Le coefficient de détermination est le carré du coefficient de corrélation:

$$r_{xy}^2 = \frac{s_{xy}^2}{s_x^2 s_y^2}.$$

Il est noté R^2 ou r^2 et prononcé "R carré", est la proportion de la variation de la variable dépendante qui est prévisible à partir de la ou des variables indépendantes.

Coefficient de détermination

Il s'agit d'une statistique utilisée dans le cadre de modèles statistiques dont l'objectif principal est soit la prédiction de résultats futurs, soit la vérification d'hypothèses, sur la base d'autres informations connexes.

Coefficient de détermination

Elle fournit une mesure de l'efficacité avec laquelle les résultats observés sont reproduits par le modèle, sur la base de la proportion de la variation totale des résultats expliquée par le modèle.

Le coefficient de détermination est normalement compris entre 0 et 1.

3.6

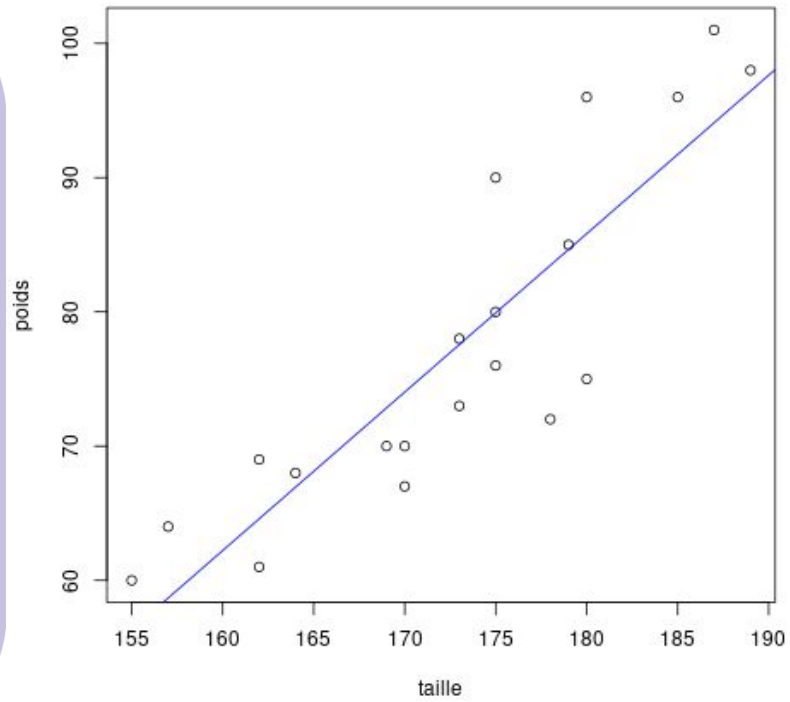
Régression linéaire

Droite de régression

La droite de régression est la droite qui ajuste au mieux un nuage de points dans le plan.

Si on considère que la variable *X* est explicative et que la variable *Y* est

dépendante, l'équation d'une droite est: $y = ax + b$.



Estimation de la droite de régression

La droite de régression peut être estimée par la méthode du maximum de vraisemblance, la méthode des moindres carrés, la méthode des moments ou encore par des méthodes bayésiennes.

Dans le cadre de ce cours nous explorerons la méthode des moindres carrés.

Estimation de la droite de régression

Dans le cas le plus standard, où les termes d'erreurs sont indépendants et identiquement distribués, l'estimateur des moindres carrés ordinaires est le plus efficace des estimateurs linéaires sans biais.

Pour déterminer la valeur des coefficients *a* et *b* on utilise le principe des moindres carrés qui consiste à chercher la droite qui minimise la somme des carrés des résidus.

Estimation de la droite de régression

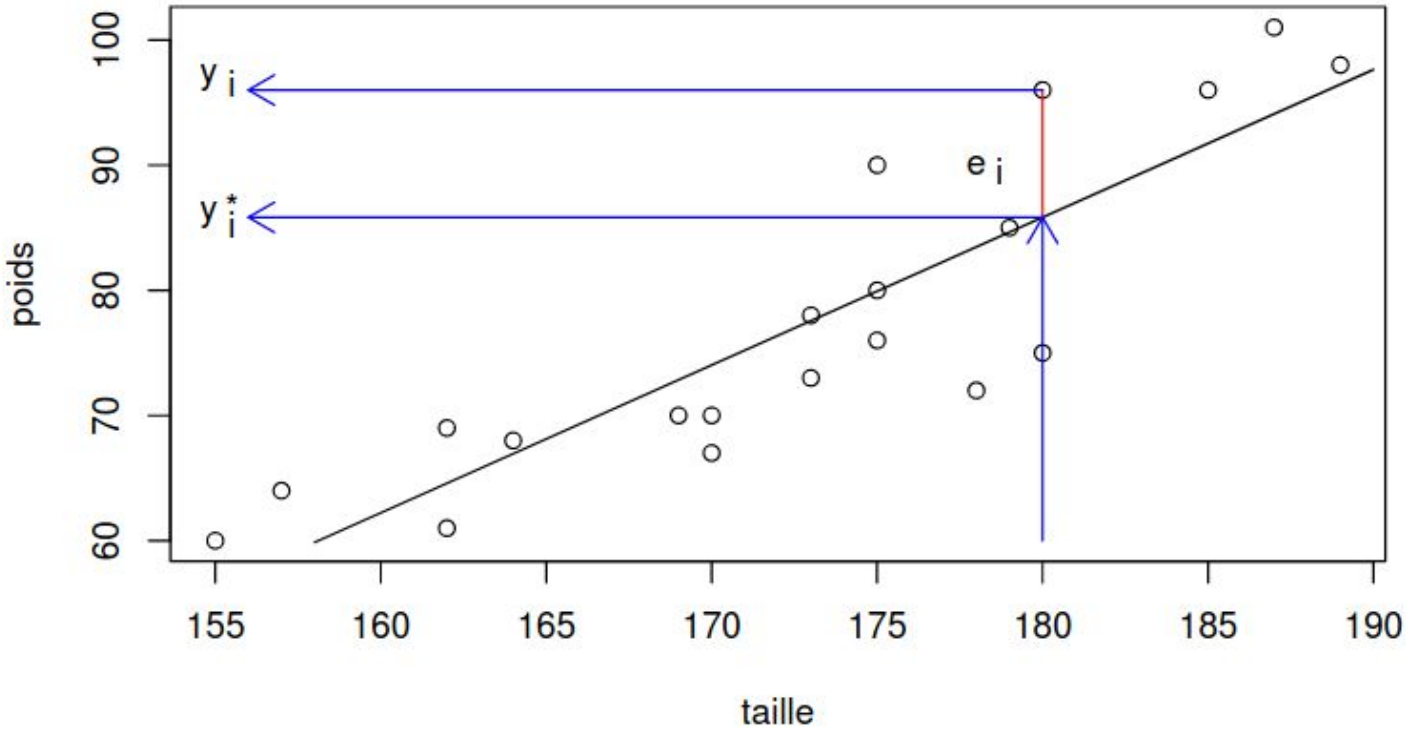
En d'autres termes, il s'agit de minimiser la somme des carrés des écarts entre les valeurs prédites et les valeurs observées par rapport aux deux paramètres a et b (b est l'ordonnée à l'origine et a est la pente de la droite de régression) :

$$S = Argmin_{a,b} \sum_{i=1}^{n} e_i^2 = Argmin_{a,b} \sum_{i=1}^{n} (y_i - ax - b)^2$$

Estimation de la droite de régression

Le résidu e_i est l'erreur que l'on commet en utilisant la droite de régression pour prédire y_i à partir de x_i. Les résidus peuvent être positifs ou négatifs.

Estimation de la droite de régression



Estimation de la droite de régression

Les coefficients *a* et *b* qui minimisent le critère des moindres carrés sont donnés par $a = \frac{s_{xy}}{s_x^2}$ et $b = \overline{y} - a\overline{x}$

La droite de régression de *y* en *x* n'est pas la même que la droite de régression de *x* en *y*.

Résidus et valeurs ajustées

Les *valeurs ajustées* sont obtenues au moyen de la droite de régression:

$$y_i^* = a + bx_i.$$

Les valeurs ajustées sont les **prédictions** des y_i, réalisées au moyen de la variable *x* et de la droite de régression de *y* en *x*.

Résidus et valeurs ajustées

Les *résidus* sont les différences entre les valeurs observées et les valeurs ajustées de la variable dépendante.

$$e_i = y_i - y_i^*.$$

Les résidus représentent la partie inexpliquée des y_i par la droite de régression.

Sommes de carrés et variances

On appelle *somme des carrés totale* la quantité $\mathrm{SC}_{TOT} = \sum_{i=1}^{n}(y_i - \bar{y})^2$

La variance marginale peut alors être définie par $s_y^2 = \frac{\mathrm{SC}_{TOT}}{n} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2.$

On appelle *somme des carrés de la régression* la quantité:

$$\mathrm{SC}_{REGR} = \sum_{i=1}^{n}(y_i^* - \bar{y})^2.$$

Sommes de carrés et variances

La *variance de régression* est la variance des valeurs ajustées:

$$s^2_{y^*} = \frac{1}{n}\sum_{i=1}^{n}(y_i^* - \bar{y})^2.$$

On appelle *somme des carrés des résidus* (ou résiduelle) la quantité:

$$\mathrm{SC}_{RES} = \sum_{i=1}^{n} e_i^2.$$

Sommes de carrés et variances

La *variance résiduelle* est la variance des résidus:

$$s_e^2 = \frac{\mathrm{SC}_{RES}}{n} = \frac{1}{n}\sum_{i=1}^{n} e_i^2.$$

La somme des carrés totale est la somme des carrés de la régression et de la somme des carrés résiduels.

$$\mathrm{SC}_{TOT} = \mathrm{SC}_{REGR} + \mathrm{SC}_{RES}.$$

Sommes de carrés et variances

La relation précédente stipule que la variation totale des valeurs observées de *Y* autour de leur moyenne peut être décomposée en deux parties : une attribuable à la droite de régression et l'autre à des facteurs aléatoires car toutes les valeurs de *Y* observées ne sont pas sur la droite de régression ajustée.

Sommes de carrés et variances

Le coefficient de détermination r^2 est une mesure de la qualité de l'ajustement d'une droite de régression. r^2 mesure la proportion de la variation totale de *Y* expliquée par le modèle de régression.

$$r^2 = SC_{REGR} / SC_{TOT}$$

Sommes de carrés et variances

Il faut noter que:

- r^2 est non négatif;
- $0 \leq r^2 \leq 1$: $r^2 = 1$ correspond à un ajustement parfait alors que $r^2 = 0$ dénote l'absence de relation entre la variable dépendante et les variables indépendantes.

Décomposition de la variance

La *variance de régression* peut également s'écrire

$$s_{y^*}^2 = s_y^2 r^2,$$

où r^2 est le coefficient de détermination.

La *variance résiduelle* est la variance des résidus:

$$s_e^2 = \frac{1}{n} \sum_{i=1}^{n} e_i^2.$$

Décomposition de la variance

La variance résiduelle peut également s'écrire en fonction du coefficient de détermination r^2.

$$s_e^2 = s_y^2(1 - r^2),$$

La variance marginale est la somme de la variance de régression et de la variance résiduelle.

$$s_y^2 = s_{y^*}^2 + s_e^2.$$

3.7

Correlation *vs* régression

Corrélation *vs* régression

- Corrélation et régression diffèrent conceptuellement
- **Corrélation**:
  - on suppose que les 2 variables *X1* et *X2* sont liées entre elles, qu'elles varient de concert, mais pas que l'une explique l'autre
  - les rôles de *X1* et *X2* sont interchangeables
  - c'est une analyse exploratoire pour déterminer s'il existe une relation de covariance entre *X1* et *X2*

Corrélation *vs* régression

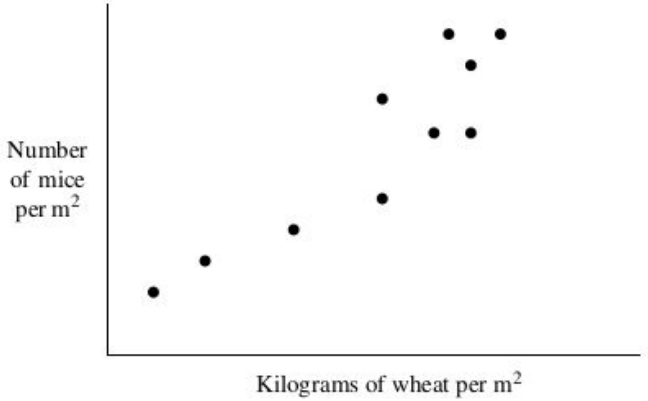
- **Régression:**
  - on suppose une relation fonctionnelle d'une variable facteur *X* qui explique/détermine une variable réponse *Y*
  - Les rôles de *X* et *Y* ne sont pas interchangeables: *X* explique *Y* mais *Y* n'explique pas *X*

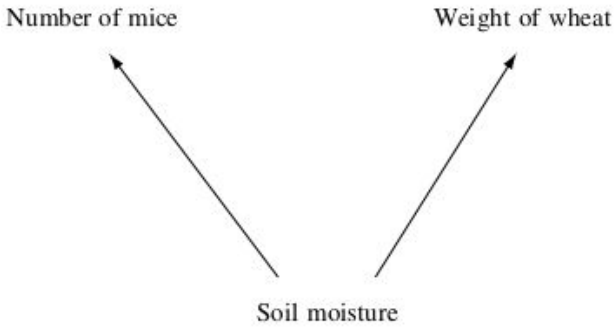
Relation n’est pas cause!

- Ce n’est pas parce que *X* **est corrélé** à *Y* que *X* **cause** *Y*
- Ce n’est pas parce que *X* **explique** *Y* que *X* **cause** *Y*
- Exemple:
  - l’usure des dents de koalas (*Y*) est expliquée par l’âge (*X*)
  - l’inverse n’est pas vrai: l’âge n’est pas expliqué par l’usure des dents
  - L’usure des dents n’est pas causée par l’âge, mais par la mastication, qui elle est fonction de l’âge

Corrélation n’est pas cause!

- On en a déjà parlé dans l’introduction du cours: pour établir un lien de cause à effet (*X* cause *Y*), il faut modifier *X* et voir si *Y* est modifié





04

Deux variables qualitatives

Données observées

Si les deux variables x et y sont qualitatives, alors les données observées sont une suite de couples de variables.

$$(x_1, y_1), \ldots, (x_i, y_j), \ldots, (x_n, y_n)$$

Chacune des deux variables prend comme valeurs des modalités qualitatives.

Tableau de contingence

Les données observées peuvent être regroupées sous la forme d'un *tableau de contingence*.

	y_1	$\cdots$	y_k	$\cdots$	y_K	total
x_1	n_{11}	$\cdots$	n_{1k}	$\cdots$	n_{1K}	$n_{1.}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	
x_j	n_{j1}	$\cdots$	n_{jk}	$\cdots$	n_{jK}	$n_{j.}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	
x_J	n_{J1}	$\cdots$	n_{Jk}	$\cdots$	n_{JK}	$n_{J.}$
total	$n_{.1}$	$\cdots$	$n_{.k}$		$n_{.K}$	n

Tableau de contingence

Les $n_{j.}$ et $n_{.k}$ sont appelés les effectifs marginaux. Dans le tableau précédent:

- $n_{j.}$ représente le nombre de fois que la modalité x_j apparaît;
- $n_{.k}$ représente le nombre de fois que la modalité y_k apparaît;
- n_{jk} représente le nombre de fois que les modalités x_j et y_k apparaissent ensemble.

Tableau de contingence

Représenter le tableau de contingence des deux variables quantitatives suivantes:

Sexe	**Couleur des yeux**
Homme	Bleu
Homme	Marron
Homme	Bleu
Femme	Vert
Femme	Marron

Tableau de fréquences

Le tableau de fréquences s'obtient en divisant tous les effectifs par la taille de l'échantillon.

	y_1	$\cdots$	y_k	$\cdots$	y_K	total
x_1	f_{11}	$\cdots$	f_{1k}	$\cdots$	f_{1K}	$f_{1.}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	
x_j	f_{j1}	$\cdots$	f_{jk}	$\cdots$	f_{jK}	$f_{j.}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	
x_J	f_{J1}	$\cdots$	f_{Jk}	$\cdots$	f_{JK}	$f_{J.}$
total	$f_{.1}$	$\cdots$	$f_{.k}$		$f_{.K}$	1

Tableau de fréquences

Représenter le tableau de fréquences à partir du tableau de contingence obtenu précédemment.

Profils lignes et profils colonnes

Un tableau de contingence s'interprète toujours en comparant des fréquences en lignes ou des fréquences en colonnes (appelés aussi profils lignes et profils colonnes).

Profils lignes et profils colonnes

Les profils colonnes sont définis par:

$$f_k^{(j)} = \frac{n_{jk}}{n_{j.}} = \frac{f_{jk}}{f_{j.}}, k = 1, \ldots, K, j = 1, \ldots, J,$$

Et les profils lignes sont définis par:

$$f_j^{(k)} = \frac{n_{jk}}{n_{.k}} = \frac{f_{jk}}{f_{.k}}, j = 1, \ldots, J, k = 1, \ldots, K.$$

Profils lignes et profils colonnes

Calculez les profils lignes et profils colonnes du tableau précédent.

Effectifs théoriques et khi-carré

On cherche souvent une interaction entre des lignes et des colonnes, un lien entre les variables. Pour mettre en évidence ce lien, on construit un tableau d'***effectifs théoriques*** qui représente la situation où les variables ne sont pas liées (indépendance).

Effectifs théoriques et khi-carré

Ces effectifs théoriques sont construits de la manière suivante:

$$n^*_{jk} = \frac{n_{j.} n_{.k}}{n}.$$

Les effectifs observés n_{jk} ont les mêmes marges que les effectifs théoriques n^*_{jk}.

Enfin, les *écarts à l'indépendance* sont définis par

$$e_{jk} = n_{jk} - n^*_{jk}.$$

Effectifs théoriques et khi-carré

La dépendance du tableau se mesure au moyen du khi-carré défini par:

$$\chi^2_{obs} = \sum_{k=1}^{K} \sum_{j=1}^{J} \frac{(n_{jk} - n^*_{jk})^2}{n^*_{jk}} = \sum_{k=1}^{K} \sum_{j=1}^{J} \frac{e^2_{jk}}{n^*_{jk}}.$$

C'est-à-dire la somme des différences entre les effectifs observés et les effectifs théoriques au carré divisé par les effectifs théoriques.

Effectifs théoriques et khi-carré

Le khi-carré peut être normalisé pour ne plus dépendre du nombre d'observations. On définit le phi-deux par:

$$\phi^2 = \frac{\chi^2_{obs}}{n}.$$

Le phi-deux ne dépend plus alors du nombre d'observations.

Effectifs théoriques et khi-carré

Le V de Cramer est définit par

$$V = \sqrt{\frac{\phi^2}{\min(J-1, K-1)}} = \sqrt{\frac{\chi^2_{obs}}{n \min(J-1, K-1)}}.$$

Le V de Cramer est compris entre 0 et 1. Il ne dépend ni de la taille de l'échantillon ni de la taille du tableau. Si V est proche de 0, les deux variables sont indépendantes. Si V = 1, il existe une relation fonctionnelle entre les variables.

Effectifs théoriques et khi-carré

Calculer les effectifs théoriques et les valeurs à l'écart du tableau précédent.

Effectifs théoriques et khi-carré

Exercice: La variable *X* est le niveau d'instruction du fils par rapport au père (plus élevé, égal, inférieur), et la variable *Y* est le statut professionnel du fils par rapport au père (plus élevé, égal, inférieur). Tester la dépendance entre ces deux variables.

Niveau d'instruction du fils par rapport au père	Statut professionnel du fils par rapport au père			
	Plus élevé	Egal	inférieur	total
plus élevé	134	96	61	291
égal	23	33	24	80
inférieur	7	16	22	45
total	164	145	107	416

Effectifs théoriques et khi-carré

Exercice: La consommation de crèmes glacées par individus a été mesurée pendant 30 périodes. L'objectif est déterminé si la consommation dépend de la température. Les données sont dans le ci-dessous.

Effectifs théoriques et khi-carré

consommation y	température x	consommation y	température x	consommation y	température x
386	41	286	28	319	44
374	56	298	26	307	40
393	63	329	32	284	32
425	68	318	40	326	27
406	69	381	55	309	28
344	65	381	63	359	33
327	61	470	72	376	41
288	47	443	72	416	52
269	32	386	67	437	64
256	24	342	60	548	71

Effectifs théoriques et test du khi-carré

$$\sum_{i=i}^{n} y_i = 10783, \qquad \sum_{i=i}^{n} x_i = 1473,$$

$$\sum_{i=i}^{n} y_i^2 = 4001293, \qquad \sum_{i=i}^{n} x_i^2 = 80145,$$

$$\sum_{i=i}^{n} x_i y_i = 553747,$$

Effectifs théoriques et test du khi-carré

1. Donnez les moyennes marginales, les variances marginales et la covariance entre les deux variables.
2. Donnez la droite de régression, avec comme variable dépendante la consommation de glaces et comme variable explicative la température.
3. Donnez la valeur ajustée et le résidu pour la première observation du tableau.