Statistiques descriptives : analyse univariée et bivariée

Juste Raimbault [1]

2025-2026

[1]LaSTIG, IGN-ENSG-UGE

Introduction

Motivation

L'analyse univariée permet de décrire la forme et de quantifier les caractéristiques de la répartition des valeurs d'une variable.

- Notion de distribution
- Visualisation (Histogramme, densité, boxplots, . . .)
- Moments, Quantiles, CV

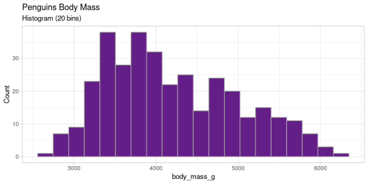
Analyse Univariée

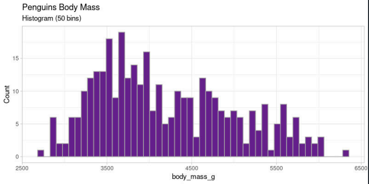
Histogramme

Histogramme d'une variable

Représentation graphique des effectifs associés à des classes de valeurs d'une variable numérique

Le nombre de classes peut varier !





Loi de probabilité

La loi de probabilité peut être définie comme une fonction qui donne la probabilité qu'un individu x pris au hasard soit dans l'ensemble de valeurs V_x pour la variable V :

$$\mathbb{P}_V(V_x) = \mathbb{P}(X \in V_x), \forall V_x \in \Omega_V$$

Avec Ω_V l'ensemble des valeurs que peut prendre V : l'univers de V

Pour une variable aléatoire à valeur réelles

Théorème de Transfert :

$$\mathbb{E}\left[\varphi(V)\right] = \int_{\Omega_V} \varphi(V(\omega))\mathbb{P}(\mathrm{d}\omega) = \int_{\mathbb{R}} \varphi(x)\mathbb{P}_V(\mathrm{d}x)$$

Fonction de répartition :

$$F_V(x) = \mathbb{P}(V \leq x)$$

Densité de probabilité :

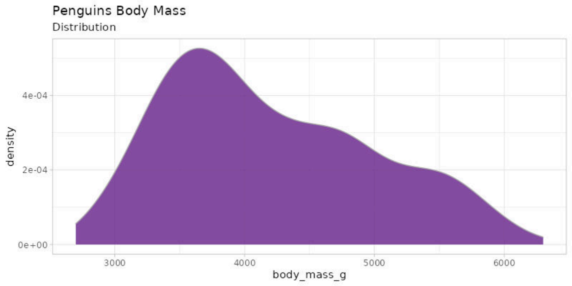
$$F_V(x) = \int_{-\infty}^{x} f_V(t)\mathrm{d}t$$

Distribution

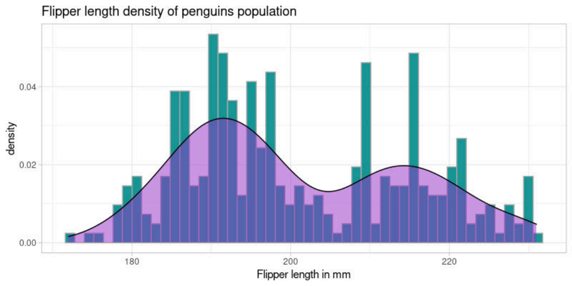
Synonymes: distribution empirique, distribution des fréquences, distribution statistique, densité de probabilité

Tableau ou graphique qui associe les (classes de) valeurs à leur fréquence d'apparition

≈ "Histogramme des fréquences en continu"



Distribution et histogramme



N.B. En toute rigueur, représenter une courbe de distribution de probabilité par dessus un histogramme est impropre : il faudrait deux graphiques distincts, ou au moins deux axes des ordonnées: un pour l'histogramme, représentant un effectif, l'autre pour la distribution , représentant une probabilité.

Parfois , les distributions empiriques ressemblent à celles de densités de probabilités connues.

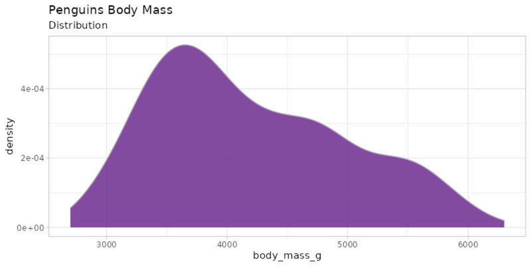
$\rightarrow$ on peut alors modéliser la variable par une variable aléatoire de densité fixée

$\rightarrow$ les paramètres de cette densité doivent être déterminés (par ajustement).

Décrire une distribution

La forme d'une distribution donne beaucoup d'informations :

- "pics" : valeurs les plus représentées dans la population
- présence de valeurs extrêmes : la courbe de la distribution est tirée à gauche ou à droite du graphique
- symétrie : les individus se répartissent équitablement de part et d'autre du pic
- aplatissement : la population est plus ou moins resserrée, ou autour de certaines valeurs
- ...



Décrire une distribution : mesures de tendance centrale

Tendance Centrale

La tendance centrale est une valeur qui résume une série de valeurs (quantitative)

- Moyenne
- Médiane
- Mode

Moyenne

$$\bar{x} = \frac{1}{n}\sum_{i=0}^{n} x_i$$

Avantages

Chaque valeur compte

Inconvénients

- sensibilité aux valeurs extrêmes
- pas de signification sur les valeurs discrètes (e.g. 2.5 enfants par foyer)

Pour y remédier (parfois):
→ exclure les outliers
→ utiliser un autre estimateur (médiane)
→ étudier la distribution des valeurs (e.g. cas bimodal) et opérer une classification

Moyenne géométrique

$$\bar{x}_{geom} = \sqrt[n]{\prod_{i=0}^{n} x_i}$$

Moins sensible que la moyenne classique aux valeurs extrêmes.

Mode

Le mode d'une variable est la valeur la plus fréquente (d'effectif maximum) d'une variable.

Avantages

- peu sensible aux valeurs extrêmes
- interprétation simple : cas le plus fréquent

Inconvénients

Ne dépend pas de toutes les observations (ce qui explique sa robustesse aux valeurs extrêmes)

Calcul du mode

Si la variable est quantitative et continue :

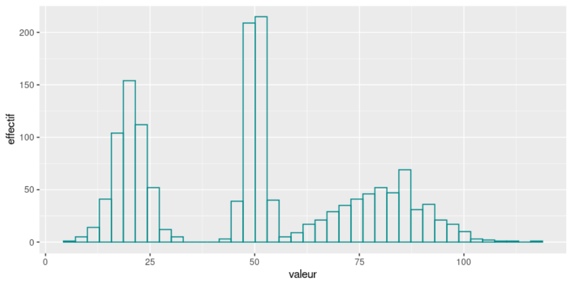
- découper l'étendue de la variable ($max - min$) en intervalle égaux
- compter les effectifs de chaque intervalle
- le mode est la moyenne des valeurs des bornes de l'intervalle de plus grand effectif.

(C'est exactement ce que fait un histogramme graphiquement !)

Modes

Par définition, le mode est unique, mais on peut appeler modes les valeurs des autres pics d'une distribution.
On parle de distribution bi-modale ou tri-modale lorsqu'une distribution présente deux ou trois pics.

Les valeurs modales d'une distribution sont les valeurs correspondant à ces pics.



Médiane

La médiane est la valeur qui sépare une population en deux classes d'égal effectif.
C'est la valeur la plus proche de toutes les autres.

Avantages

- Souvent plus pertinente que la moyenne
- les valeurs extrêmes ne modifient pas sa valeur
- interprétation facile: un individu sur deux a une valeur inférieure (respectivement supérieure) à la médiane.

Inconvénients

Comme le mode , la médiane ne dépend pas de toutes les observations.

N.B. la robustesse de la médiane est bien utile dans le cas de distribution particulièrement asymétriques, où la moyenne est dégradée par les valeurs extrêmes, à droite (valeurs très élevées) ou à gauche (valeurs très faibles).

Moyenne et médiane

Que peut on dire d'une population dont la médiane est inférieure à la moyenne ?

Exemple : revenus mensuels en équivalent temps plein en France en 2016.

Revenu mensuel net moyen 2 238 €
Revenu mensuel net médian 1 789 €

source `https://www.insee.fr/fr/statistiques/4277680?sommaire=4318291`

Un salaire mensuel net équivalent temps plein de 2000€ est-il un bon salaire ?

- 2000€ < moyenne : on peut le considérer comme trop bas pour être «bon»
- 2000€ > médiane : supérieur à (au moins) la moitié des salaires du pays, on peut le considérer comme un «bon» salaire.

Double interprétation & «instinctivement» on imagine une dispersion symmétrique, où la moyenne est proche de la médiane

Décrire une distribution : mesures de dispersion

Dispersion

La dispersion décrit la tendance des valeurs d'une variable à se disperser plus ou moins largement autour des valeurs des tendances centrales.

Variance et écart type

La variance est la somme des écarts carrés à la moyenne rapporté à l'effectif

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

Avec :

- X une variable
- x_i les valeurs de la variables
- $\bar{x}$ la moyenne de X
- n l'effectif

Variance et écart type

L'écart type est la racine carrée de la variance

$$\sigma_X = \sqrt{var_X}$$

Variance et écart-type sont sensibles aux valeurs extrêmes et toujours positifs.

Si $var_X = 0$ ou $\sigma_X = 0$, alors X est constante.
Un écart-type faible indique que les valeurs sont réparties de façon homogène autour de la moyenne.

Quantiles

La médiane sépare une population en deux classes d'égal effectif

Les quantiles séparent une population en n classes d'égal effectif.

Quartiles

Les quartiles d'une population selon une variable X sont trois valeurs, Q_1, Q_2, Q_3 qui séparent la population en quatre classes d'égal effectif.

- 25% des valeurs de X sont strictement inférieures à Q_1
- 50% des valeurs de X sont strictement inférieures à Q_2 (médiane)
- 75% des valeurs de X sont strictement inférieures à Q_3

Déciles

Les déciles sont les 9 quantiles $Q_1, Q_2, \ldots, Q_9$ qui séparent une population 10 classes d'égal effectif.

Écarts inter-quartiles et inter-déciles

Écart inter-quartile: $Q_3 - Q_1$, capture 50% des valeurs de la population les plus proches de la médiane

Écart inter-décile: $Q_9 - Q_1$, capture 80% des valeurs de la population les plus proches de la médiane

Quantiles

Avantages

Peu sensibles aux distributions aplaties et aux valeurs extrêmes

L'écart inter-quantile est plus robuste que l'écart-type

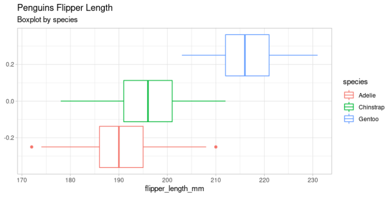
Inconvénients

Parfois délicat pour les variables quantitatives discrètes

Les écarts inter-quantiles négligent l'influence des valeurs extrêmes sur la distribution

Les boîtes à moustaches (boxplots)

Représentation courante de la dispersion d'une variable à l'aide de quartiles



- La marque centrale de la boîte est la médiane
- Les bords de la boîte sont les quartiles Q_1 et Q_3
- Les moustaches vont jusqu'à la plus grande (resp. la plus petite) valeur inférieure (resp. supérieure) à 1.5 fois l'écart interquartile
- Les valeurs qui dépassent les moustaches sont affichées sous formes de points

Le coefficient de variation

Le coefficient de variation (CV) est une autre mesure de dispersion.

C'est le ratio entre l'écart-type σ_x et la moyenne $\bar{x}$ d'une variable quantitative X.

$$CV(X) = \frac{\sigma_x}{\bar{x}}$$

Plus il est important , plus la dispersion est grande.

Plus il est proche de 0, plus les données sont homogènes.

Inconvénients similaires à ceux de $\bar{x}$ et σ_x : sensibilité aux valeurs extrêmes.

Comparaison de dispersion de deux distributions

Exemple : deux communes versent des aides aux entreprises locales, qu'on suppose distribuées suivant une loi normale.

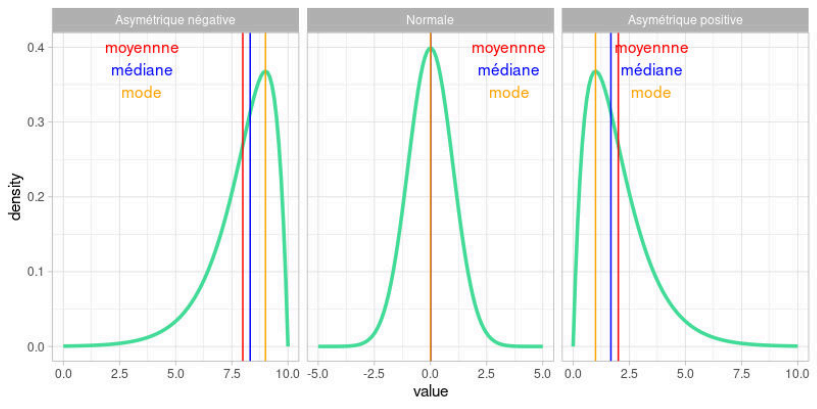
Commune A : moyenne = 390 euros, $\sigma = 30$ euros

Commune B : moyenne = 152 euros, $\sigma = 8$ euros

Pour quelle commune les aides sont les plus homogènes?

Décrire une distribution : asymétrie et aplatissement

Asymétrie (ou skewness)



Coefficients d'asymétrie de Pearson

Deux moyens simples d'estimer l'asymétrie

$$C_1 = \frac{\bar{x} - mode(X)}{\sigma_x}$$

$$C_2 = \frac{3(\bar{x} - mediane(X))}{\sigma_x}$$

- coefficient nul : la distribution est symétrique
- coefficient négatif : la distribution est déformée à gauche de la médiane (sur-représentation de valeurs faibles, à gauche)
- coefficient positif : la distribution est déformée à droite de la médiane (sur-représentation de valeurs fortes, à droite)

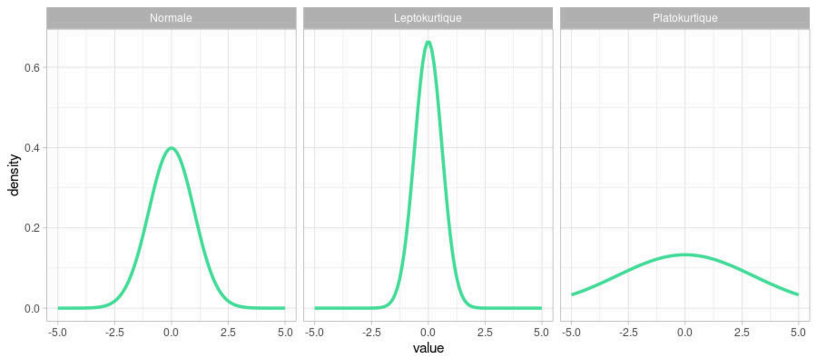
Coefficient d'asymétrie de Fischer

Ce coefficient est le moment d'ordre 3 de la variable X (de moyenne μ et d'écart-type σ) centrée réduite

$$skewness' = \mathbb{E}\left[\left(\frac{X-\mu}{\sigma}\right)^3\right] = \frac{\sum_{i=0}^{n}(x_i - \bar{x})^3}{n\sigma^3}$$

Interprétation similaire aux coefficients de Pearson

Aplatissement (ou Kurtosis)



Courbe piquée: Peu de variation, distribution relativement homogène, beaucoup de valeurs égales ou proches de la moyenne.

Courbe aplatie: Variations importantes, distribution relativement hétérogène, beaucoup de valeurs s'éloignent de la moyenne.

Coefficient de Kurtosis

$$K = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^4}{n\sigma^4}$$

Si la distribution est normale, $K = 3$

Si $K > 3$, la distribution est plus aplatie

Si $K < 3$, la distribution est moins aplatie

On normalise parfois en considérant $K' = K - 3$ (quantifie l'excès d'aplatissement)

Bilan

Pour décrire une distribution :

- Tendance centrale
- Dispersion
- Asymétrie
- Aplatissement

Analyse Bivariée

Analyse bi-variée

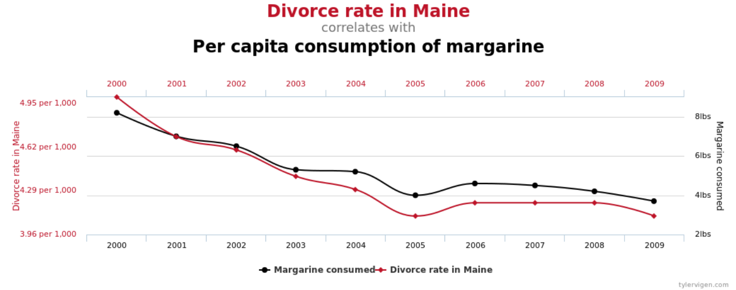
Étude de la relation entre deux variables :

- quantitatives : corrélation, régression linéaire
- qualitatives : test d'indépendance du «Chi deux» / χ^2

Pour le lien entre une variable quantitative et une variable qualitative, on fera simplement un graphique.

Corrélation $\neq$ Causalité

Une liaison, même très forte, entre deux variables, n'indique pas la causalité.



Erreur très courante , très tentante.

Analyse bivariée avec des données spatiales

ENSG Géomatique

Données «spatiales»

Individus restreints spatialement (sélection spatiale)

Variables “géographique” (e.g. lieu de résidence) renseignées pour les individus

Prise en compte des distances → modèle(s) gravitaire(s) (hors programme)

Données localisées: vu par la suite en analyse spatiale

Auto-corrélation spatiale (Moran’s I, Geary Index)

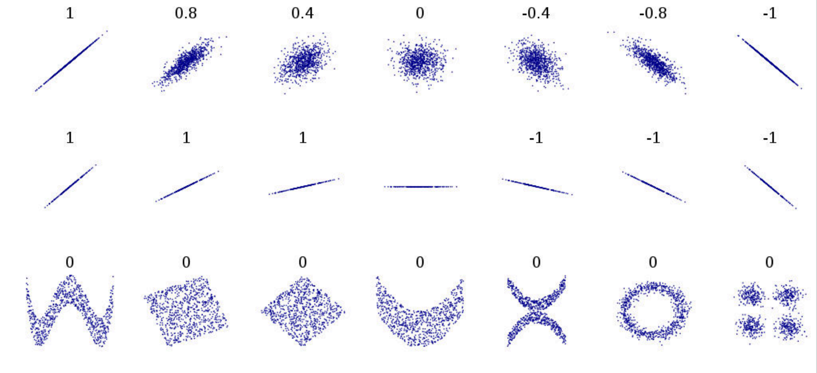
Geographically Weighted Regression (GWR) ≈ régression linéaire avec prise en compte de la distance entre individus

Variogrammes

Corrélation

Avant toute chose

Toujours afficher les données, avant de faire quoi que ce soit.



Corrélation de Pearson

La corrélation indique l'intensité du lien linéaire entre deux variables quantitatives.

$$cor(x, y) \in [-1; 1]$$

- $cor(x, y) \approx 0$: pas de relation (linéaire) entre les deux variables
- $cor(x, y) < 0$: les deux variables ont des sens de variations opposés
- $cor(x, y) > 0$: les deux variables varient conjointement
- $cor(x, y) = \pm 1$: variables parfaitement linéairement (anti-)corrélées, i.e. fonction affine l'une de l'autre.

Formule du coefficient de corrélation de Pearson

$$r = \frac{cov(x,y)}{\sigma_x \sigma_y} = \frac{E\left[(x - E(x))(y - E(y))\right]}{\sigma_x \sigma_y}$$

Avec :

- r (parfois ρ) le coefficient de corrélation
- x et y deux variables quantitatives
- $E(x)$ l'espérance d'une variable x
- σ_x l'écart-type d'une variable x
- $cov(x,y)$ la covariance de deux variables x et y

Corrélation et indépendance

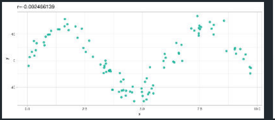
Deux variables indépendantes ont un coefficient de corrélation nul :

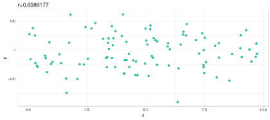
$$x \perp\!\!\!\perp y \implies cor(x, y) = 0$$

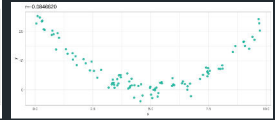
MAIS une corrélation nulle n'implique pas l'indépendance des variables !

$$cor(x, y) = 0 \not\Longrightarrow x \perp\!\!\!\perp y$$

D'autres liaisons sont possibles :







Corrélation avec R

Fonction `cor(x,y)` pour obtenir la valeur du coefficient,

Fonction `cor.test(x,y)` pour obtenir la p-value et l'intervalle de confiance.

Résultat :

```
##
##  Pearson's product-moment correlation
##
## data:  iris$Petal.Length and iris$Petal.Width
## t = 43.387, df = 148, p-value < 2.2e-16
## alternative hypothesis: true correlation
## is not equal to 0
## 95 percent confidence interval:
##  0.9490525 0.9729853
## sample estimates:
##       cor
## 0.9628654
```

R donne le coefficient de Pearson par défaut, l'argument `method` de la fonction cor() permet de spécifier deux autres coefficients : Kendall et Spearman.

Matrice de Corrélation

Fonction `cor()` appliquée à plusieurs variables de type `numeric`

e.g. `cor(iris[,1:4])`

Résultat:

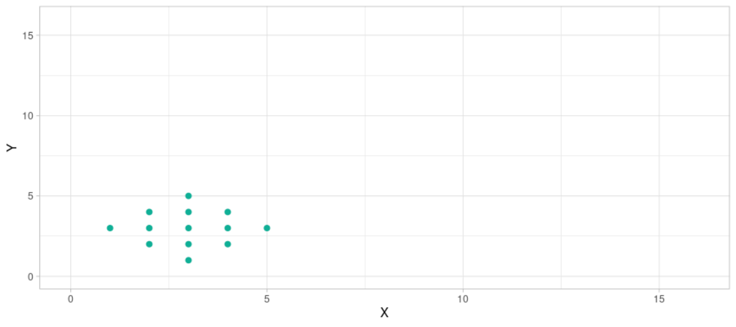
```
##              Sepal.Length Sepal.Width Petal.Length Petal.Width
## Sepal.Length    1.0000000  -0.1175698    0.8717538   0.8179411
## Sepal.Width    -0.1175698   1.0000000   -0.4284401  -0.3661259
## Petal.Length    0.8717538  -0.4284401    1.0000000   0.9628654
## Petal.Width     0.8179411  -0.3661259    0.9628654   1.0000000
```

Présentation des corrélations entre les variables quantitatives d'un tableau, pour tous les couples de variables.

La matrice de corrélation est symétrique, et sa diagonale est constituée de 1.

Sensibilité aux outliers

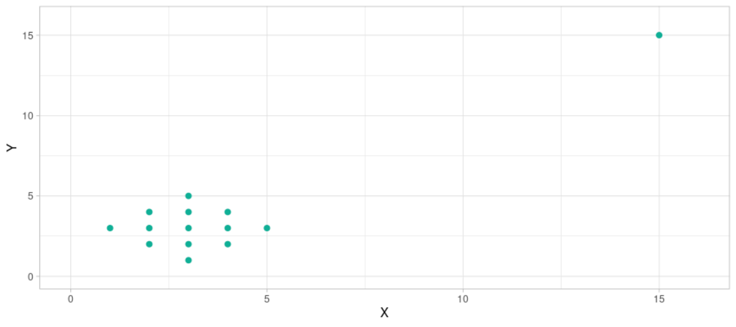
```
X <-  c(3,2,3,4,1,2,3,4,5,2,3,4,3)
Y <-  c(1,2,2,2,3,3,3,3,3,4,4,4,5)
plot(X, Y, xlim = c(0,16), ylim= c(0,16))
```




```
>cor.test(X,Y)$estimate
## cor
##   0
```

Sensibilité aux outliers

```
X <-  c(3,2,3,4,1,2,3,4,5,2,3,4,3,15)
Y <-  c(1,2,2,2,3,3,3,3,3,4,4,4,5,15)
plot(X, Y, xlim = c(0,16), ylim= c(0,16))
```




```
>cor.test(X,Y)$estimate
## cor
##   0.9052224
```

Outliers

Outlier : observation “anormale”, par ses valeurs extrêmes, comparées aux autres.

La corrélation (et la régression linéaire)) sont très sensibles aux outliers.

$\rightarrow$ s’interroger sur la nécessité de nettoyer/filtrer les données et les conséquences

$\rightarrow$ ne pas faire d’épuration brutale et aveugle

Coefficient de corrélation de Spearman

Quand les deux variables semblent corrélées , de façon monotone mais non linéaire,

$\rightarrow$ utiliser le coefficient de Spearman, basé sur le rang des individus.

$$\rho_S = \frac{cov(rg_x, rg_y)}{\sigma_{rg_x}\sigma_{rg_y}}$$

Avec :

- rg_x le rang des individus selon la variable x (en cas d'ex-aequo on prend le rang moyen)
- $cov()$ la fonction de covariance
- σ_{rg_x} l'écart-type du rang rg_x

Régression linéaire

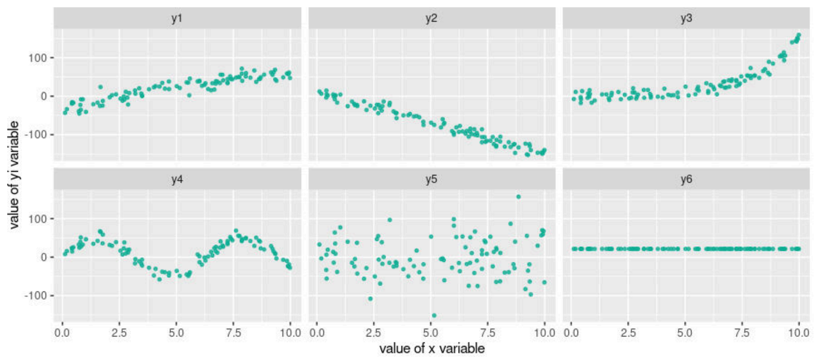
Rappel (encore): Toujours afficher les données, avant de faire quoi que ce soit.

Quand le nuage de points semble «suffisamment» linéaire , on peut tenter de décrire la relation statistique linéaire en proposant un modèle linéaire

$$\hat{y} = \alpha x + \beta$$

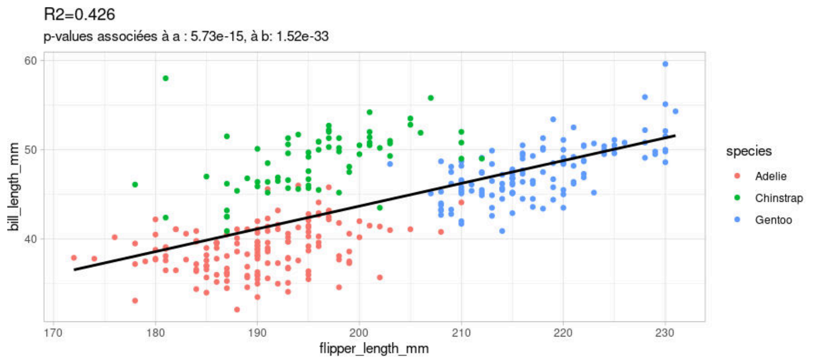
Le modèle retenu doit passer au mieux (i.e. en minimisant une certaine erreur) dans le nuage de points.

Diverses formes de dépendances



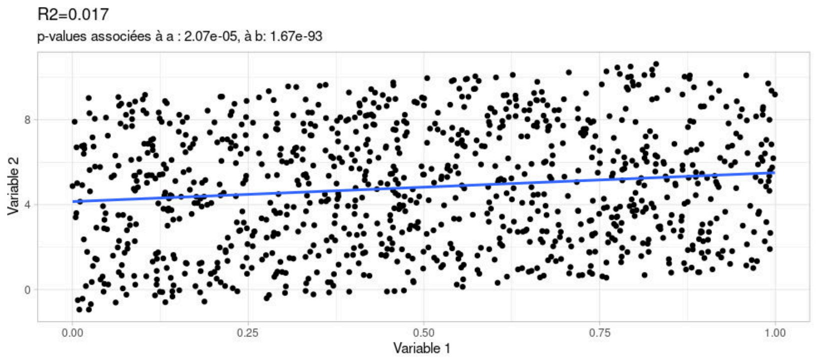
En pratique , les formes sont beaucoup moins régulières.

Exemple



R2=0.426
p-values associées à a : 5.73e-15, à b: 1.52e-33
bill_length_mm
60
50
40
170
180
190
200
210
220
230
flipper_length_mm
species
Adelie
Chinstrap
Gentoo

Exemple


R2=0.017
p-values associées à a : 2.07e-05, à b: 1.67e-93
Variable 2
8
4
0
0.00
0.25
0.50
0.75
1.00
Variable 1

Réaliser une régression linéaire

Si la forme du nuage de points s'y prête, On cherche la droite qui «passe au mieux» (=ajustée) dans le nuage de points de deux variables quantitatives V_1 et V_2. On pourra alors quantifier :

- l'intensité du lien / de la dépendance : points proche de la droite ou non ?
- la forme de la dépendance : linéaire ou non ?
- le sens de la dépendance : nulle, positive ou négative ?

Vocabulaire de la regression

L'équation de la droite est un modèle linéaire de la relation statistique qui lie V_1 et V_2;

Ici le modèle est : $\hat{V}_2 = aV_1 + b$

Si le modèle est retenu, alors pour un individu i dont on connaît V_1, on infère la valeur V_2 par le modèle : $\hat{V}_{2i} = aV_{1i} + b$

On dit aussi que V_1 explique V_2 , ou que le modèle prédit V_2 à partir de V_1 (on note les valeurs prédites $\hat{V}_2$)

Réaliser une régression linéaire avec R

La fonction `lm()` réalise une régression linéaire entre deux (ou plusieurs) vecteurs numériques de même taille.

```
result <- lm(penguins$flipper_length_mm ~ penguins$body_mass_g)
```

L'objet résultat comporte plusieurs attributs, notamment :

- `$coefficients` les coefficients du modèle linéaire
- `$residuals` les résidus

La fonction `summary()` sur l'objet synthétise les résultats

Format des résultats obtenus avec summary

```
Call:
lm(formula = penguins$flipper_length_mm ~ penguins$body_mass_g)

Residuals:
     Min       1Q   Median       3Q      Max
-23.7626  -4.9138   0.9891   5.1166  16.6392

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)          1.367e+02  1.997e+00   68.47   <2e-16 ***
penguins$body_mass_g 1.528e-02  4.668e-04   32.72   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '␣' 1

Residual standard error: 6.913 on 340 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared:  0.759,     Adjusted R-squared:  0.7583
F-statistic:  1071 on 1 and 340 DF,  p-value: < 2.2e-16
```

Qualité d'une regression

Comment évaluer la validité du modèle linéaire ?

Il faut réunir deux conditions :

- des coefficients avec des p-values associées faibles (e.g. <0.05) $\leftrightarrow$ «on a peu de chances de se tromper»
- un R^2 élevé $\leftrightarrow$ «le modèle prédit bien les observations»

Le coefficient de détermination linéaire , noté R^2 décrit la qualité de prédiction de la régression

Il est défini par :

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

- $R^2 \in [0; 1]$
- Plus le R^2 est proche de 1, meilleure est la qualité.

$$R^2 \approx 1 - \frac{residus^2}{variance\ de\ y\ *\ n}$$

qu'on peut lire comme :

$$R^2 \approx 1 - (erreur\ commise\ normalisee\ par\ la\ variation\ de\ la\ variable)$$

également :

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

«la variation expliquée (numérateur) sur la variation totale (dénominateur)»

La p-value peut s'interpréter comme «la probabilité d'avoir un résultat de régression identique avec deux variables véritablement indépendantes»

La p-value est associée à la notion d'hypothèse nulle. Ici , l'hypothèse nulle H_0 est: «les deux séries sont indépendantes».

la p-value est grosso-modo le pourcentage de chances de se tromper en rejetant l'hypothèse nulle,

l'Hypothèse nulle

H_0 : «les deux variables sont indépendantes»

- conserver H_0 : considérer les deux variables comme indépendantes
- rejeter H_0 : considérer les deux variables comme dépendantes i.e. ayant une relation statistique, un lien.

La regression est-elle bonne ?

```
Call:
lm(formula = penguins$flipper_length_mm ~ penguins$body_mass_g)

Residuals:
     Min       1Q   Median       3Q      Max
-23.7626  -4.9138   0.9891   5.1166  16.6392

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)          1.367e+02  1.997e+00   68.47   <2e-16 ***
penguins$body_mass_g 1.528e-02  4.668e-04   32.72   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.913 on 340 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared:  0.759,     Adjusted R-squared:  0.7583
F-statistic:  1071 on 1 and 340 DF,  p-value: < 2.2e-16
```

Critère de significativité du lien sur les résidus

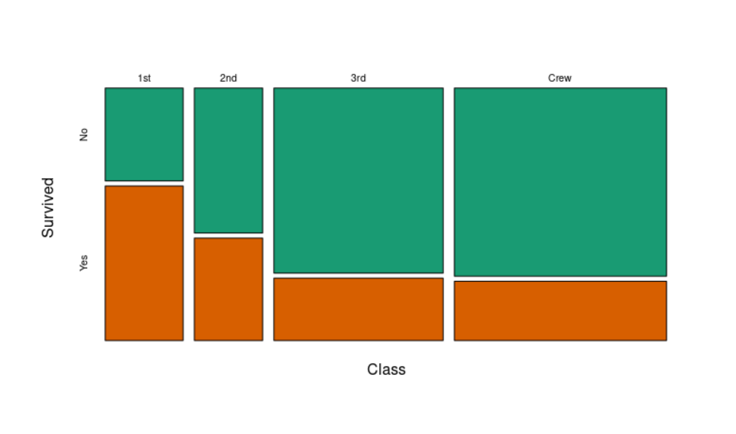
Quand "tout se passe bien" , les résidus ϵ_i doivent:

- être indépendants : covariance nulle ou très faible $cov(x_i, \epsilon_i) = 0$
- être distribués selon un loi normale de moyenne nulle $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon)$
- être distribués de façon homogène (homoscédasticité), i.e. de variance constante $var(\epsilon_i) = \sigma_\epsilon^2$, indépendante de l'observation

Lien entre deux variables qualitatives

Représentation graphique

Pas de nuages de points, ni de droite de régression, mais on peut représenter la table de contingence (fonction `mosaicplot()` de R)



Le test d'indépendance du χ^2 mesure l'écart, la différence, entre deux distributions de variables qualitatives

Il répond à la question : «Existe-t-il un lien statistique entre deux séries de valeurs qualitatives ? »

(La réponse est de type OUI/NON , le χ^2 ne donne pas l'intensité du lien)

- Hypothèse nulle H_0 : les deux distributions sont indépendantes.
- «faire le test» permet de conserver ou de rejeter H_0

Principe du χ^2

- On génère une population théorique à laquelle on va comparer la population observée .
- Cette distribution théorique reflète ce qui se passerait si on suppose que H_0 est vraie
- À l'issue de la comparaison, on pourra rejeter ou conserver H_0.

La construction de cette distribution se fait à partir du tableau de contingence

Tableau de contingence

C'est un tableau à double entrée qui croise deux variables qualitatives.

Dans une case on trouve l'effectif (= le nombre) des individus caractérisés par la conjonction des modalités en ligne et en colonnes.

Exemple sur des formes géométriques de couleurs :

	blanc	*noir*
carré	22	12
rond	10	30
triangle	26	5

Dans R : fonction `table()`.

Construction de la distribution théorique.

On commence par sommer les effectifs selon les modalités (en ligne et en colonne)

	blanc	*noir*	`total`
carré	22	12	34
rond	10	30	40
triangle	26	5	31
`total`	58	47	105

On appelle les sommes en lignes et en colonnes sommes marginales, elles sont mises dans les "marges" du tableau.

Construction de la distribution théorique

En divisant par la taille de la population, on obtient les fréquences observées.

	blanc	*noir*	total
carré	0.20952381	0.11428571	0.3238095
rond	0.09523810	0.28571429	0.3809524
triangle	0.24761905	0.04761905	0.2952381
total	0.552381	0.447619	1

On obtient les pourcentages de l'effectif dans les cases du tableau. C'est également la probabilité , qu'un individu de la population observée soit caractérisé par les modalités en ligne et en colonne.

	blanc	*noir*	total
carré	0.20952381	0.11428571	**0.3238095**
rond	0.09523810	0.28571429	**0.3809524**
triangle	0.24761905	0.04761905	**0.2952381**
total	**0.552381**	**0.447619**	1

De la même façon, les fréquences marginales (marges divisées par la taille de la pop.), donnent la probabilité d'observer un individu de la modalité correspondant à la ligne ou à la colonne considérée.

Exemple : dans cette population , j'ai 29.5% de chances de tirer un triangle, et 55% de chances de tirer une pièce blanche.

Construction de la distribution théorique

Rappel :

Probabilité conjointe de deux évènements A et B indépendants

$$P(A \cap B) = P(A) \times P(B)$$

Si on suppose H_0 (l'indépendance des variables) , on obtient pour chaque couple de modalités, sa probabilité théorique, par le produit des fréquences marginales

Exemple : Si H_0 est vraie, la probabilité d'observer un triangle noir est donnée par:

$P(triangle \cap noir) = P(triangle) \times P(noir)$

$P(triangle \cap noir) = 0.447619 \times 0.2952381 = 0.1321542$

La probabilité théorique d'observer un triangle noir est de 13,2%

Construction de la distribution théorique

On crée un second tableau, dont chaque case vaut le produit des fréquences marginales calculées sur le tableau des observations.

	blanc	*noir*	total
carré	0.1788662	0.1449433	**0.3238095**
rond	0.2104309	0.1705215	**0.3809524**
triangle	0.1630839	0.1321542	**0.2952381**
total	**0.552381**	**0.447619**	1

C'est le tableau des fréquences théoriques.

On l'obtient en multipliant les fréquences théoriques par la taille de la population observée (ici 105)

	blanc	*noir*
carré	18.78095	15.21905
rond	22.09524	17.90476
triangle	17.12381	13.87619

N.B. Il n'est pas nécessaire d'arrondir les effectifs théoriques

Calcul du χ^2

C'est la somme, pour chaque couple de modalités, des écarts carrés entre effectif observé et effectif théorique, divisés par l'effectif théorique.

Soient T^{obs} le tableau des effectifs observés, T^{theo} le tableau des effectifs théoriques,

$$\chi^2 = \sum_{i,j} \frac{(T_{i,j}^{obs} - T_{i,j}^{theo})^2}{T_{i,j}^{obs}}$$

Dans notre exemple : $\chi^2 = 26.30329$

Interprétation du χ^2

on compare la valeur du χ^2 calculée avec la valeur critique qu'on trouve dans une table de loi de Student.

C'est un tableau à double entrée : une valeur de quantile, et un degré de liberté.

La valeur de quantile est le pourcentage d'erreur qu'on s'autorise. On prend souvent 5% : ou, en fonction des tables , la valeur 1-0.05 = 0.95

Le degré de liberté est obtenu en calculant la valeur $(nb\_lignes - 1) * (nb\_colonnes - 1)$.

Dans notre exemple , le degré de liberté est 2*1 = 2

Interprétation du χ^2

En se référant au valeurs de référence de la loi de Student , la valeur critique pour un test avec 5% de chances de se tromper est un degré de liberté de 2 vaut 4.303.

Si la valeur calculée du χ^2 est supérieure à la valeur critique, on rejette H_0.

Pour notre exemple: On rejette H_0, i.e. les deux variables sont dépendantes, car $\chi^2 \approx 26 > 4.303$

Interprétation: «la forme est liée à la couleur dans cette population, nous pouvons l'affirmer avec un risque d'erreur de 5%»

Les étapes du χ^2

1. Tableau de contingence
2. Sommes marginales
3. Calcul des fréquences observées
4. Calcul des fréquences théoriques
5. Tableau d'effectifs théoriques
6. Calcul de la valeur du test
7. Comparaison avec les valeurs de la table de Student

Fonction `chisq.text()`

R calcule pour nous une p-value du test de Student associé au χ^2
$\rightarrow$ on utilise la p-value directement pour conserver ou non H_0 si elle est inférieur à la valeur de rejet désirée (e.g. 5%).

```
>chisq.test(penguins$species, penguins$island)
#
#       Pearson's Chi-squared test
#
#data:  penguins$species and penguins$island
#X-squared = 299.55, df = 4, p-value < 2.2e-16
```

Plusieurs attributs de l'objet résultat du test

```
#>mon_resultat <- chisq.test(X, Y)
```

- `mon_resultat$statistic` : la statistique du Chi2.
- `mon_resultat$parameter`: le nombre de degrés de libertés.
- `mon_resultat$p.value` : la p-value.
- `mon_resultat$observed`: la distribution observée
- `mon_resultat$expected` : la distribution théorique