Chapitre 1 : Rappels de statistique descriptive

Auteur : Sandrine Casanova

Ce document constitue des notes de cours illustrées sur deux jeux de données (fichier "Régions" et fichier "Employés"). Chaque concept est complété par un exemple qui contient des commentaires de sorties obtenues avec le logiciel R (package Rcmdr). Ces sorties sont repérées par des numéros et se trouvent à la fin du polycopié.

1 Rappels sur les notions d'individus et de variables en statistique

- individu (ou observation) : entité de base en statistique (Ex : étudiants, entreprises, régions, années,...)
- population : ensemble des individus
- effectif total : nombre d'individus de la population (noté n)
- variable : caractéristique mesurée sur des individus (Ex: âge, nombre de salariés, PNB, région,...).
- Modalités = valeurs observées de la variable

Nature de la variable : on distingue 4 types de variable

- variable quantitative discrète = variable numérique à peu de modalités

  exemple : nombre d'enfants
- variable quantitative continue = variable numérique à beaucoup de modalités

  exemple : salaire mensuel en euros, taille en cm
- variable qualitative nominale= variable non numérique, pas d'ordre dans les modalités

  exemple sexe (2 modalités : homme, femme), catégorie socio-professionnelle
- variable qualitative ordinale= variable non numérique, ordre dans les modalités

  exemple : finition d'un produit (3 modalités : moyen, bien, très bien)

2 Tableaux de données

2.1 Tableau de données brutes

Il se présente sous la forme d'un tableau individus / variables.

une ligne = un individu, une colonne = une variable.

On considère n individus et p variables.

Var Ind	Sexe	Âge	$\dots$	X^j	$\dots$	X^p
1						
2	H	25				
$\vdots$						
X_i				X_i^j		
$\vdots$						
X_n						

- tableau noté X,
- individu i noté X_i : $i$ème ligne transposée du tableau, $X_i \in \mathbb{R}^p$(vecteur à p coordonnées)

$$\mathbb{R}^p = \textbf{Espace des individus}$$

- variable j notée X^j : $j$ème colonne du tableau
- $X^j \in \mathbb{R}^n$(vecteur à n coordonnées)

$$\mathbb{R}^n = \textbf{Espace des variables}$$

Exemple 1 *Tableau de données des régions : Voir tableau 1*

Population : 21 régions françaises

Ce fichier contient uniquement des variables quantitatives :

- *POPUL : population de la région (en milliers d'individus)*
- *TACT : taux d'activité (population active / population totale de la région) en pourcentage*
- *SUPERF : superficie de la région (en kilomètres carrés)*
- *NENTR : nombre d'entreprises de la région*
- *NBBREV : nombre de brevets déposés au cours de l'année*
- *CHOM : taux de chômage (en pourcentage)*
- *TELEPH : nombre de lignes téléphoniques en place dans la région (en milliers)*

2.2 Tableau de contingence

- Forme : *variable / variable.*
- Tableau variable/variable avec 2 variables qualitatives X et Y à resp. L et C modalités notées X_1,..., X_L, Y_1,..., Y_C.

Y \ X	Y_1	$\ldots$	Y_j	$\ldots$	Y_C	Total $n_{i.}$
X_1						
$\vdots$						
X_i			n_{ij}			
$\vdots$						
X_L						
$n_{.j}$						n

- n_{ij} : effectif conjoint
- $n_{i.} = \sum_{j=1}^{C} n_{ij}$ et $n_{.j} = \sum_{i=1}^{L} n_{ij}$

Fichier 2 : Employés

Il concerne 474 employés d'une banque américaine et contient, entre autre, 3 variables qualitatives :

- Sexe (Féminin, Masculin)
- Stat-pro : statut profesionnel (3 modalités : employé de bureau, agent de sécurité, manager)
- National : nationalité (2 modalités : américain ou non américain)

Exemple 2 *Tableau de contingence croisant le sexe et le statut professionnel : voir tableau 6*

3 Analyses statistiques pour une variable (ou univariées)

- X variable étudiée sur une population d'effectif n,
- x_i modalité de X pour le ième individu de la population.
- Distribution d'une variable : répartition de la population selon les modalités de la variable

3.1 Calculs sur la distribution

Ces calculs dépendent du type de la variable

3.1.1 Variable qualitative

mode ou calculs sur effectifs (tri à plat ou tableau de distribution): on suppose que X prend K modalités distinctes.

Défitions :

- n_k effectif associé à la modalité x_k= nombre d'individus de la population pour lesquels X prend cette modalité
- f_k fréquence associée à la modalité x_k= proportion d'individus de la population pour lesquels X prend cette modalité

Sous Rcmdr : Statistiques → Résumés → Distributions de fréquences

Exemple 3 *Fichier "Employés" : Tableaux de tri à plat de chaque variable qualitative : voir tableau 5*

3.1.2 Variable quantitative

On définit des résumés numériques (indicateurs) de X

1. Indicateurs de position
   Ils donnent une idée globale de l'ordre de grandeur de la variable et s'expriment dans l'unité de la variable.

   a) Paramètres de tendance centrale

      * Moyenne

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

      Propriétés de la moyenne

      (a) $\sum_{i=1}^{n}(x_i - \overline{x}) = 0$

      (b) $\overline{x}$ est le réel a qui minimise $\sum_{i=1}^{n}(x_i - a)^2$

      (c) si $x_i = a \forall i = 1, \cdots, n$ alors $\overline{x} = a$

      (d) si $y_i = ax_i \forall i = 1, \cdots, n$ alors $\overline{y} = a\overline{x}$

      (e) si $y_i = a + x_i \forall i = 1, \cdots, n$ alors $\overline{y} = a + \overline{x}$

      (f) si $z_i = x_i + y_i \forall i = 1, \cdots, n$ alors $\overline{z} = \overline{y} + \overline{x}$

      L'inconvénient de $\bar{x}$ est qu'elle est sensible aux valeurs extrêmes. On dit que $\bar{x}$ n'est pas robuste.

* Médiane
C'est la valeur de X, notée Me qui partage l'effectif en 2. On ordonne les modalités de X. On note $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$ les modalités ordonnés de X.
$\hookrightarrow$ 2 cas :
- si n impair, Me$=x_{(\frac{n+1}{2})}$
- si n pair, Me$=\frac{x_{(n/2)}+x_{(n/2+1)}}{2}$
* Mode
$=$ valeur (non nécéssairement unique) de X la plus représentée dans la population

b) Autres
* Maximum $=$ valeur maximum de X
* Minimum $=$ valeur minimum de X
* Quantiles
Ils généralisent la notion de médiane
Définition : pour $\alpha \in [0,1]$, le quantile d'ordre α est la valeur q_α de X tel que $\alpha \times 100\%$ de la population ait une caractéristique inférieure ou égale à q_α.
Cas particuliers :
- Me$=q_{0.5}$
- $Q_1 = q_{0.25}=$ 1er quartile$=$ valeur de X tel que 25% de la population ait une caractéristique inférieure ou égale à Q_1. (et 75% ait une caractéristique au dessus)
- $Q_3 = q_{0.75}=$ 1er quartile$=$ valeur de X tel que 75% de la population ait une caractéristique inférieure ou égale à Q_1. (et 25% ait une caractéristique au dessus)

2. Indicateurs de dispersion

a) Autour de la moyenne
* Variance
C'est la moyenne des carrés des écarts à la moyenne.
$$V(X) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$$
La variance s'exprime dans le carré de l'unité de la variable. Elle permet de comparer la dispersion de variables qui ont la même moyenne.
Propriétés de la variance
(a) $\forall a \in R, \sum_{i=1}^{n}(x_i - a)^2 = \sum_{i=1}^{n}(x_i - \overline{x})^2 + n(\overline{x} - a)^2$
(b) $V(X) = \overline{x^2} - \overline{x}^2$

(c) si $x_i = a \forall i = 1, \cdots, n$ alors $V(X) = 0$

(d) si $y_i = ax_i \forall i = 1, \cdots, n$ alors $V(Y) = a^2 V(X)$

(e) si $y_i = a + x_i \forall i = 1, \cdots, n$ alors $V(Y) = V(X)$

* Ecart-type

$$\sigma_X = \sqrt{V(X)}$$

σ_X s'exprime dans l'unité de X

* Coefficient de variation

$$CV(X) = \frac{\sigma_X}{\overline{x}}$$

$CV(X)$ n'a pas d'unité. Il permet de comparer la dispersion de 2 variables de moyennes différentes. La variable la plus dispersée est celle qui a le plus grand coefficient de variation.

De plus, pour mesurer la dispersion d'une variable, on compare $CV(X)$ à 1.

Règle :

$\hookrightarrow$ Si $CV(X)$»1 alors σ_X»$\overline{x}$: série très dispersée

$\hookrightarrow$ Si $CV(X)$«1 alors σ_X«$\overline{x}$: série peu dispersée

b) Autour de la médiane : Ecart inter-quartile I

$$I = Q_3 - Q_1$$

I s'exprime dans l'unité de X. On peut diviser I par Me pour avoir un indicateur sans unité. Dans l'intervalle $[Q_1, Q_3]$, il y 50% des observations (celles autour de la médiane).

c) Autres : Etendue ("range" en anglais)

Etendue=$x_{max} - x_{min}$

Elle s'exprime dans l'unité de X

3. Indicateurs de forme

a) Coefficient d'asymétrie : coefficient d'asymétrie de Fisher ("skewness" en anglais)

$$\gamma_1 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^3}{\sigma_X^3}$$

Pour une distribution symétrique, $\sum_{i=1}^{n}(x_i - \overline{x})^3 = 0$, donc $\gamma_1 = 0$

Si $\gamma_1 > 0$, alors la distribution est asymétrique et étalée à droite. Si $\gamma_1 < 0$, alors la distribution est asymétrique et étalée à gauche.

b) Coefficient d'aplatissement : coefficient d'aplatissement de Fisher ("kurtosis" en anglais)

$$\gamma_2 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^4}{\sigma_X^4} - 3$$

Propriétes :

- $\gamma_2 > -2$
- γ_2 mesure l'importance des queues de distribution. Si $\gamma_2 > 0$, alors il y a des queues de distribution (des valeurs extrêmes) et la distribution est peu aplatie.

Exemple 4 *Fichier "Régions" : voir tableau 2*

3.2 Représentations graphiques

Pour représenter graphiquement la distribution d'une variable, il faut au préalable avoir défini son type (variable qualitative ou variable quantitative discrète ou variable continue)

3.2.1 Variable qualitative

- Diagramme en barres

  Exemple 5 *Fichier "Employés" : voir graphique 5*

- Diagramme en secteurs (ou circulaire) ou "camembert"

  Exemple 6 *Fichier "Employés" : voir graphique 6*

3.2.2 Variable quantitative

• Variable discrète
$\hookrightarrow$ Diagramme en bâtons :

- en abscisse : les modalités
- en ordonnée : les effectifs associés (ou les fréquences associées)

• Variable continue
= beaucoup de modalités
Pour simplifier, on regroupe ces modalités dans des classes ou intervalles ce qui entraîne une perte d'information
Notations

- K classes (K arbitraire) de la forme $[b_{k-1}, b_k[$
- n_k= effectif associé à la $k$ième classe = nombre d'individus pour lesquels $X \in [b_{k-1}, b_k[$
- f_k= fréquence associée à la $k$ième classe = proportion d'individus pour lesquels $X \in [b_{k-1}, b_k[$
- la classe $[b_{k-1}, b_k[$ est caractérisée par son amplitude $A_k = b_k - b_{k-1}$ son centre $c_k = \dfrac{b_k + b_{k-1}}{2}$ et sa densité $d_k = \dfrac{n_k}{A_k}$ (ou $d_k = \dfrac{f_k}{A_k}$)

$\hookrightarrow$ Histogramme :

- en abscisse : les classes
- en ordonnée : les densités associées

Si les classes sont de même amplitude, il est équivalent de représenter les effectifs ou les fréquences en ordonnée.
Lorsque l'histogramme est représenté en densité, l'aire d'un rectanngle correspond à un effectif ou à une fréquence.

Exemple 7 *Fichier "Régions" : voir graphique 1*

$\hookrightarrow$ Boîte à moustaches (boxplot en anglais)
5 paramètres de position représentés :

- Max(Minimum observé, Minimum théorique = $Q_1 - 1.5(Q_3 - Q_1)$)
- 1er quartile
- Médiane
- 3ème quartile
- Min (Maximum observé, Maximum théorique =$Q_3 + 1.5(Q_3 - Q_1)$)

Les observations dont la caractéristique est supérieure au maximum théorique ou inférieure au minimum théorique sont appelées observations atypiques ou abérrantes et sont représentées sur la boîte à moustaches.

Exemple 8 *Fichier "Régions" : voir graphique 2*

4 Analyses statistiques pour deux variables (bivariées)

Objectif : étude de la liaison entre ces variables
Exemples de problématiques : le salaire à l'embauche dépend-il du salaire courant, le sexe et le statut professionnel sont-ils liés, le salaire moyen diffère-t-il suivant le sexe dans une entreprise ?

4.1 Graphiques

- 2 variables quantitatives : nuage de points (en abscisse variable explicative, en ordonnée variable à expliquer),

  Exemple 9 *Fichier "Régions" : Voir graphiques 3 et 4*

- 2 variables qualitatives : diagramme en barres juxtaposées,

Exemple 10 *Fichier "Employés" : Voir graphique 7*

- 1 variable quantitative, 1 variable qualitative : boîtes à moustaches juxtaposées

Exemple 11 *Fichier "Employés" : on dispose en outre de la variable salaire. Voir graphique 8*

4.2 Indicateurs numériques

4.2.1 Liaison entre 2 variables quantitatives

* Covariance

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

La covariance mesure le sens de la liaison entre 2 variables

- $cov(X,Y) = 0$: X et Y non corrélées linéairement
- $cov(X,Y) > 0$: X et Y varient dans le même sens
- $cov(X,Y) < 0$: X et Y varient dans le sens opposé

* Coefficient de corrélation linéaire

$$r(X,Y) = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$$

$r(X,Y) \in [-1, 1]$

- $r = 0$: X et Y non corrélées linéairement
- r proche de 0 : X et Y faiblement corrélées linéairement
- $|r|$ proche de 1 : X et Y fortement corrélées linéairement
- $|r| = 1$: liaison linéaire exacte entre X et Y

Exemple 12 *Fichier "Régions" : voir tableau 3*

* Test du coefficient de corrélation linéaire de Pearson à 0 (statistique inférentielle)

  X et Y sont deux variables aléatoires que l'on suppose de loi normale.

  Coefficient de corrélation linéaire de Pearson : $\rho = \dfrac{cov(X,Y)}{Var(X)Var(Y)}$

  avec $Cov(X,Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$,$Var(X) = \mathbb{E}(X - \mathbb{E}(X))^2$ et $Var(Y) = \mathbb{E}(Y - \mathbb{E}(Y))^2$

  Hypothèse nulle : H0 : $\rho = 0$ (absence de liaison linéaire entre X et Y)

  Sous H0, $\sqrt{n-2}\dfrac{R}{\sqrt{1-R^2}} \sim T(n-2)$

  (loi de Student de paramètre $n-2$)

  où R est le coefficient de corrélation linéaire empirique de X et Y , calculé sur un échantillon de taille n. R est une variable aléatoire de réalisation r .

  Règle de décision : Rejet H0 si $|tobs| > t_{\alpha/2}(n-2)$ avec $\alpha=$ probabilité de rejeter H0 alors qu'elle serait vraie = niveau du test

  Quasi-universellement $\alpha = 5\%$

  ou Utilisation du "petit p" donné par le logiciel

  p = niveau de signification empirique = p-value

  C'est la probabilité, sous H0, d'observer une valeur de la statistique de test plus éloignée de H0 que celle qu'on a effectivement observée.

  Règle d'utilisation :

  • si $p < \alpha$ on rejette H0,

  • si $p > \alpha$, on accepte H0.

  Sous Rcmdr : Statistiques → Résumés → Test de corrélation

  Exemple 13 *Fichier "Régions" : voir tableau 4*

4.2.2 Liaison entre 2 variables qualitatives

On représente la distribution jointe par le tableau de contingence (=tri croisé)
A partir de ce tableau on calcule :

* Profils-lignes : vecteurs des distributions en fréquence de Y dans les sous-populations définis par une modalité fixée de X. Il y a L profils-lignes qui ont chacun C composantes.

* Profils-colonnes : vecteurs des distributions en fréquence de X dans les sous-populations définis par une modalité fixée de Y. Il y a C profils-colonnes qui ont chacun L composantes.

Exemple 14 *Fichier "Employés" : voir tableaux 7 et 8*

Exemples de commentaires : 95.4% *des femmes sont employées de bureau*

56.7% *des employés de bureau sont des femmes*

* Les effectifs attendus dans chaque cellule si indépendance de X et Y : $e_{ij} = \dfrac{n_{i.}n_{.j}}{n}$

Exemple 15 *Fichier "Employés" : voir tableau 10*

La statistique du χ^2 pour mesurer l'écart à l'indépendance : $\chi^2 = \displaystyle\sum_{i=1}^{L}\sum_{j=1}^{C}\frac{(n_{ij}-e_{ij})^2}{e_{ij}}$

* **Proposition 1** $\chi^2 = 0 \Leftrightarrow$ *variables indépendantes (la distribution en fréquence de la première variable ne dépend pas de la sous-population définie par une modalité fixée de l'autre variable)*

On a l'inégalité suivante :

$$\frac{\chi^2}{n} \leq \min\{L-1, C-1\}$$

Définition 1 *Le coefficient C de Cramer est défini par :*

$$C = \sqrt{\frac{\chi^2}{n\min\{L-1, C-1\}}}$$

Proposition 2 $C \in [0,1]$

$C = 0$ *lorsqu'il y a indépendance.*

C *proche de 0 signifie faible liaison entre X et Y*

C *proche de 1 signifie forte liaison entre X et Y*

Remarque 1 *En pratique, le calcul du χ^2 a un sens lorsque les effectifs théoriques sont supérieurs ou égaux à 5.*

- Test d'indépendance du χ^2

  Soit un échantillon aléatoire de n observations prélevées dans une population définie selon deux variables aléatoires qualitatives X et Y avec respectivement L et C réalisations.

  1. La distance d'indépendance du Khi-deux est la variable aléatoire $\chi^2 = \displaystyle\sum_{i=1}^{L}\sum_{j=1}^{C}\frac{(N_{ij}-N_{i.}N_{.j}/n)^2}{N_{i.}N_{.j}/n}$

2. Si les variables X et Y sont indépendantes et si les réalisations des effectifs théoriques sont >5 alors χ^2 suit approximativement (pour n assez grand) une loi de Khi-deux de paramètre $k = (L-1)(C-1)$.
3. La réalisation de χ^2 est $\chi^2_{obs} = \displaystyle\sum_{i=1}^{L}\sum_{j=1}^{C} \frac{(n_{ij} - n_{i.}n_{.j}/n)^2}{n_{i.}n_{.j}/n}$

D'où la règle de décision : Rejet H0 si $\chi^2_{obs} > \chi^2_{\alpha}(k)$

ou utilisation du p fourni par le logiciel

Sous Rcmdr : Statistiques $\rightarrow$ Tables de contingence $\rightarrow$ Tableau à double entrée

Exemple 16 *Fichier "Employés" : voir tableau 9*

$p < 2.2e-16 < 5\%$ *donc on rejette H0 : le sexe et le statut professionnel sont liés.*

S'il y a liaison alors

* Étude des contributions : contribution de la cellule (i, j)

$ct(i,j) = \dfrac{(n_{ij} - e_{ij})^2}{e_{ij}}$,

Exemple 17 *Fichier "Employés" : voir tableau 11*

On repère les "fortes contributions" (couples de modalités pour lesquels l'effectif observé est très éloigné de l'effectif attendu).

Dans cet exemple, les fortes contributions sont obtenues pour les couples (Femme, Manager) et (Homme, Manager).

* Signe du résidu $= n_{ij} - n_{i.}n_{.j}/n$ pour les associations (positif sur-représentation de la cellule (i, j) par rapport à la représentation si indépendance, négatif sous-représentation)

Exemple 18 *Fichier "Employés"*

Pour les fortes contributions, on compare effectif observé et effectif attendu.

Pour le couple (Femme, Manager), obs<attendu donc il y a sous-représentation des femmes manager.

Pour le couple (Homme, Manager), obs>attendu donc il y a sur-représentation des hommes manager.

4.2.3 Liaison entre une variable qualitative X (facteur) et une variable quantitative Y

On note $\bar{Y}_1, \cdots, \bar{Y}_K$ les moyennes de Y dans chacune des classes (groupes) définies par les K modalités (niveaux) de X, d'effectifs respectifs $n_1, \cdots, n_K$. On note $\sigma_1^2, \cdots, \sigma_K^2$ les variances de Y dans chacun des groupes

Variance-inter$=\frac{1}{n}\sum_{k=1}^{K} n_k(\bar{Y}_k - \bar{Y})^2=$ mesure de la dispersion de Y "entre les classes" (dispersion due au facteur)

Variance-intra$=\frac{1}{n}\sum_{k=1}^{K} n_k\sigma_k^2=$ mesure de la dispersion de Y "à l'intérieur des classes" (dispersion due au hasard)

Équation d'analyse de la variance :
$V(Y)=$ Variance-intra + Variance-inter

Rapport de corrélation=Variance-inter/V(Y)

• Le rapport de corrélation $\in [0, 1]$

• rapport proche de 1 : les moyennes conditionnelles de Y sont très différentes suivant les niveaux de X (effet du facteur X sur Y)

• rapport proche de 0 : les moyennes conditionnelles de Y diffèrent très peu suivant les niveaux de X (pas d'effet du facteur X sur Y)

Exemple 19 *Fichier "Employés" : voir tableau 12*
Le rapport de corrélation est égal à 0.6485, assez proche de 1.
Donc le salaire dépend du statut professionnel.

Notes de cours : Statistique descriptive sur le fichier des régions avec Rcmdr

Fichier 1 : Régions

Tableau 1

NOM	REGION	POPUL	TACT	SUPERF	NBENTR	NBBREV	CHOM	TELEPH
A	Alsace	1624	39.14	8280	35976	241	5.2	700
Q	Aquitain	2795	36.62	41308	85531	256	10.2	1300
U	Auvergne	1320	37.48	26013	40494	129	9.3	600
N	Bas-Norm	1390	38.63	17589	35888	91	9.0	600
O	Bourgogn	1600	38.26	31582	40714	223	8.1	750
B	Bretagne	2795	36.62	27208	73763	296	9.5	1300
C	Centre	2370	38.78	39151	56753	229	7.9	1100
E	Champ-Ar	1340	37.85	25606	24060	155	9.3	550
F	Fr-Comte	1090	37.27	16202	27481	159	7.1	450
H	Hte-Norm	1730	37.80	12317	37461	181	10.8	750
I	Ile-de-F	10660	46.04	12012	273604	6722	7.3	5800
G	Lang-Rou	2110	32.12	27376	62202	179	13.2	1000
S	Limousin	720	38.06	16942	21721	73	7.9	350
L	Lorraine	2300	34.34	23547	48353	185	8.6	950
M	Midi-Pyr	2430	37.14	45348	78771	237	9.0	1100
P	Nord-PdC	3960	32.05	12414	78504	278	12.6	1600
Y	Pays-Loi	3060	37.93	32082	72027	339	9.6	1300
D	Picardie	1810	34.39	19399	36285	139	9.8	750
T	Poit-Cha	1590	36.82	25809	44598	133	10.1	750
Z	Pr-Cte-A	4260	34.96	31400	132552	610	11.0	2300
R	Rh-Alpes	5350	39.44	48698	159634	1474	7.4	2500

Tableau 2

	mean	sd	0%	25%	50%	75%	100%	n
CHOM	9.18	1.845071	5.20	7.90	9.30	10.10	13.20	21
NBBREV	587.09	1436.473700	73.00	155.00	223.00	278.00	6722.00	21
NBENTR	69827.23	58161.016814	21721.00	36285.00	48353.00	78504.00	273604.00	21
POPUL	2681.14	2151.172454	720.00	1590.00	2110.00	2795.00	10660.00	21
SUPERF	25727.76	11348.954775	8280.00	16942.00	25809.00	31582.00	48698.00	21
TACT	37.22	2.906537	32.05	36.62	37.48	38.26	46.04	21
TELEPH	1261.90	1178.548340	350.00	700.00	950.00	1300.00	5800.00	21

	IQR	cv	skewness	kurtosis
CHOM	2.20	0.20086306	0.2143556	0.2444974
NBBREV	123.00	2.44674732	3.9827664	14.4800815
NBENTR	42219.00	0.83292736	2.3247451	5.3732296
POPUL	1205.00	0.80233414	2.6188804	7.1712356
SUPERF	14640.00	0.44111706	0.4034929	- 0.6775245
TACT	1.64	0.07807874	0.7949018	2.6224746
TELEPH	600.00	0.93394397	2.9266909	8.6961389

Tableau 3

	CHOM	NBBREV	NBENTR	POPUL	SUPERF	TACT
CHOM	1.00000000	-0.2565763	-0.07804957	-0.07313003	0.062058491	-0.69854149
NBBREV	-0.25657627	1.0000000	0.89160714	0.92137414	-0.163957955	0.70845007
NBENTR	-0.07804957	0.8916071	1.00000000	0.98101936	0.149291848	0.51571338
POPUL	-0.07313003	0.9213741	0.98101936	1.00000000	0.024369703	0.51376438
SUPERF	0.06205849	-0.1639580	0.14929185	0.02436970	1.000000000	-0.05925506
TACT	-0.69854149	0.7084501	0.51571338	0.51376438	-0.059255061	1.00000000
TELEPH	-0.09833108	0.9444463	0.98290899	0.99391186	0.004764791	0.55526402

	TELEPH
CHOM	-0.098331084
NBBREV	0.944446274
NBENTR	0.982908993
POPUL	0.993911864
SUPERF	0.004764791
TACT	0.555264016
TELEPH	1.000000000

Tableau 4

Pearson's product-moment correlation

data: regions$NBBREV and regions$NBENTR
t = 8.5829, df = 19, p-value = 5.806e-08
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.7477104 0.9555193
sample estimates:
      cor
0.8916071

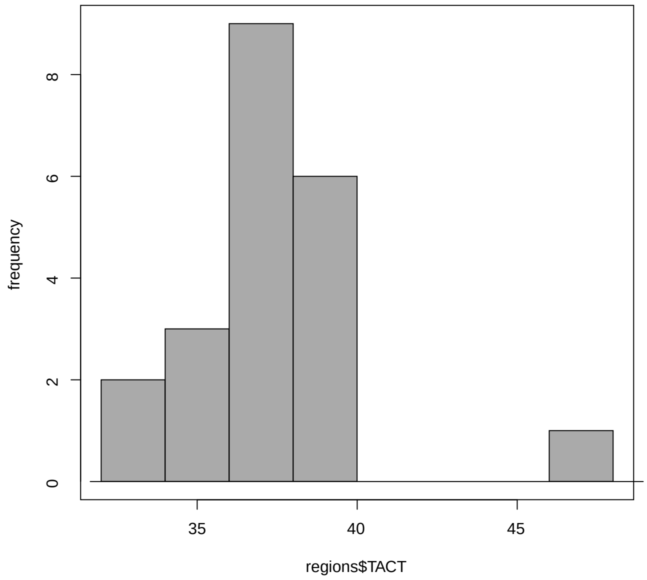


FIGURE 1 – Histogramme de la variable TACT

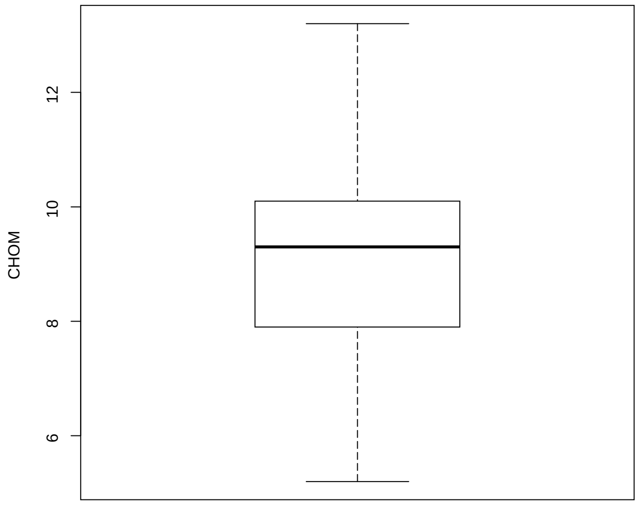


FIGURE 2 – Boîte à moustaches de la variable CHOM

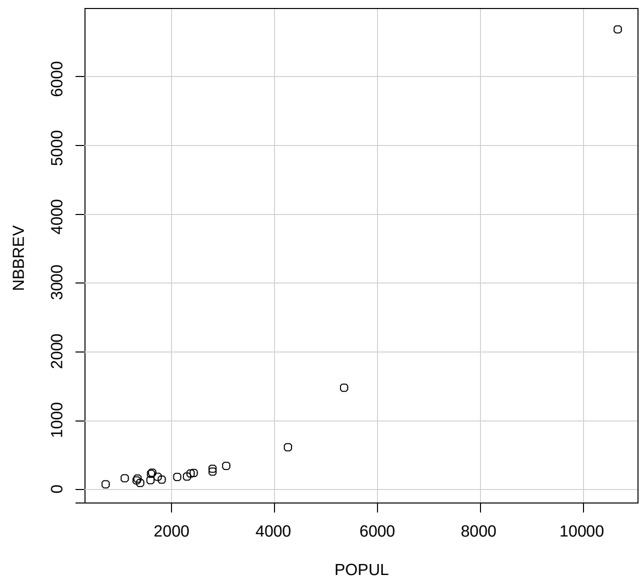


FIGURE 3 – Diagramme de dispersion

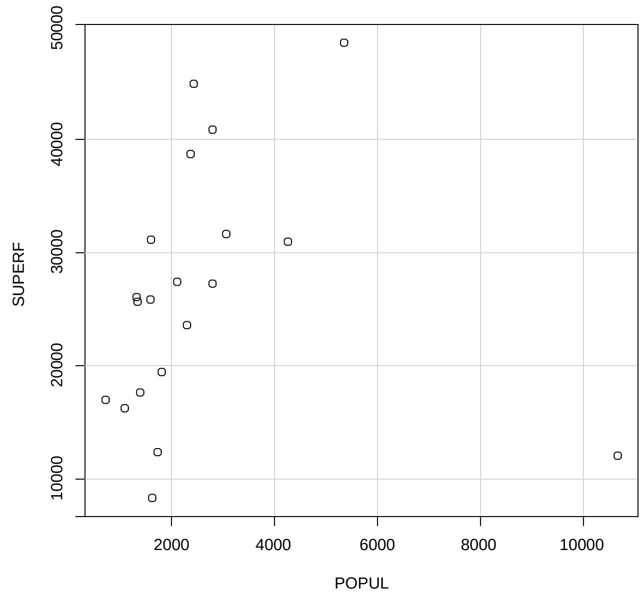


FIGURE 4 – Diagramme de dispersion

Fichier 2 : Employés

Tableau 5 :

counts for national

non américain	américain
370	104

percentages for national

non américain	américain
78.06	21.94

counts for sexe

F	M
216	258

percentages for sexe

F	M
45.57	54.43

counts for stat_pro

employé de bureau	agent de sécurité	manager
363	27	84

percentages for stat_pro

employé de bureau	agent de sécurité	manager
76.58	5.70	17.72

Tableau 6

stat_pro

sexe	employé de bureau	agent de sécurité	manager
F	206	0	10
M	157	27	74

Percentage of Total

	employé de bureau	agent de sécurité	manager	Total
F	43.5	0.0	2.1	45.6
M	33.1	5.7	15.6	54.4
Total	76.6	5.7	17.7	100.0

Tableau 7

Row Percentages

stat_pro

sexe	employé de bureau	agent de sécurité	manager	Total	Count
F	95.4	0.0	4.6	100.0	216
M	60.9	10.5	28.7	100.1	258

Tableau 8

Column Percentages

stat_pro			
sexe	employé de bureau	agent de sécurité	manager
F	56.7	0	11.9
M	43.3	100	88.1
Total	100.0	100	100.0
Count	363.0	27	84.0

Tableau 9

Pearson's Chi-squared test

data: .Table

X-squared = 79.2771, df = 2, p-value < 2.2e-16

Tableau 10

Expected Counts

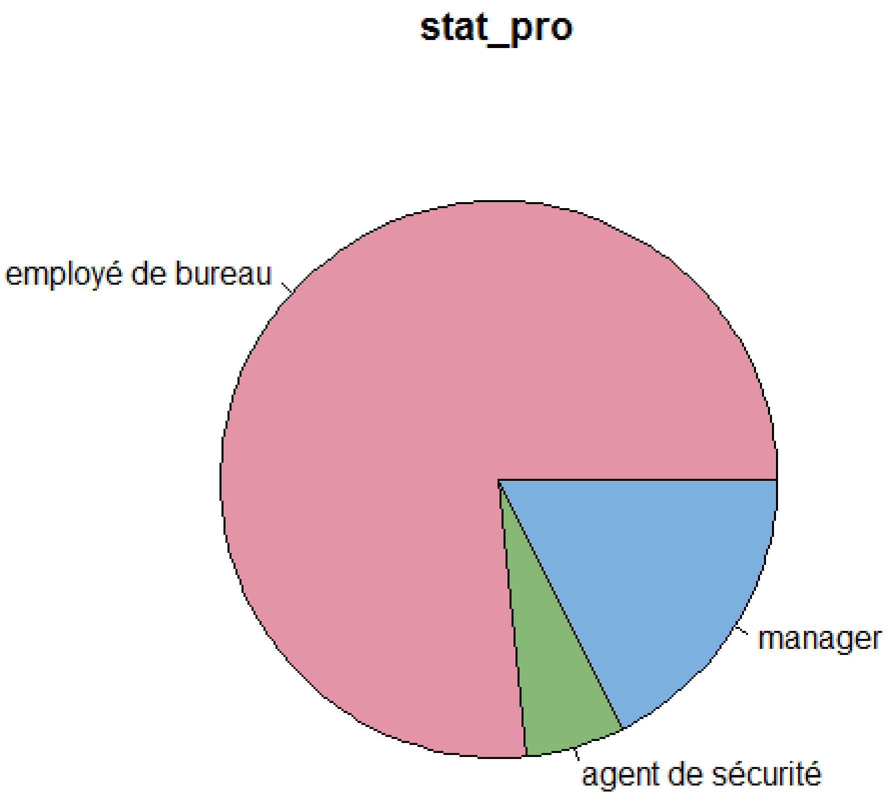
stat_pro			
sexe	employé de bureau	agent de sécurité	manager
F	165.4177	12.3038	38.27848
M	197.5823	14.6962 4	5.72152

Tableau 11

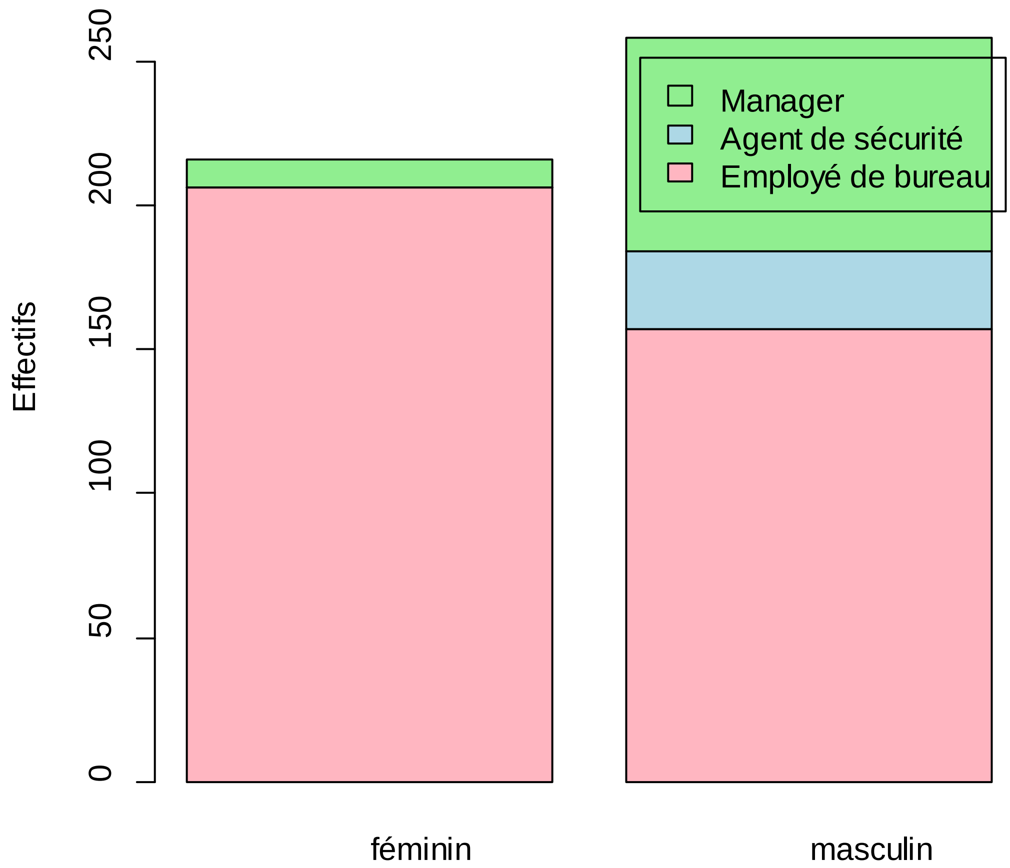
Chi-square Components

stat_pro			
sexe	employé de bureau	agent de sécurité	manager
F	9.96	12.30	20.89
M	8.34	10.30	17.49

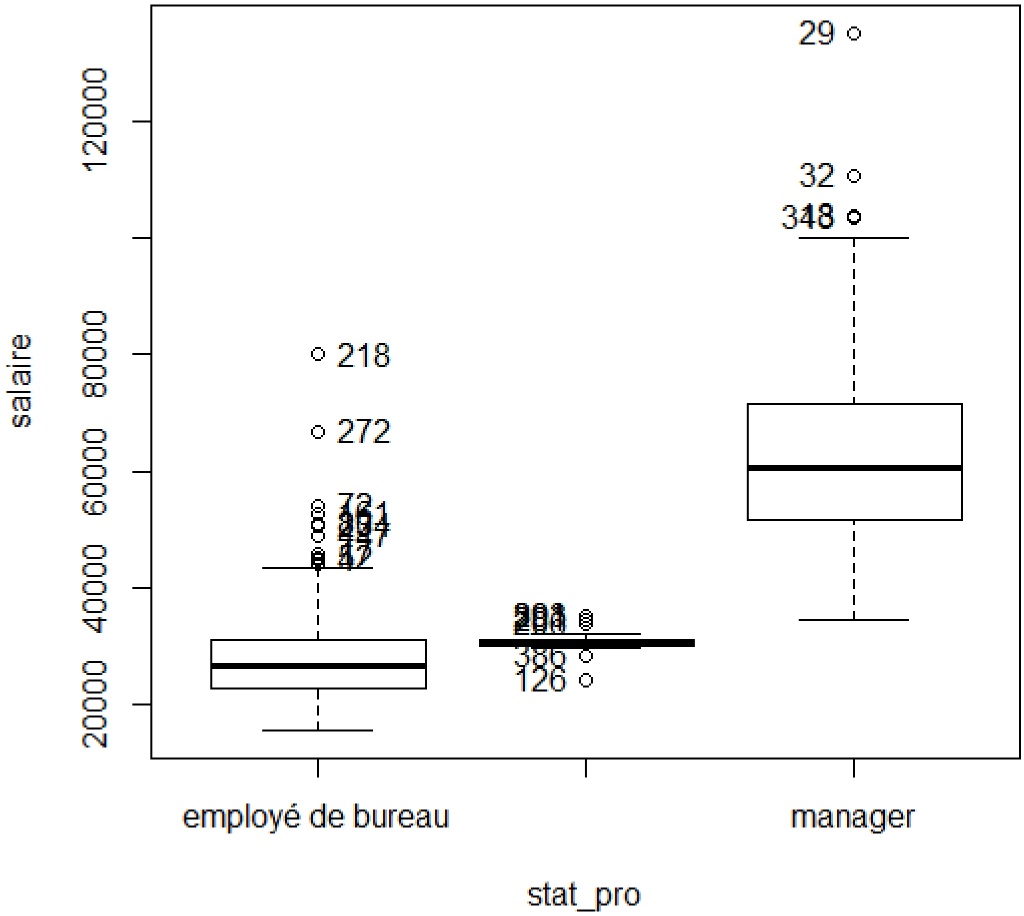
Graphique 5 : Diagramme en barres du statut professionnel



Graphique 6 : Diagramme en secteurs du statut professionnel



Graphique 7 : Diagramme en barres du statut professionnel en fonction du sexe



Graphique 8 : Boîtes à moustaches juxtaposées du salaire selon le statut professionnel

Tableau 12

Residuals:

Min	1Q	Median	3Q	Max
-29568	-5339	-1139	3551	71022

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	27838.5	532.5	52.280	<2e-16 ***
stat_pro[T.agent de sécurité]	3100.3	2023.8	1.532	0.126
stat_pro[T.manager]	36139.3	1228.4	29.421	<2e-16 ***

Residual standard error: 10150 on 471 degrees of freedom

Multiple R-squared: 0.6485, Adjusted R-squared: 0.647

F-statistic: 434.5 on 2 and 471 DF, p-value: < 2.2e-16