Probabilité et statistique I (partim statistique descriptive)

Bachelier en sciences informatiques
Correction de l'examen du 20 juin 2017

QCM

1.
   - Soit n le nombre d'individus dans l'électorat. Puisqu'il y a 65% de femmes et que 40% d'entre elles ont voté oui, le nombre de femmes ayant voté oui vaut

     $$n_{FO} = 0.40 \times 0.65 \times n = 0.26\ n.$$

     De la même manière, puisqu'il y a 35% d'hommes parmi lesquels 70% ont voté oui, le nombre d'hommes ayant voté oui vaut

     $$n_{HO} = 0.70 \times 0.35 \times n = 0.245\ n.$$

     Il y a donc moins d'hommes ayant voté oui que de femmes ayant voté oui.
   - **Vrai.** Le nombre de oui est égal à

     $$n_O = n_{FO} + n_{HO} = (0.26 + 0.245)\ n = 0.505\ n.$$

2.
   - **Faux.** Le mode de la série est égal à 4.
   - **Vrai.** Le 1er décile de cette série correspond à la plus petite valeur pour laquelle l'effectif cumulé est supérieur ou égal à $50 \times \frac{1}{10} = 5$. Il s'agit donc de la valeur 2 (dont l'effectif cumulé vaut 7).

     Le 9ème décile de cette série correspond à la plus petite valeur pour laquelle l'effectif cumulé est supérieur ou égal à $50 \times \frac{9}{10} = 45$. Il s'agit donc de la valeur 9 (dont l'effectif cumulé vaut 47).
   - Le graphique le plus adéquat pour représenter la distribution des fréquences cumulées de cette série est le graphique de droite puisque la variable est une variable quantitative discrète.

     Aucun des deux graphiques n'est adéquat pour représenter la distribution des effectifs cumulés puisque les deux graphiques sont basés sur les fréquences.

3.
   - Médiane : Tendance centrale
   - Coefficient $\gamma_1 = \frac{m_3}{s^3}$: Dissymétrie.
   - Ecart interquartile : Dispersion.

4.
   - Si la série des températures est transformée en degrés Farenheit, la covariance est multipliée par $1 \times 1.8 = 1.8$.
   - Si la série des précipitations est transformée en mètres, le coefficient de corrélation est multiplié par $\frac{10^{-3} \times 1}{|10^{-3}| \times |1|} = 1$.
   - Si les deux séries sont transformées, la pente de la droite de régression de Y en X est multipliée par $\frac{10^{-3}}{1.8} = \frac{1}{1800}$, i.e. est divisée par 1800.
   - Si la série des températures est transformée en degrés Farenheit, le coefficient de corrélation est multiplié par $\frac{1.8 \times 1}{|1.8| \times |1|} = 1$.
   - Si la série des précipitations est transformée en mètres, la covariance est multipliée par $1 \times 10^{-3} = 0.001$.

Exercices

1. (a)
   - Puisque la somme des effectifs vaut 218, on a
   $$218 = 15 + 40 + 44 + n_4 + 64 + 15 \Leftrightarrow n_4 = 40.$$
   - La médiane est la valeur $\tilde{x}$ telle que
   $$N(\tilde{x}) = \frac{218}{2} = 109,$$
   où $N(x)$ est l'effectif cumulé de x. En calculant les effectifs cumulés des différentes classes, on déduit que la classe médiane est la classe $]110, e_4]$. Dès lors, dans l'ogive des effectifs cumulés, les points $(110, 99), (112.7, 109)$ et $(e_4, 139)$ sont alignés. Ces points sont situés sur la droite d'équation
   $$y - 99 = \frac{109 - 99}{112.7 - 110}(x - 110).$$
   Puisque le point $(e_4, 139)$ appartient à cette droite, on a
   $$139 - 99 = \frac{109 - 99}{112.7 - 110}(e_4 - 110) \Leftrightarrow e_4 = 120.8.$$
   Arrondir cette borne à l'unité la plus proche semble peu correspondre à la philosophie suivie dans la construction des autres classes. Le calcul de la moyenne à l'aide des centres des classes est un calcul qui peut être partiellement perturbé par des erreurs d'arrondis. Il semble donc logique de prendre $e_4 = 120$ (la réponse $e_4 = 121$ a été également considérée correcte lors de la correction).

   (b) Puisque les amplitudes des classes sont différentes, il faut construire un histogramme d'aire unitaire. Pour calculer la hauteur des rectangles, on divise les fréquences des classes par leur amplitude. Ainsi, l'aire totale de l'histogramme est égale à 1 et ce sont les aires des rectangles que l'on interprète et non les hauteurs. L'histogramme construit à partir de la table ci-dessous est donné en Figure 1. Le polygone correspondant y est ajouté en rouge.

Classe	Amplitude	Fréquence	Fréquence ajustée
[20, 90]	70	0.0688	0.0010
]90,100]	10	0.1835	0.0183
]100,110]	10	0.2018	0.0202
]110,120]	10	0.1835	0.0183
]120,150]	30	0.2936	0.0098
]150,200]	50	0.0688	0.0014

   Le polygone nous permet facilement de repérer un pic aux environs de la valeur 115 autour de laquelle la masse diminue relativement rapidement, plus rapidement à gauche qu'à droite, ce qui induit une certaine dissymétrie.

   (c) On a
   $$\begin{aligned} s &= \sqrt{s^2} \\ &= \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2} \\ &= \sqrt{\frac{1}{218}3030225 - 115.7^2} \\ &= 22.6. \end{aligned}$$

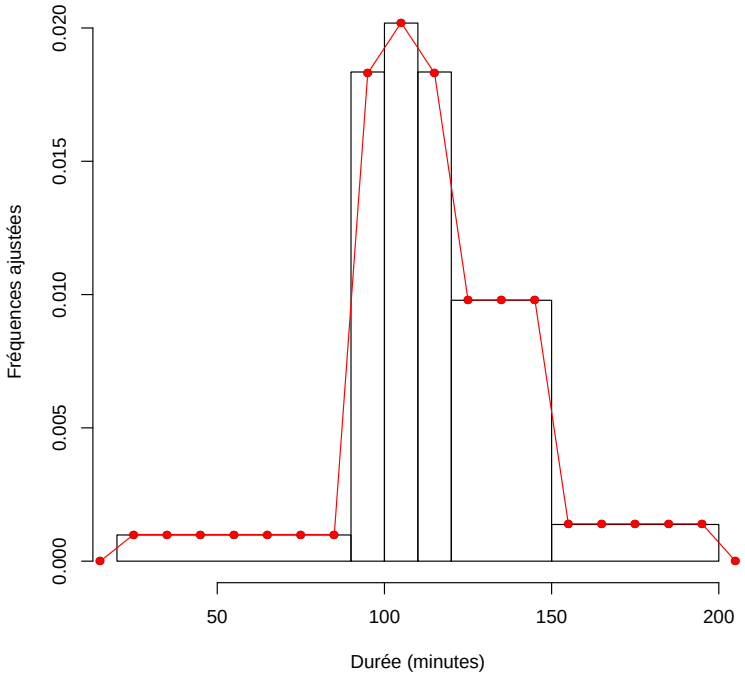

Figure 1: Histogramme et polygone des fréquences de la variable Durée

Dès lors,

$$\bar{x} - 2s = 70.38 \text{ et } \bar{x} + 2s = 161.02.$$

Le nombre d'observations situées entre ces deux bornes est donné par $N(161.02) - N(70.38)$, où $N(x)$ est l'effectif cumulé de x.
Dans l'ogive des effectifs cumulés, les points $(20, 0), (70.38, N(70.38))$ et $(90, 15)$ sont alignés. Ils sont situés sur la droite d'équation

$$y - 0 = \frac{15 - 0}{90 - 20}(x - 20).$$

Puisque le point $(70.38, N(70.38))$ appartient à cette droite, on a

$$N(70.38) = \frac{15}{70}(70.38 - 20) = 10.80.$$

Par ailleurs, les points $(150, 203), (161.02, N(161.02))$ et $(200, 218)$ sont eux aussi alignés dans l'ogive des effectifs cumulés. Ils sont situés sur la droite d'équation

$$y - 203 = \frac{218 - 203}{200 - 150}(x - 150).$$

Puisque le point $(161.02, N(161.02))$ appartient à cette droite, on a

$$N(161.02) = \frac{15}{50}(161.02 - 150) + 203 = 206.31.$$

On en tire que la proportion d'observations situées entre les bornes 70.38 et 161.02 est donnée par $\dfrac{206.31 - 10.80}{218} \approx 90\%$.

L'inégalité de Tchebychev affirme que la proportion d'observations situées dans l'intervalle $[\bar{x} - 2s, \bar{x} + 2s]$ est supérieure à $\frac{1}{2^2} = 25\%$. Cette inégalité est donc bien vérifiée.

(d) On a

$$\begin{aligned} s_{xy} &= \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y} \\ &= \frac{1}{218}186974 - 115.7 \times 7.4 \\ &= 1.50. \end{aligned}$$

La covariance est positive. En effet, le diagramme de dispersion de la Figure 2 peut être découpé en 4 quadrants à partir du centre de gravité $(\bar{x}, \bar{y})$. Les observations des quadrants 2 et 4 apportent une contribution positive à la covariance alors que les observations des quadrants 1 et 3 lui apportent une contribution négative. Ces dernières étant moins nombreuses (peu d'observations dans le quadrant 3) que les premières, le signe de la covariance est positif. Cela signifie que la relation entre les deux variables est croissante : lorsque la durée est importante, la cote a tendance à être meilleure.

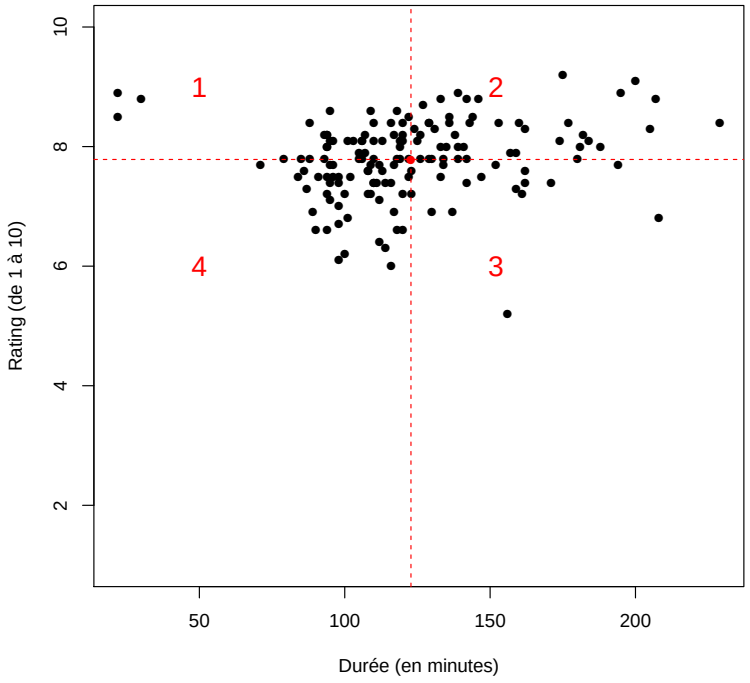

Figure 2: Diagramme de dispersion de la variable `Rating` en fonction de la durée du film

Parmi les outils utiles pour déterminer si l'appréciation d'un film dépend ou pas de la durée du film il y a :

- Le coefficient de corrélation, qui mesure l'intensité du lien linéaire entre les deux variables, toujours compris entre -1 et 1. Dans ce cas précis, celui-ci sera relativement faible car le lien n'est pas très marqué (voir diagramme de dispersion).
- La pente de la droite de régression des moindres carrés. Dans ce cas-ci, cette droite

sera quasiment horizontale au vu de la dispersion des données. Cela signifie que la durée n'influence pas beaucoup la cote.

- Le coefficient de détermination, i.e. la part de la variance totale expliquée par la régression linéaire. Encore une fois, dans ce cas-ci, le coefficient de détermination sera relativement faible.

2. (a) On sait que la moyenne des durées des 218 films vaut 115.7 minutes. La somme des observations de la série initiale est donc égale à 218×115.7. Afin de calculer la moyenne de la série des 219 observations, il suffit d'ajouter à cette somme partielle la valeur de la nouvelle observation, avant de diviser par l'effectif 219. On obtient

$$\bar{x}' = \frac{218 \times 115.7 + 5520}{219} = 140.38$$

En ce qui concerne la médiane, son approximation basée sur la distribution groupée est égale à 112.7 et provenait du résultat d'une approximation linéaire au sein de la 4ème classe. Vu l'ajout de l'observation 5520, la distribution groupée va être perturbée, mais uniquement via sa dernière classe (dont la borne supérieure doit être adaptée et dont l'effectif doit être augmenté d'une unité). Toujours au sein de la classe $]110, 120]$, le calcul de la médiane se fait par interpolation linéaire à partir des points $(110, 99)$ et $(120, 139)$ tout en prenant un effectif total égal à 219. On obtient

$$\tilde{x}' = 110 + \frac{1}{4}\left(\frac{219}{2} - 99\right) = 112.6$$

On voit donc que la médiane reste globalement inchangée, tandis que la moyenne a beaucoup augmenté.

Concernant la médiane, on pouvait également travailler (moins précisément) à partir des données brutes. Pour la série initiale, l'effectif étant pair, la médiane correspond à la moyenne entre les observations de rangs 109 et 110. Avec l'ajout d'une observation dans la queue droite de la distribution, l'effectif devient impair et la médiane coincide avec l'observation initiale de rang 110 (elle doit donc être assez proche de la médiane de départ puisqu'elle reste située dans la même classe).

(b) Afin de modéliser l'impact de la nouvelle observation, notons $\bar{x}$ (resp. $\tilde{x}$) la moyenne (resp. la médiane) initiale et $\bar{x}'$ (resp. $\tilde{x}'$) la nouvelle moyenne (resp. médiane) et considérons uniquement le cas où l'observation x ajoutée est plus grande ou égale à l'ensemble des observations $x_1, \ldots, x_n$.

Pour le calcul de la moyenne, en suivant une démarche similaire à celle expliquée ci-dessus dans le cas particulier de la série des durées, on obtient

$$\bar{x}' = \frac{n \times \bar{x} + x}{n+1}$$

puisque $n \times \bar{x} + x$ correspond à la somme des données initiales et l'observation ajoutée et $n+1$ est l'effectif de la nouvelle série.

Pour le calcul de la médiane, il convient de discuter en fonction de la parité de n et d'ordonner les observations de la plus petite $(x_{(1)})$ à la plus grande $(x_{(n+1)} = x)$.

Si n est pair (égal à $2k$), alors la médiane de départ est égale à

$$\tilde{x} = \frac{x_{(k)} + x_{(k+1)}}{2}$$

et la nouvelle médiane (égale à l'observation du milieu sachant que $n = 2k+1$) correspond à l'observation $x_{(k+1)}$.

Si n est impair (égal à $2k+1$), alors la médiane de départ est égale à l'observation $x_{(k+1)}$, tandis que la nouvelle médiane (correspondant à la moyenne des deux observations centrales sachant que le nombre d'observations est $2k+2$) vaut

$$\tilde{x}' = \frac{x_{(k+1)} + x_{(k+2)}}{2}$$

Dans chacune des expressions de $\tilde{x}'$, la valeur x de l'observation ajoutée n'intervient pas!

(c) La moyenne $\bar{x}'$ s'exprime comme une fonction linéaire en x. Son comportement, en fonction de x, est celui d'une droite de pente $1/(n+1)$ (droite qui passe par $\bar{x}$ en $x = \bar{x}$). La médiane (quel que soit n) est indépendante de x. Elle reste donc constante pour toute valeur de x supérieure ou égale à $x_{(n)}$.

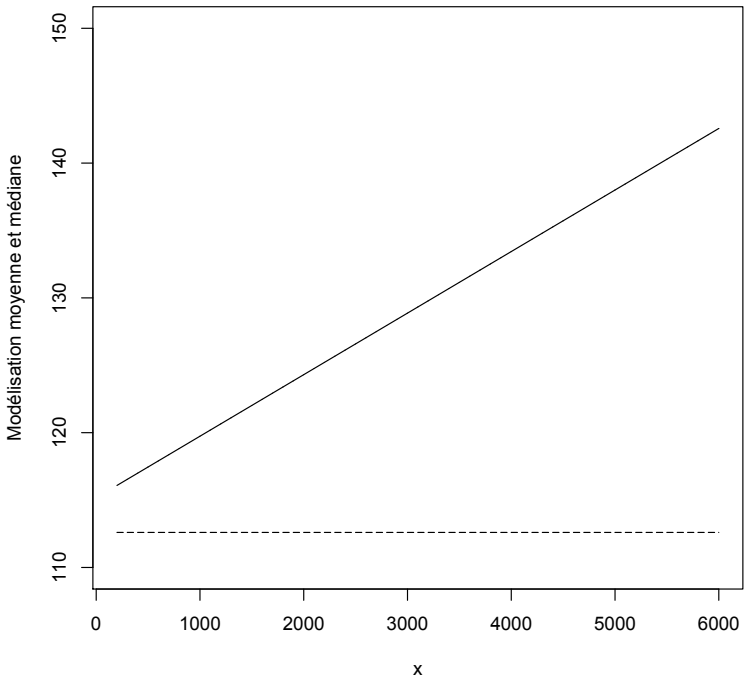

Figure 3: Représentation de $\bar{x}'$ (traits pleins) et $\tilde{x}'$ (traits discontinus) en fonction de x

Les deux fonctions sont représentées à la Figure 3 (en utilisant les durées; mais la modélisation pourrait aussi être visualisée sur un exemple fictif). On constate que la moyenne est susceptible d'"exploser" si l'observation ajoutée devient très grande, tandis que la médiane reste imperturbable. Ce paramètre est clairement plus robuste que la moyenne.

SAS

Groupe a

1. Voir Figure 4.

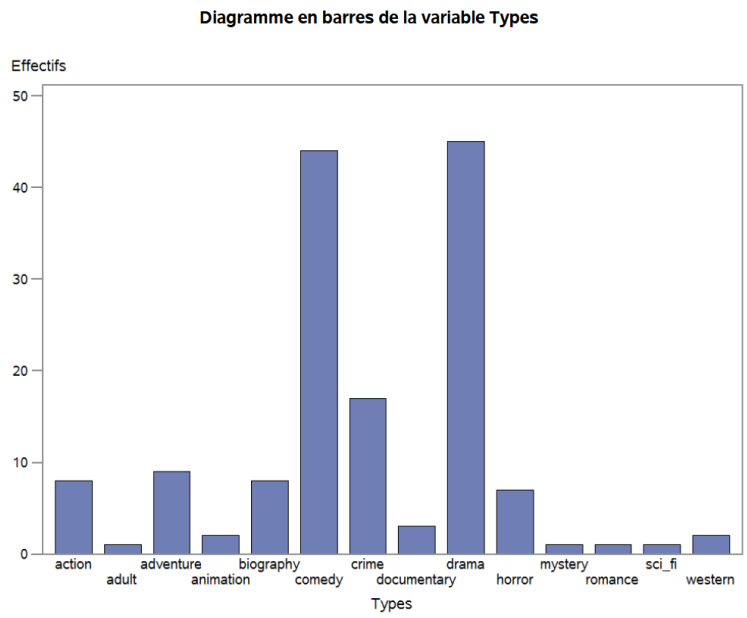

Figure 4: Distribution des effectifs de la variable `Types`

2. Drama.
3. Effectif conjoint : 24.
   Fréquence conditionnelle : 0.53.
4. Voir Figure 5.
5. Voir graphique de gauche de la Figure 6.
6.

Distr. Cond.	Médiane	Etendue
`comedy`	106	150
`horror`	109	57

7. (a) $\hat{b} = 7.38795$.
   (b) 2.7%.
   (c) Caligula.
   (d) Voir Figure 7.

Groupe b

1. Voir Figure 8.
2. 10.
3. Effectif conjoint : 3.
   Fréquence conditionnelle : 0.33.

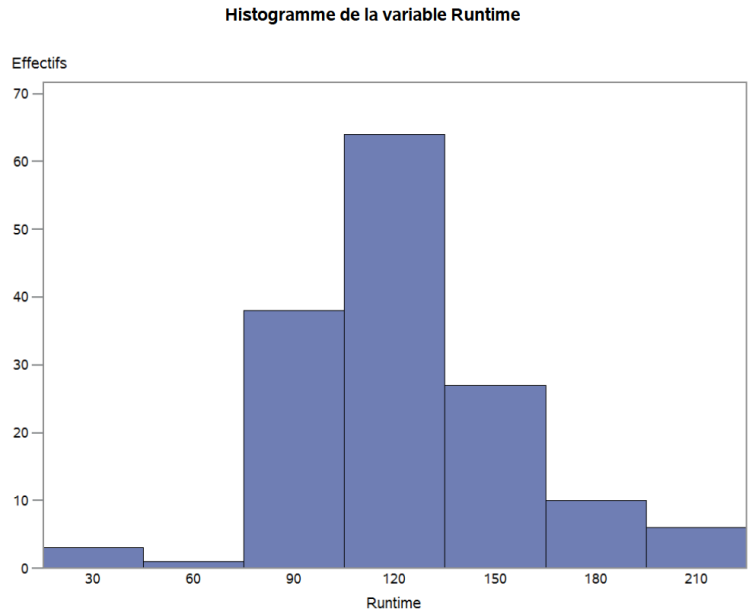

Figure 5: Histogramme de la variable `Runtime`

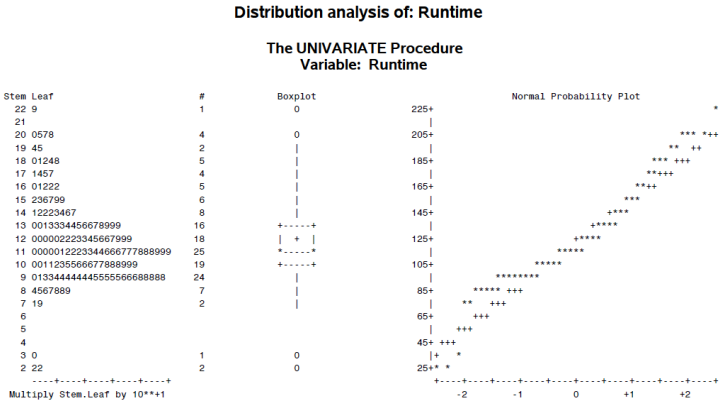

Figure 6: Diagramme en tiges et feuilles de la variable `Runtime`

4. Voir Figure 9.

5. Voir Figure 10.

6.

Distr. Cond.	Médiane	Etendue
1994	17 055	777 053
1998	115 202	370 893

7. (a) $\hat{b} = 7.51056$.
   (b) 33.4%.
   (c) Caligula.
   (d) Voir Figure 11.

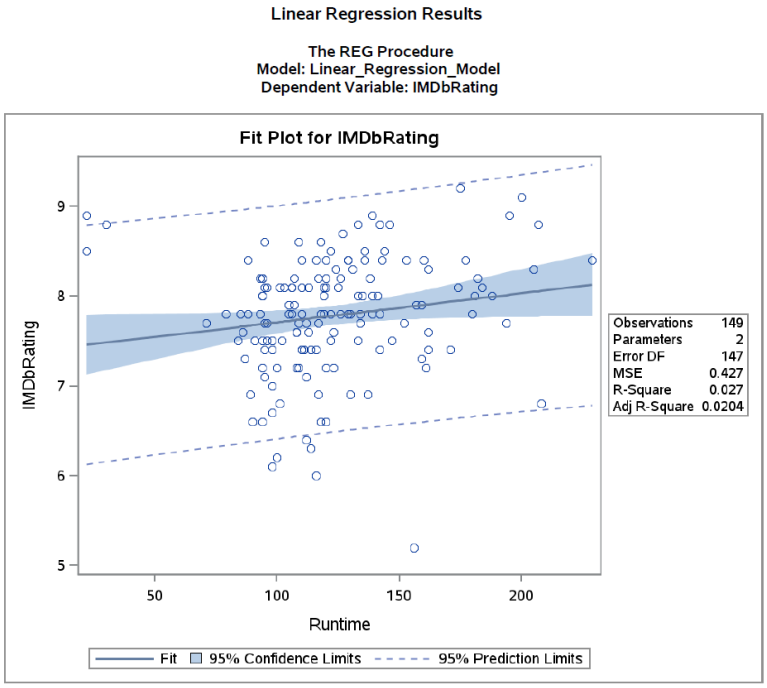

Figure 7: Diagramme de dispersion de la variable `IMDbRating` en fonction de la variable `Runtime` et droite de régression

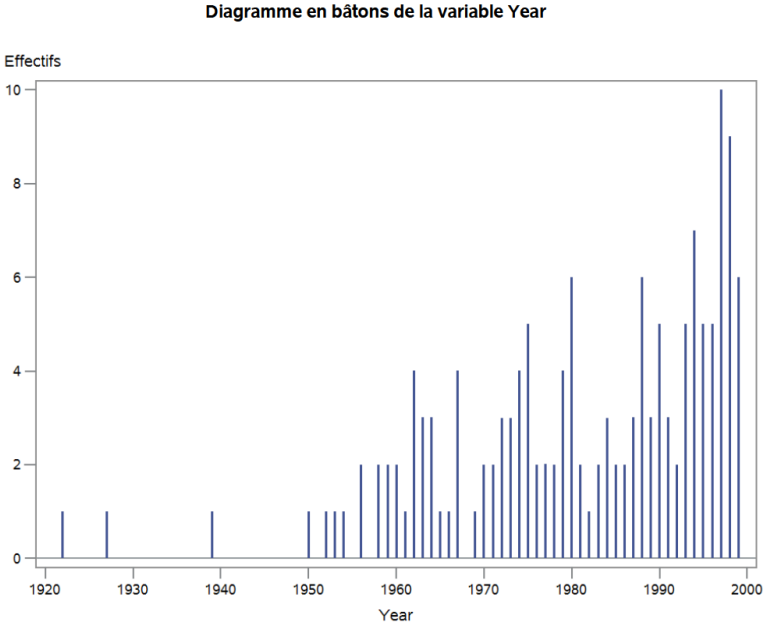

Figure 8: Distribution des effectifs de la variable `Year`

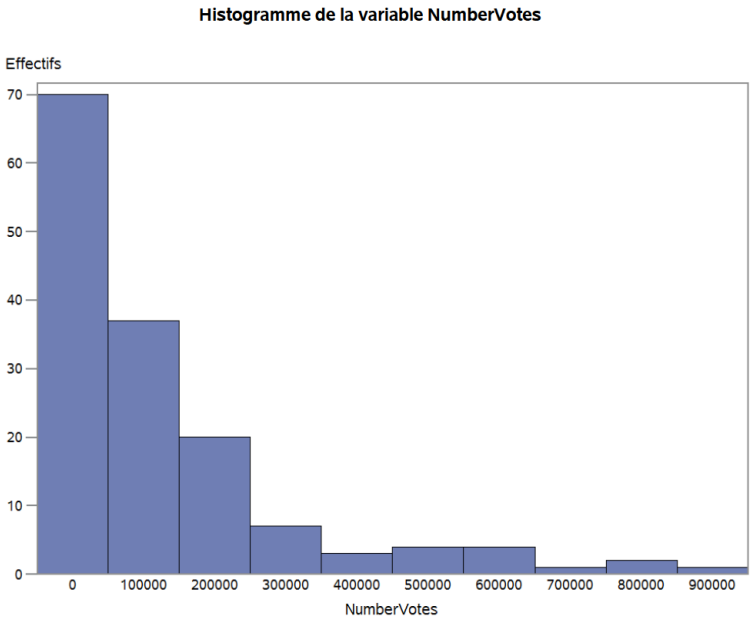

Figure 9: Histogramme de la variable `NumberVotes`

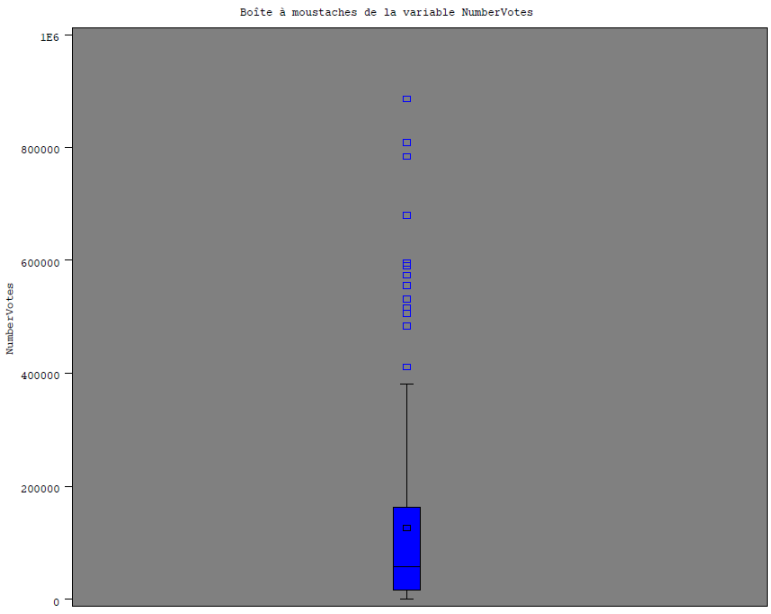

Figure 10: Boîte à moustaches de la variable `NumberVotes`

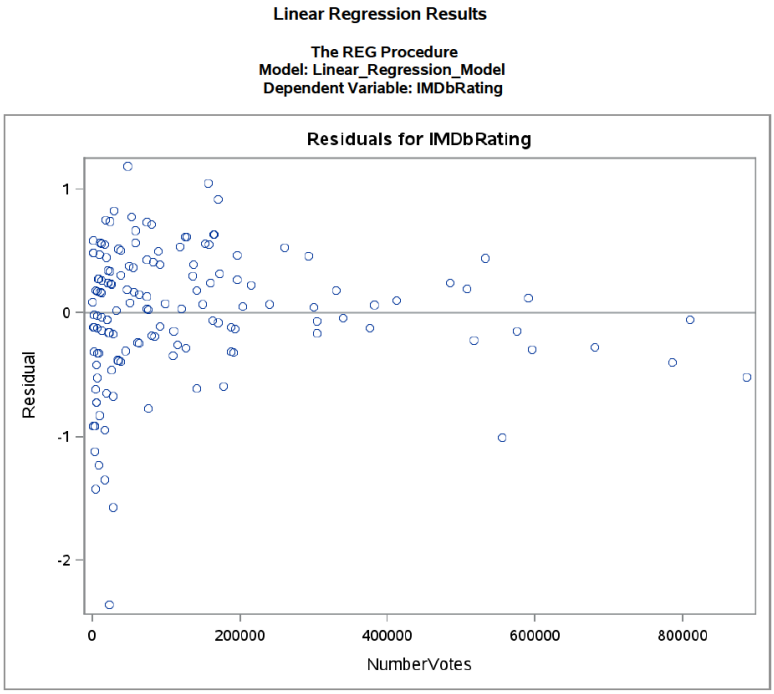

Figure 11: Résidus de la régression de la variable `IMDbRating` en fonction de la variable `Runtime`