affect the exactitude of the test of the null hypothesis that they are drawn from the same population. Such a case as you mention, supposing t to be non-significant on 'Student's' test, would seem to be nicely treated by the supplementary z which I suggest for the differences between the variances, for this would show that the variances were very significantly different, although on the data it could not be said that the means were different.

Now in practice this would advance the investigation as much as in the standard case; for, supposing the test were made between two varieties of the plant, the fact of a real difference in the variances shows that in some circumstances one variety is the better and in other circumstances that it is the worse. The situation is thus proved to be more complicated than perhaps the experimenter originally thought. He will see now that it is useless to compare the means unless he has some specification of the circumstances, or range of circumstances, in which the test is to be made. His preliminary enquiry 'Are the samples from the same population?' is answered definitely in the negative and, this being so, it will depend entirely on the circumstances of the case whether any comparison of the means is desired at all.

The point has, I think, received the rather large amount of theoretical attention that it has chiefly through lack of contact with the practical experimental situation. Some years ago Behrens published a test of significance appropriate to the rather academic question 'Might these samples have been drawn from different normal populations having the same mean?', and more recently Sukhatme has been preparing tables needed for Behrens' test. I have, however, always doubted whether the test has any real importance.

Fisher to G. Darmois: 5 *June* 1936

I write to thank you most heartily for your excellent little book on *Methods of Estimation*, which you were kind enough to send to me. I do not know any other publication of the same length which gives so good an introduction to the subject.

There is one point to which perhaps I may draw your attention, in respect of the propositions you ascribe on p. 19 to J.L. Doob. I have little doubt that the estimate obtained by the method of maximum likelihood always converges, and, if so, it certainly is consistent, or correct. It is not, however, always true that its limiting sampling distribution, when the sample is increased, is Gaussian. This is only true in general when the amount of information, g_{ii} in your notation, is finite, i.e. neither zero nor infinity. For particular values of the parameter it may always be made zero or infinity by an appropriate transformation of the parameter, but such failures of the measure of information are only trivial. When, however, the likelihood is a discontinuous function, as it is for the probability distribution

$$e^{-(x-\mu)}dx$$

for which the least observation in the sample supplies a sufficient estimate of μ, we have, nevertheless, no measurement of the amount of information, and the limiting distribution is not Gaussian. These cases, therefore, where the measurement of information is not a convergent integral for all values of the parameter, deserve special study to ascertain whether any measure of the value of the estimate can be found by which such cases might be compared.

On another point, on which I think you do not touch, I believe it must be true that the maximum likelihood estimate has always a higher intrinsic precision than any other estimate which can be made, but I have only proved this as true in the limit for large samples, where the loss of information is certainly minimised by using the method of maximum likelihood.

I notice you use the term 'exhaustive estimation' for what I speak of as 'sufficient statistics'. I should like, if you approve, to use the term 'exhaustive estimation' in a somewhat wider sense, including not only the cases in which sufficient statistics exist, but also those cases described in my paper 'Two new properties of mathematical likelihood' [*CP* 108], in which the whole of the ancillary information may be utilised by taking account of characteristics of the sample, which, by themselves, yield no information respecting the parameter sought. Estimation in these cases also may be said to be exhaustive, and it seems to me likely that further research may enlarge the field of application of this method. Indeed exhaustive estimation in this sense would be possible if it were always true:- that whenever $x_1, \ldots, x_n$ have a simultaneous distribution depending on a parameter θ it is possible to find $n-1$ functionally independent functions of the variates, the distribution of each of which is independent of θ. It would be interesting to know if you could form an intuitive judgement of the probable truth of this proposition.

Fisher to G. Darmois: 2 *April* 1940

I am very glad to have your letter. . . .

I should be very much surprised in a practical case to get a divergent series[1] by the process which I think is easiest in principle, namely:

If $m_r(\theta_1, \ldots \theta_p)$, when $r = 1, \ldots s$, is the expected frequency in class r for given values of the parameters θ_1 to θ_p, let

$$I_{jk} = \sum_{r=1}^{s} \frac{1}{m_r} \frac{dm_r}{d\theta_j} \frac{dm_r}{d\theta_k}$$

and let V_{jk} be the corresponding element in the reciprocal matrix, then if a_r is the observed frequency, let

$$\sum_{r=1}^{s} \frac{a_r}{m_r} \frac{dm_r}{d\theta_j} = A_j$$

then

$$\theta_j' = \theta_j + \sum_{k=1}^{p} V_{jk} A_k$$

gives a second approximation. The matrices are determined at the trial values, and for more precise values of the variances and co-variances of the values fitted may usefully be redetermined from the improved values. From my own experience I should expect in practice that this would fail to give a convergent series of approximations only if the trial values were exceptionally unfortunate. I think, however, I should have time in the near future to try my hand at any batch of data which has been giving trouble. An approximation based on percentiles is often quicker to obtain than one based on moments, and is likely to be more accurate if the data are heavily grouped. If I ran into a divergent series I should more readily suspect that the theoretical form chosen was unsuitable to the data, i.e. that the goodness of fit was very unsatisfactory, than that my starting point, supposing this to be ordinarily plausible, was the cause of the trouble. . . .

[1] Darmois had asked if the method of successive approximations could lead to a divergent series when used with maximum likelihood estimation.

Fisher to G. Darmois: 27 *July* 1940

I am sending herewith some of the material you want . . .

We are much looking forward to seeing you on Tuesday and certainly hope that you will meet with no official obstructions, though anything is possible in these days. I am enclosing also what I have just written out, in the hope that you may be interested, an example of ancillary information which I think you have seen before, though perhaps not so fully worked out. I have not, however, delayed to enlarge upon the question of its relation to the problem of the Nile. I believe, however, that this problem is equivalent to the finding of appropriate ancillary statistics in general.

[Enclosure]

Example of ancillary information supplied by characteristics of the data other than the total number of observations

Consider a distribution of two observable values x, y such that the element of frequency is

$$df = e^{-(\theta x + y/\theta)} dx\, dy$$

where x and y each may take positive values from 0 to ∞. The parameter θ is unknown but may be estimated from the observations.

Estimation

If we have a sample of n observations, and apply the Method of Maximal Likelihood, we find

$$\begin{aligned} \log L &= -\theta S(x) - S(y)/\theta, \\ \frac{\partial}{\partial\theta} \quad {''} &= -\ S(x) + S(y)/\theta^2. \end{aligned}$$

Equating this to zero, we have the Equation of Estimation

$$\theta^2 = S(y)/S(x), \qquad (1)$$

and may define our estimate T by the equation

$$T^2 = S(y)/S(x); \qquad (2)$$

we shall then be interested in the distribution of T for given θ.

Distribution

If ancillary information is ignored, we consider the distribution of T for different samples of n observations each. If for brevity we write X for $S(x)$ and Y for $S(y)$, it is easy to see that the simultaneous distribution of X and Y is

$$df = \frac{1}{(n-1)!^2} X^{n-1} Y^{n-1} e^{-\theta X - Y/\theta} dX\, dY. \qquad (3)$$

Substituting $Y = T^2 X$,
$dY = X d(T^2)$,

we have

$$df = \frac{1}{(n-1)!^2} \quad T^{2n-2} d(T^2) . X^{2n-1} e^{-X(\theta + T^2/\theta)} dX \qquad (4)$$

and integrating from 0 to ∞ with respect to X

$$df = \frac{(2n-1)!}{(n-1)!^2} \frac{T^{2n-2}\, d(T^2)}{(\theta + T^2/\theta)^{2n}} \qquad (5)$$

which is the distribution of T given θ and n.

Amount of information available, and amount utilised.

From the original distribution

$$df = e^{-\theta x - y/\theta} dx dy$$

we ascertain the mean value or expectation of

$$\{d(\log f)/d\theta\}^2.$$

This is the mean value of

$$(x - y/\theta^2)^2$$

which by simple integration is found to be $2/\theta^2$.

The amount of information about θ contained in n observations is therefore

$$I = 2n/\theta^2. \qquad (6)$$

For the amount available in the distribution of T for given θ and n, we require the mean value of

$$4n^2\left(\frac{1 - T^2/\theta^2}{\theta + T^2/\theta}\right)^2;$$

using the Eulerian integral of the first kind it appears that

$$E\left\{\frac{1}{(\theta + T^2/\theta)^2}\right\} = \frac{n(n+1)}{2n(2n+1)} \cdot \frac{1}{\theta^2},$$

$$E\left\{\frac{T^2/\theta^2}{(\theta + T^2/\theta)^2}\right\} = \frac{n^2}{2n(2n+1)} \cdot \frac{1}{\theta^2},$$

$$E\left\{\frac{T^4/\theta^4}{(\theta + T^2/\theta)^2}\right\} = \frac{n(n+1)}{2n(2n+1)} \cdot \frac{1}{\theta^2},$$

giving in all

$$4n^2/(2n+1)\theta^2. \qquad (7)$$

This is less than the total information available by one part in $2n + 1$; or, when $n = 1$, only 2/3 of the available information is utilised.

Ancillary statistic

Suppose now we propose to ignore n, and consider in its place the ancillary statistic N defined by

$$N^2 = XY.$$

The simultaneous distribution of T and N for given n is found by substituting

N/T for X and NT for Y;

then $$\mathrm{d}X\mathrm{d}Y = (2N/T)\mathrm{d}N\mathrm{d}T$$

and $$\mathrm{d}f = \frac{1}{(n-1)!^2} N^{2n-1}\mathrm{d}N\, \mathrm{e}^{-N(\theta/T+T/\theta)}\frac{2\mathrm{d}T}{T}. \qquad (8)$$

Now $$\int_0^\infty \mathrm{e}^{-N(\theta/T+T/\theta)}\frac{\mathrm{d}T}{T} = \int_0^\infty \mathrm{e}^{-N(z+1/z)}\frac{\mathrm{d}z}{z}$$

is independent of θ, but is some function $K(N)$ of N. Hence the distribution of N is

$$\mathrm{d}f = \frac{2}{(n-1)!^2} N^{2n-1}K\mathrm{d}N \qquad (9)$$

and the distribution of T for given N is

$$\mathrm{d}f = \frac{1}{K}\,\mathrm{e}^{-N(\theta/T+T/\theta)}\frac{\mathrm{d}T}{T} \qquad (10)$$

which is independent of n. Now when a sample is observed we know n and equally we know N. If we prefer it we have equally good grounds for regarding (10) as the sampling distribution of T as for so regarding equation (5).

It may now be shown that the information supplied by (10) is the whole of that available.

Since by definition

$$K = \int_0^\infty \mathrm{e}^{-N(z+1/z)}\frac{\mathrm{d}z}{z},$$

$$\frac{\mathrm{d}^2K}{\mathrm{d}N^2} = \int_0^\infty \mathrm{e}^{-N(z+1/z)}(z + 1/z)^2\frac{\mathrm{d}z}{z}. \qquad (11)$$

But $I_{T,N}$ is the expectation of $\{N(1/T - T/\theta^2)\}^2$ which is

$$\frac{N^2}{\theta^2K}\int_0^\infty \mathrm{e}^{-N(z+1/z)}(z - 1/z)^2\frac{\mathrm{d}z}{z}$$

$$= \frac{N^2}{\theta^2K}\left(\frac{\mathrm{d}^2K}{\mathrm{d}N^2} - 4K\right). \qquad (12)$$

We now evaluate the mean value of this expression for samples having fixed n,

$$\mathrm{d}f = \frac{2}{(n-1)!^2}\, N^{2n-1}\,K\,\mathrm{d}N.$$

Now $N^{2n}K = 0$ when $N = 0$ and when N is ∞;

hence $$\int_0^\infty N^{2n}\frac{\mathrm{d}K}{\mathrm{d}N}\,\mathrm{d}N = -\,2n\int_0^\infty N^{2n-1}\,K\,\mathrm{d}N.$$

Similarly

$$\int_0^\infty 2n(2n+1)N^{2n-1}K\mathrm{d}N + 2\int_0^\infty (2n+1)N^{2n}\frac{\mathrm{d}K}{\mathrm{d}N}\,\mathrm{d}N + \int_0^\infty N^{2n+1}\frac{\mathrm{d}^2K}{\mathrm{d}N^2}\,\mathrm{d}N = 0.$$

But the mean value of $I_{T,N}$

$$\frac{1}{\theta^2}\frac{2}{(n-1)!^2}\int_0^\infty N^{2n+1}\frac{\mathrm{d}^2K}{\mathrm{d}N^2}\mathrm{d}N-\frac{1}{\theta^2}\frac{8}{(n-1)!^2}\int_0^\infty N^{2n+1}K\mathrm{d}N$$

$$=\frac{1}{\theta^2}\frac{2}{(n-1)!^2}\left\{-2(2n+1)\int_0^\infty N^{2n}\frac{\mathrm{d}K}{\mathrm{d}N}\mathrm{d}N-2n(2n+1)\int_0^\infty N^{2n-1}K\mathrm{d}N\right\}-\ldots$$

$$=\frac{1}{\theta^2}\left\{2n(2n+1)-4n^2\right\}$$

$$=\frac{2n}{\theta^2} \tag{13}$$

which is equal to the whole of the information available.

Whereas the estimate T is not sufficient, on the *convention* that n only is used in its distribution, the pair of values N and T supplies exhaustive information, N being an ancillary statistic capable in this case of replacing the sample number n.

For the same number n the value of N will be sometimes greater sometimes less; the change of convention which makes the estimation exhaustive is that we judge of the precision of the estimate from the value of N observed, and can then totally ignore the size of the sample on which the estimate is based.

In other cases both the size of the sample n and some other characteristics of it will be required as ancillary information. The phrase 'conditionally sufficient' has been used by Bartlett and others for such a statistic as T in this case. I do not think this helpful, since it gives the impression that the sufficiency referred to is an intrinsic property of the estimate *per se*, whereas I do not know that almost any estimate has not the same property when interpreted in relation to some chosen set of ancillary statistics.

The choice of N in this case is suggested by the Likelihood function. It is justifiable from the fact that the distribution of N for given (n,θ) is independent of θ. That is, the series of rectangular hyperbolas specified by constant values of N does in fact divide up the field of variation of x and y, or of X and Y, in such a way that the total frequency expected between any two contours is independent of θ. This is the connection with the Problem of the Nile.

G. Darmois to Fisher: 20 *August* 1940

Je vous envoie quelques remarques sur le problème du Nil et les résumés exhaustifs. Ce que je dis aux pages 1 et 2 est fait depuis assez longtemps, j'y avais réfléchi après votre passage à Paris et votre Conférence à la Société de Biotypologie [*CP* 156]. Le reste provient de la lecture des pages que vous m'avez envoyées. . . .

[Enclosure]

Problème du Nil — Résumés exhaustifs

Le problème du Nil se pose pour une loi de probabilité à deux (ou plusieurs) variables, dépendant d'un (ou plusieurs) paramètres. Soit, pour fixer les idées $f(xy,\theta)\ d(xy)$. Il faut voir si une fonction convenablement choisie $X(xy)$ a une loi de probabilité indépendante de θ. Bien entendu, on peut construire tout de suite toutes les lois ayant cette propriété. Il suffit de prendre

$$A(X)\mathrm{d}X\,B_X(Y,\theta)\mathrm{d}Y$$

$A(X)$ étant la diversité de probabilité marginale de X, indépendante de θ, $B_X(Y,\theta)$ est une fonction de X, Y, θ, qui est la densité de probabilité liée de Y quand X est fixé. Ces deux fonctions peuvent être les plus générales de leur définition. Il suffit ensuite de remplacer X par une fonction arbitraire de xy, et de remplacer Y par y. Dans ces conditions, les courbes $X(xy) = \mathrm{C}^{\mathrm{te}}$ limitent des aires où la probabilité totale est indépendante de θ.

Si l'on envisage l'information qu'une telle loi de probabilité peut apporter, relativement au paramètre θ, on sait qu'on peut le représenter par:

$$i=E\left\{\frac{\partial}{\partial\theta}\log f\right\}^2=-E\left\{\frac{\partial^2}{\partial\theta^2}\log f\right\}.$$

D'autre part un théorème général relatif à une décomposition de la loi de probabilité sous la forme

$$A(X,\theta)\mathrm{d}X\,B_X(Y,\theta)\mathrm{d}Y$$

où X et Y sont deux fonctions distinctes de xy, indique que l'information i est la somme de l'information qu'on peut déduire de la loi marginale de X, et de l'information qu'on peut déduire de la loi de probabilité liée de Y.

On voit que, dans le cas du problème du Nil, pour toute solution de ce problème, l'information totale est fournie par la seule loi de probabilité de Y quand X est connu (espérance mathématique dans la loi totale). La forme générale de solution que nous avons donnée plus haut permet de former aisément une infinité de solutions particulières. En voici une très simple:

$$\frac{1}{\sqrt{2\pi}}\exp\left(-X^2/2\right)\mathrm{d}X\,\frac{1}{\sqrt{2\pi\sigma}(X,\theta)}\exp\left(-\{Y-m(X,\theta)\}^2/2\sigma^2(X,\theta)\right)\mathrm{d}Y.$$

La loi liée est gaussienne si, pour simplifier encore, on suppose σ indépendant de θ ou aux

$$i=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\frac{1}{\sigma^2(X)}\left\{\frac{\partial m}{\partial\theta}(X,\theta)\right\}^2\exp\left(-X^2/2\right)\mathrm{d}X.$$

On peut trouver des formes explicites de solutions par un procédé moins général, mais qui donne un mécanisme pour l'étalement de la probabilité dans la bande X, $X + \mathrm{d}X$.

Un moyen très simple d'obtenir le résultat est en effet de soumettre les points de cette bande à une transformation ponctuelle, dépendant de θ, mais laissant invariants, à la fois la fonction X et l'élément de probabilité. On voit qu'on est amené à considérer un groupe de transformations qui échange les éléments de probabilité en conservant leur valeur.

Si l'on prend le groupe de transformations:

$$\begin{aligned} X &= x, \\ Y - t &= y, \end{aligned}$$

on voit que l'élément de probabilité a la forme

$$F(X, Y - t)\,\mathrm{d}(XY).$$

Le problème du Nil est résolu par les bandes $X = \mathrm{C}^{\text{te}}$, et les éléments de probabilité se correspondant dans une translation t.

L'exemple de R.A. Fisher. L'élément a pour expressions

$$\mathrm{e}^{-x\theta - y/\theta}\mathrm{d}x\,\mathrm{d}y \qquad \begin{cases} 0 \leqslant x < \infty, \\ 0 \leqslant y < \infty, \\ 0 \leqslant \theta < \infty. \end{cases}$$

Il est clair que si, au point x_1, y_1, on fait correspondre x_2, y_2 tels que:

$$\begin{aligned} x_1\theta_1 &= x_2\theta_2, \\ y_1/\theta_1 &= y_2/\theta_2. \end{aligned}$$

On laisse invariants à la fois la grandeur xy et l'élément de probabilité. C'est la transformation ponctuelle indiquée, laissant invariantes les hyperboles $xy = \mathrm{C}^{\text{te}}$.

On peut mettre l'exemple de R.A. Fisher sous la *forme canonique.*

Posons:

$$xy = \mathrm{e}^X, \; y/x = \mathrm{e}^Y, \; \theta = \mathrm{e}^{t/2}.$$

La probabilité élémentaire devient

$$\tfrac{1}{2}\exp - \{\mathrm{e}^{(X-Y+t)/2} + \mathrm{e}^{(X+Y-t)/2}\}.\,\mathrm{e}^X\,\mathrm{d}X\,\mathrm{d}Y.$$

C'est bien la forme $F(X, Y - t)\,\mathrm{d}X\,\mathrm{d}Y$.

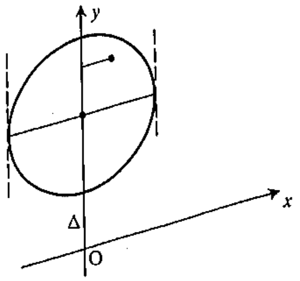
On y trouve en évidence la loi marginale de X

$$\mathrm{d}X \textstyle\int F(X, u)\,\mathrm{d}u$$

où, évidemment, t ne figure plus. L'exemple de R.A. Fisher possède, du point de vue de l'estimation, d'autres propriétés sur lesquelles nous reviendrons.

Quelques exemples. Une loi de Gauss à deux variables, où la forme de l'ellipse de probabilité est connue, mais où le centre est inconnu sur une droite, peut évidemment se mettre sous la forme

$$K \exp - \tfrac{1}{2}\left\{\frac{x^2}{a^2} + \frac{(y-t)^2}{b^2}\right\}\mathrm{d}(xy).$$



La direction de Ox est conjuguée de la direction de la droite Δ, lieu du centre. Il est clair qu'ici l'information est tout entière dans y. x est inutile.

Exemple de rotation. Une loi de Gauss circulaire, dont le centre seul est inconnu sur un cercle de centre O:

$$K \exp - \tfrac{1}{2}\{(x - r\cos\theta)^2 + (y - r\sin\theta)^2\}\,\mathrm{d}(xy),$$

$$x = \rho\cos\omega, \; y = \rho\sin\omega,$$

$$K \exp - \tfrac{1}{2}\{\rho^2 + r^2 - 2\rho r\cos(\omega - \theta)\}\rho\mathrm{d}\rho\mathrm{d}\omega.$$

On voit bien la forme caractéristique

$$F(\rho, \omega - \theta)\,\mathrm{d}\rho\mathrm{d}\omega.$$

Ici, les bandes qui donnent la solution du problème du Nil sont des anneaux circulaires de centre O.

Information, résumés exhaustifs, estimation. Supposons maintenant que nous considérons n couples x_1y_1, x_2y_2, . . . suivant une même loi de probabilité. Nous voulons en extraire tout ou partie de l'information relative à un paramètre θ. Il y a lieu d'introduire la notion générale de résumé d'une série d'observations, et particulièrement de *résumé exhaustif.* Un résumé sera constitué par une suite d'éléments aléatoires fonctions des observations. Par exemple, centre de gravité des points observés, moments . . ., point de plus grande abscisse . . ., etc.

Par exemple, pour la loi à une variable

$$A\,\mathrm{e}^{-(x-\theta)^4}\mathrm{d}x$$

aux observations $x_1, x_2, \ldots x_n$, nous ferons correspondre le résumé:

$$\Sigma x_i, \Sigma x_i^2, \Sigma x_i^3,$$

formé par les trois premiers moments.

Dans le cas le plus général, le résumé possède une certaine loi de probabilité, et le reste des observations possède une loi de probabilité liée (où le résumé est supposé fixé).

L'information totale est toujours égale à la somme de l'information fournie par le résumé et de l'information contenue dans la loi de probabilité liée.

Si cette dernière loi ne contient plus le paramètre θ, nous dirons que le résumé est exhaustif. Il contient alors toute l'information. Il est clair que ces notions sont absolument générales et que, soit dans l'espace des observations, soit dans l'espace des paramètres, le nombre des coordonnées peut être absolument quelconque.

Exemple. La loi considérée plus haut donne, pour n observations:

$$A^n \exp - (n\theta^4 - 4\theta^3 S_1 + 6\theta^2 S_2 - 4\theta S_3 + S_4)\, \mathrm{d}x_1\, \mathrm{d}x_2 \ldots \mathrm{d}x_n.$$

Il est clair que si l'on prend comme nouvelles coordonnées:

$S_1, S_2, S_3, \xi_4, \ldots \xi_n$, on aura

$$A^n \{\exp - (n\theta^4 - 4\theta^3 S_1 + 6\theta^2 S_2 - 4\theta S_3)\}\, \mathrm{e}^{-S_4}\, \Delta \mathrm{d}S_1\, \mathrm{d}S_2\, \mathrm{d}S_3\, \mathrm{d}\xi_4 \ldots \mathrm{d}\xi_n.$$

L'intégrale $\underbrace{\int \ldots \int}_{n-3} \Delta\, \mathrm{e}^{-S_4}\, \mathrm{d}\xi_4 \ldots \mathrm{d}\xi_n$

étendu à la multiplicité où S_1, S_2, S_3 sont fixés, est une fonction de S_1, S_2, S_3, $\phi(S_1, S_2, S_3)$, et la loi de probabilité liée est donc

$$\frac{\mathrm{e}^{-S_4}\, \Delta(S_1, S_2, S_3, \xi_4, \ldots \xi_n)\, \mathrm{d}\xi_4 \ldots \mathrm{d}\xi_n,}{\phi(S_1, S_2, S_3)}$$

ou θ a disparu.

Ainsi, l'ensemble S_1, S_2, S_3 constitue un résumé exhaustif, pour tout ce qui regarde la connaissance du paramètre θ.

Estimation. On a évidemment $E(x_i) = \theta$,

$$E(x_i - \theta)^2 = a_2, \quad E(x_i - \theta)^3 = 0.$$

On pourrait construire diverses estimations de θ. La plus simple est S_1/n, dont la précision est donnée par a_2/n. $a_2 \sim 1/3$. L'efficacité $1/12a_2^2$ vaut en gros 70/100.

Ainsi, pour n très grand, l'estimation de l'optimum donne, à une quantité finie près, toute l'information. Le premier moment perd environ 30/100 des observations.

Pour n quelconque, l'information tout entière est comprise dans le résumé

S_1, S_2, S_3. L'estimation de l'optimum serait fournie par:

$$\theta^3 - 3\theta^2 \frac{S_1}{n} + 3\theta \frac{S_2}{n} - \frac{S_3}{n} = 0.$$

Il est facile de voir que cette équation n'a qu'une racine réelle, $\hat{\theta}(S_1, S_2, S_3)$. Vraisemblablement, elle fournit une très grande partie de l'information. Peut être les dérivées suivantes, prise pour la valeur $\hat{\theta}$

$$\left(\hat{\theta} - \frac{S_1}{n}\right)^2 + \frac{S_2}{n} - \left(\frac{S_1}{n}\right)^2,$$

$$\hat{\theta} - \frac{S_1}{n},$$

qui certainement fournissent avec $\hat{\theta}$ toute l'information, en fournissent celles des parties de plus en plus petites.

Forme des lois à résumés exhaustifs. On peut trouver, sous certaines hypothèses, la forme générale de ces lois, qui comprennent évidemment les lois à estimations exhaustives. J'ai donné cette forme dans une note aux *Comptes Rendus T.200* – 1$^{\text{er}}$ semestre p. 1265 et dans une communication à la session d'Athènes de l'Institut International de Statistique, en Septembre 1936.

Cas de la loi de R.A. Fisher, solution du problème du Nil

L'élément de probabilité a la forme

$$\mathrm{d}f = \mathrm{e}^{\alpha(xy)a(\theta) + \beta(xy)b(\theta) + \gamma(xy) + c(\theta)}\mathrm{d}(xy)$$

qui convient aux résumés exhaustifs. Le résumé est, dans ce cas

$$\rho_1 = \Sigma\, \alpha(x_i y_i),$$
$$\rho_2 = \Sigma\, \beta(x_i y_i).$$

Dans le cas spécial $\rho_1 = \Sigma x_i$, $\rho_2 = \Sigma y_i$,

ces variables ρ_1, ρ_2 fournissent donc un résumé exhaustif. Si on les appelle ξ, η, leur loi de probabilité est de la forme:

$$\mathrm{d}\phi = \frac{1}{\{(n-1)!\}^2}\; \xi^{n-1}\eta^{n-1}\mathrm{e}^{-\theta\xi - \eta/\theta}\mathrm{d}\xi\mathrm{d}\eta.$$

Toute l'information est donc contenue dans cette loi de probabilité.

Or, il se trouve que la loi en ξ, η est aussi une solution du problème du Nil, puisque la transformation

$$\theta_2\xi_2 = \theta_1\xi_1, \quad \eta_2/\theta_2 = \eta_1/\theta_1$$

laisse invariante la fonction $\xi\eta$ et l'élément d'intégrale. Par conséquent, le résumé exhaustif, mis sous la forme

$$X = \xi\eta, \qquad Y = \eta/\xi$$

est tel que toute l'information est fournie par la loi de probabilité de Y liée.
On voit que cette loi très simple possède *à la fois* les propriétés de résumés exhaustifs et de solution du problème du Nil.

Recherche d'autres solutions. Il suffit de trouver une solution de la forme

$$e^{\phi(x,y-t)}dxdy$$

où la fonction ϕ permette un résumé exhaustif. On peut en trouver très aisément de deux types:
1) $\phi(x,y-t)$ est un polynome en $y-t$, dont les coefficients sont des fonctions de x. Par exemple

$$A \exp-(x^4+x^2y^2+y^4)\, d(xy)$$

fournit une solution en remplacant y par $y-t$.

2) $\phi(x,y-t) \equiv \Sigma A_i(x) \dfrac{e^{\alpha_i y}}{e^{\alpha_i t}}$.

On voit que la solution de R.A. Fisher appartient à ce deuxième type.
La solution générale du problème qui consiste à mettre $\phi(x,y-t)$ sous une forme capable de résumé exhaustif ne présente aucune difficulté.
Il est sans doute plus difficile de s'arranger pour que le résumé lui-même soit solution du problème du Nil. Toutefois, l'exemple de la page [73] tiré de la rotation, est une solution. En effet, sous la forme

$$\frac{1}{2\pi}\exp-\tfrac{1}{2}\{(x-r\cos\theta)^2+(y-r\sin\theta)^2\}\,dxdy$$

on aperçoit que cette loi comporte un résumé exhaustif Σx, Σy. En posant $\Sigma x/n = \xi$, $\Sigma y/n = \eta$, la loi de ξ,η est évidemment

$$\frac{1}{2\pi}\exp\left(-\frac{n}{2}\{(\xi-r\cos\theta)^2+(\eta-r\sin\theta)^2\}\right) n\,d\xi\,d\eta.$$

Par conséquent, elle a la même propriété que la loi primitive

$$\xi = R\cos\Omega,$$
$$\eta = R\sin\Omega,$$
$$\frac{n}{2\pi}\exp\left(-\frac{n}{2}\{R^2+r^2-2Rr\cos(\Omega-\theta)\}\right) R\,dR\,d\Omega.$$

La loi de R est donc

$$\frac{n}{2\pi}\exp\left\{-\frac{n}{2}(R^2+r^2)\right\} RdR\int_0^{2\pi}\exp(nRr\cos u)\,du.$$

Elle ne dépend pas de θ; par conséquent, la loi de probabilité liée de Ω fournit toute l'information.

On pourrait peut être, sans trop de peine, trouver d'autres exemples.
Il est à remarquer que l'exemple de R.A. Fisher a une autre propriété, c'est que la loi de probabilité liée qui fournit toute l'information est indépendante de n, nombre des observations.

Fisher to G. Darmois: 26 August 1940

I have been reading your letter of August 20th and its extremely interesting enclosure. I am extremely sorry to hear of your accident with the automobile, which must have been most troublesome, though, from your letter, I am glad to see that you can now write perfectly.

With respect to your enclosure, of which I hope you have a copy, I believe it would be helpful as a matter of exposition to bring the size of the sample, let us say n, into the foreground at an early stage. Thus for the amount of information available on the average from a sample of n we have

$$i = E_n\{\partial(\log f)/\partial\theta\}^2$$

but after observing the ancillary statistic X the actual amount of information which our sample contains may be more or less than this expectation, namely

$$E_{n,X}\{\partial(\log f)/\partial\theta\}^2.$$

At this stage I think it is recognisable that the more general expectation is less appropriate to our particular experience than the more limited expectation, and that it was only by a thoughtless convention, or perhaps *faute de mieux*, through our not having discovered the ancillary function X, that the general expectation ever came to be adopted.

If the distribution of our estimate for given X still depends also on n, then n has been replaced by the pair of values n and X. Sometimes, as in the example I gave, and indeed not infrequently, X can be chosen so as to absorb n altogether, so that the knowledge of precision for which we formerly relied on n is now supplied by X only.

In elementary work this occurs for example when a regression coefficient is calculated by the familiar formula

$$b = S\{y(x-\bar{x})\}/S(x-\bar{x})^2.$$

For any set of samples having given values $x_1, \ldots x_n$ it is easily shown that if y is normally distributed with variance v for each value of x, the sampling variance of b is

$$v/S(x-\bar{x})^2;$$

b is normally distributed about the true regression β as mean, with this variance.

From this it follows that if we extend our aggregate of samples to include all having the same value of $S(x-\bar{x})^2$ as we have observed, the same distribution is true of the estimates b we should obtain from them. Thus the number of pairs of values (x,y) observed in the sample is irrelevant, so soon as we have taken note of the value of $S(x - \bar{x})^2$. This is true whatever may be the sampling distribution of the variate x. Even if this were normal, however, the question—'What is the distribution of b for samples of a given size n?' would be not only more complex, but less useful, to answer than the question of the distribution of b for a given value of $S(x - \bar{x})^2$. This quantity completely replaces or supplants n in determining the precision of our estimate. The pair of values (b,n) is not an exhaustive résumé (i.e. b is not a Sufficient estimate), but the pair $\{b, S(x - \bar{x})^2\}$ constitutes an exhaustive résumé. I suggest that whenever n is required it should appear explicitly in the specification of a *résumé*.

In your pretty example of rotation I imagine that $\sqrt{S^2(x) + S^2(y)}$ replaces n in the same way.

The example you take with the quartic exponent

$$\mathrm{d}f = (1/A) \exp - (x - 0)^4 \,\mathrm{d}x$$

where $$A \text{ is } \tfrac{1}{4}(-\tfrac{3}{4})!.$$

This is of the general form $\phi(x - \theta)\,\mathrm{d}x$ for which complete ancillary information is always supplied by the set of differences between successive observations when these are arranged in order of magnitude, these differences having a distribution jointly or severally independent of θ.

In the case you take, the variance of the mean $\bar{x}$ is

$$(-\tfrac{1}{4})!/n(-\tfrac{3}{4})!$$

whereas the average information from a sample of n is

$$12n(-\tfrac{1}{4})!/(-\tfrac{3}{4})! = 4.1138n$$

giving an efficiency 0.70753. This is the fraction utilized by the mean for large samples. If we take the maximal likelihood estimate $T = \bar{x} + u$ where u is the real root of the cubic equation

$$u^3 + 3m_2 u - m_3 = 0$$

in which

$$m_2 = S(x - \bar{x})^2/n, \quad m_3 = S(x - \bar{x})^3/n,$$

then T will be efficient and will utilize a fraction of the information tending to unity for large samples. T is not sufficient, the résumé (n,T) supplying on the average less than the total amount of information available, the limiting amount of the deficiency for large samples being given by the formula

$$\frac{S\left\{\dfrac{1}{m}\left(m'' - \dfrac{m'^2}{m}\right)^2\right\}}{S\left(\dfrac{m'^2}{m}\right)} - \frac{1}{n}S\left(\frac{m'^2}{m}\right) - \frac{S^2\left\{\dfrac{m'}{m}\left(m'' - \dfrac{m'^2}{m}\right)\right\}}{S^2\left(\dfrac{m'^2}{m}\right)}$$

where m is the expectation in any class, $m' = \mathrm{d}m/\mathrm{d}\theta$, $m'' = \mathrm{d}^2m/\mathrm{d}\theta^2$, which comes to 4.6231, nearly 1/8th more than the amount of information in one observation. The loss of information in such a résumé is not, therefore, very formidable in practice; but it is, I think, of great theoretical interest to consider the procedure by which it may be recovered.

For samples having a given configuration u is constant, and the amount of information supplied by T is the same as that supplied by $\bar{x}$. For such samples the distribution of $\bar{x}$ is

$$\mathrm{d}f = (1/\phi) \exp\left(- n\{(\bar{x} - \theta)^4 + 6m_2(\bar{x} - \theta)^2 + 4m_3(\bar{x} - \theta)\}\right) \mathrm{d}\bar{x}$$

where $$\phi(n,m_2,m_3) = \int_{-\infty}^{\infty} \exp\{- n(t^4 + 6m_2t^2 + 4m_3t)\}\,\mathrm{d}t;$$

the amount of information supplied by $\bar{x}$ (together with the configuration) is therefore

$$i = 12n\, E\{(\bar{x} - \theta)^2 + m_2\}$$

of which the average value is seen without difficulty to be

$$12n\,(-\tfrac{1}{4})!/(-\tfrac{3}{4})!.$$

In this case, although you might properly have considered the distribution of T in samples having absolutely the same configuration, it is clear that the sampling distribution depends only on the two elements m_2 and m_3, which together with n and T supply an exhaustive résumé. The particular amount of information supplied by such a résumé may be written explicitly as

$$12n\left\{m_2 + \frac{\int_{-\infty}^{\infty} t^2 \exp\{- n(t^4 + 6m_2t^2 + 4m_3t)\}\,\mathrm{d}t}{\int_{-\infty}^{\infty} \exp\{- n(t^4 + 6m_2t^2 + 4m_3t)\}\,\mathrm{d}t}\right\}$$

If we know anything about the function ϕ, it might be interesting to write this as

$$12nm_2 - \frac{2}{\phi}\frac{\partial\phi}{\partial m_2}, \text{ or } 12nm_2 + \frac{3}{4n\phi}\frac{\partial^2\phi}{\partial m_3^2}.$$

Indeed this function is full of interesting points.

Fisher to G. Darmois: 22 *March* 1955

. . . I am writing to you now primarily to seek your assistance in a literary reference, for I recall about 1946 when I was on a visit to Paris, at your invitation, that you were good enough to show me a paper of Kolmogoroff's in French translation; in particular you wished to show me the curious axiom which Kolmogoroff had excogitated in wrestling with the problem of fiducial inference.

As I am now engaged in setting out more fully, from the point of view of mathematical logic, the bases, as I understand them, of scientific inference, I am wanting to recover the reference of this little essay in axiomatics.

For a time, as perhaps you know, I was doubtful whether Sir Harold Jeffreys and others were not perhaps right in thinking that the form of reasoning, to which I gave the name 'fiducial', required some special axiom, but I am now fully satisfied that this is not so, but that the matter turns on the fact that the word 'probability' was framed by our predecessors in the 17th and 18th centuries not only with abstract and deductive inferences in view, but with the intention of actually applying the idea to the real world, for example, in the advice given to gamblers, and that in consequence the true meaning of the word includes both the specification of what is known, which enters readily into deductive processes, but also a specific requirement as to what is unknown, which is a type of datum we constantly have to use in inductive reasoning, but which is not easily accommodated to the canons of deduction.

All this confirms me in my belief that what is inferred, using the method of fiducial probability, is a classical probability just as conceived by de Moivre or Montmort, and not in any sense a special kind, or species, of probability, as has been diligently insinuated. In fact there are a great many cases in which fiducial inferences could be experimentally verified to any degree of accuracy.

It is in trying to make it clear that I am myself introducing no new axiom that I want to refer to the attempts of Jeffreys and Kolmogoroff to cope with the problem in this way. I hope you will recall the reference without trouble.

Fisher to W.E. Deming: 25 *September* 1934

Many thanks for sending me your paper with R.T. Birge from *Reviews of Modern Physics.*[1] I think the paper will be found most valuable. It is, I believe, the first attempt to give to physicists, or even to astronomers, a comprehensive account of the ways in which quite modern work has extended and revolutionised the classical theory of errors. You ask me for criticisms, but really I have found very little in substance to criticise. I think the discussion on page 135 is somewhat hard on 'Student's' z test. (By the way, since 1925 'Student' has adopted the transformation I suggested, $t = z\sqrt{n}$, so that he uses the t test as much as I do.) I would not myself admit that 'Student's' test is ever misleading, and it can only be called hazardous in the strict and nonpolemical sense that it lays down and accepts a certain definite hazard. It is the u test which requires guesswork and is, therefore, exposed to objection by those who want their inferences to flow from the data only.

As I expect you know, up to well within the last 15 years writers on statistics were accustomed to be extremely careless in confusing that which is estimated with our estimate of it. The same terms and the same symbols were used for both without distinction. In 1921, in a paper of the *Phil. Trans.* [*CP* 18], aimed at clarifying some of the contradictions and paradoxes of the subject, I introduced two new terms, intended to be antithetical, namely, 'parameter', used to specify the parent population, and 'statistic', calculated from the observed sample. I was quite deliberate in choosing unlike words for these ideas which it was important to distinguish as clearly as possible. That work has now been largely done, so far as concerns the better writers on the subject, and certainly there is no confusion in your paper, where I think you systematically use Greek and Latin letters to distinguish these two classes of quantity; but, perhaps by a slip, you do (Section 3e, line 7) use the expression 'corresponding parameter of a sample', which on consideration you may agree is rather a dangerous one for some classes of student. A population is completely specified by its one or two or more parameters. A sample of n would need n different statistics if these were to be used to specify it. They are, in fact, not used for this purpose at all, but essentially for estimation. To each statistic there corresponds a particular parameter or parametric function to which the value of the statistic tends, as the sample is increased indefinitely, but to each parameter there 'corresponds' in this sense as many different statistics as a cat can have kittens. In fact there is no 1:1 correspondence as suggested by your clause and I am sure it is better not to use the word parameter for one of the fluctuating quantities obtained from samples, which one may call statistical estimates, or something of the kind if that is preferred to the word statistics.

I may say in this connection that I think your exposition in Section 3e of the fiducially related values of σ and s is altogether excellent; the only thing I should add on the logical side is that the statements of fiducial probability obtained should only be taken from distributions, such as s for given σ, where the problem of estimation of the parameters has been completely and therefore uniquely solved, i.e. where s is known to contain the whole of the information contained by the sample. One can see the necessity for this stipulation by considering what would happen if, like the astronomers, we used an estimate of σ based on the mean error, rather than on the mean square error. If s_1 is the estimate, then the distribution clearly will be a function of s_1/σ only and there is nothing but hard work to prevent a misguided astronomer from tabulating the percentile points of the distribution for different sizes of sample. Then, given s_1, it would be possible, apparently, to state the fiducial 5 per cent and 95 per cent points for σ and these would not, of course, agree exactly with the values derived by the mean square method from the same sample. The use of fiducial probability in this precipitate way would, in fact, have led to a definite numerical contradiction, of a kind not unlike those which brought discredit on the use of inverse probability, based on some form of doctrine of insufficient reason.

In the light of the theory of estimation the logical contradiction is easily resolved. An inductive statement (unlike a deductive one) is only true if it is the whole truth; suppression of part of the data and the treatment of the remainder as though it were the whole, although these data are really true and the method of treatment unexceptional if applied to the whole, will, as all statisticians know, lead to very false results. What the theory of estimation is capable of showing is that a definite portion of the information supplied by the sample is omitted or thrown away in using an estimate based on the mean error, but that the whole is retained or conserved in any estimate based on the mean square error. Consequently when faced with such contradictory statements, apparently equally well founded on the fiducial argument, we can with the theory of estimation behind us say that one statement is true and the other is false and why. It is for this reason that I think it worth while to emphasise that the theory of fiducial probability is only an outgrowth or branch of the theory of estimation and that the attempt which Neyman and Pearson have made to make it stand alone without regard to the quantity of information utilised is bound to lead to contradictions and confusion.

Again let me congratulate you most heartily on the completion of a very fine enterprise.

[1] Deming, W.E. and Birge, R.T. (1934). On the statistical theory of errors. *Rev. Mod. Phys.* **6**, 119–61.

Fisher to W.E. Deming: 19 *September* 1935

I am very glad to see from your letter[1] that you only wish to suggest that when σ is known the traditional theory of errors procedure (the u contours of your paper) is appropriate. I got the impression from what you had written that you considered that there was ground for choosing u rather than t, other than the possession of definite knowledge of the value of σ.

If I remember right, 'Student' in putting forward his new test was perfectly clear that he regarded it as a correction of the test traditional up to his time, needed especially for small samples owing to our uncertainty of the true value to be ascribed to the variance of the population.

There is a good deal in the approach chosen by Neyman and Pearson that I disagree with, but so far it seems to have led to nothing more than the conclusion that the tests of significance which I and those who agree with me had previously put forward were the best possible for their purpose; in fact, to use their terminology, the u regions are uniformly the best possible in relation to one class of alternative hypotheses, the population variance being given, while the t contours are uniformly the best possible for another class of alternative hypotheses, the variance being unknown. It is, however, in my

opinion, a pity that these writers have introduced the concept of 'errors of the second kind', i.e. of accepting an hypothesis when it is false, seeing that until the true hypothesis is specified, such errors are undefined both in magnitude and in frequency. Their phraseology also encourages the very troublesome fallacy that when a deviation is not significant the hypothesis tested should be accepted as true.

[1] Deming's letter was in response to Fisher's comments on his manuscript that E.B. Wilson had sent to Fisher for an opinion. (See Fisher's letter of 20 May 1935 to Wilson (p. 237).) In his letter, Deming suggested that Neyman and Pearson's papers were the source of his belief that it is *not* a matter of complete indifference as to which samples are rejected — a view that Fisher had criticized.

W.E. Deming to Fisher: 27 *July* 1937

Recently a certain question has come up involving maximum likelihood. It may seem trivial, but the fact is that there is disagreement among people who ought to be able to agree. I appeal to you as the only person who can set us all at ease, hoping that this encroachment on your time will not be irksome.

For the likelihood L of n normal variates having values $x_1, x_2, \ldots, x_n$ we have

$$L = (\sigma\sqrt{2\pi})^{-n}\exp\{-\Sigma(x_i-\mu)^2/2\sigma^2\}. \tag{1}$$

Now $\mathrm{d}L/\mathrm{d}\mu = 0$ gives $\mu = \bar{x}$ (2)

$\mathrm{d}L/\mathrm{d}\sigma = 0$ gives $\sigma^2 = s^2 + (\bar{x}-\mu)^2$ (3)

(This s is the S.D. of the sample, as you perceive; not your s.)

Now the question is, what is the maximum likelihood estimate of σ made from the sample of the n observations? Professor Birge and I have taken the estimate of σ^2 to be $s^2 + (\bar{x}-\mu)^2$ as given on the right-hand side of Eq. (3). We say that if you don't know μ then you must go to the distribution of s and apply maximum likelihood to it, the result being of course $s^2n/(n-1)$ for the estimate of σ^2, n being the number of observations. The distinction between the two estimates

$$s^2 + (\bar{x}-\mu)^2 \text{ and } s^2n/(n-1)$$

is that the former contains more information than the latter; knowledge of μ gives us a trifle better estimate than s alone can give.

The point of disagreement comes when some people insist that Eqs. (2) and (3) are to be taken as *simultaneous* in μ and σ. If one does that, he replaces μ by $\bar{x}$ in Eq. (3) and gets simply s^2 for the estimate of σ^2. To Professor Birge and me this seems a highly arbitrary and destructive procedure.

Have you ever published anything on the simultaneous estimation of several parameters, where the estimates by maximum likelihood involve other parameters, as the estimate of σ involves μ? Has anything ever come out in print on this subject? . . .

Fisher to W.E. Deming: 7 August 1937

I have your letter of July 27. It is curious that the point you raise is one of the first that attracted my attention to Statistics. Indeed, I wrote a juvenile paper upon it in 1912 or 1913. It has no particular merit for your purpose.

I notice that with the definition

$$s'^2 = S(x-\bar{x})^2/n$$

you call s' *the* S.D. of the sample. I think this is a very arbitrary nomenclature, since, whereas the standard deviation of the population sampled is unequivocally defined, a sample can provide innumerable different estimates of varying merit, and, without discussing the relative advantages of these, any one of these might equally be called the S.D. of the sample. When the theory of estimation was developed one of the points which I found most surprising is that bias in an estimate, at least bias of the order of $1/n$, where n is the size of the sample, is of no practical importance, whereas its variance and, in the theory of small samples, the form of its distribution curve, is all important. The reason is that, in estimating the value of an unknown parameter θ, you are equally estimating the value of any given function of θ, and if one such estimate is chosen to be unbiassed, the others will generally be found to have positive or negative biases of order $1/n$.

Now there is no denying the fact that, when there is more than one unknown parameter, maximising the likelihood provides simultaneous equations of estimation and these are solved by taking as the estimated variance the statistic s'^2, as defined above. This is not the estimate I use, but it might be used, at some later algebraic inconvenience, to lead to identically the same tests of significance. In fact, in these tests of significance we deliberately make exact allowance for the sampling distribution of s' or s, and avoid the older practices of assuming the true standard deviation to be equal to s' or s; and since s' is a known function of s, to make exact allowance for the sampling distribution of one is to do so equally for the other.

My reason for using an unbiassed estimate, s^2, of the variance, apart from the fact that it simplifies the algebra, of testing the significance of the differences of two variances drawn from samples of different sizes, is that variances are things which one often wants to sum or to average. If this were equally true of standard deviations, or of invariances, it would be equally desirable to use unbiassed estimates of these; and for these, of course, s and $1/s^2$ are not unbiassed estimates. The indifference of bias is brought out most clearly by the property of sufficiency. The fact that the likelihood function involves, apart from a constant factor peculiar to the sample, only the statistics s^2 and $\bar{x}$, shows that these are jointly sufficient for the estimation of the mean and variance of the population. The relation is essentially a joint one. We cannot infer from it in general that $\bar{x}$ is a sufficient estimate for μ and s^2 for the variance. Indeed, as you note, if μ is known, the variance is properly estimated by $s'^2+(\bar{x}-\mu)^2$. This happens also to be a sufficient estimate. But if, instead of μ being known, there was known only some more general functional relationship between the mean and the variance, the maximum likelihood estimates of these will generally involve both statistics, and not generally be sufficient. The joint sufficiency, however, of $\bar{x}$ and s^2 implies equally the joint sufficiency of any two independent functions of n, $\bar{x}$, and s^2, if such functions can be regarded as estimates at all.

In fact, I think the distinction you are drawing is one without an essential difference, one's choice of an unbiassed estimate being arbitrary in the sense that it is only justifiable by the use to which the estimate is intended to be put.

Click here for next section