O.R.S.T.O.M

Service Hydrologique

Note technique n° 3

Diffusion interne

Le TEST du χ^2 de PEARSON

par

Y. BRUNET-MORET

Novembre 1966

Cette note développe nos réflexions sur la signification du test du χ^2 et sur les modalités de découpages en classes les plus aptes à maximiser la puissance de ce test. Dans certains cas, la méthode des agrégats et les tests des suites et des signes peuvent améliorer la signification attribuable au χ^2.

Sans être définitivement codifiée, la méthode de calcul du χ^2 proposée en conclusion paraît aujourd'hui la plus sûre.

I. DEFINITION de χ^2 -

Le mieux est de se reporter à l'ouvrage "Méthode Statistique de MORICE et CHARTIER p. 238" pour un exposé complet sur la théorie du χ^2. Si nous avons un certain nombre $n = \sum_1^k n_i$ d'observations rangées en K classes, n_i étant l'effectif observé de la classe i, p_i la probabilité d'apparition d'une observation dans cette classe i d'après la loi de répartition choisie ($n\ p_i$ étant l'effectif théorique de la classe i) le test qui permet de vérifier l'adéquation de la loi choisie aux observations se calcule à partir de :

$$\chi^2 = \sum_1^k \frac{(n_i - n\ p_i)^2}{n\ p_i}$$

Tables ou abaques permettent de déterminer la probabilité d'apparition de la valeur de χ^2 d'après le nombre ν de degrés de liberté.

Ce nombre de degrés de liberté est égal au nombre de classes moins le nombre de liaisons entre la distribution théorique et la distribution observée :

- l'égalité des effectifs globaux $\sum n_i = \sum n\ p_i$ compte pour une liaison (cette égalité n'est pas toujours remplie)
- le calcul de chaque paramètre de la répartition théorique à partir des observations (ou par minimisation du χ^2) compte aussi pour une liaison.

Le test du χ^2 ne peut s'employer que si :

a) l'échantillon, prélevé dans la population, n'est pas exhaustif, et est tiré au hasard (pas d'autocorrélation dans la population, attention aux suites chronologiques)

b) aucune des probabilités p_i n'est trop voisine de zéro

c) les produits $n\ p_i$ sont tous supérieurs à 5, ou même 10 si p_i est petit.

II. DISTRIBUTION de χ^2 -

Nous donnons deux graphiques de distribution de χ^2 pour les degrés de liberté ν de 1 à 10.

Si ν est égal ou supérieur à 10, nous pouvons utiliser l'approximation suivante :

$$3\sqrt{\frac{\chi^2}{\nu}}$$

est distribué normalement avec comme moyenne $1 - \frac{2}{9\nu}$ et écart type $\frac{1}{3}\sqrt{\frac{2}{\nu}}$

Si ν est égal ou supérieur à 30, nous pouvons considérer que

$$\sqrt{2\chi^2} - \sqrt{2\nu - 1}$$

est distribué normalement avec comme moyenne zéro et écart type 1. Cette approximation est toujours moins serrée que la précédente.

Nota important sur le calcul de la probabilité de rang r

Soit N observations rangées en ordre croissant (ou décroissant), la probabilité expérimentale liée à l'observation de rang r est, au non-dépassement (ou au dépassement)

$\frac{r}{N+1}$. On peut le démontrer.

Soit N observations rangées en ordre croissant 1, 2, ... r ... N, P étant la probabilité théorique de la valeur x_i au non-dépassement, et Q = 1 - P la probabilité théorique de la valeur x_i au dépassement.

La probabilité pour qu'il y ait r valeurs $\leqslant x_i$ ou N - r valeurs $\geqslant x_i$ dans le tirage de l'échantillon est (loi binomiale, r variant de 0 à N)

$$\varphi = \frac{N\ !}{r!\ (N-r)!} \quad P^r \quad Q^{N-r}$$

(C'est la probabilité pour que la valeur x_i de l'échantillon soit au rang r du rangement de l'échantillon).

La probabilités moyenne de l'observation du rang r est au non-dépassement :

$$\sum_1^r \int_0^1 \varphi \, dP = \frac{r}{N+1}$$

III. SIGNIFICATION de χ^2 -

Le calcul du χ^2 sert usuellement à déterminer, compte tenu du nombre de degrés de liberté, la probabilité P de dépassement de la valeur du χ^2.

a) Cas d'une population connue :

Si nous avons calculé le χ^2 à partir des valeurs d'un échantillon tiré d'une population dont la répartition est connue par avance (et non déterminée par les valeurs de l'échantillon, le χ^2 étant calculé à l'aide des paramètres de la population et non de l'échantillon), la probabilité P est celle du dépassement de la valeur calculée de χ^2 par le simple effet du hasard du tirage de l'échantillon.

Ainsi, si d'une même population (de répartition connue par avance) nous tirons au hasard plusieurs séries d'un millier d'échantillons (tirés au hasard) chacune, nous trouverons en moyenne par série 10 χ^2 correspondant à une probabilité P au dépassement de 0,99 ou plus et 10 χ^2 correspondant à une probabilité P de 0,01 ou moins. La densité de la répartition de la probabilité P est constante : la médiane et la moyenne de toutes les probabilités P déduites de tous les χ^2 des tirages ci-dessus (et calculés, nous le répétons, en utilisant les paramètres, connus par avance, de la répartition de la population) sera de 0,50. Nous trouverons autant de χ^2 correspondant à des probabilités P comprises entre 0,60 et 0,50 que de χ^2 correspondant à des probabilités P comprises entre 0,10 et zéro. La probabilité P calculée d'après un seul tirage d'échantillon (et les paramètres connus par avance de la population mère) a donc 50 % de chances d'être comprise entre 0,25 et 0,75, 90 % de chances d'être

comprise entre 0,05 et 0,95, 99 % de chances d'être comprise entre 0,005 et 0,995.

b) Cas d'un échantillon de population inconnue :

Si les paramètres de la population sont calculés à partir de l'échantillon dont on déduit aussi la probabilité P (d'après la valeur du χ^2, et un nombre de degrés de liberté obtenu en retranchant le nombre de liaisons ayant servi au calcul des paramètres), cette probabilité P représente l'adéquation de la loi choisie et des paramètres calculés à la répartition de l'échantillon : elle a donc 50 % de chances de n'être pas comprise entre 0,25 et 0,75, 10 % de chances de n'être pas comprise entre 0,05 et 0,95, 1 % de chances de n'être pas comprise entre 0,005 et 0,995 (intervalles de confiance de la probabilité P). Les χ^2 trop petits sont aussi troublants que les χ^2 trop grands.

On se trouve généralement dans ce cas et le test du χ^2 est alors un test d'adéquation à une loi choisie a priori (hypothèse à vérifier).

Le rejet ou l'adoption de l'hypothèse dépendent d'un seuil de probabilité P (χ^2) que l'on se fixe a priori (voir fin de la note).

IV. Le DECOUPAGE de l'ECHANTILLON en CLASSES -

Ce problème est habituellement éludé par les auteurs qui signalent ou utilisent (en exemple) le test du χ^2. Nous y reviendrons plus loin (Cf. V).

Ce problème est important comme le montre la figure sur laquelle on trace d'une part, la loi de répartition de la population et, d'autre part, la courbe obtenue en joignant les différents points observés. Si l'on choisit comme valeurs des limites de classes celles pour lesquelles les deux courbes se coupent, le χ^2 sera nul et la probabilité P de 1,00. Si l'on choisit comme valeurs des limites celles pour lesquelles les tangentes des deux courbes sont parallèles, le χ^2 atteindra sa valeur maximale.

Prenons par exemple la distribution des pluviométries annuelles de 45 années à ZIGUINCHOR, étudiée suivant une loi de PEARSON III avec 2 paramètres. Le découpage en 7 classes peut conduire de $\chi^2 = 0$ soit $P = 1,00$ à $\chi^2 = 14,25$ soit $P = 0,005$! Les découpages en classes de probabilités théoriques égales conduisent aux valeurs suivantes : 9 classes (5 unités théoriques dans chaque classe) $\chi^2 = 8,40$ d'où $P = 0,21$ - 8 classes $\chi^2 = 8,19$ d'où $P = 0,15$ - 7 classes $\chi^2 = 0,578$ d'où $P = 0,96$ - 6 classes $\chi^2 = 2,34$ d'où $P = 0,50$ - 5 classes $\chi^2 = 3,56$ d'où $P = 0,18$ - 4 classes (11,25 unités théoriques dans chaque classe, 1 seul degré de liberté) $\chi^2 = 1,31$ d'où $P = 0,25$.

Cet exemple montre bien que si l'on choisit les limites de classes, l'on peut trouver le χ^2 que l'on désire et que, même si l'on effectue un découpage en classes théoriquement égales, la probabilité déduite du χ^2 peut varier dans de larges limites.

La règle communément donnée (et appliquée ci-dessus) est de ne pas créer de classes ayant moins de 5 (ou de 10 ?) unités tant "théoriques" qu'"observées". La restriction relative au nombre d'unités "observées" est d'ailleurs inutile, car dans une classe,

$$\Delta\chi^2 = \frac{(\text{nbre théorique} - \text{nbre observé})^2}{\text{nbre théorique}}$$

le nombre observé n'intervient qu'au numérateur et par la valeur absolue de sa différence avec le nombre théorique. Nous spécifions tout de suite que, le nombre théorique n'a pas besoin d'être un entier, mais que si nous ne choisissons pas ce nombre théorique entier, le $\Delta\chi^2$ de la classe ne peut être que $\geq$ zéro, et non nul. La valeur de χ^2 en est sûrement biaisée, et quelquefois peut-être d'une quantité sensible : ainsi pour reprendre l'exemple donné plus haut "ZIGUINCHOR, 45 ans, 6 classes théoriquement égales" de 7,5 unités par classe : le $\Delta\chi^2$ minimal par classe est de :

$$\frac{0,5^2}{7,5}$$

soit un χ^2 minimal total de 0,20 (3 degrés de liberté P = 0,978). Si nous enlevons cette valeur minimale au χ^2 de 2,34 réellement trouvé, P passe de 0,50 à 0,545. Ceci peut avoir des conséquences lorsqu'on applique un test d'agrégat d'un certain nombre de probabilités indépendantes (Cf. § VI).

Le fait de découper l'échantillon, en créant des classes d'au moins 5 unités "théoriques" aux deux extrémités notamment, restreint terriblement la valeur du test qui renseigne seulement sur la possibilité qu'a la loi choisie (avec ses paramètres calculés) de représenter la distribution dans sa zone de forte densité de probabilité. Ainsi, nous pouvons admettre que la distribution observée, et rangée, des pluviométries annuelles de 45 années à ZIGUINCHOR est bien représentée par une loi de PEARSON III du rang 5 au rang 40, c'est-à-dire pour des probabilités de $\frac{1}{9}$ à $\frac{8}{9}$, donc pour des temps de récurrence inférieurs à 9 ans. L'adéquation de la loi choisie aux fréquences rares ou temps de récurrence élevés n'est absolument pas testée, quelle que soit la valeur de χ^2. Et la plupart du temps, ce sont ces fréquences rares qui nous intéressent. Nous reviendrons sur ce point.

V. Le CHOIX des CLASSES -

Ce qui suit est une traduction abrégée de KENDALL and STUART("The advanced theory of statistics" vol 2 p. 430 et suivantes) avec quelques remarques personnelles entre parenthèses.

Le nombre de classes et leur contenu peuvent être imposés par les données : tableau de contingence par exemple, mais souvent et le nombre de classes et leurs frontières sont à choisir. La facilité arithmétique est quelquefois utilisée pour donner la solution suivante : les classes sont choisies pour couvrir d'égaux intervalles de variation de la variable aléatoire, sauf aux extrémités où les intervalles peuvent devenir infinis (exemple de la répartition de la pluviométrie journalière de 10 en 10 mm, la classe la plus élevée allant par exemple de 120 mm à l'infini).

Nous devrions choisir, pour un nombre K donné de classes, les frontières qui maximisent la puissance du test. Le problème n'est pas encore résolu. Il faut chercher une méthode pour esquiver le fait déplaisant qu'il y a une multiplicité de choix, chacun donnant des résultats différents (Cf. § IV) (Par exemple, combien de libertés se donne-t-on - donc de degrés de liberté à retrancher au nombre théorique de degrés de liberté du χ^2 -en choisissant une à une les limites de classes ?). Il est proposé, comme règle plausible et pratique, de <u>choisir des classes égales en probabilités théoriques</u>.

Cette méthode peut ne pas accroître la puissance du test, car l'hypothèse d'adéquation est surtout vulnérable aux extrémités de la répartition et la méthode peut amener une perte de sensibilité si K n'est pas assez grand (dans ce cas, les probabilités des frontières internes des deux classes extrêmes sont éloignées de 0 ou de 1, mais si K est très grand on risque de noyer des $\Delta\chi^2$ élevés des classes extrêmes dans une masse de $\Delta\chi^2$ peu élevés tout en augmentant le nombre de degrés de liberté de χ^2). Cette méthode a l'avantage de ne pas donner de valeurs biaisées de χ^2 (à condition que le nombre théorique dans chaque classe soit entier) et de permettre de déterminer le nombre K de classes, n étant l'effectif de l'échantillon. En effet, pour maximiser la puissance du test, on prend K proportionnel à $n^{2/5}$, le facteur de proportionnalité variant avec la valeur de χ^2. Retenons que pour $n = 200$ et prob.$\chi^2 = 0{,}05$, on obtient

$$\frac{n}{K} \# 6 \quad (K = 30)$$

et que pour prob.$\chi^2 = 0{,}01$, on obtient

$$\frac{n}{K} \# 8 \quad (K = 27).$$

Cependant, K peut être divisé par deux sans provoquer une sérieuse diminution de la puissance du test. D'autre part, une limite supérieure de K est fournie par le fait que l'approximation multinormale à la distribution multinomiale n'est plus satisfaisante si le produit $n\,p_i$ est trop petit. La

règle communément appliquée est de prendre :

$$n\ p_i = \frac{n}{K} \geqslant 5$$

D'autres critiques ont été faites au test de χ^2 : tout d'abord, on perd de l'information en groupant les observations en classes, et il est possible que cette perte soit plus grande lorsque l'on teste une distribution continue. Ensuite, comme l'on travaille sur des carrés, le test est insensible aux arrangements des signes des différences. Contre cette seconde objection, on peut utiliser le test des suites des différences entre les observations rangées et les valeurs - aux mêmes probabilités respectives - déduites de la loi théorique. Le test des suites et le test du χ^2 sont indépendants lorsque l'effectif de l'échantillon est grand.

VI. Les AGREGATS -

Du paragraphe III peut se dégager l'impression qu'un χ^2 isolé calculé à partir d'un seul échantillon n'a pas une signification bien précise, et du § IV, celle que le test de χ^2 escamote les fréquences rares (impressions pessimistes ?).

Si nous disposons de probabilités P_i indépendantes - provenant d'échantillons indépendants tirés de la même population - nous pouvons vérifier que la densité de la répartition des probabilités P_i est constante (Cf. § III). Nous pouvons aussi utiliser le test de Fischer (p. 91) de "l'agrégat d'un certain nombre de probabilités indépendantes" : soit N probabilités P_i, le χ^2 global à 2 N degrés de liberté est égal à $-2 \sum L_e P_i$. Ou bien, si les probabilités P_i sont déduites de N tests de χ_i^2 (non biaisés) à ν_i degrés de liberté, le χ^2 global à $\sum \nu_i$ degrés de liberté est égal à $\sum \chi_i^2$. Nous pouvons utiliser ces méthodes pour tester l'adéquation d'une loi de répartition à un ensemble d'échantillons indépendants, provenant de populations différentes. Par exemple, nous pouvons vérifier que la loi de PEARSON III s'applique aux distributions des pluviométries annuelles des stations d'une même zone climatique et géographique : le χ^2 est calculé pour chaque station en utilisant les paramètres propres de cette

station. Il nous semble nécessaire pour l'application de ces méthodes, et pour éviter les distorsions signalées du χ^2 (exemple du § III), d'utiliser pour chaque station un découpage en classes égales en probabilités théoriques et de plus, soit d'égaliser les probabilités des classes pour toutes les stations, soit d'égaliser les effectifs de toutes les classes pour les stations (ce qui n'est possible qu'en travaillant sur des échantillons de même longueur).

Contre l'escamotage des fréquences rares, nous ne voyons que la méthode des stations-années. Encore faut-il qu'il n'y ait pas corrélation entre les stations ni autocorrélation à l'intérieur de l'échantillon de chaque station. Cette condition d'indépendance des échantillons est certainement plus stricte ici qu'au paragraphe précédent. Voici la méthode utilisée pour vérifier que la loi de PEARSON III s'applique aux distributions des pluviométries journalières, de fréquence rare, des stations d'une même zone climatique et géographique. Pour chaque station, observée pendant m_i années, nous avons noté le nombre observé de dépassements des hauteurs journalières calculées de fréquence annuelle, une fois en 2, 5, 10, 20, 50 et 100 ans (calculées pour chaque station en fonction de ses paramètres propres). L'ensemble des stations nous donne le nombre de stations-années $\sum m_i$ et les nombres globaux de dépassements pour les récurrences 1, 2, 5, 10, 20, 50 et 100 ans. Du nombre de stations-années, nous déduisons les effectifs théoriques des 7 classes dont les bornes sont les récurrences 1, 2, 5, 10, 20, 50 et 100 ans. Des nombres globaux de dépassements, nous déduisons les effectifs observés de ces classes. D'où le calcul d'un χ^2 à 7 classes et 7 degrés de liberté (la somme des effectifs théoriques n'est pas égale à la somme des effectifs observés).

VII.A) TEST des SUITES (d'après "Méthode statistique" de MORICE et CHARTIER - p. 253) -

Ce test n'offre d'intérêt complémentaire de celui du χ^2 que si on l'applique à de grands échantillons ($n>100$?) car il en est alors indépendant.

Nous supposons les observations rangées (en ordre croissant ou décroissant, peu importe) et affectons à chaque rang le signe de la différence entre la valeur observée et celle déduite de la loi de probabilité (ce signe peut être trouvé sur un graphique). On appelle suite la succession d'un ou plusieurs signes identiques, succession précédée et suivie de signes de l'autre sens (sauf aux deux extrémités du rangement où les suites ne sont bornées que d'un seul côté). Soit n le nombre total d'observations égal à n_1 (nombre de signes +) plus n_2 (nombre de signes -), on démontre que le nombre R de suites a pour valeur moyenne

$$\overline{R} = \frac{2\, n_1\, n_2}{n} + 1$$

et variance $\sigma^2_{(R)} = \frac{2\, n_1\, n_2\, (2 n_1\, n_2 - n)}{n^2\, (n - 1)}$

Dans le cas qui nous intéresse, la répartition de R peut être considérée comme normale pour $n \geqslant 25$ (car n_1 et n_2 sont peu différents de $\frac{n}{2}$, et l'on peut utiliser les tables de la loi de GAUSS en prenant comme variable réduite

$$t = \frac{1}{\sigma(\overline{R})} \left(R + \frac{1}{2} - \overline{R}\right)$$

expression qui tient compte du caractère discontinu de R (nombre entier). On en tire la valeur de la probabilité de R, qui peut être comparée à un seuil choisi a priori pour conclure au rejet ou à l'adoption de l'hypothèse : adéquation d'une loi et d'un échantillon.

B) TEST des SIGNES -

L'on peut aussi utiliser les signes affectés ci-dessus à chaque rang dans un tableau de contingence de 2 lignes (+ et -) et K colonnes. Chaque colonne correspond à un groupe d'au moins 10 éléments consécutifs du rangement.

Le test est effectué en calculant le χ^2 (à K-1 degrés de liberté) de ce tableau.

Si le découpage en K colonnes correspond à K classes de probabilité théoriques égales, le nombre de degrés de liberté du χ^2 est 2 K - nombre de liaisons. Et dans ce cas, on combine le test des signes avec celui du χ^2 en un seul test, supposé ainsi plus puissant (?).

CONCLUSION PROVISOIRE pour le CALCUL du χ^2

Etant donné qu'en général nous travaillons sur de petits échantillons, nous pouvons fixer la taille minimale de l'effectif théorique d'une classe à 5 unités.

L'échantillon sera divisé en classes d'effectifs théoriques égaux (1), le χ^2 calculé et la probabilité déduite de la valeur de χ^2 en tenant compte du nombre de degrés de liberté.

Si nous faisons toutes les divisions possibles en classes d'effectifs égaux, depuis celle qui donne un effectif théorique de 5 unités par classe (et le plus grand nombre de classes) jusqu'à celle qui donne le nombre de classes minimal pour que le χ^2 soit calculé avec 1 seul degré de liberté, nous obtiendrons un χ^2 minimal, mais par suite de cette condition de minimum, le χ^2 minimal ainsi choisi possède un degré de liberté de moins que celui qui correspond à son nombre de classes (diminué des liaisons).

Le test des signes est toujours applicable, mais n'est peut-être pas souvent très significatif. Le test des suites semble demander, pour être indépendant de celui du χ^2, peut-être une centaine d'observations.

Nous rappelons que la probabilité P déduite du χ^2 représente l'adéquation de la loi choisie et des paramètres calculés à la répartition de l'échantillon. 50 % des valeurs de P doivent se trouver dans l'intervalle 0,25-0,75. Une probabilité P très forte est aussi rare qu'une probabilité P très faible. Dans le premier cas, elle incite à revoir les valeurs des observations et à recommencer le calcul ; dans

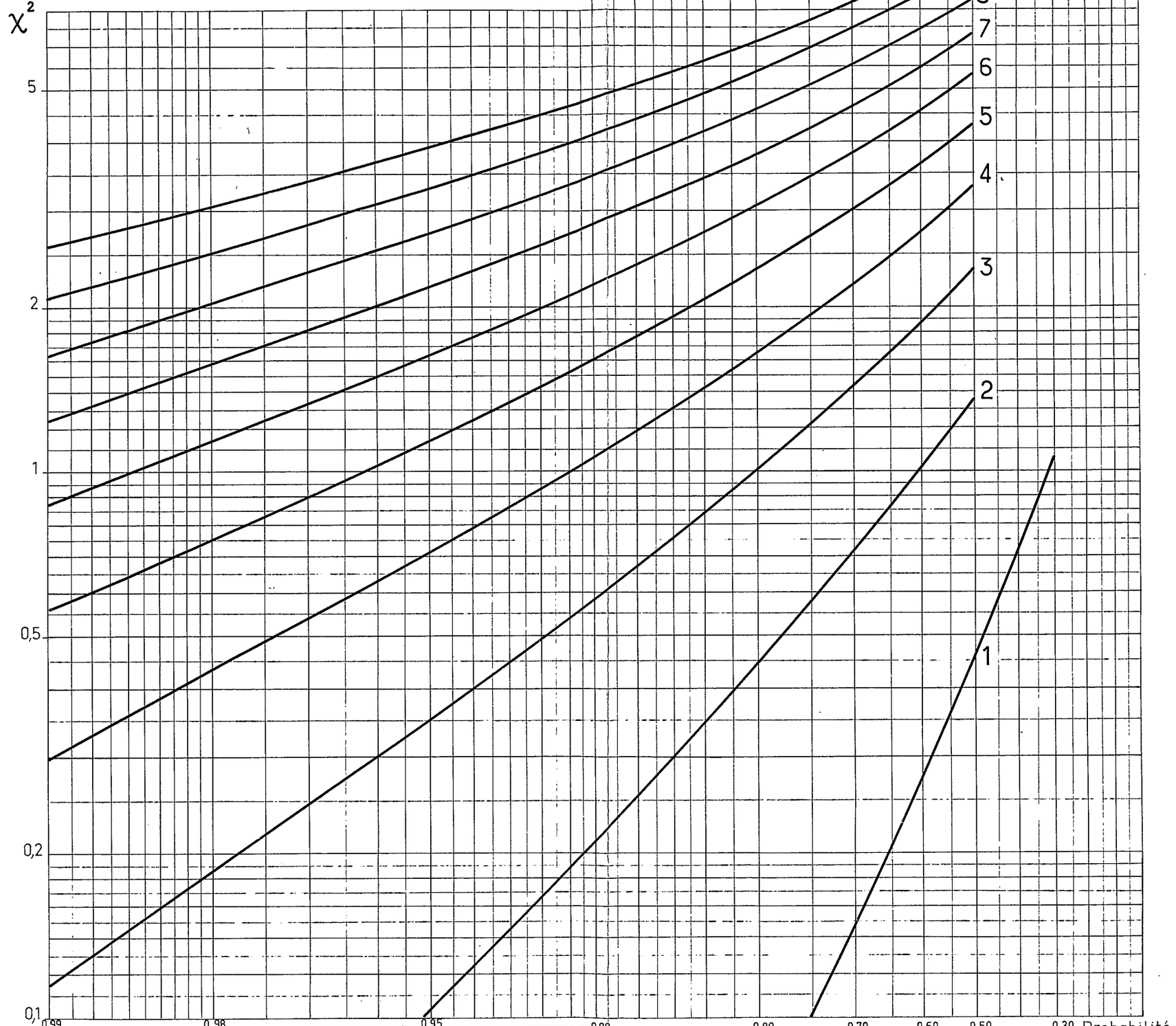
---

(1) On admet que le biaisage du χ^2 résultent d'effectifs théoriques non entiers est négligeable devant l'amélioration de la signification du test.

le second, à accepter la loi choisie avec ses paramètres calculés comme représentative de la distribution de l'échantillon et de la population si P est supérieur à un certain seuil, par exemple 0,05 ou à rejeter cette représentation si P est inférieur à 0,01. On peut d'ailleurs faire varier ces seuils avec la qualité des observations.

L'exemple de calcul d'un χ^2 donné page 6 sur les pluviométries annuelles de ZIGUINCHOR est déjà clair. On trouvera un calcul complet développé d'un test de χ^2 dans la note technique n° 1 "Utilisation d'une loi de PEARSON III pour un échantillon de taille connue", effectué sur les pluviométries annuelles de ABENGOUROU.

b) Probabilités supérieures à 0,50



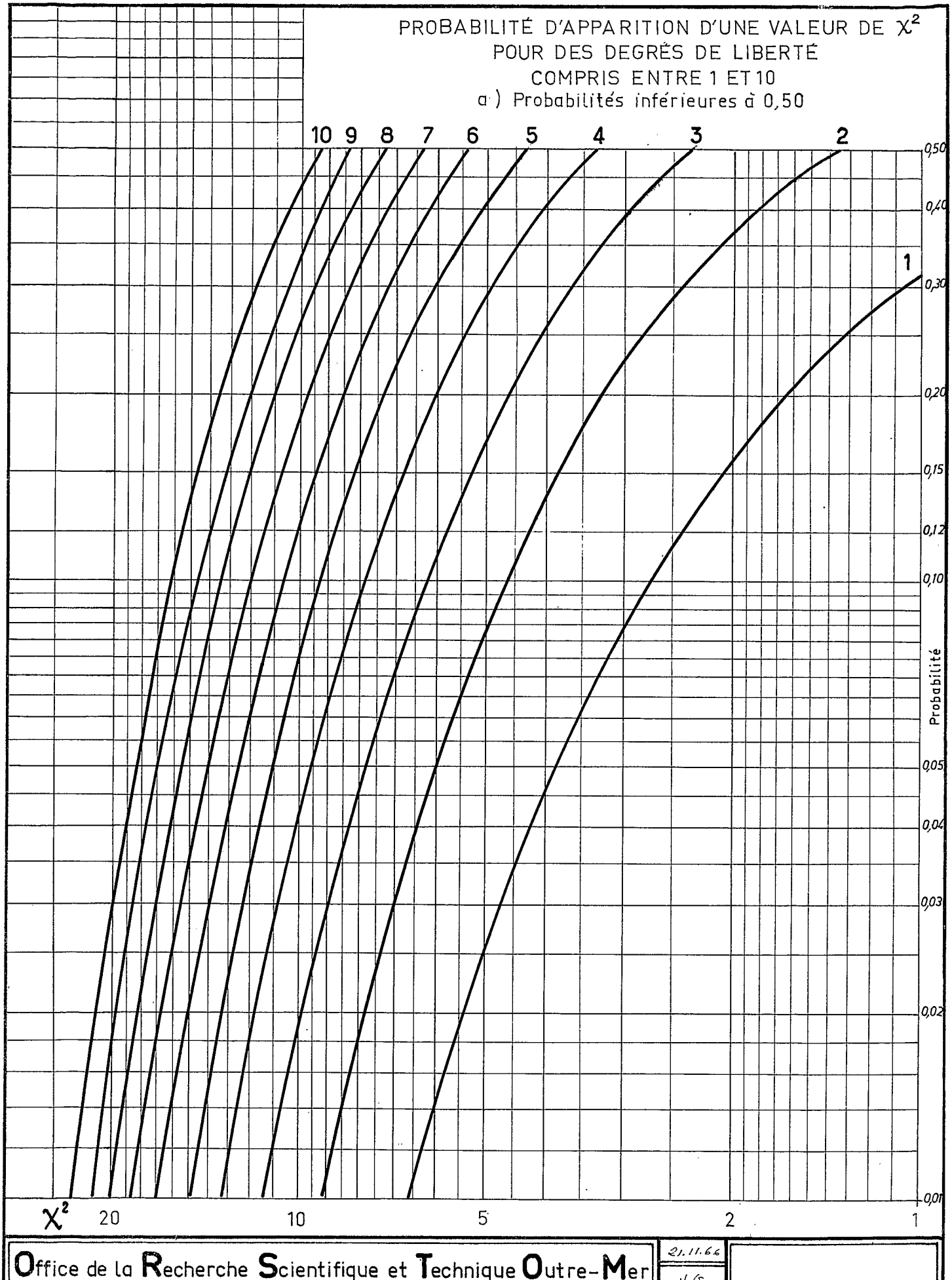
Office de la Recherche Scientifique et Technique Outre-Mer

21.11.66

V.G.

PROBABILITÉ D'APPARITION D'UNE VALEUR DE χ^2 POUR DES DEGRÉS DE LIBERTÉ COMPRIS ENTRE 1 ET 10

a) Probabilités inférieures à 0,50



Office de la Recherche Scientifique et Technique Outre-Mer

21.11.66

J.G.