

Compte-rendu en *Principes et méthodes statistiques*

Analyse de signaux oculométriques

Aurélien PEPIN, Léo DESBUREAUX, Julien LABOURÉ — Grenoble INP - Ensimag

5 mai 2017

1 Analyse d'échantillons de loi binomiale négative

QUESTION 1. On suppose dans un premier temps le paramètre r connu. L'estimateur obtenu par la méthode des moments est noté \tilde{p}_n .

Comme on suppose les variables X_1, X_2, \dots, X_n de l'échantillon indépendantes et de même loi binomiale négative $\mathcal{BN}(r, p)$, on a $\mathbb{E}[X] = \frac{r}{p}$. Alors l'estimateur des moments est :

$$\tilde{p}_n = \frac{r}{\overline{X}_n}$$

où \overline{X}_n désigne la moyenne empirique de l'échantillon.

Une deuxième méthode d'estimation ponctuelle est l'estimation par maximum de vraisemblance. On note maintenant l'estimateur trouvé \hat{p}_n .

$$\mathcal{L}(p; x_1, \dots, x_n) = P(X_1 = x_1, \dots, X_n = x_n; p) = \prod_{i=1}^n P(X = x_i; p)$$

Plutôt que de dériver directement ce produit, on préfère maximiser le logarithme de la fonction de vraisemblance \mathcal{L} , c'est la **log-vraisemblance** :

$$\begin{aligned} \ln \mathcal{L}(p; x_1, \dots, x_n) &= \ln \prod_{i=1}^n P(X = x_i; p) = \sum_{i=1}^n \ln P(X = x_i; p) \\ &= \sum_{i=1}^n \ln \left(\binom{x_i-1}{r-1} (1-p)^{x_i-r} p^r \right) \end{aligned}$$

On cherche désormais à obtenir \hat{p}_n , valeur qui maximise cette log-vraisemblance. On dérive pour cela l'expression précédente :

$$\begin{aligned} \frac{\partial}{\partial p} \ln \mathcal{L}(p; x_1, \dots, x_n) &= \frac{\partial}{\partial p} \left(\sum_{i=1}^n \ln \binom{x_i-1}{r-1} \right) + \sum_{i=1}^n (x_i - r) \frac{\partial}{\partial p} (\ln(1-p)) + \sum_{i=1}^n r \frac{\partial}{\partial p} (\ln p) \\ &= 0 - \sum_{i=1}^n \left(\frac{x_i - r}{1-p} \right) + \sum_{i=1}^n \frac{r}{p} \end{aligned}$$

Cette expression s'annule sous les conditions suivantes :

$$\begin{aligned}
-\sum_{i=1}^n \left(\frac{x_i - r}{1-p} \right) + \sum_{i=1}^n \frac{r}{p} = 0 &\iff \sum_{i=1}^n \frac{-px_i + pr + (1-p)r}{p(1-p)} = 0 \\
&\iff \sum_{i=1}^n -px_i + r = 0 \\
&\iff \sum_{i=1}^n \frac{r}{p} = \sum_{i=1}^n x_i \\
&\iff n \frac{r}{p} = \sum_{i=1}^n x_i \\
&\iff \frac{r}{p} = \overline{X}_n
\end{aligned}$$

Finalement, on retrouve donc le résultat précédent :

$$\hat{p}_n = \frac{r}{\overline{X}_n} = \tilde{p}_n$$

Les cas aux limites où $p = 0$ ou $p = 1$ correspondent à des situations triviales où tous les X_i sont identiques, il n'y a aucune part d'aléatoire.

QUESTION 2. Pour une suite de variables aléatoires i. i. d. $\{X_n\}_{n \geq 1}$, le **théorème central limite** exprime la convergence suivante :

$$Z_n = \sqrt{n} \frac{\overline{X}_n - E[X]}{\sigma(X)} \xrightarrow{n \rightarrow +\infty} \mathcal{N}(0, 1)$$

Soient X_1, \dots, X_n les variables de l'échantillon de loi $\mathcal{BN}(r, p)$. Par définition, on a :

$$\mathbb{E}[X] = \frac{r}{p} \quad \text{et} \quad \sigma(X) = \frac{\sqrt{r(1-p)}}{p}$$

La suite X_1, \dots, X_n satisfait les conditions du théorème. On définit ainsi :

$$\begin{aligned}
Z_n &= p\sqrt{n} \frac{\overline{X}_n - \frac{r}{p}}{\sqrt{r(1-p)}} \\
&= \frac{\sqrt{nr}}{\sqrt{1-p}} \left(\frac{p}{\hat{p}_n} - 1 \right) \xrightarrow{n \rightarrow +\infty} \mathcal{N}(0, 1)
\end{aligned}$$

Sachant que $\mathbb{P}(|Z_n| > u_\alpha) = \alpha$, on peut construire un intervalle de confiance :

$$\begin{aligned} \frac{\sqrt{nr}}{\sqrt{1-p}} \left| \frac{p}{\hat{p}_n} - 1 \right| > u_\alpha &\iff \sqrt{nr} |p - \hat{p}_n| > u_\alpha \hat{p}_n \sqrt{1-p} \\ &\iff rn (p - \hat{p}_n)^2 > u_\alpha^2 \hat{p}_n^2 (1-p) \\ &\iff rn p^2 + (-2rn \hat{p}_n + u_\alpha^2 \hat{p}_n^2) p + \hat{p}_n^2 (rn - u_\alpha^2) > 0 \\ &\iff p^2 + \hat{p}_n (-2 + \frac{u_\alpha^2 \hat{p}_n}{rn}) p + \hat{p}_n^2 (1 - \frac{u_\alpha^2 \hat{p}_n}{rn}) > 0 \end{aligned}$$

Posons $\lambda = \frac{u_\alpha^2 \hat{p}_n}{rn}$ et réécrivons l'équation précédente :

$$p^2 + p\hat{p}_n(-2 + \lambda \hat{p}_n) + \hat{p}_n^2 (1 - \lambda) > 0$$

C'est un polynôme en p dont on déduit le discriminant Δ :


$$\Delta = (\hat{p}_n^2 \lambda)^2 + 1 + \frac{4(1 - \hat{p}_n)}{\lambda \hat{p}_n}$$

À partir du discriminant, on peut calculer les racines du polynôme qui sont les bornes de l'intervalle de confiance qu'on cherche à déterminer :

$$IC_p = \left[\hat{p}_n - \frac{1}{2} \lambda \hat{p}_n^2 \left(1 + \sqrt{1 + \frac{4(1 - \hat{p}_n)}{\lambda \hat{p}_n}} \right) ; \hat{p}_n - \frac{1}{2} \lambda \hat{p}_n^2 \left(1 - \sqrt{1 + \frac{4(1 - \hat{p}_n)}{\lambda \hat{p}_n}} \right) \right]$$

QUESTION 3. Pour tracer le graphe de probabilités de la loi géométrique (qui correspond au cas $r = 1$), on cherche en premier lieu des fonctions h, α, g, β telles que :

$$h[F(k)] = \alpha(p) g(k) + \beta(p)$$

 Se référer à : P1_Q3_Graphe_Probabilites.r

La fonction de répartition de la loi géométrique est :

$$\begin{aligned} F_G(k) = 1 - (1-p)^k &\iff 1 - F_G(k) = (1-p)^k \\ &\iff \ln(1 - F_G(k)) = k \ln(1-p) \end{aligned}$$

Par identification, on établit les correspondances suivantes :

- $h[F_G(k)] = \ln(1 - F_G(k))$
- $\alpha(p) = \ln(1 - p)$
- $g(k) = k$
- $\beta(p) = 0$

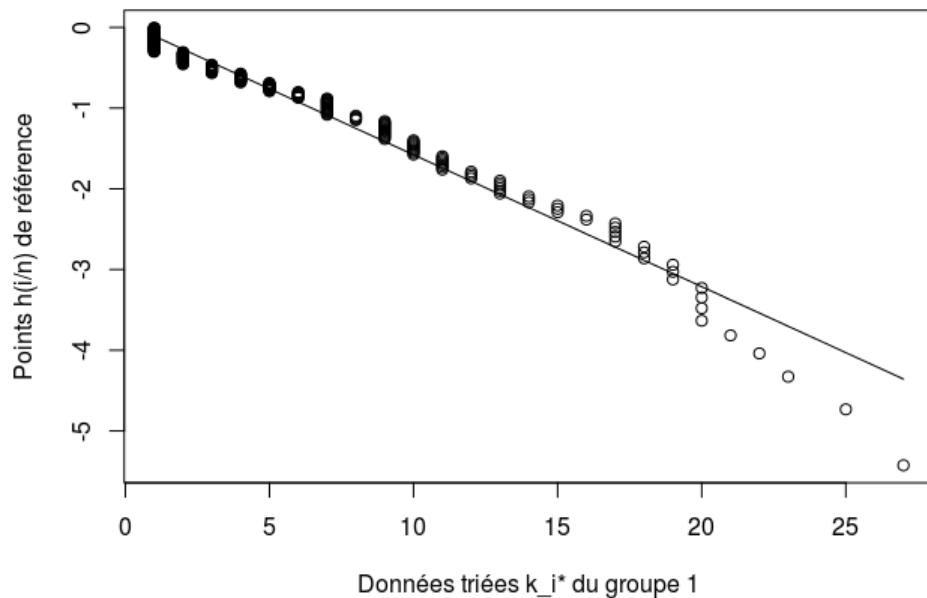
Le graphe de probabilités de $F_G(k)$ est donc le nuage de points :

$$\left(g(k_i^*) ; h\left(\frac{i}{n}\right) \right) = \left(k_i^* ; \ln\left(1 - \frac{i}{n}\right) \right) \quad \forall i \in \llbracket 1; n-1 \rrbracket$$

Ce graphe permet de mesurer visuellement l'adéquation entre une distribution et la loi qu'elle est supposée suivre. On vérifie pour cela que les points affichés sont alignés.

Par une régression linéaire, on peut aussi ajouter la droite des moindres carrés de y en x . Appliquons ces outils aux groupes de données dont on dispose ainsi qu'à un échantillon aléatoire.

Graphe de probabilités de la loi géométrique sur le groupe 1



Sur le **groupe 1**, on remarque que les points sont bien alignés en eux, sauf pour quelques mesures aux extrémités. La loi géométrique semble donc être ici un modèle plausible.

Pour estimer le paramètre p , on peut utiliser la régression linéaire. La droite des moindres carrés a pour coefficient directeur $m = \alpha(p) = \ln(1 - p)$.

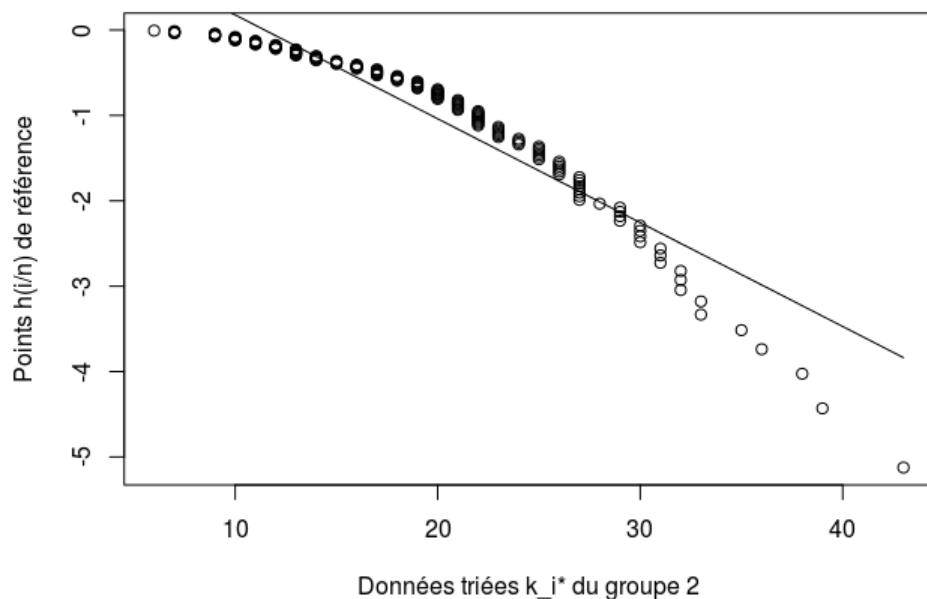
$$m = \ln(1 - p) \iff p = 1 - e^m$$

En **R**, le coefficient directeur est donné par l'attribut `coefficients` de la régression linéaire. Dans le cas du **groupe 1**, on estime ainsi :

$$p = p_{g_1} \simeq 0.1508$$

Si la loi géométrique semble un modèle plausible pour le **groupe 1**, il n'en est pas de même pour le groupe de données suivant.

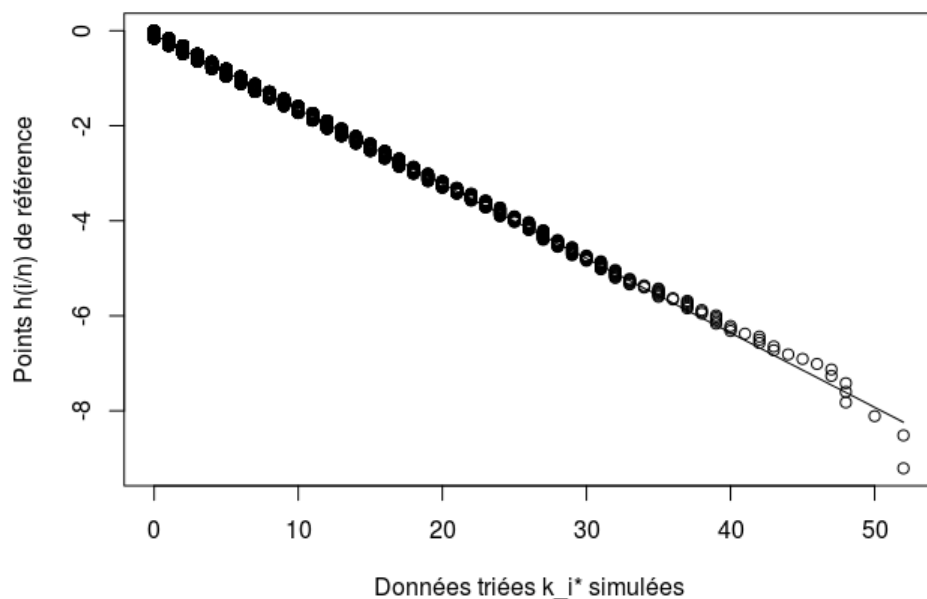
Graphique de probabilités de la loi géométrique sur le groupe 2



Dans le cas du **groupe 2**, les points sont peu alignés. Ils semblent plutôt suivre la trajectoire d'une courbe, ce qui laisse penser que la loi géométrique n'est pas un modèle adapté pour cet échantillon. Il est donc inutile de chercher à approcher un paramètre p .

Pour s'assurer que le graphe de probabilités modélise bien ce qu'on cherche, on utilise enfin un dernier graphique. On simule puis on trie 10 000 réalisations d'une loi géométrique. Comme les points sont alignés de manière quasi-parfaite, on en déduit que le graphe de probabilités est bien celui de la loi géométrique.

Vérification du graphe de probabilités



QUESTION 4. (a). Soit $X \hookrightarrow \mathcal{BN}(r, p)$.

$$\begin{aligned} \frac{P(X = x)}{P(X = x + 1)} &= \frac{\binom{x-1}{r-1} (1-p)^{x-r} p^r}{\binom{x}{r-1} (1-p)^{x+1-r} p^r} \\ &= \frac{\frac{(x-1)!}{(r-1)! (x-r)!} (1-p)^{x-r} p^r}{\frac{x!}{(r-1)! (x-r+1)!} (1-p)^{x+1-r} p^r} \\ &= \frac{x-r+1}{x(1-p)} \end{aligned}$$

(b). À partir du calcul précédent, on définit la fonction $g(x)$ telle que :

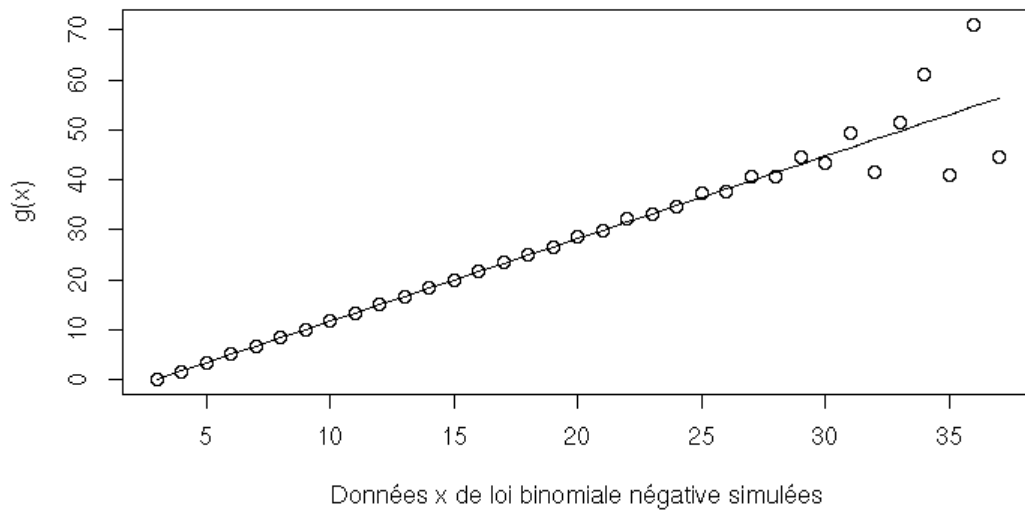
$$g(x) = x \frac{P(X = x)}{P(X = x + 1)} = \frac{1}{1-p} x + \frac{1-r}{1-p}$$

On affiche le nuage de points $(x, g(x))$. Grâce à une régression linéaire, on trace la droite au sens des moindres carrés $y = ax + b$ pour obtenir un coefficient directeur et une ordonnée à l'origine à partir desquels on peut retrouver p :

$$\begin{aligned} a = \frac{1}{1-p} &\iff p_{g_2} := p = 1 - \frac{1}{a} \\ b = \frac{1-r}{1-p} &\iff p_{g_3} := p = 1 - \frac{1-r}{b} \end{aligned}$$

(c). Vérifions si les deux méthodes d'estimation graphique de la question précédente sont efficaces sur un jeu de données simulées.

Vérification de la méthode 1.4 par simulation



On lance 10 000 000 simulations d'une loi binomiale négative $\mathcal{BN}(r, p)$ de paramètres $r = 4$ et $p = 0.4$. Appliquées à $g(x)$, ces mesures sont plutôt alignées. En traçant la droite des moindres carrés, on peut estimer les paramètres p_{g_3} et p_{g_2} comme expliqué ci-dessus :

$$p_{g_3} \simeq 0.3953 \quad \text{et} \quad p_{g_4} \simeq 0.3784$$

Il faut noter que les valeurs maximales de l'échantillon ont été supprimées avant de tracer la droite de la régression linéaire. Avec la commande `head`, on enlève ainsi le dernier cinquième des données triées. En effet, avec la binomiale négative, les valeurs sont concentrées autour de la moyenne. Pour autant, la droite linéaire minimise l'écart pour tous les points, et les valeurs maximales isolées perturbent l'alignement.

L'approximation du paramètre p pourrait d'ailleurs être encore meilleure si on enlevait plus de valeurs à droite.

QUESTION 5. On suppose désormais que le paramètre r est inconnu. On va donc chercher à estimer simultanément les deux paramètres (r, p) via la méthode des moments.

Pour cela, on rappelle les égalités suivantes :

$$E[X] = \frac{r}{p} \quad \text{et} \quad \text{Var}(X) = \frac{r(1-p)}{p^2}$$

$$\begin{aligned} \frac{E[X]}{\text{Var}(X)} = \frac{p}{1-p} &\iff E[X](1-p) = \text{Var}(X)p \\ &\iff E[X] = p(E[X] + \text{Var}(x)) \end{aligned}$$

On en déduit ainsi que :

$$\tilde{p}_n = \frac{\overline{X}_n}{\overline{X}_n + S_n^2} \quad \text{et} \quad \tilde{r}_n = \frac{\overline{X}_n^2}{\overline{X}_n + S_n^2}$$

QUESTION 6. On propose l'algorithme suivant :

1. On détermine $z = \min x_i$ pour trouver une première borne à r .
En effet, la valeur de r ne peut pas être supérieure à la valeur des x_i .
2. Pour chaque valeur de r possible, donc pour tout $i \in \llbracket 1; z \rrbracket$, on calcule sur le modèle de la **question 1** l'estimateur par maximum de vraisemblance $\hat{p}_{n,i} = \frac{i}{\overline{X}_n}$.
3. Pour chaque couple $(r_i, p_i) = (i, \hat{p}_{n,i})$, on calcule sa vraisemblance.
4. On trouve $(\hat{r}_n, \hat{p}_n) = \underset{i}{\operatorname{argmax}} \mathcal{L}(r_i, p_i; x_1, \dots, x_n)$ parmi les vraisemblances de l'étape 3.

2 Analyse des deux jeux de données

QUESTION 1. Dans le cadre de ce projet, on s'intéresse à deux groupes de lecteurs. Le premier compte 227 individus tandis que le deuxième en compte 168. Pour chaque individu, on s'intéresse au nombre de fixations relevées pendant la lecture des textes, différents d'un groupe à l'autre.

 Se référer à : P2_Q1_Statistique_Descriptive.r

Les données relevées sont des réalisations entières de variables quantitatives. Les indicateurs suivants aident à décrire les deux échantillons :

Indicateur	Groupe 1	Groupe 2
Moyenne empirique	6.498	19.738
Valeurs extrêmes	1 ; 31	6 ; 50
Mode	1	22
Médiane empirique	5	20
Variance empirique	36.003	59.979
Écart-type empirique	6	7.75
Coefficient de variation empirique	0.92	0.39

Quartiles du groupe 1 (via summary)

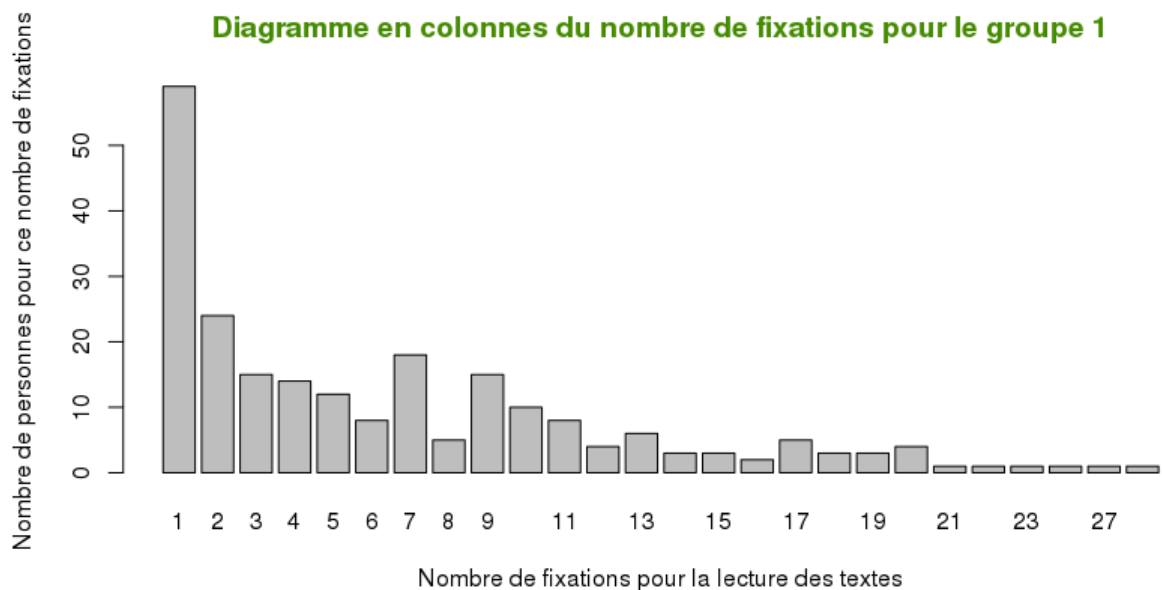
Min.	1 ^{er} quart.	Médiane	Moy. emp.	3 ^e quart.	Max.
1	1	5	6.5	9.5	31

Quartiles du groupe 2 (via summary)

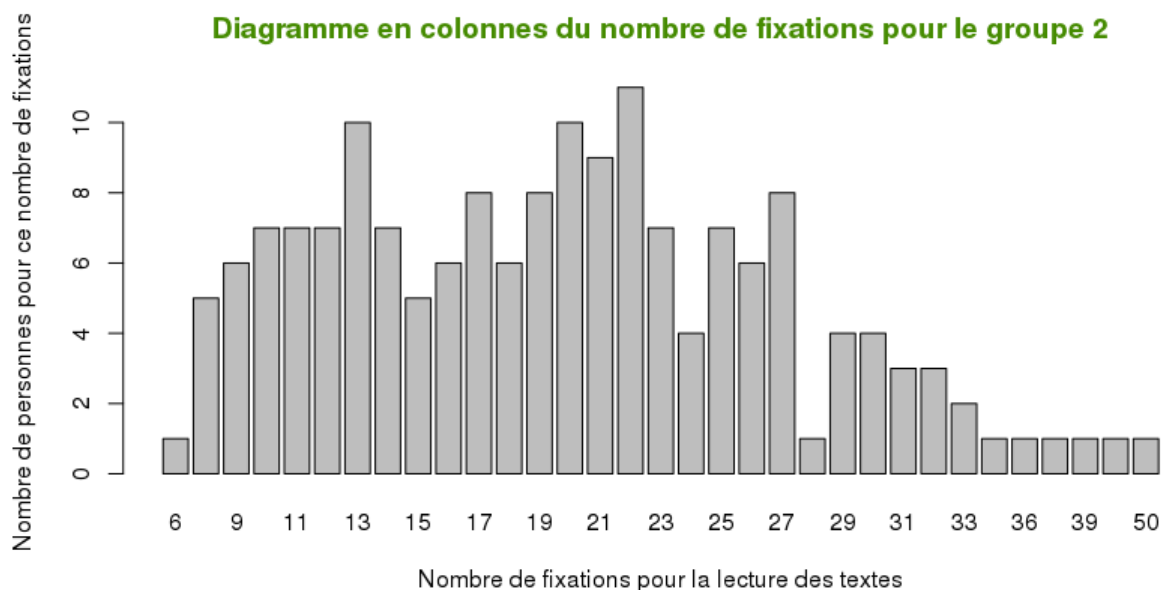
Min.	1 ^{er} quart.	Médiane	Moy. emp.	3 ^e quart.	Max.
6	13	20	19.74	25	50

À partir de ces indicateurs de tendance centrale et de dispersion pour les deux groupes de données, on peut déjà tirer quelques enseignements :

- Les personnes du **groupe 1** ont généralement besoin de beaucoup moins de fixations pour lire leurs textes que celles du **groupe 2**. Parmi elles, un très grand nombre n'a d'ailleurs eu besoin que d'une seule fixation.
- Le coefficient de variation empirique est moins importante chez les personnes du **groupe 2** que chez les personnes du **groupe 1**. Cela signifie que la moyenne empirique est un meilleur résumé de l'échantillon dans ce groupe que dans l'autre.
- Dans le **groupe 2**, la valeur minimale est 6, bien supérieure au mode du **groupe 1**. Deux hypothèses (non mutuellement exclusives) peuvent l'expliquer. Soit la série de textes du **groupe 2** était plus compliquée que celle du **groupe 1** et obligeait *forcément* à arrêter son regard sur certains mots ; soit les individus du **groupe 2** présentent un caractère commun comme une difficulté à lire.



Le mode de ce diagramme semble indiquer qu'il y a eu un mot en particulier sur lequel les individus du **groupe 1** se sont fixés massivement.




Ici au contraire, il n'y a pas réellement de tendance générale. Les données sont assez bien réparties autour de la moyenne empirique.

Dans la section suivante, on essaie de trouver des lois de probabilités pour modéliser l'évolution de ces réalisations. Parmi elles, deux se distinguent :

1. La loi géométrique $\mathcal{G}(p)$;
2. La loi binomiale négative $\mathcal{BN}(r, p)$.

QUESTION 2. Dans la **question 1.3**, nous avons appliqué le graphe de probabilités de la fonction de répartition de la loi géométrique $\mathcal{G}(p)$ aux deux groupes de données. Nous en avons déduit que $\mathcal{G}(p)$ était un modèle plausible pour le **groupe 1**, mais pas pour le **groupe 2**.

Poursuivons cette démarche dans cette section.

 Se référer à : P2_Q2_Determination_Estimateur.r

Groupe 1.

La loi géométrique (cas particulier de la loi binomiale négative pour $r = 1$) est un modèle suffisamment plausible pour ce groupe. Par la méthode des moments, on trouve l'estimateur :

$$\tilde{p}_{n_1} = \frac{1}{\bar{X}_n} = \frac{1}{6.498} \simeq 0.1538$$

Puisque la loi géométrique est un cas particulier de la loi binomiale négative, la méthode des moments conçue à la **question 1.5** s'applique aussi à ce groupe. Elle donne une estimation de :

$$\tilde{p}_{n_2} = \frac{\bar{X}_n}{\bar{X}_n + S_n^2} \simeq 0.1528$$

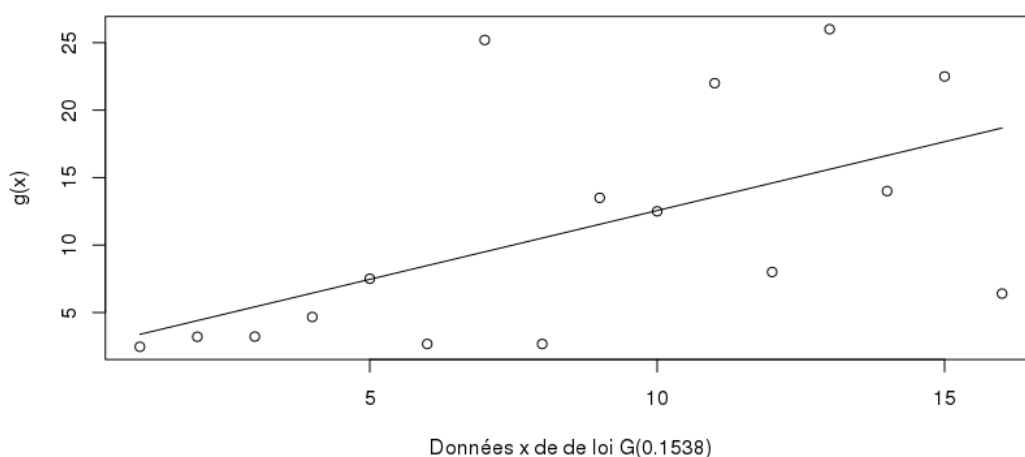
Pour rappel, le paramètre trouvé par le graphe de probabilités $(\tilde{p}_n)_{g1}$ dans la première section est 0.1508, soit une valeur à peu près égale. À chaque fois qu'un individu de l'échantillon fixe un mot, il y a donc un peu plus de 15% de chances qu'il en déduise le thème.

Au seuil de $\alpha = 5\%$, on réutilise l'intervalle de confiance déterminé dans la **question 1.2** et appliqué au cas particulier de la loi géométrique qui donne :

$$[0.1510331; 0.1567018]$$

On en déduit que la régression linéaire du graphe de probabilités a légèrement sous-estimé le paramètre de la loi géométrique du **groupe 1** avec une probabilité de 95%.

Vérification du paramètre p par simulation



Le paramètre obtenu par la régression linéaire (tronquée) est $p_n \simeq 0.23$. Cette méthode de vérification perd de son intérêt car l'échantillon est réduit et sensible aux valeurs aberrantes.

Groupe 2.

Cette fois-ci, le modèle géométrique n'est pas suffisant comme l'a montré la régression linéaire appliquée à la **question 1.2**. On s'intéresse à sa version généralisée avec un paramètre r en plus : c'est la loi binomiale négative $\mathcal{BN}(r, p)$.

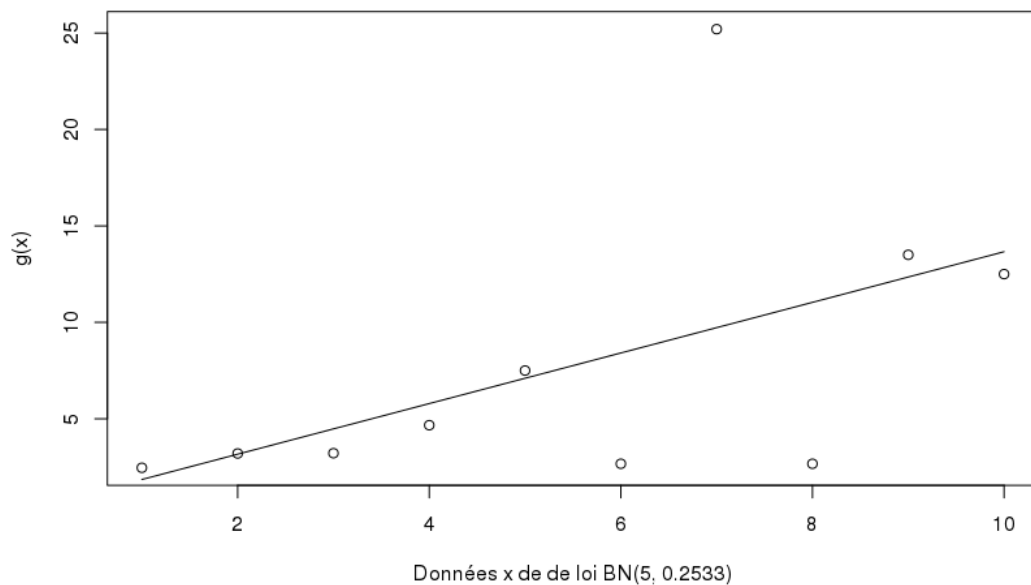
Le diagramme en colonnes possède d'ailleurs l'allure de la fonction de masse de la loi binomiale négative avec une moyenne environ égale à 19. Via la méthode de la **question 1.5**, on approche les paramètres r et p avec les estimateurs des moments \tilde{r}_n et \tilde{p}_n :

$$\tilde{r}_n = \left\lceil \frac{\overline{X_n^2}}{\overline{X_n} + S_n^2} \right\rceil \simeq \lceil 4.88 \rceil = 5 \quad \text{et} \quad \tilde{p}_{n_1} = \frac{\overline{X_n}}{\overline{X_n} + S_n^2} \simeq 0.2476$$

On arrondit \tilde{r}_n à l'entier le plus proche car c'est un paramètre entier. Pour avoir des paramètres « plus cohérents entre eux », on peut aussi estimer p à partir de \tilde{r}_n :

$$\tilde{p}_{n_2} = \frac{\tilde{r}_n}{\overline{X_n}} = \frac{5}{19.738} \simeq 0.2533$$

Vérification du paramètre p par simulation



Le paramètre obtenu par la régression linéaire (tronquée) est $p_n \simeq 0.2379$. Là encore, on a enlevé le dernier tiers des valeurs que la division rend aberrantes. Pour autant, la taille de l'échantillon rend là aussi la méthode peu fiable.

QUESTION 3. On admet ici que les variables $\{X_i\}_{1 \leq i \leq n}$ du **groupe 1** suivent une loi géométrique et que les variables $\{Y_i\}_{1 \leq i \leq n}$ du **groupe 2** suivent une loi binomiale négative :

$$X_i \hookrightarrow \mathcal{G}(0.1538) \quad \text{et} \quad Y_i \hookrightarrow \mathcal{BN}(5, 0.2533)$$

Grâce à l'expression respective de l'espérance de ces deux lois, on en déduit que :

$$\mathbb{E}[X_i] = \frac{1}{p} = \frac{1}{0.1538} \simeq 6.502 \quad \text{et} \quad \mathbb{E}[Y_i] = \frac{r}{p} = \frac{5}{0.2533} \simeq 19.739$$

Conclusion.

Dans le **groupe 1**, il faut donc en moyenne 6.5 fixations pour identifier un type de texte tandis que le **groupe 2** a besoin de 19.7 fixations en moyenne, c'est-à-dire environ trois fois plus.

Les modèles étendent les réalisations des échantillons et permettent de connaître la probabilité que le nombre de fixations soit supérieur à 10.

Calculons cette probabilité pour les deux lois.

$$\begin{aligned} P(X_i > 10) &= (1 - p)^{10} \\ &= (1 - 0.1538)^{10} \\ &= 0.8462^{10} \\ &= 0.1882479 \end{aligned}$$

$$\begin{aligned} P(Y_i > 10) &= 1 - \sum_{k=r}^{10} \binom{k-1}{r-1} p^r (1-p)^{k-r} \\ &= 1 - \sum_{k=5}^{10} \binom{k-1}{4} 0.2533^5 0.7467^{k-5} \\ &= 0.67554 \end{aligned}$$

Dans le **groupe 1**, la probabilité que le nombre de fixations soit supérieur à 10 est faible. Cela s'explique notamment par le fait que la majorité des réalisations est concentrée au début de la série. La médiane empirique du **groupe 1** vaut d'ailleurs 5. Ainsi, ce groupe n'a pas eu de problèmes majeurs pour identifier les thèmes des textes qui lui ont été soumis.

Au contraire, dans le **groupe 2**, la probabilité que le nombre de fixations soit supérieur à 10 est assez élevé.

QUESTION 2. à voir en fonction de la q4. **QUESTION 3.** SIMULATION.