REVUE DE STATISTIQUE APPLIQUÉE

ANIS ABI FARAH

Un test pour le contrôle de la qualité du travail dans un recensement

Revue de statistique appliquée, tome 22, nº 1 (1974), p. 67-86

<http://www.numdam.org/item?id=RSA_1974__22_1_67_0>

UN TEST POUR LE CONTROLE DE LA QUALITÉ DU TRAVAIL DANS UN RECENSEMENT

Anis ABI FARAH
Université Libanaise, Faculté des Sciences, Hadeth-Beyrouth Liban

1 – INTRODUCTION

Dans un recensement, la qualité du travail peut être mesurée par la proportion d'unités statistiques recensées effectivement ; ou encore par la proportion p d'unités statistiques oubliées ; en général p varie de 2 à 6 %.

L'objet de cet article est de présenter une méthode de contrôle de qualité. On compare entre eux plusieurs modèles probabilistes, afin de choisir le meilleur modèle qui servira pour la construction d'un test classique ou séquentiel. Ce test porte sur les valeurs que peut prendre la proportion p, il nous permet alors de contrôler la qualité du travail au fur et à mesure qu'il est fourni.

2 –

2.1 – **Description de la méthode de contrôle.**

Supposons qu'on veuille recenser les logements dans une ville donnée V. La ville V est considérée comme étant quadrillée par un quadrillage aussi fin qu'on le désire ; (on pourrait se servir des tournées déterminées par les rues), de telle sorte que chaque logement appartienne à un pâté de logements et à un seul.

Un enquêteur E_1, est envoyé pour recenser les logements dans une région R_1 de V comprenant f_1 pâtés, en remplissant un questionnaire approprié pour chaque logement.

Un enquêteur E_2 est envoyé pour recenser indépendamment de E_1 les logements d'une sous-région R_2 de R_1 formée de f_2 pâtés choisis au hasard parmi les f_1 pâtés de R_1. On a bien $R_2 \subseteq R_1$ et $f_2 \leqslant f_1$.

Soit X l'ensemble des logements de R_2 recensés par E_1, et soit Y l'ensemble des logements recensés par E_2 dans la même région R_2. On peut écrire la relation suivante :

$$R_2 = X \Delta Y + X \cap Y + C_{R_2}(X \cup Y)$$
$$= (X - Y) + (Y - X) + X \cap Y + C_{R_2}(X \cup Y).$$

Les ensembles au second membre sont deux à deux disjoints.

Remarque : L'indépendance entre E_1 et E_2 peut être assurée en les choisissant au hasard parmi l'ensemble des recenseurs, et en veillant à ce qu'ils ne soient pas courant du fait qu'ils recensent la même région R_2.

2.2. – **Modèles mathématiques.**

Soient p_1 et p_2 les probabilités d'oubli d'un logement quelconque par les enquêteurs E_1 et E_2 respectivement. Une expérience aléatoire consiste à considérer chaque logement de R_2 ; il peut appartenir (indépendamment des autres logements) à $X - Y$ ou à $Y - X$ ou à $X \cap Y$ ou enfin à $R_2 - X \cup Y$.

Posons :

$$N_1 = \text{card}\,(X - Y) \qquad N_2 = \text{card}\,(Y - X)$$
$$N_3 = \text{card}\,(X \cap Y) \quad \text{et} \quad N_4 = \text{card}\,(R_2 - X \cup Y)$$

Ayant pour réalisations n_1, n_2, n_3 et n_4 respectivement.

Si n est le nombre théorique de logements dans R, la loi de probabilité de (N_1, N_2, N_3, N_4) est multinomiale ; ses paramètres sont $[n ; (1 - p_1)\,p_2,\ (1 - p_2)\,p_1,\ (1 - p_1)(1 - p_2),\ p_1 p_2]$.$N_1$, N_2, N_3 et N_4 sont quatre variables aléatoires liées par la relation $N_1 + N_2 + N_3 + N_4 = n$. Autrement dit, ce sont les coordonnées d'un point aléatoire d'un sous-espace linéaire à trois dimensions.

2.3. – **Loi conditionnelle de** (N_1, N_2, N_3).

Un inconvénient du modèle décrit au § 2.2. est qu'on ne connait pas n ; sinon le problème serait partiellement résolu. L'information fournie par la réalisation de N_1, N_2 et N_3 devrait nous permettre de définir un autre modèle utilisable par la suite pour la construction du test.

On considère pour cela la loi conditionnelle de (N_1, N_2, N_3) sachant que $N_1 + N_2 + N_3 = l$.

Soit (N_1, N_2, N_3, N_4) de loi multinomiale $[n ; r_1, r_2, r_3, r_4]$

En posant :

$$r_1 = (1 - p_1)\,p_2,\ r_2 = (1 - p_2)\,p_1,\ r_3 = (1 - p_1)(1 - p_2) \quad \text{et} \quad r_4 = p_1 p_2.$$

On a :

$$P[N_1 = n_1, N_2 = n_2, N_3 = n_3/N_1 + N_2 + N_3 = l]$$

$$= \frac{P[N_1 = n_1, N_2 = n_2, N_3 - n_3, N_1 + N_2 + N_3 = l]}{P[N_1 + N_2 + N_3 = l]}$$

$$= \frac{\dfrac{n!}{n_1!\, n_2!\, n_3!\, (n - n_1 - n_2 - n_3)!}\; r_1^{n_1}\; r_2^{n_1}\; r^{l-n_1-n_2}}{\dfrac{n!}{l!\,(n-l)!}\; (r_1 + r_2 + r_3)^l\, r_4^{n_4}}$$

$$= \binom{l}{n_1\; n_2\; n_3} \left(\frac{r_1}{1 - r_4}\right)^{n_1} \left(\frac{r_2}{1 - r_4}\right)^{n_2} \left(\frac{r_3}{1 - r_4}\right)^{l-n_1-n_2}$$

avec $n_i \geqslant 0$, $\sum_{i=1}^{3} n_i = l$; et $r_i \geqslant 0$, $\sum_{i=1}^{3} r'_i = 1$, $r'_i = r_i/(1 - r_4)$.

Donc $[N_1, N_2, N_3/N_1 + N_2 + N_3 = l]$ suit une loi multinomiale de paramètres

$$\left[l\;;\frac{r_1}{1 - r_4}\quad \frac{r_2}{1 - r_4}\quad \frac{r_3}{1 - r_4}\right]$$

3. –

3.1 – **Modèle à retenir pour faire le test portant sur les valeurs de** p_1

Un critère pour retenir un modèle M_1 plutôt qu'un autre M_2 consiste à comparer les quantités d'information au sens de Fisher des deux modèles au sujet du paramètre p_1. Celui qui a une quantité d'information plus grande sera préféré à l'autre.

Quelle est alors la quantité d'information $I(M_i)$ fournie par une réalisation d'une variable aléatoire Z au sujet du paramètre p_1, si Z a comme loi de probabilité M_i :

1/ M_1 : Z suit la loi décrite dans le § 2.2.

2/ M_2 : Z suit la loi décrite dans le § 2.3.

3/ M_3 : On considère la variable aléatoire $N_1 + N_2/N_1 + N_2 + N_3 = l$ elle suit une loi binomiale $\left(l\;;\dfrac{r_1 + r_2}{1 - r_4}\right)$

On a :

$$1/\ I(M_1) = -E\left(\frac{\partial^2 \log L_1}{\partial p_1^2}\right) \text{ avec } L_1 = \left(\frac{n!}{n_1!\, n_2!\, n_3!\, n_4!}\right) r_1^{n_1}\, r_2^{n_2}\, r_3^{n_3}\, r_4^{n_4}$$

Ce qui donne $I(M_1) = \dfrac{n}{p_1(1-p_1)}$.

2/ $I(M_2) = -E\left(\dfrac{\partial^2 \log L_2}{\partial p_1^2}\right)$ avec $L_2 = \left(\dfrac{l\,!}{n_1!\,n_2!\,n_3!}\right)$

$$\left(\frac{r_1}{1-r_4}\right)^{n_1}\left(\frac{r_2}{1-r_4}\right)^{n_2}\left(\frac{r_3}{1-r_4}\right)^{n_3}$$

Le calcul donné en Annexe A_1 conduit à :

$$I(M_2) = \frac{l}{1-p_1p_2}\left[\frac{1}{1-p_1} + \frac{1-p_2}{p_1}\right] - \frac{l\,p_2^2}{(1-p_1p_2)}$$

3/ $I(M_3) = -E\,\dfrac{\partial^2 \log L_3}{\partial p_1^2}$ avec $L_3 = \binom{n}{k}$

$$\left[\frac{(1-p_1)\,p_2 + (1-p_2)\,p_1}{1-p_1p_2}\right]^k \left[\frac{(1-p_1)(1-p_2)}{1-p_1p_2}\right]^{l-k}$$

Le calcul donné en Annexe A_2 donne :

$$I(M_3) = \frac{l}{1-p_1p_2}\left[\frac{(1-2p_2)^2}{(1-p_1)\,p_2 + (1-p_2)\,p_1} + \frac{1-p_2}{1-p_1}\right] - \frac{l\,p_2^2}{(1-p_1p_2)^2}$$

Remarque : Dans le modèle M_1, E_2 ne nous renseigne par sur la valeur du paramètre p_1.

3.2. – Comparaison de $I(M_1)$ et $I(M_2)$.

Si $p_1 p_2$, p_1^2 et p_2^2 sont négligeables à côté de p_1 ou de p_2 on a :

$$I(M_2) \simeq l\,\frac{1+p_1-p_2}{p_1(1-p_1)}$$

$$\frac{I(M_2)}{I(M_1)} \simeq \frac{l}{n}(1+p_1-p_2)$$

3.3. – Comparaison de $I(M_1)$ et de $I(M_3)$

En négligeant p_1p_2 ; p_1^2 et p_2^2 à côté de p_1 ou de p_2 comme on a fait au § précédent on obtient :

$$I(M_3) \simeq l\left[\frac{1-4p_2}{p_1p_2} + \frac{1-p_2}{1-p_1}\right]$$

$$= l\left[\frac{1-3\,p_2}{(p_1+p_2)(1-p_1)}\right]$$

D'où :

$$\frac{I(M_3)}{I(M_1)} \simeq \frac{l}{n} p_1 \left[\frac{1 - 3\,p_2}{p_1 + p_2}\right]$$

Exemple numérique : $p_1 = 0{,}02$ et $p_2 = 0{,}06$ donne $\dfrac{I(M_3)}{I(M_1)} \simeq \dfrac{l}{n}\, 0{,}2050$.

Il y a environ 4/5 de l'information qui est perdue si l'on adopte le modèle M_3 au lieu du modèle M_1. Or n n'est pas connu et l'on ne peut prendre dans ce cas que le modèle M_2.

3.4. – **Etude du cas** $p_1 = p_2 = p$.

Si les deux enquêteurs E_1 et E_2 fournissent théoriquement un travail de même qualité, en d'autres termes si $p_1 = p_2 = p$, les quantités d'information fournies par les modèles M_1, M_2 et M_3 au sujet du paramètre p, sont respectivement $I'(M_1)$, $I'(M_2)$ et $I'(M_3)$ où :

1/ $$I'(M_1) = -E \frac{\partial^2 \log L_1}{\partial p^2}$$

avec $$L_1 = \binom{n}{n_1\ n_2\ n_3\ n_4} [p\,(1-p)]^{n_1+n_2} \times [(1-p)^2]^{n_3}\,(p^2)^{n_4}$$

Le calcul (cf Annexe A_3) conduit à

$$I'(M_1) = \frac{2n}{p(1-p)}$$

2/ $$I'(M_2) = -E\left(\frac{\partial^2 \log L_2}{\partial p^2}\right)$$

avec $$L_2 = \binom{l}{n_1\ n_2\ n_3} \left[\frac{p(1-p)}{1-p^2}\right]^{n_1} \times \left[\frac{p(1-p)}{1-p^2}\right]^{n_2} \left[\frac{(1-p)^2}{1-p^2}\right]^{n_3}$$

Le calcul (cf Annexe A_4) donne

$$I'(M_2) = \frac{2l}{p(1-p)} \cdot \frac{1}{(1+p)^2}$$

3/ $$I'(M_3) = -E\left(\frac{\partial^2 \log L_3}{\partial p^2}\right) \text{ avec } L_3 = \binom{l}{k}\left(\frac{2p}{1+p}\right)^k \left(\frac{1-p}{1+p}\right)^{l-k}$$

Le calcul (cf Annexe A_5) donne encore :

$$I'(M_3) = \frac{2l}{p(1-p)} \cdot \frac{1}{(1+p)^2}$$

Remarque 1 – On a :

$$\frac{I'(M_1)}{I(M_1)} = 2\,\frac{n}{p(1-p)} \cdot \frac{p_1(1-p_1)}{n}$$

$$= 2\,\frac{p_1(1-p_1)}{p(1-p)}$$

La quantité d'information fournie par le modèle M_1 au sujet de p se trouve alors doublée.

Remarque 2 – De même si l'on compare $I'(M_1)$ à $I'(M_2)$ on obtient :

$$\frac{I'(M_1)}{I'(M_2)} = \frac{2n}{p(1-p)} \cdot \frac{p(1-p)(1+p)^2}{2l}$$

$$= \frac{n}{l}\,(1+p)^2$$

$$\simeq \frac{n}{l}\,(1+2p)$$

Il vaudrait mieux utiliser le modèle M_1 plutôt que le modèle M_2. Mais, vu qu'on ne peut pas obtenir une réalisation de N_4 on est obligé d'utiliser le modèle M_2 ; il en résulte une perte d'information relative de l'ordre de 2p si n est à peu près égal à *l*.

Remarque 3 – Les quantités d'information $I'(M_2)$ et $I'(M_3)$ sont égales : les second et troisième modèles sont équivalents, en ce sens qu'ils apportent la même quantité d'information au sujet de p. Ceci est dû à la propriété énoncée au § 3.5.

3.5 – **Proposition.**

$N_1 + N_2/_{N_1 + N_2 + N_3 = l}$ *est un résumé exhaustif au sujet de p.*

Démonstration – En effet la fonction de vraisemblance L est :

$$L = \binom{l}{n_1\ n_2\ n_3}\left(\frac{p}{1+p}\right)^{n_1+n_2}\left(\frac{1-p}{1+p}\right)^{n_3}$$

$$L = \binom{l}{n_1+n_2}\left(\frac{2p}{1+p}\right)^{n_1+n_2}\left(\frac{1-p}{1+p}\right)^{n_3}\cdot\binom{n_1+n_2}{n_1}\frac{1}{2}^{n_1+n_2}$$

$$= f(n_1+n_2, p)\,.\,\phi(n_1, n_2)$$

$$= p\,[N_1+N_2 = n_1+n_2]\,.\,\phi(n_1, n_2/_{N_1+N_2 = n_1+n_2})$$

avec $\phi(n_1, n_2/_{N_1+N_2 = n_1+n_2})$ est indépendante du paramètre p.

3.6. – **Généralisation de la propriété donnée dans le § 3.5.**

Proposition – Si $(N_1, N_2, \ldots N_{h-1}, N_h)$ *est de loi multinomiale* $(n\,;\, p_1, p_2, \ldots, p_h)$ *et si* $p_1 = \alpha_1\, g(p),\ p_2 = \alpha_2\, g(p), \ldots, p_{h-1} = \alpha_{h-1} g(p)$, *alors* $N_1 + N_2 + \ldots + N_{h-1}$ *est un résumé exhaustif au sujet de p.*

Démonstration – La fonction de vraisemblance L est donnée par :

$$L = \binom{n}{n_1\ n_2 \ldots n_h} p_1^{n_1}\, p_2^{n_2} \ldots p_h^{n_h}$$

$$= \binom{n}{n_1 + n_2 + \ldots n_{h-1}} (p_1 + p_2 + \ldots + p_{h-1})^{n_1 + n_2 + \ldots + n_{h-1}}\, p_h^{n_h}$$

$$\times \binom{n_1 + n_2 + \ldots n_{h-1}}{n_1\ n_2 \ldots n_{h-1}} \left(\frac{p_1}{p_1 + \ldots + p_{h-1}}\right)^{n_1} \ldots \left(\frac{p_{h-1}}{p_1 + \ldots + p_{h-1}}\right)^{n_h}$$

et en remplaçant les p_i par leurs valeurs on obtient :

$$L = \binom{n}{n_1 + n_2 + \ldots + n_{h-1}} [(\alpha_1 + \alpha_2 + \ldots + \alpha_{h-1})\, g(p)]^{n_1 + \ldots + n_{h-1}}\, p_h^{n_h}$$

$$\times \binom{n_1 + \ldots + n_{h-1}}{n_1\,, \ldots, n_{h-1}} \left(\frac{\alpha_1}{\alpha_1 + \ldots + \alpha_{h-1}}\right)^{n_1} \ldots \left(\frac{\alpha_{h-1}}{\alpha_1 + \ldots + \alpha_{h-1}}\right)^{n_{h-1}}$$

$$= P\,[N_1 + N_2 + \ldots + N_{h-1} = l,\ p] \times \phi\,(n_1, n_2, \ldots, n_h)$$

où P est la loi de probabilité de $N_1 + N_2 + \ldots + N_{h-1}$ qui dépend du paramètre p ; et ϕ est la loi de probabilité de

$$N_1, N_2, \ldots, N_{h-1}/N_1 + N_2 + \ldots + N_{h-1} = l$$

ϕ étant fonction des n_i et non de p.

Remarque 4 – Dans le cas où $\alpha_1 = \alpha_2 = \ldots = \alpha_{h-1} = 1$,

$$g(p) = \frac{p(1-p)}{1-p^2} = \frac{p}{1+p}$$

et h = 3, on obtient le cas décrit au § 3.5.

3.7. – **Généralisation de la méthode décrite au § 2.1.**

On suppose qu'on a envoyé m enquêteurs $E_1, E_2, \ldots, E_m$ pour recenser indépendamment les uns des autres les logements de la région R_2. R_2 se trouve alors partitionnée en 2^m parties disjointes

$A_j\,;\, j = 1, 2, \ldots, 2^m$. On a : $R_2 = \bigcup_{i=1}^{2^m} A_i$ et $\cap A_i A_j = \phi$, $i \neq j$.

Soit A_1 la région des logements oubliés par tous les enquêteurs sans exception, et N_1 = card A_1 (l'homologue du N_4 de 2.2.). Si p_1 est la probabilité d'oubli de E_1, p_2 celle de $E_2, \ldots, p_m$ celle de E_m, alors la probabilité qu'un logement appartienne à A_1 est $p_1, p_2, \ldots, p_m$.

De même, soit A_2 la région des logements oubliés par tous les enquêteurs sauf le 1^{er}. et soit N_2 = card A_2.

On a :

$$p(\text{logement} \in A_2) = (1 - p_1) \prod_{i \neq 1} p_i$$

et ainsi de suite pour les autres régions, la dernière région est A_{2m} dont le cardinal est N_{2m} et telle que $P[\text{logement} \in A_{2m}] = \prod_{i=1}^{m} (1 - p_i)$. On constate que les probabilités respectives des régions A_j sont les termes du développement du produit : $[p_1 + (1 - p_1)]\,[p_2 + (1 - p_2)] \ldots [p_m + (1 - p_m)]$.

Alors : $(N_1, N_2, \ldots, N_{2m})$ est de loi multinomiale

$$\left[n\ ;\ \prod_{i=1}^{m} p_i, (1 - p_1) \quad p_i, \ldots, \prod_{i=1}^{m} (1 - p_i)\right].$$

Autrement dit $(N_1, N_2, \ldots, N_{2m})$ est de loi mult. $[n\ ; r_1, \ldots, r_j, \ldots, r_{2m}]$ en posant $r_1 = \prod_{i=1}^{m} p_i$, $r_2 = (1 - p_1) \prod_{i=1}^{m} p_i, \ldots, r_{2m} = \prod_{i=1}^{m} (1 - p_i)$. Tout r_j s'écrit sous la forme $p_1\ r'_k$ ou $(1 - p_1)\ r'_k$; les r'_k étant les termes du developpement : $[p_2 + (1 - p_2)]\,[p_3 + (1 - p_3)] \ldots [p_m + (1 - p_m)]$. Ce qui fait que $\sum_{k=1}^{2m-1} r'_k = 1$.

En calculant la quantité d'information $I(p_1)$ on trouve alors :

$$I(p_1) = \sum_{k=1}^{2m-1} E\left[\frac{n_k}{p_1^2} + \frac{n - n_k}{(1 - p_1)^2}\right]$$

$$= \sum_{k=1}^{2m-1} \left[\frac{np_1\ r_k}{p_1^2} + \frac{n(1 - p_1)\ r_k}{(1 - p_1)^2}\right]$$

$$= \frac{n}{p_1\,(1 - p_1)}$$

Remarque – La quantité d'information au sujet de p_1 est seulement fournie par E_1 ; les autres enquêteurs n'apportent aucune information supplémentaire au sujet de p_1.

3.8. – **Etude du cas où $p_1 = p_2 = \ldots = p_m$**

On a alors : $(N_1, N_2, \ldots, N_{2m})$ qui suit une loi multinomiale dont les paramètres sont :

$$[n ; \underbrace{p^m, (1-p)\, p^{m-1}}_{\text{répétée } \binom{m}{1} \text{ fois}}, \ldots, \underbrace{p^{m-h}(1-p)^h}_{\text{répétée } \binom{m}{h} \text{ fois}}, \ldots, (1-p)^m]$$

La quantité d'information I(p) fournie alors au sujet de p par une réalisation de $(N_1, N_2, \ldots, N_{2m})$ est :

$$I(p) = -E\left(\frac{\partial^2 \log L}{\partial p^2}\right) = \frac{n.m}{p(1-p)}$$

L étant la fonction de vraisemblance de ce modèle, (cf Annexe A_6)

3.9. – **Deuxième méthode.**

Soit h_1 le nombre de logements oubliés par $E_1, \ldots, h_m$ le nombre de logements oubliés par E_m. $h_1, h_2, \ldots h_m$ sont des variables aléatoires indépendantes en probabilité et h_i suit une loi binomiale de paramètres n et p quelque soit $i = 1, 2, \ldots, m$.

La fonction de vraisemblance est alors donnée par :

$$L = \prod_{i=1}^{m} C_n^{h_i}\, p^{h_i}\, (1-p)^{n-h_i}$$

d'où

$$\frac{\partial^2 (\log L)}{\partial p^2} = \sum_{i=1}^{m} \left[-\frac{h_i}{p^2} - \frac{n-h_i}{(1-p)^2}\right]$$

et

$$I(p) = -E\left(\frac{\partial^2 \log L}{\partial p^2}\right) = \sum_{i=1}^{m} \left[\frac{np}{p^2} + \frac{n(1-p)}{(1-p)^2}\right] = \frac{n.m}{p^2}.$$

Remarque – On constate que les deux modèles 3.8. et 3.9. donnent la même quantité d'information au sujet de p.

3.10. – **Proposition.**

$$(N_2 m, \ldots, N_2 / N_2 m + N_2 m_{-1} + \ldots + N = l)$$

suit une loi multinomiale ; et

$$I(p) = \frac{ml}{p(1-p)} \frac{1}{(1-p^m)^2} [(1-p)(1+p+\ldots+p^{m-1}) - mp^{m-1}]$$

Cette proposition est démontrée en Annexe A_7.

Remarque – Si m est suffisamment grand, on a :

$$I(p) \simeq \frac{m.l}{p(1-p)} .$$

Au contraire, si m = 2, on a :

$$I(p) = \frac{2l}{p(1-p)} \cdot \frac{1}{(1+p)^2}$$

formule déjà trouvée au § 3.4.

4.

4.1. –

On considère les hypothèses des § 2.3., et 3.4. On a vu que dans le cas où $p_1 = p_2 = p$ les modèles M_2 et M_3 sont équivalents et que

$$(N_1 + N_2 / N_1 + N_2 + N_3 = l)$$

est un résumé exhaustif au sujet de p (§ 3.5) p étant la probabilité d'oubli des recenseurs E_1 et E_2. Donc on a intérêt à ce que les deux enquêteurs E_1 et E_2 aient la même probabilité p afin de pouvoir se servir du modèle M_3 pour construire le test portant sur la valeur de ce paramètre p.

4.2. – **Test de l'hypothèse H_0 : $p_1 = p_2$ contre H_1 : $p_1 \neq p_2$.**

La loi de probabilité de N_1 sachant que $N_1 + N_2 = l$ est binomiale ; de paramètres

$$\left(l, \frac{p_2 (1-p_1)}{1 - p_1 p_2 - (1-p_1)(1-p_2)}\right) = (l, q)$$

Si l'hypothèse nulle est vraie $p_1 = p_2 \Longrightarrow q = \frac{1}{2}$; et si l'hypothèse alternative est vraie. $q \neq \frac{1}{2}$: $0 \leqslant q \leqslant 1$.

D'ailleurs on voit facilement que $q > 1/2$ signifie $p_2 > p_1$, et $q < 1/2$ l'inverse.

D'où la région critique w du test est définie par :

$\frac{L_0}{L_1} < c$ à l'intérieur de la région critique, c'est-à-dire :

$$\frac{\binom{l}{k} \left(\frac{1}{2}\right)^l}{\binom{l}{k} q^k (1-q)^{l-k}} < c$$

où k désigne N_1 ou N_2 (notations de 2.2. et 2.3).

$$l \log \frac{1}{2} - k \log q - (l - k) \log (1 - q) < c'$$

$$k\left(\log \frac{1 - q}{q}\right) < c''$$

si $\dfrac{1 - q}{q} > 1$ c'est-à-dire, $q < \dfrac{1}{2} \Longrightarrow$ w est définie par $k < c_1$

et si $\dfrac{1 - q}{q} < 1$ c'est-à-dire, $q > \dfrac{1}{2} \Longrightarrow$ w est définie par $k > c_2$

Donc on ne rejette pas l'hypothèse nulle si $c_2 > k > c_1$ et on la rejette dans le cas contraire.

Donc on obtient un test bilatéral symétrique.

Remarque – Si l'hypothèse nulle H_0 : $p_1 = p_2$ n'est pas rejetée on adopte le modèle M_3 du § 3. 4, c'est-à-dire, la loi de $N_1 + N_2/N_1 + N_2 + N_3 = l$ avec $p_1 = p_2 = p$ et on procède à la construction d'un test portant sur la valeur de p. Mais dans le cas où l'hypothèse nulle H_0 est rejetée, il faut adopter le modèle M_2 du § 3.1, en considérant la loi de $(N_1, N_2, N_3/N_1 + N_2 + N_3 = l)$ avec $p_1 \neq p_2$; et l'on construit alors un test portant sur les valeurs du paramètre p_1 en supposant que p_2 est connu.

Dans ce qui suit, on suppose que H_0 n'est pas rejetée et l'on procède à la construction d'un test classique et un test séquentiel portant sur la valeur de p. (Modèle M'_2 du § 3.4).

Dans le second cas, deux tests analogues peuvent être construits facilement utilisant le modèle M_2 de § 3.1.

4.3. – **Test classique de l'hypothèse nulle** H_0 :

$p \leqslant P_0$ contre l'hypothèse alternative H_1 : $p \geqslant P_1$

En adoptant le modèle M'_3 du § 3.4., on a :

$$L = P\,[N_1 + N_2 = k/N_1 + N_2 + N_3 = l]$$

$$= \binom{l}{k}\left(\frac{2p}{1 + p}\right)\left[\frac{(1 - p)}{1 + p}\right]^{l-k} \quad k = 0, 1, 2, \ldots, l$$

$N_1 + N_2$ est la somme des nombres des logements oubliés par E_1 et recensés par E_2 et ceux, oubliés par E_2 et retrouvés par E_1.

On a : $\dfrac{L_0}{L_1} < c$ dans la région critique w.

Ce qui donne dans le cas où $P_0 < P_1$, une région critique définie par $k >$ constante C.

4.4. – **Application numérique et calcul de la puissance du test.**

On admet en général dans les recensements une proportion d'unités oubliées allant de 2 à 6 %. On va donner dans ce § une liste des tests de l'hypothèse $p \leqslant P_0$ contre l'hypothèse alternative $p \geqslant P_1$, où $P_0 = 0{,}02$ et P_1 prenant successivement les valeurs 0,03 ; 0,04 ; 0,05 ; 0,06 ; 0,07 ; et 0,08. (Table n° 1) On va indiquer dans cette table la taille l de l'échantillon, la région critique, l'erreur α de première espèce et la puissance du test.

Table N° 1

Puissances des tests, $1 - \beta$

Taille l	Valeur Crit. C	Risque α	p_1					
			0,03	0,04	0,05	0,06	0,07	0,08
10	1	0,056	0,1120	0,1765	0,2454	0,3153	0,3834	0,4487
20	2	0,042	0,1077	0,1958	0,2958	0,3984	0,4964	0,5858
50	4	0,046	0,1652	0,3396	0,5234	0,6818	0,8002	0,8809
100	7	0,043	0,2282	0,5070	0,7464	0,8908	0,9589	0,9862
150	10	0,035	0,2598	0,6092	0,8550	0,9595	0,9909	0,9983
200	12	0,053	0,3840	0,7730	0,9490	0,9922	0,9991	0,9999
300	17	0,051	0,4844	0,8896	0,9850			
400	22	0,045	0,5585					

4.5. – **Test séquentiel de l'hypothèse nulle** H_0

$p \leqslant P_0$ contre H_1 : $p \geqslant P_1$ avec $P_0 < P_1$

On considère ici aussi la variable aléatoire

$$N_1 + N_2/N_1 + N_2 + N_3 = l$$

qui suit une loi binomiale.

Si H_0 est vraie :

$$N_1 + N_2/N_1 + N_2 + N_3 = l$$

a une loi binomiale $\left[l \,;\, r_0 = \dfrac{2P_0}{1 + P_0}\right]$

et si H_1 est vraie, elle a une loi binomiale de paramètres

$$\left[l \,;\, r_1 = \frac{2P_1}{1 + P_1}\right]$$

En prenant pour erreur α et β (de première et deuxième espèce respectivement) on obtient :

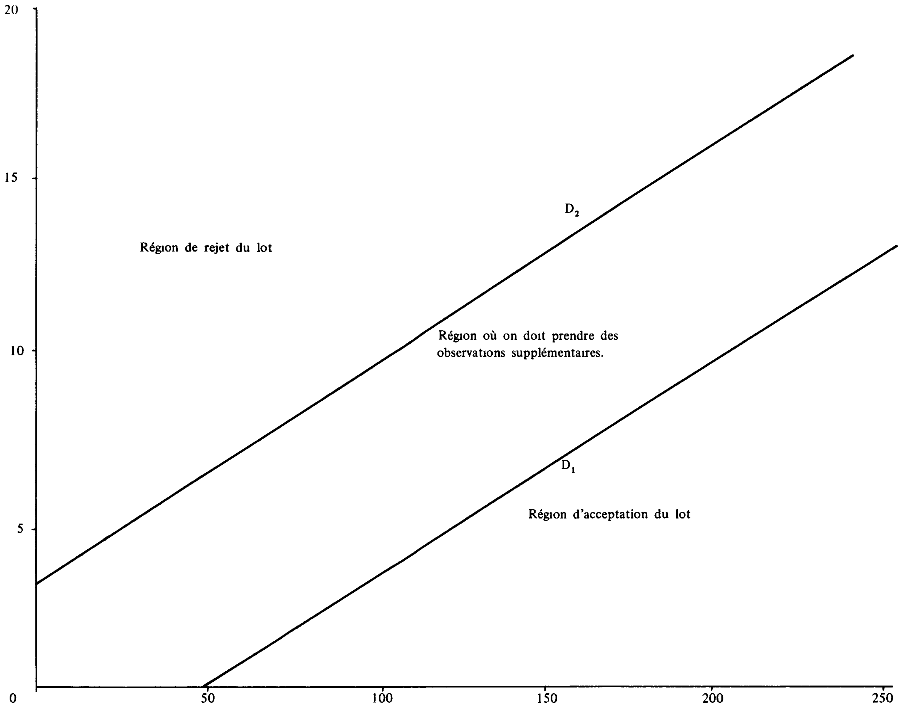

Figure 1 – Graphiques des droites D_1 et D_2 définissant les trois régions du tests.

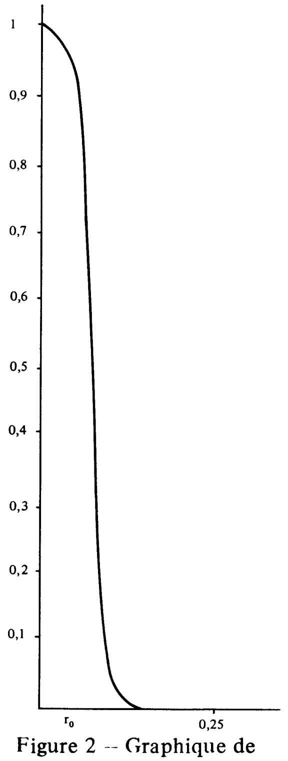

Figure 2 – Graphique de l'O.C.L(ρ) du test.

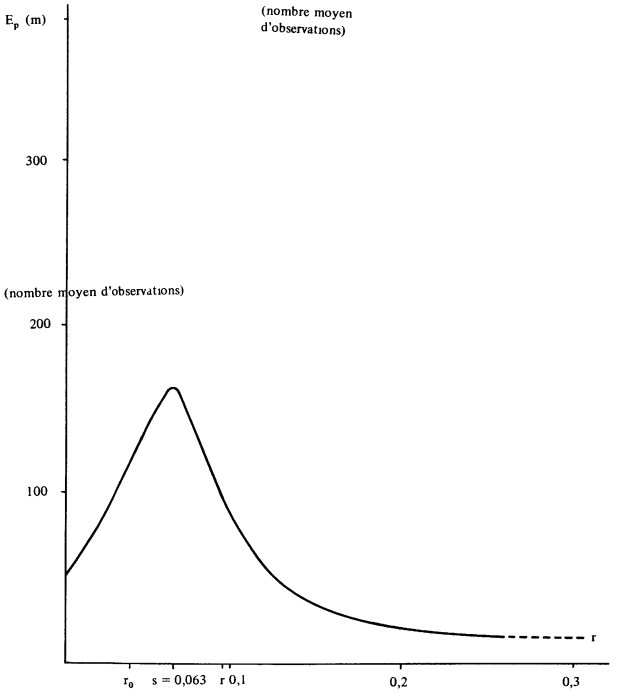

Figure 3 –

$$D_1 = \frac{\log \dfrac{\beta}{1-\alpha}}{\log \dfrac{r_1}{r_0} - \log \dfrac{1-r_1}{1-r_0}} + l \, \frac{\log \dfrac{1-r_0}{1-r_1}}{\log \dfrac{r_1}{r_0} - \log \dfrac{1-r_1}{1-r_0}} \leqslant k \leqslant$$

$$\frac{\log \dfrac{1-\beta}{\alpha}}{\log \dfrac{r_1}{r_0} - \log \dfrac{1-r_1}{1-r_0}} + l \, \frac{\log \dfrac{1-r_0}{1-r_1}}{\log \dfrac{r_1}{r_0} - \log \dfrac{1-r_1}{1-r_0}} = D_2$$

L'enquêteur E_2 continue à recenser tant que k varie entre les deux limites indiquées dans l'inégalité double : $D_1 < k < D_2$; et lorsqu'on a $k \leqslant D_1$ ou $k \geqslant D_2$, on arrête l'échantillonnage pour ne pas rejeter ou pour rejeter l'hypothèse nulle (respectivement).

4.6. – **Application numérique.**

Donnons aux paramètres du § 4.5, les valeurs numériques suivantes :

$$P_0 = 0{,}02 \ ; P_1 = 0{,}05 \ ; \alpha = 0{,}05 \qquad \text{et} \qquad \beta = 0{,}05.$$

D'où

$$r_0 = 0{,}0392 \qquad \text{et} \qquad r_1 = 0{,}0952$$

D_1 et D_2 ont pour équations respectives.

$$D_1 = -\,3{,}109 + 0{,}063\, l < k < 3{,}109 + 0{,}063\, l = D_2 .$$

Voir figure 1, 2 et 3 où l'on donne respectivement les graphes de D_1 et D_2, l'O.C du test et la courbe du nombre moyen d'observations necessaires pour le test.

5. CONCLUSION

On peut retenir le modèle M'_3 pour contrôler la qualité du recensement, à condition que les deux enquêteurs E_1 et E_2 fournissent un travail de même qualité. Cette condition peut être satisfaite en choisissant E_1 et E_2 au hasard dans un groupe homogène d'enquêteurs. On peut chaque fois tester si p_1 est égale à p_2 en faisant le test du § 4.2.

Un test séquentiel nous permet alors de décider après chaque observation faite par le second enquêteur (ou un groupe d'observations).

Le test fournit une réduction dans le nombre d'observations nécessaires par rapport au test classique (α et β étant les mêmes pour les deux tests).

Cette réduction est de l'ordre de 15 % au moins. C'est-à-dire que, dans le cas où $r = 0{,}063 = s$ et $\alpha = 0{,}05$ et $\beta = 0{,}05$, on a besoin de 200 observations pour le test classique et de 170 observations environ pour le test sequentiel. (voir Tableau n° 1 et figure n° 3).

Cependant, si le test du § 4.2 nous conduit à rejeter l'hypothèse nulle $p_1 = p_2$, alors on devra utiliser le modèle M_2 pour faire le test portant sur les valeurs de p_1, en supposant p_2 connu. On pourra construire d'une manière analogue un test classique et un autre séquentiel pour contrôler la qualité du recensement.

BIBLIOGRAPHIE

[1] DIRECTION CENTRALE DE LA STATISTIQUE AU LIBAN – Enquête par sondage sur la population active 1970 volume I.

[2] FELLER, W. – An introduction to probability and its applications. Wiley.

[3] KENDALL, M. STUART, A. – The advanced theory of statistics, London Griffin.

[4] KOHN, A. – A cost effectiveness model for air pollution control with a single stochastic variable. Journal of the American Statistical Association, March 1972.

[5] MINTON, G. – Vérification error in single sampling inspection plans for processing survey data. Journal of the American Statistical Association, March 1972.

[6] WALD, A. – Sequential analysis. John Wiley and Sons inc. Eighth Printing, October 1966.

[7] U.S. BUREAU OF THE CENSUS – United States Census of Population and Housing 1960 ; quality of preparatory operations, microfilming and coding, Washington, D.C. : U.S. Government Printing Office.

ANNEXES

A_1 Calcul de $I(M_2)$ du 3.1/2 :

$$\mathrm{Log}\, L_2 = \log \binom{l}{n_1\ n_2\ n_3} + n_1 \log r_1 + n_2 \log r_2 + n_3 \log r_3$$

$$- (n_1 + n_2 + n_3) \log (1 - r_4).$$

6

$$\frac{\partial \log L_2}{\partial p_1} = - \frac{n_1}{(1-p_1)\,p_2}\,p_2 + \frac{n_2\,(1-p_2)}{(1-p_2)\,p_1} - \frac{n_3\,(1-p_2)}{(1-p_1)\,(1-p_2)}$$

$$+ \frac{(n_1 + n_2 + n_3)}{1 - p_1\,p_2}\,p_2$$

$$\frac{\partial^2 (\log L_2)}{\partial p_1^2} = - \frac{n_1}{(1-p_1)^2} - \frac{n_2}{p_1^2} - \frac{n_3}{(1-p_1)^2} + \frac{(n_1 + n_2 + n_3)\,p_2^2}{(1-p_1\,p_2)^2}$$

$$I(M_2) = \frac{l}{(1-p_1)^2}\left[\frac{(1-p_1)\,p_2 + (1-p_1)\,(1-p_2)}{1 - p_1 p_2}\right]$$

$$+ \frac{l\,p_1}{p_1^2}\;\frac{(1-p_2)}{(1-p_1\,p_2)} - \frac{p_2^2}{(1-p_1\,p_2)}\;2.l$$

$$= \frac{l}{(1-p_1)} \times \frac{1}{1-p_1 p_2} + l\left[\frac{1-p_2}{p_1\,(1-p_1 p_2)}\right] - l\,\frac{p_2^2}{(1-p_1 p_2)}$$

D'où :

$$I(M_2) = \frac{l}{1 - p_1 p_2}\left[\frac{1}{1-p_1} + \frac{1-p_2}{p_1}\right] - \frac{l\,p_2^2}{(1-p_1 p_2)^2}$$

A_2 **Calcul de $I(M_3)$. du 3.1/3 :**

$$\text{Log } L_3 = \text{Log}\binom{n}{k} + k \log\,[(1-p_1)\,p_2 + (1-p_2)\,p_1]$$

$$+ (l - k)\,[\log\,(1-p_1) + \log(1-p_2)] - l\,(1 - p_1 p_2).$$

$$\frac{\partial \log L_3}{\partial p_1} = \frac{k(1-2p_2)}{p_2\,(1-p_1) + (1-p_2)\,p_1} - \frac{l-k}{1-p_1} + \frac{l\;p_2}{1-p_1 p_2}$$

$$\frac{\partial^2 (\log L_3)}{\partial p_1^2} = - \frac{k(1-2\,p_2)^2}{[(1-p_1)\,p_2 + (1-p_2)\,p_1]^2} - \frac{l-k}{(1-p_1)^2} + \frac{l\,p_2^2}{(1-p_1 p_2)^2}$$

$$- E\left(\frac{\partial^2 \log L_3}{\partial p_1^2}\right) = \frac{(1-2p_2)^2}{[(1-p_1)\,p_2 + (1-p_2)\,p_1]^2} \times \frac{l[1-p_1)\,p_2 + (1-p_2)\,p_1]}{1 - p_1 p_2}$$

$$+ \frac{l\,(1-p_1)\,(1-p_2)}{(1-p_1 p_2)}\cdot\frac{1}{(1-p_1)^2} - \frac{l\,p_2^2}{(1-p_1\,p_2)^2}$$

$$= \frac{l}{1-p_1 p_2}\left[\frac{(1-2\,p_2)^2}{(1-p_1)\,p_2 + (1-p_2)\,p_1} + \frac{1-p_2}{1-p_1}\right] - \frac{l\,p_2^2}{(1-p_1\,p_2)^2}$$

A_3. **Calcul de $I'(M_1)$ du 3.4/1 :**

$$\log L_1 = \log \binom{n}{n_1 \; n_2 \; n_3 \; n_4} + (n_1 + n_2) [\log p + \log (1-p)]$$
$$+ 2n_3 \log(1-p) + 2n_4 \log p.$$

$$= \log \binom{n}{n_1 \; n_2 \; n_3 \; n_4} + (\log p)(n_1 + n_2 + 2n_4)$$
$$+ [\log (1-p)] (n_1 + n_2 + 2n_3)$$

$$\frac{\partial \log L_1}{\partial p} = \frac{(n_1 + n_2 + 2n_4)}{p} - \frac{(n_1 + n_2 + 2n_3)}{1-p}$$

$$\frac{\partial^2 (\log L_1)}{\partial p^2} = - \frac{(n_1 + n_2 + 2n_4)}{p^2} - \frac{(n_1 + n_2 + 2n_3)}{(1-p)^2}$$

D'où

$$I'(M_1) = \frac{n}{p^2} [2p(1-p) + 2p^2] + \frac{n}{(1-p)^2} [2p(1-p) + 2(1-p^2]$$

$$= 2. \frac{n}{p(1-p)}.$$

A_4. **Calcul de $I'(M_2)$ du 3.4/2 :**

$$L_2 = \binom{l}{n_1 \; n_2 \; n_3} \left[\frac{p}{1+p}\right]^{n_1 + n_2} \left[\frac{1-p}{1+p}\right]^{n_3}$$

$$\log L_2 = \log \binom{l}{n_1 \; n_2 \; n_3} + (n_1 + n_2) [\log p - \log (1+p)]$$
$$+ n_3 [\log (1-p) - \log (1-p)]$$

$$\frac{\partial \log L_2}{\partial p} = (n_1 + n_2) \left[\frac{1}{p} - \frac{1}{1+p}\right] + n_4 \left[- \frac{1}{1-p} - \frac{1}{1+p}\right]$$

$$\frac{\partial^2 (\log L_2)}{\partial p^2} = (n_1 + n_2) \left[- \frac{1}{p^2} + \frac{1}{(1-p)^2}\right] + n_3 \left[- \frac{1}{(1-p)^2} + \frac{1}{(1+p)^2}\right]$$

D'où :

$$I'(M_2) = \frac{2lp}{1+p}\left[\frac{1}{p^2} - \frac{1}{(1+p)^2}\right] + \frac{l(1-p)}{1+p}\left[\frac{1}{(1-p)^2} - \frac{1}{(1+p)^2}\right]$$

$$= \frac{l}{1+p}\left[\frac{2}{p} - \frac{2p}{(1+p)^2}\right] + \frac{l}{1+p}\left[\frac{1}{1-p} - \frac{1-p}{(1+p)^2}\right]$$

$$= \frac{l}{1+p}\left[\frac{2-2p+p}{p(1-p)} - \frac{1}{1+p}\right]$$

$$= \frac{2l}{p(1-p)} \cdot \frac{1}{(1+p)^2}$$

A_5. **Calcul de $I'(M_3)$ du 4.4/3 :**

$$\log\ L_3 = \log\binom{l}{k} + k\,[\log 2 + \log p - \log(1+p)]$$

$$+ [\log(1-p) - \log(1+p)]\,(l-k)$$

$$\frac{\partial^2 \log L_3}{\partial p^2} = k\left[-\frac{1}{p^2} + \frac{1}{(1+p)^2}\right] + (l-k)\left[\frac{1}{(1-p)^2} + \frac{1}{(1+p)^2}\right]$$

$$I'(M_3) = \frac{2lp}{1+p}\left[\frac{1}{p^2} - \frac{1}{(1+p)^2}\right] + l\,\frac{1-p}{1+p}\left[\frac{1}{(1-p)} - \frac{1}{(1+p)^2}\right]$$

$$= \frac{2l}{p(1-p)} \cdot \frac{1}{(1+p)^2}$$

A_6. **Calcul de $I(P)$ du 3.8. :**

On a :

$$\log L = \log\binom{n}{n_1, \ldots, n_{2m}} + n_0 \log p^m + \left[n_1 + n_2 + \ldots n_{\binom{m}{1}}\right]$$

$$\log p^{m-1}(1-p) + \ldots + \left[n_{k_1} + n_{k_2} + \ldots + n_{k\binom{m}{k}}\right] \log p^{m-k}(1-p)^k$$

$$+ \ldots + n_{2m}\ \log(1-p)^m$$

$$\frac{\partial \log L}{\partial p} = \sum_{k=0}^{m}\left[n_{k_1} + n_{k_2} + \ldots + n_{k\binom{m}{k}}\right]\left[\frac{m-k}{p} - \frac{k}{1-p}\right]$$

$$\frac{\partial^2 (\log L)}{\partial p^2} = \sum_{k=0}^{m}\left[n_{k_1} + n_{k_2} + \ldots + n_{k\binom{m}{k}}\right]\left[-\frac{m-k}{p^2} - \frac{k}{(1-p)^2}\right]$$

D'où

$$I(p) = \sum_{k=1}^{m} n\binom{m}{k} p^{m-k} (1-p) \left[\frac{m-k}{p^2} + \frac{k}{(1-p)^2}\right]$$

$$= n \left[\frac{m - m(1-p)}{p^2} + \frac{m(1-p)}{(1-p)^2}\right]$$

$$= \frac{n \, . \, m}{p(1-p)}$$

A_7. **Démonstration de la proposition 3.10.**

Considérons la loi de probabilité du § 3.8. On peut montrer, comme on l'a fait au § 2.3., que la loi conditionnelle de

$$(N_{2m}, \ldots, N_2 / N_{2m} + \ldots + N_2 = l)$$

est multinomiale, avec les paramètres

$$\left[l \, ; \frac{(1-p)^m}{1-p^m}, \ldots, \frac{p^{m-k}(1-p)}{1-p^m}, \ldots, \frac{p^{m-1}(1-p)}{1-p^m}\right]$$

répétés $\binom{m}{k}$ fois

On a :

$$\log L = \text{constante} + \Sigma \left(n_{k_1} + \ldots + n_{\binom{m}{k}}\right) [(m-k) \log p + k \log (1-p)]$$

$$- l \log (1 - p^m)$$

$$\frac{\partial^2 \log L}{\partial p^2} = \sum_{k=1}^{m} \left[n_{k_1} + \ldots + n_{\binom{m}{k}}\right] \left[-\frac{m-k}{p^2} - \frac{k}{(1-p)^2}\right]$$

$$+ l \frac{m(m-1) p^{m-2} (1-p^m) + (m p^{m-1})^2}{(1-p^m)^2}$$

D'où :

$$I(p) = \sum_{k=1}^{m} l \frac{\binom{m}{k} p^{m-k} (1-p)^k}{1-p} \left[\frac{m-k}{p^2} + \frac{k}{(1-p)^2}\right]$$

$$- l \frac{m(m-1) p^{m-2} (1-p^m) + (m p^{m-1})^2}{(1-p^m)^2}$$

$$= \left[\frac{m\,l}{p} + \frac{m\,l}{1-p}\right] \cdot \frac{1}{1-p^m} - l\,\frac{m^2\,p^{m-2}(1-p^m) + m^2\,p^{2m-2}}{(1-p^m)^2}$$

$$= \frac{m\,l}{p(1-p)} \cdot \frac{1}{(1-p^m)^2}\,[1 - p^m - m\,p^{m-2}(p-p^2)]$$

$$= \frac{m\,l}{p(1-p)} \cdot \frac{1}{(1-p^m)}\,[(1-p)(1+p+\ldots+p^{m-1}) - m\,p^{m-1} + m\,p^m]$$