BULLETIN

DE

L'INSTITUT INTERNATIONAL

DE STATISTIQUE

TOME XXII - 1ère LIVRAISON

ROMA
PROVVEDITORATO GENERALE DELLO STATO
LIBRERIA
1926

TROISIÈME PARTIE.

Rapports, Communications et Mémoires présentés à la XVIème Session de l'Institut International de Statistique de Rome (à suivre dans la 2ème et 3ème Livraison);

RÉSUMÉ DU MÉMORANDUM SUR L'ÉVALUATION DE LA PRÉCISION OBTENUE PAR LE CHOIX D'UN ÉCHANTILLON.

Par A. L. Bowley.

I. *Choix au hasard.* Un certain nombre de personnes ou d'objets est choisi dans une totalité soigneusement définie, de manière que chaque personne (objet) ait la même chance d'être compris. Il convient de regarder séparément trois cas spéciaux: (1) la fréquence d'une qualité que peuvent posséder les personnes; (2) la distribution d'après l'importance d'une qualité ou d'après des qualités alternatives; (3) la moyenne d'une quantité variable caractérisant chaque personne. Un exemple du premier cas serait le nombre proportionnel de personnes payant l'impôt sur le revenu; du second cas: le nombre proportionnel de personnes dont le revenu s'élève respectivement à moins de £200, £200–£500, etc.; du troisième cas: le revenu moyen.

Les proportions et les quantités résultant du choix peuvent être établies, dans le premier et dans le troisième cas, dans des limites déterminées par des erreurs moyennes auxquelles s'applique le tableau de probabilités de la loi normale des erreurs, abstraction faite de certaines réserves et modifications qui sont discutées dans des termes mathématiques. Dans le second cas, où il s'agit d'un nombre de classes, la probabilité peut être déterminée comme étant au-dessus d'une certaine fonction des erreurs des classes séparées.

Ces résultats généraux sont, dans leur forme la plus élémentaires, les suivants:

1*er cas*: n unités sont choisies au hasard parmi N, et $p \times n$ signifie le nombre d'unités possédant la qualité en question; dans ces conditions la proportion de la totalité (N) est exprimée par la formule suivante:

$$p \pm \sqrt{p\,(1-p)\left(\frac{1}{n}-\frac{1}{N}\right)},$$

l'expression après $\pm$ indiquant l'erreur moyenne, et non pas "l'erreur probable," terme souvent employé. Si p est insignifiant la formule se modifie.

2*e cas*: n unités sont choisies au hasard parmi N, et les proportions de certains groupes alternatifs (dont le nombre est de c) sont parmi les N de $P_1, P_2 \ldots P_c$, et parmi les n de $p_1, p_2 \ldots p_c$; dans ces conditions il est à supposer que les proportions de la totalité diffèrent de celles de l'échantillon de sorte que χ^2 n'est pas au-dessus du chiffre de $(c-2)$, alors qu'il est très peu probable qu'il s'élève au-dessus du chiffre de $2c$, étant donné:

$$\chi^2 = \left\{\frac{(P_1-p_1)^2}{p_1} + \frac{(P_2-p_2)^2}{p_2} + \ldots + \frac{(P_c-p_c)^2}{p_c}\right\} : \left(\frac{1}{n}-\frac{1}{N}\right).$$

3*e cas:* n unités sont choisies au hasard parmi N, et la quantité en question qui les caractérise est de $x_1, x_2 \ldots x_n$; alors à condition que

$$\bar{x} = \frac{1}{n}(x_1 + x_2 + x_3 + \ldots x_n) \text{ et que } s^2 = \frac{1}{n}\{(\bar{x} - x_1)^2 + (\bar{x} - x_2)^2 + \ldots + (\bar{x} - x_n)^2\}$$

la moyenne de la quantité en question, pour la totalité (N), peut être exprimée dans la manière suivante:

$$\bar{x} \pm s\sqrt{\frac{1}{n} - \frac{1}{N}}.$$

Dans chacun des cas la précision du résultat s'augmente, si l'échantillon, au lieu d'être choisi exclusivement au hasard, est établi de sorte qu'on classifie, d'abord, la totalité dans un certain nombre de groupes (méthode dite "stratification"), et choisit, dans une proportion égale, un nombre d'unités dans chacun des groupes, le choix étant pour chaque groupe dirigé par le hasard. L'augmentation de la précision n'est, dans des circonstances ordinaires, que faible, mais elle peut être considérable dans des cas exceptionnels, qui présentent de grandes variations des qualités ou des quantités en question.

J'ai souligné la difficulté logique à établir, à la base de l'échantillon, une conclusion quantitative au sujet de la totalité (problème différent du problème direct d'évaluer la précision d'un échantillon dans des circonstances où la totalité est connue), et j'ai tiré l'attention sur une méthode permettant de surmonter ladite difficulté.

Parmi les facteurs déterminant la précision le plus important est, dans tous les cas, $\frac{1}{\sqrt{n}}$; *la précision augmente avec la racine carrée du nombre choisi, mais non pas dans une proportion directe.*

II. *Choix raisonné.* L'unité faisant l'objet du choix est un district ou un groupe, dont chaque membre est compris dans l'échantillon. Le choix est établi de sorte que l'ensemble des districts choisis donne les mêmes résultats que la totalité pour ce qui concerne certaines quantités ("caractères de contrôle"), dont on a déjà connaissance, et pour les districts choisis et pour la totalité, et qui présentent une corrélation avec les proportions et quantités inconnues qui font l'objet de l'étude.

J'ai démontré que la moyenne ou la proportion de la totalité s'accorde à celle que présente l'ensemble des districts choisis (après une faible correction dont le calcul est possible), dans des limites déterminées par une erreur moyenne contenant 4 facteurs, à savoir:

$$s_x \cdot \sqrt{1 + \frac{s_a^2}{a^2}} \cdot \sqrt{R'} \cdot \frac{1}{\sqrt{n}}.$$

Dans cette formule s_x signifie l'erreur moyenne du groupe de fréquence formé par les diverses valeurs de x (la quantité ou la proportion étudiée), constatées dans les divers districts; a signifie l'importance moyenne du district (nombre de personnes ou d'objets y appartenant), et s_a l'erreur

moyenne du groupe de fréquence formé par les chiffres indiquant l'importance des divers districts. R' dépend de la corrélation entre la quantité étudiée et les "caractères de contrôle" et de la corrélation entre l'un et l'autre de ces caractères; n signifie le nombre des districts.

S'il y a seulement un "caractère de contrôle," on obtient:

$$R' = 1 - r^2,$$

r signifiant le coefficient de corrélation entre le "caractère de contrôle" et la quantité étudiée. L'avantage obtenu par l'augmentation du nombre des "caractères de contrôle" n'est que faible, particulièrement s'il existe entre eux une corrélation. *En effet, l'emploi de "caractères de contrôle" n'augmente, dans des problèmes ordinaires, pas considérablement la précision, et la précision dépend plutôt d'une grande valeur de n, le nombre des districts, et d'une petite valeur de s_x, celui mesurant la variation entre les districts de la quantité ou de la proportion étudiée.*

Si l'étude porte sur une classification dans un certain nombre de groupes alternatifs, au lieu de se rapporter à une seule qualité, les mêmes considérations s'y appliquent. Cependant, il y a quelquefois une faible augmentation de la précision, due à la corrélation entre les proportions des diverses classes d'un district.

Je tiens à reconnaître l'assistance précieuse que m'a rendue M. E. C. Rhodes (Agrégé en statistique à l'université de Londres) en fournissant des suggestions, des critiques et des vérifications.

SUMMARY OF THE MEMORANDUM ON MEASUREMENT OF THE PRECISION ATTAINED IN SAMPLING

By A. L. Bowley

I. *Random selection.* A number of persons or things is selected in such a way that every one in a carefully defined universe has an equal chance of inclusion. We have to consider three cases: first, the prevalence of one attribute which the persons may possess; second, the distribution in grades or among alternative attributes; third, the average of a variable magnitude associated with each person. An example of the first would be the proportion of men who paid income-tax; of the second, the relative numbers whose incomes were less than £200, £200 to £500 etc.; of the third, the average income.

The resulting proportions and quantities can be stated in the first and third cases with standard deviations to which the table of probabilities of the normal law of error applies, with certain limitations and modifications which are discussed mathematically. In the second case, where we deal with a number of classes, the probability can be assigned of not exceeding a certain function of the errors in the separate classes.

The general results in their simplest form are as follows:

Case 1. If n things are at random selected out of N, and $p \times n$ are found to possess the attribute in question, the proportion in the universe (N) may be written

$$p \pm \sqrt{\left\{p\,(1-p)\left(\frac{1}{n}-\frac{1}{N}\right)\right\}},$$

where the expression after $\pm$ is the standard deviation, not the "probable error" which is often used. If p is very small the formula is modified.

Case 2. If n things are selected at random out of N, and among the N the proportions in certain alternative groups (c in all) are $P_1, P_2 \ldots P_c$, and among the n the proportions are $p_1, p_2 \ldots p_c$, then we may expect the proportions in the universe (N) to differ from $p_1, p_2 \ldots p_c$ in such a way that χ^2 does not exceed $(c-2)$, while it is very unlikely that it will exceed $2c$, where

$$\chi^2 = \left\{\frac{(P_1-p_1)^2}{p_1}+\frac{(P_2-p_2)^2}{p_2}+\ldots+\frac{(P_c-p_c)^2}{p_c}\right\} \div \left(\frac{1}{n}-\frac{1}{N}\right).$$

Case 3. If n things are selected at random out of N, and the magnitudes associated with them are $x_1, x_2 \ldots x_n$, then, if

$$\bar{x} = \frac{1}{n}(x_1 + x_2 + \ldots x_n), \text{ and } s^2 = \frac{1}{n}\{(x_1-\bar{x})^2 + (x_2-\bar{x})^2 + \ldots + (x_n-\bar{x})^2\},$$

the average of the magnitudes in the universe (N) may be written

$$\bar{x} \pm s\sqrt{\left(\frac{1}{n}-\frac{1}{N}\right)}.$$

In every case the precision of the result is improved, if, instead of selecting purely at random, we first divide the universe into a number of

classes (the method called "stratification") and select equal proportions at random from each. The improvement in ordinary cases is but slight; but there are exceptional cases, when the qualities or magnitudes vary considerably from one class to another, where it is important.

Emphasis is laid on the logical difficulty of making a quantitative inference from the sample to the universe (as contrasted with the direct problem of measuring the precision of a sample, when the universe is known), and a method is suggested by which this difficulty can be overcome.

In every case the most important factor in the precision is $1 \div \sqrt{n}$*; the precision increases, but not in direct proportion, with the square root of the number selected.*

II. *Purposive selection.* Here the unit of selection is a district or group, every member of which is included in the sample. The selection is so made that the aggregate of the districts gives the same results as the universe in respect of certain quantities (called "controls") which are known in the districts and universe, and which are correlated with the unknown proportions or quantities which are the subject of investigation.

It is shown that the average or proportion in the universe is that found in the aggregate of the districts (after a small calculable correction) subject to a standard deviation containing four factors, viz.:

$$s_x \cdot \sqrt{\left(1 + \frac{{s_a}^2}{a^2}\right)} \cdot \sqrt{R'} \div \sqrt{n}.$$

Here s_x is the standard deviation of the frequency group formed by the various values of x (the quantity or proportion under investigation) in the different districts. a is the average size of (number of persons or things in) the district, and s_a the standard deviation of the frequency group formed by these sizes. R' depends on the correlation between the quantity under investigation and the "controls" and on the correlation between the controls. n is the number of districts.

If there is only one control, $R' = 1 - r^2$, where r is the correlation coefficient between the control and the quantity under investigation. Little improvement is obtained by increasing the number of controls, especially if they are correlated with each other. *In fact, in ordinary problems the use of controls does not increase the precision greatly, and reliance must be placed rather on the greatness of* n*, the number of districts, and the smallness of* s_x *which measures the variation of the quantity or proportion investigated between the districts.*

If we are concerned with the distribution in grades, instead of with a single attribute, the same considerations apply, but in some cases there is a slight increase of precision, owing to correlation between the proportions in the grades within a district.

I desire to acknowledge the valuable help given by Mr E. C. Rhodes (Reader in Statistics in the University of London) in suggestion, criticism and verification.

MEASUREMENT OF THE PRECISION ATTAINED IN SAMPLING

By A. L. Bowley

I. RANDOM SELECTION

Summary

PAGE

Distinction between attributes and variable. In selection the fundamental rule is that of equal chances, which may be obtained by pure chance or by stratification 6

For attributes the precision depends mainly on the square root of the number in the sample divided by the proportion found. A more complex formula is involved if a distribution in grades is in question 8

For variables the precision depends mainly on the square root of the number in the sample divided by the standard deviation of the distribution 9

The formulae may be expressed in terms of the constants of the universe (A. Direct problem) or in terms of the results of the sample (B. Inverse problem). The latter is the practical problem and it involves the theory of inverse probability:

One attribute. A. There are four cases. 1. Pn is large, where P is the proportion in the universe, n the total number in the sample. 2. P is small, but Pn finite. 3. n is small. 4. The selection is stratified 10

B. The formulae are expressed in terms of p, the proportion found in the sample 12

Inference is drawn from the sample as to the proportion in the universe . . 15

Distribution of attributes. A formula showing the chance of exceeding a certain function of the errors is given, and its use illustrated 16

Magnitude of an average. A. Full and approximate formulae are given for the precision of an average, for random and for stratified selection 18

Examples are worked to illustrate the improvement obtained by stratification . 19

B. The inverse problem can only be solved if n is large 20

The mathematical analysis leading to the preceding conclusions is given in a series of notes 22

Notation used 23

INTRODUCTION

The problem is divided throughout into two sections, in one of which we consider the *prevalence* in the population of a characteristic or *attribute* which is either present or absent independently of its intensity, in the other the average *magnitude* of some *variable* quantity which is universally present. An example of the former is the male sex, of the latter is age. Mr Yule's nomenclature is followed in speaking of the first section as the problem of "attributes," of the second that of "variables."

In both cases the first necessity is to define exactly the population or

"universe" in question. Only those populations can be treated in which there exists or can be made an adequate directory or list of members, every one of which is theoretically accessible to observation. The attribute or variable must also be adequately defined.

It being decided, from considerations discussed below, how many persons or things should be observed, a number of them is selected at random in such a way that *à priori* every person or thing has an equal chance of being selected. The universe which is sampled is in fact limited by this condition. If, for example, observations of children of school age were to be made, the universe might be either children present on a certain day in state-supported schools, or in any schools, or children on the register of schools whether present or absent, or all children in the country between certain ages whether on a school register or not. Which of these universes is in fact represented depends on the answers to the questions:— for which was it that we had or could make a list, and from which was it that we selected children at random, each with an equal chance of inclusion?

The selection may be made in either of two ways. The first corresponds to drawing for prizes in a lottery. A number is assigned to each member of the population, and there are chosen at random from an independent list (which contains the same total of entries) sufficient numbers for the pre-determined size of the sample. The persons whose numbers correspond win the prize, in this case the privilege of being selected for observation. In the second way, distinguished in the sequel as the method of stratification, a list of persons or things is taken which may be grouped in classes or districts, and in each district the same proportion is selected at random for observation. Especially in the second method, the selection may be made by taking e.g. the first, the twenty-first, the forty-first etc. throughout the list, comprising all districts, if the sample is to be one in twenty (and corresponding intervals for other proportions), unless the order is in any way correlated with the attribute.

Minute precautions are necessary to ensure that the method of selection is completely uncorrelated with the presence of the attribute or the size of the variable.

The selection being made, every person or thing selected must be observed, if possible. Where observation is impossible or inaccurate the resulting unknown element must be retained and exposed in the final report.

Any breach of these conditions, however slight, introduces an unknown element of error in the result, and destroys the relevance of the formulae. It is naturally to be understood that very small departures from the rule in large samples cannot have any great effects, but in general the magnitudes of the resulting errors cannot be estimated.

A common and very injurious departure from the rules is to ignore persons or things in which observation is difficult, e.g. when no one is

present at a selected house when the investigator calls. Another and even more obvious mistake is to define the universe loosely, and to be content with answers from people who happen to be willing to give them.

Complete formulae are given in the following section, and the mathematical analysis on which they depend in a concluding section. Here the more important results are summarized.

Attributes.

A sample containing n persons or things is selected from a universe containing N. Then $p \times n$ among the persons are found to possess a certain attribute. P is the unknown proportion in the universe.

The difference $P \sim p$ is subject to a standard error

$$\sqrt{\left\{\frac{P\,(1-P)}{n}\left(1-\frac{n}{N}\right)\right\}},$$

for which we may write

$$\sqrt{\left\{\frac{p\,(1-p)}{n}\left(1-\frac{n}{N}\right)\right\}}.$$

The table of the normal curve of error applies to this expression*, if pn is sufficiently large. The odds are about 2 to 1 against this error being exceeded, and about 21 to 1 against twice and about 370 to 1 against three times this error being exceeded.

As a rough guide we may say that pn should exceed 100. The result is approximately true when pn is less than 100 but exceeds 20, but then there is a tendency for the proportion to be under rather than over-estimated.

If pn is less than 20, the form of the standard error is unchanged, but the normal table should be replaced by the table for small numbers (p. 36).

A convenient approximate expression for the standard error, which always exaggerates it, is $\frac{1}{n} \cdot \sqrt{(pn)}$.

If for example we find 100 cases with the attribute in a sample containing 500 cases, we might write

$$100 \pm \sqrt{100} = 100 \pm 10,$$

for the number we ought to have found; or, for p, $\cdot 2 \pm \cdot 02$.

If we were content to know that the proportion was between $\cdot 14$ and $\cdot 26$ this would be sufficient. If not, we must increase the size of the sample.

If in this case there were 5000 persons in the universe, the more correct statement for the standard error of p would be

$$\sqrt{\left\{\frac{\cdot 2 \times \cdot 8}{500} \times \left(1-\frac{500}{5000}\right)\right\}} = \cdot 017.$$

If the sample is stratified the standard error is reduced in accordance

* Subject to considerations discussed on p. 15 and p. 42 below. It is believed that the conditions there named are commonly satisfied in the kind of sample here under discussion.

with the formula (38) on p. 32; but in ordinary cases the reduction is inconsiderable. The method is, however, to be recommended as giving some additional security.

It is often required to state the proportion of those in a number of alternative classes (together making up the universe) at the same time, instead of in only one class. Thus instead of simply estimating the proportion of males, we might want to know for each sex the number of infants, children, adults, etc., each term being defined. The above formula applies to each class singly, but a test of composite agreement is discussed also on pp. 16 and 37 below.

Commonly N is so great that the term $\frac{n}{N}$ in the formula for standard deviation may be ignored. Then the error depends only on p and n, and the resulting precision is independent of the size of the universe, provided always that the rule of equal chance of selection can be extended throughout the universe.

Variables.

The standard error for the average of the magnitudes of some measurable characteristic possessed by all the persons or things in a universe containing N, is

$$\frac{s}{\sqrt{n}} \cdot \sqrt{\left(1 - \frac{n}{N}\right)},$$

where n is the number in the sample, and s is the standard deviation from their average of the observed magnitudes.

The rules and considerations explained for attributes apply, except that since s itself is subject to error, n ought to be large enough to allow us to neglect $\frac{1}{\sqrt{n}}$. That n should exceed 100 is a fairly safe rule, so far as this difficulty is concerned.

To determine generally how large n should be, we must take s as well as n into consideration.

Suppose we wish to determine the earnings of coal-miners correct to one shilling per week, and that our sample gives an average 61 shillings with standard deviation 10 shillings.

The average would be stated as

$$61 \pm \frac{10}{\sqrt{n}} \cdot \sqrt{\left(1 - \frac{n}{N}\right)}.$$

If $n = 900$ the standard error is less than $\frac{1}{3}$ of a shilling, and we obtain at least the required precision, whatever N is.

The standard error is reduced by stratification to an extent given by formula (17), p. 19 below, and there illustrated by examples.

SUMMARY OF RESULTS

I. SAMPLING FOR THE PREVALENCE OF ONE ATTRIBUTE

A. Direct Problem

In a universe or population containing N persons or things PN have a certain attribute. The whole population having been numbered or otherwise indexed, n persons are selected at random. Required to find the probability that the number in the sample possessing the attribute should be pn.

Write $pn = Pn + x$, $Q = 1 - P$, $q = 1 - p$.

Then $qn = Qn - x$.

There are ${}_NC_n$ equally probable combinations of n things chosen out of N, of which ${}_{PN}C_{pn} \times {}_{QN}C_{qn}$ contain pn with, and qn without, the attribute.

Write E_x for the required probability.

$$E_x = {}_{PN}C_{pn} \times {}_{QN}C_{qn} \div {}_NC_n = {}_{PN}C_{Pn+x} \times {}_{QN}C_{Qn-x} \div {}_NC_n$$
$$= \frac{(PN)!\,(QN)!\,n!\,M!}{N!\,(Pn+x)!\,(PM-x)!\,(Qn-x)!\,(QM+x)!},$$

where $M = N - n$.

We may take $n < M$ and $P < Q$.

Case 1. *Random sample.*

Pn is so large that we may neglect $\frac{1}{Pn}$ in comparison with unity.

Then

$$E_x = \frac{1}{\sigma\sqrt{(2\pi)}} \cdot e^{-\frac{x^2}{2\sigma^2}} \cdot \left\{1 - \frac{Q-P}{2\sigma}\left(1 - \frac{2n}{N}\right)\left(\frac{x}{\sigma} - \frac{x^3}{3\sigma^3}\right)\right\}, \quad \ldots(1)$$

where $\sigma^2 = PQn\left(1 - \frac{n}{N}\right)$, Note p. 32. Formula (36)

and the chance that p should not differ from P by more than z is

$$\int_{-z}^{z} \frac{1}{\sqrt{\left\{2\pi PQ\left(\frac{1}{n} - \frac{1}{N}\right)\right\}}} \cdot e^{-\frac{z^2}{2PQ\left(\frac{1}{n} - \frac{1}{N}\right)}} \cdot dz \quad \ldots\ldots\ldots\ldots(2)$$

since the terms containing $Q - P$ disappear on integration.

Note p. 32. Formula (39)

Case 2. *P small.*

n is still so large that $\frac{1}{n}$ is negligible, but P is small and of the order $\frac{1}{n}$, so that Pn is finite and $\frac{1}{Pn}$ is not negligible.

Write $Pn + x = r$, $n = kN$, and $Pn = w$.

Then $$E_x = \frac{e^{-w} \cdot w^r}{r!}, \text{ if } k \text{ is negligible, } \ldots\ldots\ldots\ldots\ldots\ldots(3)$$

but, if k is retained and $\frac{1}{PN}$ neglected,

$$E_x = \frac{e^{-w} \cdot w^r}{r!} \cdot (1-k)^{-\frac{1}{2}} \cdot e^{-\frac{x^2 k}{2w(1-k)}}. \quad \ldots\ldots\ldots\ldots\ldots\ldots(4)$$

Note p. 35. Formulae (42) and (43)

The results given in Cases 1 and 2 are practically the same, when w is as great as 20 and n is as great as 1000. Note p. 35

Case 3. *Small sample.*

If n is small and $\frac{n}{N}$ is negligible,

$$E_x = {}_nC_{Pn+x} \cdot P^{Pn+x} Q^{Qn-x}, \quad \ldots\ldots\ldots\ldots\ldots\ldots(5)$$

the value of which can be written down with the help of a table of Binomial coefficients.

Case 4. *"Stratified" sample.*

If the population is an aggregate of the populations of d districts, in which the populations are $N_1, N_2 \ldots N_d$ and $P_1N_1, P_2N_2 \ldots P_dN_d$ persons have the attribute, and $kN_1, kN_2 \ldots kN_d$ are selected in each case at random from the various districts, then the standard deviation of the frequency group of E_x becomes

$$\sigma_d = \sqrt{\left\{(PQ - \sigma_v^2)\, n \left(1 - \frac{n}{N}\right)\right\}}$$

Note p. 32. Formula (38)

instead of $$\sigma = \sqrt{\left\{PQn\left(1 - \frac{n}{N}\right)\right\}},$$

where $$N\sigma_v^2 = N_1 (P_1 - P)^2 + N_2 (P_2 - P)^2 + \ldots + N_d (P_d - P)^2,$$

and $$PN = P_1N_1 + P_2N_2 + \ldots + P_dN_d.$$

Then $$E_x = \frac{1}{\sigma_d \sqrt{(2\pi)}} \cdot e^{-\frac{x^2}{2\sigma_d^2}} \cdot \left\{1 - \frac{\kappa_d}{2}\left(\frac{x}{\sigma_d} - \frac{x^3}{3\sigma_d^3}\right)\right\}, \quad \ldots\ldots\ldots\ldots(6)$$

where

$$\kappa_d \sigma_d^3 = n\left(1 - \frac{n}{N}\right)\left(1 - \frac{2n}{N}\right) PQ\,(Q - P) \times \left\{1 - \frac{3\sigma_v^2}{PQ} + 2 \cdot \frac{N_1 (P - P_1)^3 + N_2 (P - P_2)^3 + \ldots}{NPQ\,(Q - P)}\right\}.$$

To take an extreme case, if all the persons possessing the attribute were concentrated in district 1,

$$\sigma_d^2 = nP\left(1 - P\frac{N}{N_1}\right)\left(1 - \frac{n}{N}\right),$$

and $$E_x = \frac{1}{\sigma_d \sqrt{(2\pi)}} \cdot e^{-\frac{x^2}{2\sigma_d^2}} \cdot \left\{1 - \frac{1 - 2P_1}{2\sigma_d}\left(1 - \frac{2n}{N}\right)\left(\frac{x}{\sigma_d} - \frac{x^3}{3\sigma_d^3}\right)\right\}.$$

Note p. 32

If P_1 approaches unity σ_d approaches zero.

When P is very small, there is a perceptible chance, viz. $e^{-Pn} = e^{-w}$, of missing the attribute in question altogether. If for example w is less than ·4, e^{-w} is greater than ·018. In a stratified sample this chance is diminished a little; if all the cases having the attribute are in one district, the diminution is somewhat greater. Note p. 37. Formula (44)

Thus increased accuracy is always attained by stratification, unless the attribute is evenly distributed throughout the districts, and in some cases the improvement is considerable.

B. Inverse Problem

Given that in a sample of n persons or things, drawn at random from a universe containing N, pn possess a certain attribute, what can we infer about the prevalence of the attribute in the universe?

The answer is given in two parts; in one the chances that the sample would be drawn from various hypothetical universes are compared; in the other it is considered under what circumstances we can make any inference about the relative chances that in fact the universes contained given proportions. The second part alone involves the theory of inverse probability.

For the first part we have merely to recast our formulae so that they depend on the observed p, instead of on the P in the universe. The result is that in an unstratified sample the expectation that $pn + x$ will be found from a universe in which the proportion is P is

$$E_x = \frac{1}{\sigma' \sqrt{(2\pi)}} \cdot e^{-\frac{x^2}{2\sigma'^2}} \cdot \left\{1 - \frac{q-p}{6\sigma'}\left(2 - \frac{n}{N}\right)\frac{x^3}{\sigma'^3}\right\}, \quad \text{.....(7)}$$

Note p. 40. Formula (47)

where $$\sigma'^2 = pqn\left(1 - \frac{n}{N}\right),$$

if $\frac{1}{pn}$ is negligible.

If $\frac{1}{\sqrt{n}}$ is negligible, this reduces to

$$E_x = \frac{1}{\sigma' \sqrt{(2\pi)}} \cdot e^{-\frac{x^2}{2\sigma'^2}} \text{.....(8)}$$

If $\frac{n}{N}$ is negligible, but $\frac{1}{\sqrt{n}}$ not negligible,

$$E_x = \frac{1}{s' \sqrt{(2\pi)}} \cdot e^{-\frac{x^2}{2s'^2}} \left\{1 - \frac{q-p}{3s'} \cdot \frac{x^3}{s'^3}\right\}, \quad \text{.....(9)}$$

where $s'^2 = pqn$.

If $\frac{1}{\sqrt{n}}$ and $\frac{n}{N}$ are negligible,

$$E_x = \frac{1}{s' \sqrt{(2\pi)}} \cdot e^{-\frac{x^2}{2s'^2}}. \quad \text{.....(10)}$$

Further, if $\frac{1}{\sqrt{n}}$ is negligible, $s^2 \doteq PQn$ and $\sigma^2 = s^2\left(1 - \frac{n}{N}\right)$ do not differ significantly from s'^2 and σ'^2.

In each case E_x is the chance that pn would be found, if the proportion in the universe was $p - \frac{x}{n}$.

In a stratified sample we can only proceed definitely if $\frac{1}{\sqrt{n}}$ is negligible.

Then
$$E_x = \frac{1}{\sigma'_d\sqrt{(2\pi)}} \cdot e^{-\frac{x^2}{2\sigma_d'^2}}\, dx, \quad \text{.....(11)}$$

where

$$\sigma_d'^2 = \left(1 - \frac{n}{N}\right)\{pqn - n_1(p_1 - p)^2 - n_2(p_2 - p)^2 \ldots - n_c(p_c - p)^2\}.$$

Note p. 40. Formula (48)

In the case of small numbers, we can find the chances of obtaining $w = pn$ from a universe in which the proportion is P from the table on p. 36.

Thus if we find $w = 4$, or $w = 10$, we have

Value of Pn	c_1 = chance of obtaining 4	c_2 = chance of obtaining 10
1	·015	—
2	·090	—
3	·168	·000
4	·195	·005
5	·175	·018
6	·134	·041
7	·091	·071
8	·057	·099
9	·034	·119
10	·019	·125
11	·010	·119
12	005	·105
13	·003	·086
14	·001	·066
15	·001	·049
16	·000	·034
17	—	·023
18	—	·015
19	—	·009
20	—	·006
	$c_1 = \frac{e^{-Pn}(Pn)^4}{4!}$	$c_2 = \frac{e^{-Pn}(Pn)^{10}}{10!}$.

Twice the standard deviation when $w = 10$ is $2\sqrt{10} = 6{\cdot}32$. The chance of finding 10 when the proportion in the universe differs from 10 by as much as twice the standard deviation is very small. But if w is very small, say = 4, there is a perceptible chance that Pn exceeded w by more than twice the standard deviation, viz. $2\sqrt{w} = 4$.

To illustrate the numerical values of these formulae, we will consider the case where $N = 10{,}000$, $n = 1000$, $pn = 100$, $p = {\cdot}1$.

Then $\sigma'^2 = \cdot1 \times \cdot9 \times 1000\,(1 - \frac{1}{10})$ and $\sigma' = 9$; $s'^2 = \cdot1 \times \cdot9 \times 1000$, and $s' = 3\sqrt{10} = 9{\cdot}487$.

Chance that 100 *instances should be found*

Proportion in Universe P	Pn	x	$\frac{x}{\sigma'}$	$\frac{x}{s'}$	E_x = chance that $p = \cdot1$ Formula (7)	Formula (8)	Formula (9).
·060	60	40	4·444	4·216	·0000	·0000	·0000
·070	70	30	3·333	3·162	·0000	·0017	·0002
·080	80	20	2·222	2·108	·0026	·0038	·0034
·085	85	15	1·667	1·581	·0096	·0111	·0107
·090	90	10	1·111	1·054	·0229	·0238	·0233
·095	95	5	·556	·527	·0378	·0380	·0364
·100	100	0	0	0	·0443	·0443	·0421
·105	105	− 5	− ·556	− ·527	·0382	·0389	·0368
·110	110	−10	−1·111	−1·054	·0247	·0238	·0249
·115	115	−15	−1·667	−1·581	·0125	·0111	·0134
·120	120	−20	−2·222	−2·108	·0049	·0038	·0058
·130	130	−30	−3·333	−3·162	·0004	·0002	·0005
·140	140	−40	−4·444	−4·216	·0000	·0000	·0000

It is apparent from this table that the chance falls rapidly as x increases; but the full importance of the fall is lost, because the chance at any individual value of x is small.

A better view is obtained if we suppose that the *à priori* chance that P should have certain values is constant over small ranges and add (by integration) the chances over these ranges. We have the following table.

Chance that 100 *cases would be found*

Proportion in Universe P	Formula (7)	Formula (8)	Formula (9)	Formula (10)
·060 to ·070	·0000	·0004	·0000	·0008
·070 „ ·080	·0067	·0127	·0098	·0167
·080 „ ·085	·0278	·0346	·0328	·0394
·085 „ ·090	·0793	·0855	·0834	·0890
·090 „ ·095	·1532	·1558	·1510	·1532
·095 „ ·100	·2106	·2109	·2007	·2009
·100 „ ·105	·2111	·2109	·2010	·2009
·105 „ ·110	·1584	·1558	·1554	·1532
·110 „ ·115	·0918	·0855	·0946	·0890
·115 „ ·120	·0416	·0346	·0460	·0394
·120 „ ·130	·0187	·0127	·0237	·0167
·130 „ ·140	·0016	·0004	·0016	·0008

Here (7) is the most complete; in (8) $\frac{1}{\sqrt{n}}$ is neglected, in (9) $\frac{n}{N}$ is neglected, and in (10) $\frac{1}{\sqrt{n}}$ and $\frac{n}{N}$ are neglected. It is noticeable that the chances from universes in which P exceeds ·1 are somewhat greater than

where P is slightly less than ·1. We are in general more likely to underestimate than to overestimate P.

For many practical purposes the four formulae are equivalent.

It is now clear that if choice was made from a universe in which P was outside the limits ·081 and ·119, the chance that $p = ·1$ would be found in the sample would be very small. These limits differ from ·1 by approximately twice the standard deviation, viz.

$$2\sqrt{\left\{pq\left(\frac{1}{n}-\frac{1}{N}\right)\right\}}.$$

We have still to consider whether any more definite inference can be made from the sample to the universe. This necessitates some assumption about the *à priori* chance that in the universe from which selection was made the proportion should be P. We are *not* justified in assuming, as in the first form of Bayes' theorem, that P is equally likely to be anywhere on the scale 0 to 1.

Write $F(P)$ for the chance that the proportion in the universe *à priori* was P. Then the double chance that P was the proportion in the universe and that then p should be found in the sample is $F(P) \times E_x$, and it follows that the inverse chance that p being found P was the proportion in the universe is

$$F(P) \times E_x \div \Sigma\{F(P) \times E_x\},$$

the summation being extended over all possible values of P.

Note p. 42

If we assume that $F(P)$ is of definite (though unknown) form, is continuous and integrable, and that its change in the neighbourhood of $P = p$ is finite, then it can be shown that the chance that P does not differ from p on either side by more than $\frac{x}{n}$ is independent of F, and does not involve the unsymmetrical term. On these assumptions the chance that P is within the limits $p \pm z$, where $z = \frac{x}{n}$, is

$$\int_{-z}^{z} \frac{1}{\sqrt{\left\{2\pi pq\left(\frac{1}{n}-\frac{1}{N}\right)\right\}}} \cdot e^{-\frac{z^2}{2pq\left(\frac{1}{n}-\frac{1}{N}\right)}} dz. \quad \ldots\ldots\ldots\ldots(12)$$

This result depends essentially on the rapid fall in the value of E_x, illustrated in the table above, as $P - p$ increases. To compensate this rapid fall it would be necessary that extreme values of P should have great *à priori* probability, if they were to be regarded as competitors for inclusion.

The same formula applies, with the modified form of standard deviation, $s_d' \div n$, in the case of a stratified sample, but only if $\frac{1}{\sqrt{n}}$ is negligible.

It is to be emphasized that the inference thus formulated is based on

assumptions that are difficult to verify and which are not applicable in all cases*.

The method cannot be used in Case II where $pn = w$ is small. There dependence must be placed on such a table as on p. 36.

II. DISTRIBUTION OF ALTERNATIVE ATTRIBUTES BY SAMPLE

A. Direct Problem

It is often the case that as a result of an investigation by sample the persons or things are divided into a number of classes, e.g. as married, widowed, divorced or single. As regards any one of these classes taken by itself the former analysis applies, but we may want to regard the distribution as a whole, and then no such simple procedure is possible. It is no longer a question of the chance of the difference between a single p and the corresponding P, and there are many ways possible in which a combination of differences can be expressed. The method that lends itself to mathematical analysis is to take the function

$$\chi^2 = \frac{Nn}{N-n}\left\{\frac{(P_1-p_1)^2}{p_1} + \frac{(P_2-p_2)^2}{p_2} + \ldots + \frac{(P_c-p_c)^2}{p_c}\right\}, \ldots (13)$$

where n things are selected at random from a universe containing N things, in which $P_1, P_2 \ldots P_c$ are the proportions in c certain defined classes, which combine to make up the universe, and $p_1, p_2 \ldots p_c$ are the proportions found in these classes in the sample†.

Note p. 41. Formula (50)

This function is a measure of the complex of differences, and it can be shown that the chance of obtaining an assigned χ^2 is the same, however the errors are distributed within the function. Note p. 37

As a rough generalization it may be said that the chance is about $\frac{1}{2}$ that χ^2 will not exceed $\overline{c-2}$, where c is the number of separate classes, and more than 20 to 1 against χ^2 exceeding $2c$.

B. Inverse Problem

The proposition as it stands relates to deviations from a known universe. It can be inverted on assumptions similar to those on p. 15, and applied with discretion to the chance that an unknown universe will not differ from an observed sample by errors which make χ^2 exceed given values.

To illustrate the use of this formula we will consider two tables in

* Note that the integration on p. 14 only assumed stationariness of the chances of P over small ranges, and was given for illustration rather than to establish a definite argument.

† A more elaborate form of χ^2, in which $\frac{1}{\sqrt{n}}$ is not neglected, is given below, p. 38.

the report of the enquiry by sample made by the Ministry of Labour of the United Kingdom to ascertain the ages and other circumstances of persons claiming unemployment payments*. In November 1923 a sample consisting of 1 in 100 was examined, and we may compare it with a sample of 1 in 3 taken in the previous January, regarding the earlier sample as giving a practically true account.

Age distribution of Males claiming Unemployment payments

Age	Sample 1 in 100 p	Sample 1 in 3 P	$p-P$	$(p-P)^2 \div p$
16–	·019	·020	−·001	·00005
18–	·047	·064	−·017	·00615
20–	·165	·178	−·013	·00102
25–	·120	·130	−·010	·00083
30–	·103	·103	·000	·00000
35–	·171	·176	−·005	·00015
45–	·178	·167	+·011	·00068
55–	·082	·069	+·013	·00206
60–	·115	·093	+·022	·00421
Total	1·000	1·000	—	·01515

$N = 900{,}000$ approx., the total number of males insured in 1923.
$n = 8137$, the number in the smaller sample.

$$\chi^2 = \frac{Nn}{N-n} \Sigma \frac{(p-P^2)}{p} = 124 \text{ approx.}$$

Now the number of classes, c, is 9, and χ^2 is much greater than $2c$. Either the samples were not properly collected, or they relate to different populations, or there is some mis-statement in the table. In fact, ages in the higher age-groups in the earlier sample were understated by about a year (*loc. cit.* p. 570); but this does not account for the abnormal difference in the group 18– years.

A second table is nearly independent of age statements, and is more satisfactory, for χ^2 is less than c.

Cases in which benefit to men was authorized in respect of dependent children†

	Number of children							
	1	2	3	4	5	6	7 or 8	Totals
Sample 1 in 100, p	·359	·288	·171	·097	·055	·023	·007	1·000
Sample 1 in 3, P	·376	·273	·168	·101	·054	·021	·007	1·000
$p-P$	−·017	+·015	+·003	−·004	+·001	+·002	0	0
$(p-P)^2 \div p$	·00081	·00078	·00005	·00016	·00002	·00017	0	·00199

Here $N = 300{,}000$ approx. $n = 2526$. $c = 7$.

$$\chi^2 = \frac{Nn}{N-n} \times ·00199 = 5·1 \text{ approx.}$$

* *Statistical Journal*, July 1924. Tables, pp. 555 and 559.

† *Loc. cit.* p. 559.

III. SAMPLING FOR THE DETERMINATION OF THE MAGNITUDE OF AN AVERAGE

A. Direct Problem

Let a universe contain N measurable objects, whose magnitudes are $X_1, X_2 \ldots X_N$.

Write
$$N\bar{u} = X_1 + X_2 + \ldots + X_N,$$
$$N\mu_2 = (X_1 - \bar{u})^2 + (X_2 - \bar{u})^2 + \ldots + (X_N - \bar{u})^2,$$
$$N\mu_3 = (X_1 - \bar{u})^3 + (X_2 - \bar{u})^3 + \ldots + (X_N - \bar{u})^3,$$
$$\sigma^2 = \mu_2,$$
$$\kappa = \mu_3 \div \sigma^3.$$

Then $\bar{u}$ is the average, σ the standard deviation, μ_2 the second and μ_3 the third moment about the average, of the magnitudes in the universe.

Case I. Random sample.

Select n things at random all together from the universe, and let the average of their magnitudes be $\bar{u} + x$. Write k for $\frac{n}{N}$.

Then the standard deviation of x is given by

$$\sigma_a = \frac{\sigma}{\sqrt{n}} \cdot \sqrt{(1-k)},$$

which becomes $\frac{\sigma}{\sqrt{n}}$ when N is indefinitely large.

The chance that $\bar{u} + x$ will be found is

$$E_x = \frac{1}{\sigma_a \sqrt{(2\pi)}} \cdot e^{-\frac{x^2}{2\sigma_a^2}} \cdot \left\{1 - \frac{\kappa_1}{2}\left(\frac{x}{\sigma_a} - \frac{x^3}{3\sigma_a^3}\right)\right\}, \qquad \ldots\ldots(14)$$

where
$$\kappa_1 = \frac{1}{\sqrt{n}} \cdot \frac{1-2k}{\sqrt{(1-k)}} \cdot \kappa,$$

Note p. 29. Formula (29) if $\frac{1}{n}$ is negligible—subject to the condition that the great part of the magnitudes in the universe is contained within the range $\bar{u} \pm 3\sigma$, or more exactly that the ratio $\mu_r \div \sigma^r$ is finite for all values of r.

If n is so large that $\frac{1}{\sqrt{n}}$ is negligible, the term containing κ_1 is negligible.

If also k is so small, N so large, that $\frac{1}{2}k$ is negligible, σ_a becomes $\frac{\sigma}{\sqrt{n}} = s_a$ and we have

$$E_x = \frac{\sqrt{n}}{\sigma\sqrt{(2\pi)}} \cdot e^{-\frac{x^2 n}{2\sigma^2}}. \qquad \ldots\ldots(15)$$

From formula (14) the chance that the average in the sample differs from the average in the universe by not more than x is

$$\int_{-x}^{x} \frac{1}{\sigma_a\sqrt{(2\pi)}} \cdot e^{-\frac{x^2}{2\sigma_a^2}} \cdot dx, \qquad \ldots\ldots(16)$$

the term containing κ_1 vanishing.

Case II. Stratified sample.

If the population is divided among c districts, containing respectively $N_1, N_2 \ldots N_c$ persons, the averages in the districts differing one from another, and if the same proportion (k) is taken at random from each district, the standard deviation of the average of the sample is diminished, and the constants of its frequency curve modified.

Write ${}_as_d$ for the standard deviation in this case. Then

$$ {}_as_d^2 = \frac{1}{n}(1-k)(\sigma^2 - \sigma_v^2), $$

where $n\sigma_v^2 = n_1(\bar{x}_1 - \bar{u})^2 + n_2(\bar{x}_2 - \bar{u})^2 + \ldots + n_c(\bar{x}_c - \bar{u})^2$, $\bar{x}_1, \bar{x}_2 \ldots$ being the averages in the districts.

Formula (14) becomes

$$ E_x = \frac{1}{{}_as_d\sqrt{(2\pi)}} \cdot e^{-\frac{x^2}{2{}_as_d^2}} \cdot \left\{1 - \frac{\kappa_1'}{2}\left(\frac{x}{{}_as_d} - \frac{x^3}{3{}_as_d^3}\right)\right\}, \quad \ldots\ldots(17) $$

Note p. 30. Formulae (31) to (34) where κ_1' has a complicated relation to κ_1 shown below (see pp. 29–30).

X	Districts 1	2	3	4	Whole population	
	Population					
(*a*) 1	100	—	—	—	100	Scheme (*a*). Take $n=160$, $k=\frac{1}{10}$.
2	100	—	—	—	100	$\sigma=2{\cdot}5$. $\bar{u}=5{\cdot}5$. $\bar{x}_1=2{\cdot}5$, $\bar{x}_2=4{\cdot}5$,
3	100	100	—	—	200	$\bar{x}_3=6{\cdot}5$, $\bar{x}_4=8{\cdot}5$.
4	100	100	—	—	200	$16\sigma_v^2=80$. $\sigma_v^2=5$.
5	—	100	100	—	200	${}_as_d^2=\frac{1}{160}(1-{\cdot}1)\{(2{\cdot}5)^2-5\}$.
6	—	100	100	—	200	${}_as_d={\cdot}084$.
7	—	—	100	100	200	s_a (in an unstratified selection) $={\cdot}188$.
8	—	—	100	100	200	${}_as_d=s_a\times{\cdot}45$.
9	—	—	—	100	100	
10	—	—	—	100	100	
Total	400	400	400	400	1600	
(*b*) 1	100	—	—	—	100	Scheme (*b*). Take $n=240$, $k=\frac{1}{10}$.
2	100	100	—	—	200	$\sigma^2=\frac{25}{6}$. $\sigma_v^2=\frac{5}{4}$.
3	100	100	100	—	300	${}_as_d={\cdot}105$. $s_a={\cdot}125$.
4	100	100	100	100	400	${}_as_d=s_a\times{\cdot}84$.
5	100	100	100	100	400	
6	100	100	100	100	400	
7	—	100	100	100	300	
8	—	—	100	100	200	
9	—	—	—	100	100	
Total	600	600	600	600	2400	
(*c*) 1	100	—	—	—	100	Scheme (*c*). Take $n=280$, $k=\frac{1}{10}$.
2	100	100	—	—	200	$\sigma^2=\frac{21}{4}$. $\sigma_v^2=\frac{5}{4}$.
3	100	100	100	—	300	${}_as_d={\cdot}113$. $s_a={\cdot}130$.
4	100	100	100	100	400	${}_as_d=s_a\times{\cdot}87$.
5	100	100	100	100	400	
6	100	100	100	100	400	
7	100	100	100	100	400	
8	—	100	100	100	300	
9	—	—	100	100	200	
10	—	—	—	100	100	
Total	700	700	700	700	2800	

If the averages of the districts differ considerably from the general average, or if the standard deviations in the districts are considerably smaller than in the population as a whole*, the gain in accuracy by stratification may be considerable.

For example take the simple cases shown in tabular form on p. 19.

B. Inverse Problem

In a sample consisting of n magnitudes, drawn at random from a universe containing N magnitudes, the average is found to be $\bar{u}_1$ and the second moment about it μ_2'.

The chance that such an average would be obtained from a universe in which the average was $\bar{u}_1 - x$ is

$$E_x = \frac{1}{\sigma' \sqrt{(2\pi)}} \cdot e^{-\frac{x^2}{2\sigma'^2}}, \quad \text{.....(18)}$$

Note p. 41. Formula (51)

if terms involving $\frac{1}{\sqrt{n}}$ are negligible, where

$\sigma'^2 = \frac{\mu_2'}{n}\left(1 - \frac{n}{N}\right)$ in a restricted but unstratified sample,

$\sigma'^2 = \frac{\mu_2'}{n}$ in an unrestricted and unstratified sample,

and $\sigma'^2 = \frac{1}{n}\left(1 - \frac{n}{N}\right)(\mu_2' - \sigma_v'^2)$ in a restricted and stratified sample.

Here $n\sigma_v'^2 = \Sigma n_t (\bar{x}_t' - \bar{u}_1)^2$, $\bar{x}_t'$ being the observed average of n_t things in the tth district.

The formula cannot be extended to include the term involving $\frac{1}{\sqrt{n}}$, unless the standard deviation in the universe is known.

If in a manner similar to that on p. 15 above we assume that the chance of various values of the average in the universe is a definite function that varies continuously, then the chance that the average in the universe is within the limits $\bar{u}_1 \pm x$ is

$$\int_{-x}^{x} \frac{1}{\sigma' \sqrt{(2\pi)}} \cdot e^{-\frac{x^2}{2\sigma'^2}} \cdot dx, \quad \text{.....(19)}$$

Note p. 45. Formula (53)

where σ' has whichever is appropriate of the above written values.

* Since $N\mu_2 = N_1\{{}_1m_2 + (\bar{x}_1 - \bar{u})^2\} + N_2\{{}_2m_2 + (\bar{x}_2 - \bar{u})^2\} + \ldots,$

where ${}_1m_2$, ${}_2m_2$ are the second moments in the districts about their averages, therefore

$$\mu_2 = \frac{N_1}{N} \cdot {}_1m_2 + \frac{N_2}{N} \cdot {}_2m_2 + \ldots + \sigma_v^2.$$

Hence for a given μ_2, if σ_v^2 is large, ${}_1m_2$, ${}_2m_2$ ... must be small, and the apparent alternatives are nearly identical.

Formula (19) is strictly correct when $\frac{1}{\sqrt{n}}$ is negligible, and a rule is given in the notes, p. 45, for its application when $\frac{1}{\sqrt{n}}$ is retained and $\frac{1}{n}$ neglected.

Example.

An investigation was made in Northampton in 1913* in which details were obtained relating to 693 working-class households. k, the ratio of the number examined to the whole number, was $1 \div 22{\cdot}7$.

Number of persons in house	Number of houses
1	16
2	113
3	146
4	127
5	115
6	73
7	50
8	31
9	14
10	6
11	0
12	2
	693

Average $4{\cdot}342 = \bar{u}_1$

$\mu_2' = 3{\cdot}91$

$\sigma_1^2 = \frac{3{\cdot}91}{693}\left(1 - \frac{1}{22{\cdot}7}\right)$

$\sigma_1 = {\cdot}073$

Chance that in the aggregate of working-class houses the average was (from (19))

Outside $\bar{u}_1 \pm 3\sigma_1$,	i.e. above	4·562 or below 4·122	·0027
Between $\bar{u}+2\sigma_1$ and $\bar{u}+3\sigma_1$, or $\bar{u}-2\sigma_1$ „ $\bar{u}-3\sigma_1$,	i.e. between	4·489 and 4·562 / 4·195 „ 4·122	·0429
Between $\bar{u}+\sigma_1$ „ $\bar{u}+2\sigma_1$, $\bar{u}-\sigma_1$ „ $\bar{u}-2\sigma_1$,	i.e. „	4·415 „ 4·489 / 4·269 „ 4·195	·2718
Between $\bar{u}+\sigma_1$ „ $\bar{u}-\sigma_1$,	i.e. „	4·415 „ 4·269	·6826

These results apply to the universe "working-class houses as defined for the investigation." In the different universe defined for the whole Borough in the Census of 1911, viz. the whole population less those in large institutions, divided by the number of "families or separate occupiers," the average was 4·44.

* *Livelihood and Poverty.* London: G. Bell and Sons, 1915.

MATHEMATICAL NOTES

CONTENTS

INTRODUCTION

So far as the direct problem is concerned, the expressions for E_x in the case of purely random sampling from an infinite universe, including the unsymmetrical term that involves $1 \div \sqrt{n}$, have been known since the time of Laplace, Gauss, Bernoulli and Poisson. Also the modifications of the standard deviation when the universe is restricted (or "the balls not replaced in the urn") or when the sample is stratified were determined long ago. But, so far as I can ascertain, no one has brought together these formulae so as to give the frequency correct to the second (or $1 \div \sqrt{n}$) term, when the universe is restricted, or when the sample is stratified, or when both these conditions apply, either for variables or for attributes. To obtain these frequency curves, it is necessary to go back to first principles, and in the following pages the lines of well-known analysis are developed.

The briefest method is to combine in one analysis the cases of attributes and of variables, and since the former can be regarded as special cases of the latter, variables are taken first.

Most of the work will be recognized as a simple extension of generally accepted principles; but the problem before us is definitely to make inferences from a given sample to an unknown universe, whereas the great bulk of recent work has proceeded from an assumed universe to a sample, and we are therefore obliged to go on to the doubtful ground of inverse probability. My treatment, in which the intention has been to follow Professor Edgeworth's methods, is explained in a short article in *Metron*, Vol. II, No. 3.

NOTATION

Variables.

n frequency groups. In the tth group the average is $\bar{x}_t$, standard deviation σ_t, moments about the average ${}_t\mu_2$, ${}_t\mu_3$ ....

${}_t\kappa_{r-2} = {}_t\mu_r \div \sigma_t{}^r$.

When the groups are identical the prefix t is dropped.

For the *sum* of n unrestricted selections: average U, deviations u_s or u, standard deviation S, moments about average M_1, M_2, M_3 ...; $KS^3 = M_3$.

Standard deviation of *average* of n unrestricted selections, s_a.

Restricted selection. n_1 are selected from a group whose constants are N_1, $\bar{x}_1$, m_2, m_3 ..., and the values of n_1 form one of the frequency groups defined above.

$n_1 = kN_1$.

σ, μ_2, κ_2 etc. are used for the restricted as well as for the unrestricted elemental group; as are S, M_1 ....

σ_a is the standard deviation of the average of an n_1-fold restricted selection.

Stratified selection. The constants for the tth stratum are $\bar{x}_t$, ${}_tm_2$, ${}_tm_3$....

For all the strata merged $\bar{x}$ is average and s_1 the standard deviation.

$n\sigma_v{}^2 = \Sigma n_t\,(\bar{x}_t - \bar{x})^2$.

${}_as_d$ is the standard deviation of the average of the restricted, stratified selection.

In the universe from which the actual sample is selected the constants are written N, $\bar{u}$, σ, μ_2, κ etc.

The values of these found from the sample are $\bar{u}_1$, σ', μ_2' etc.

Thus the quantities without prefixes are used rather generally, and are defined each time in the text.

Attributes.

Numbers:—N in universe, n in sample. $M = N - n$. $k = \dfrac{n}{N}$.

Proportions:—P in universe, p in sample. $Q = 1 - P$. $q = 1 - p$.

$x = (p - P)\,n$. $z = p - P$.

Districts:—$(P_1, N_1) \ldots (P_t, N_t) \ldots (P_d, N_d)$ in universe.
$(p_1, n_1) \ldots (p_t, n_t) \ldots (p_d, n_d)$ in sample.
$N_1 + \ldots + N_t + \ldots + N_d = N.$

$$\frac{n_1}{N_1} = \ldots = \frac{n_t}{N_t} = \ldots = \frac{n_d}{N_d} = k.$$

$\sigma_0^2 = PQ$, $s^2 = PQn$, $\sigma^2 = PQn\,(1-k)$, $s'^2 = pqn$, $\sigma'^2 = pqn\,(1-k)$, $\sigma_p = \frac{\sigma}{n}$.

$n\sigma_v^2 = \Sigma n_t\,(P_t - P)^2$, $n\sigma_v'^2 = \Sigma n_t\,(p_t - p)^2$.

$\sigma_d^2 = n\,(1-k)\,(PQ - \sigma_v^2)$, $\sigma_d'^2 = n\,(1-k)\,(pq - \sigma_v'^2)$.

$m_1 = p_1 n_1 \ldots m_t = p_t n_t \ldots m_c = p_c n_c$.

$$\chi^2 = \frac{Nn}{M} \cdot \Sigma \frac{(P_t - p_t)^2}{p_t}.$$

Alternative attributes—see p. 37.

I. GENERAL. *Frequency of the sum of a number of independent variables*

Let there be n frequency groups of measurable quantities, and let $\bar{x}_t$, σ_t, ${}_t\mu_2$, ${}_t\mu_3$ ... be the average, standard deviation and second, third ... moments about the average in any, the tth, group.

Select one quantity from each group, e.g. $\bar{x}_t + {}_tu_s$ from the tth, and take their weighted sum, thus

$$U + u_s = a_1\,(\bar{x}_1 + {}_1u_s) + a_2\,(\bar{x}_2 + {}_2u_s) + \ldots + a_n\,(\bar{x}_n + {}_nu_s),$$

where $a_1, a_2 \ldots a_n$ are constants, and

$$U = a_1\bar{x}_1 + a_2\bar{x}_2 + \ldots + a_n\bar{x}_n,$$
$$u_s = a_1 \cdot {}_1u_s + a_2 \cdot {}_2u_s + \ldots + a_n \cdot {}_nu_s.$$

Assume that there is no correlation between any ${}_tu_s$ and ${}_{t'}u_s$, or between any ${}_tu_s$ and ${}_tu_{s'}$.

Required the frequency curve of u_s.

Take any small constant α, to facilitate collection of terms.

Then $\qquad e^{\alpha u_s} = e^{\alpha a_1 \cdot {}_1u_s} \times e^{\alpha a_2 \cdot {}_2u_s} \times \ldots$ to n factors.

Therefore

$$1 + \alpha u_s + \frac{\alpha^2}{2}\,u_s^2 + \frac{\alpha^3}{3!}\,u_s^3 + \ldots = \prod_{t=1}^{t=n}\left(1 + \alpha a_t \cdot {}_tu_s + \frac{\alpha^2}{2} \cdot a_t^2 \cdot {}_tu_s^2 + \ldots\right).$$

Since the factors on the right-hand side are independent of each other, the mean of their product equals the product of their means.

Write $M_1, M_2 \ldots$ for the moments of u_s, and write $M_2 = S^2$.

Take the means of both sides of the equation, observing that mean ${}_tu_s$ is zero, since the ${}_tu$'s are measured from their average.

Then

$$1 + \alpha M_1 + \frac{\alpha^2}{2}\,M_2 + \ldots = \Pi\left(1 + \frac{\alpha^2}{2} \cdot a_t^2 \cdot {}_t\mu_2 + \frac{\alpha^3}{3!} \cdot a_t^3 \cdot {}_t\mu_3 + \ldots\right)$$
$$= e^{\Sigma \log\left(1 + \frac{\alpha^2}{2} \cdot a_t^2 \cdot {}_t\mu_2 + \ldots\right)}$$
$$= e^{\frac{\alpha^2}{2}\Sigma a_t^2 \cdot {}_t\mu_2 + \frac{\alpha^3}{3!}\Sigma a_t^3 \cdot {}_t\mu_3 + \ldots - \frac{1}{2}\Sigma\left(\frac{\alpha^2}{2} \cdot a_t^2 \cdot {}_t\mu_2 + \ldots\right)^2 + \ldots}.$$

Expand, and equate the coefficients of α, α^2 and α^3.

$$M_1 = 0,$$
$$S^2 = M_2 = \Sigma {a_t}^2 \,.\, {}_t\mu_2 = \Sigma {a_t}^2 \,.\, {\sigma_t}^2,$$
$$M_3 = \Sigma {a_t}^3 \,.\, {}_t\mu_3.$$

Use these values, and collect the coefficients of α^4, α^5 ... in the index.

$$1 + \frac{\alpha^2}{2} M_2 + \frac{\alpha^3}{3!} M_3 + \ldots = e^{\frac{\alpha^2}{2} \cdot S^2} \,.\, e^{\frac{\alpha^3 S^3}{3!} \cdot \frac{M_3}{S^3}} \,.\, e^{\frac{\alpha^4 S^4}{4!} \cdot C_4} \,.\, e^{\frac{\alpha^5 S^5}{5!} \cdot C_5} \ldots,$$

where the first three C's are as follows:

$$C_4 = \frac{\Sigma {a_t}^4 \,({}_t\mu_4 - 3{\sigma_t}^4)}{(\Sigma {a_t}^2 \,.\, {\sigma_t}^2)^2} = \frac{\Sigma {a_t}^4 \,({}_t\kappa_2 - 3)\,{\sigma_t}^4}{(\Sigma {a_t}^2 \,.\, {\sigma_t}^2)^2}, \text{ where } {}_t\kappa_2 \text{ is written for } \frac{{}_t\mu_4}{{\sigma_t}^4},$$

$$C_5 = \frac{\Sigma {a_t}^5 \,({}_t\kappa_3 - 10\,{}_t\kappa_1)\,{\sigma_t}^5}{(\Sigma {a_t}^2 \,.\, {\sigma_t}^2)^{\frac{5}{2}}}, \text{ where } {}_t\kappa_1,\ {}_t\kappa_3 \text{ are written for } \frac{{}_t\mu_3}{{\sigma_t}^3},\ \frac{{}_t\mu_5}{{\sigma_t}^5},$$

$$C_6 = \frac{\Sigma {a_t}^6 \,\{{}_t\kappa_4 - 15 - 15\,({}_t\kappa_2 - 3) - 10\,{}_t{\kappa_1}^2\}\,{\sigma_t}^6}{(\Sigma {a_t}^2 \,.\, {\sigma_t}^2)^3}, \text{ where } {}_t\kappa_4 \text{ is written for } \frac{{}_t\mu_6}{{\sigma_t}^6}.$$

In the special case where $a_1\sigma_1 = a_2\sigma_2 = \ldots = a_n\sigma_n$, each of these products would $= \frac{S}{\sqrt{n}}$, and we should have

$$K = \frac{M_3}{S^3} = \frac{1}{n^{\frac{1}{2}}} \times \text{Mean } {}_t\kappa_1,$$

$$C_4 = \frac{1}{n} \times \text{Mean } ({}_t\kappa_2 - 3),$$

$$C_5 = \frac{1}{n^{\frac{3}{2}}} \times \text{Mean } ({}_t\kappa_3 - 10\,{}_t\kappa_1),$$

$$C_6 = \frac{1}{n^2} \times \text{Mean } \{{}_t\kappa_4 - 15 - 15\,({}_t\kappa_2 - 3) - 10\,{}_t{\kappa_1}^2\}.$$

Now regard n as large and S as finite.

Restrict the original frequency groups, so that $a_1\sigma_1$, $a_2\sigma_2$ ... $a_n\sigma_n$ are of the same order of magnitude, viz. $n^{-\frac{1}{2}}$, and all such quantities as ${}_t\kappa_1$, ${}_t\kappa_2$, ${}_t\kappa_3$ ... are finite.

Then the coefficients of α^2, α^3 ... in the index written above are of the orders 1, $n^{-\frac{1}{2}}$, n^{-1}, $n^{-\frac{3}{2}}$, n^{-2} ....

Neglect terms of the order n^{-1}, $n^{-\frac{3}{2}}$..., and notice that K^2 is of order n^{-1}.

Then

$$1 + \frac{\alpha^2}{2} M_2 + \frac{\alpha^3}{3!} M_3 + \ldots + \frac{\alpha^{2l}}{(2l)!} M_{2l} + \frac{\alpha^{2l+1}}{(2l+1)!} M_{2l+1} + \ldots = e^{\frac{\alpha^2}{2} S^2} \,.\, e^{\frac{\alpha^3}{6} S^3 K}$$
$$= \left\{1 + \frac{\alpha^2}{2} S^2 + \frac{1}{2}\left(\frac{\alpha^2}{2} S^2\right)^2 + \ldots + \frac{1}{l!}\left(\frac{\alpha^2}{2} S^2\right)^l + \ldots\right\}\left(1 + \frac{\alpha^3}{6} S^3 K\right).$$

Equate coefficients. Then

$$M_{2l} = \frac{(2l)!}{2^l \,.\, l!} \,.\, S^{2l}, \quad M_{2l+1} = \frac{K\,(2l+1)!}{3 \cdot 2^l\,(l-1)!} \,.\, S^{2l+1}.$$

But integrating by parts, with the help of the fundamental integral

$$\sqrt{\pi} = \int_{-\infty}^{\infty} e^{-x^2}\, dx,$$

as in the calculation of the moments of the normal curve of error, we have

$$\int_{-\infty}^{\infty} \frac{1}{S\sqrt{(2\pi)}} \cdot e^{-\frac{u^2}{2S^2}} \cdot \left\{1 - \tfrac{1}{2}K\left(\frac{u}{S} - \frac{u^3}{3S^3}\right)\right\} \cdot u^{2l} \cdot du = \frac{(2l)!}{2^l \cdot l!} \cdot S^{2l},$$

and

$$\int_{-\infty}^{\infty} \frac{1}{S\sqrt{(2\pi)}} \cdot e^{-\frac{u^2}{2S^2}} \cdot \left\{1 - \tfrac{1}{2}K\left(\frac{u}{S} - \frac{u^3}{3S^3}\right)\right\} \cdot u^{2l+1} \cdot du = \frac{K\,(2l+1)!}{3 \cdot 2^l (l-1)!} \cdot S^{2l+1}.$$

Hence the moments of u are the moments of the curve

$$y = \frac{1}{S\sqrt{(2\pi)}} \cdot e^{-\frac{u^2}{2S^2}} \cdot \left\{1 - \tfrac{1}{2}K\left(\frac{u}{S} - \frac{u^3}{3S^3}\right)\right\}, \quad \text{.....(20)}$$

and the frequency curve of u can be identified with this curve.

Here $\quad S^2 = \sum_{t=1}^{t=n} a_t^2 \cdot \sigma_t^2$, and $KS^3 = \sum_{t=1}^{t=n} a_t^3 \cdot {}_t\mu_3$.

The conditions are

Independence in selection of items.

That $a_1\sigma_1$, $a_2\sigma_2$ ... are of the same order, viz. that of $Sn^{-\frac{1}{2}}$,

That $\frac{{}_t\mu_r}{\sigma_t{}^r} = {}_t\kappa_{r-2}$ is finite for all values of r and of t.

That $\frac{1}{n}$ is negligible in comparison with unity.

If $\frac{1}{\sqrt{n}}$ is also negligible, the curve becomes

$$y = \frac{1}{S\sqrt{(2\pi)}} \cdot e^{-\frac{u^2}{2S^2}}. \quad \text{.....(21)}$$

If all the frequency groups have the same moments, or if all selections are made with replacement from one group, and if $a_1 = a_2 = \ldots = a_n = 1$, then writing $\sigma = \sigma_1 = \sigma_2 \ldots$, and $\mu_3 = {}_1\mu_3 = {}_2\mu_3 = \ldots$, and $\kappa_1 = {}_1\kappa_1 = {}_2\kappa_1 = \ldots$, we have $S^2 = n\sigma^2$, $\kappa S^3 = n\mu_3$, and $\kappa = \frac{1}{\sqrt{n}} \cdot \kappa_1$.

So far we have considered the sum of n magnitudes. We frequently require the frequency of the unweighted average of n items selected from n groups or from one group with replacement. This we obtain by dividing U and u_s by n throughout, in the simplified case just described.

Write $u = nx$. The standard deviation for x is given by

$$s_a = \frac{1}{n} \cdot S = \frac{\sigma}{\sqrt{n}} \quad \text{.....(22)}$$

and κ is unaltered.

The frequency curve of x is then

$$y = \frac{1}{s_a\sqrt{(2\pi)}} \cdot e^{-\frac{x^2}{2s_a^2}} \cdot \left\{1 - \frac{\kappa}{2}\left(\frac{x}{s_a} - \frac{x^3}{3s_a^3}\right)\right\}. \quad \text{.....(23)}$$

Note. A theoretical difficulty arises from the consideration that in formula (20) or (23) y becomes negative, if (κ being positive) $\frac{u}{s}$ is less than a certain negative quantity. κ is, however, so small that κ^2 is negligible, and if κ is as great as $\cdot 1$, u must be less than $-4s$ before y becomes negative, and in this region y is negligible. It is recognized, however, that the terms neglected in the approximation, involving multiples of $\frac{1}{n}$, have effect as $\frac{u}{s}$ increases and that the extremities of the curve are indefinite. See for example Edgeworth in the *Statistical Journal*, 1906, p. 512.

II. Restricted Universe. *Moments of the sum of n_1 selected quantities*

Let a group of n_1 things be chosen at random, without replacement, from a frequency group consisting of N_1 measurable quantities, whose average is $\bar{x}_1$ and moments about the average $m_2, m_3 \ldots$.

Write $n_1 = kN_1$.

Required the moments $\mu_1, \mu_2 \ldots$ of the sum of the n_1 things, all possible selections being made.

Write the N_1 quantities as $\bar{x}_1 + u_1, \bar{x}_1 + u_2 \ldots \bar{x}_1 + u_{N_1}$.

Then $u_1 + u_2 + \ldots + u_{N_1} = 0$.

There are ${}_{N_1}C_{n_1}$ possible selections of sums such as

$$n_1\bar{x}_1 + u_1 + u_2 + \ldots u_{n_1}.$$

Evidently $\mu_1 = n_1\bar{x}_1$.

${}_{N_1}C_{n_1} \cdot \mu_2 = \Sigma\,(u_1 + u_2 + \ldots + u_{n_1})^2$, the summation being extended over the ${}_{N_1}C_{n_1}$ choices.

Each term such as u_1^2 occurs $\frac{n_1}{N_1} \cdot {}_{N_1}C_{n_1}$ times, and each term such as $2u_1u_2$ occurs $\frac{n_1\,(n_1-1)}{N_1\,(N_1-1)} \cdot {}_{N_1}C_{n_1}$ times.

Therefore

$$\mu_2 = \frac{n_1}{N_1}\Sigma u^2 + \frac{2n_1\,(n_1-1)}{N_1\,(N_1-1)}\Sigma uu, \text{ where } \Sigma uu \text{ is the sum of all possible pairs}$$

$$= \frac{n_1}{N_1}\left(1 - \frac{n_1-1}{N_1-1}\right)\Sigma u^2, \text{ since } \Sigma u^2 + 2\Sigma uu = (\Sigma u)^2 = 0,$$

$$= \frac{n_1\,(N_1-n_1)}{N_1-1}\,m_2, \text{ since } N_1m_2 = \Sigma u^2,$$

$$= n_1\,(1-k)\,m_2, \text{ where } k = \frac{n_1}{N_1}, \text{ if } \frac{1}{N_1} \text{ is neglected.} \quad \text{.....(24)}$$

Similarly $$\mu_3 = \frac{n_1}{N_1}\Sigma u^3 + 3\frac{n_1 C_2}{N_1 C_2}\Sigma u^2 u + 6\frac{n_1 C_3}{N_1 C_3}\Sigma uuu$$
$$= n_1(1-k)(1-2k)\,m_3, \text{ if } \frac{2}{N_1} \text{ is neglected,} \quad \ldots\ldots(25)$$

for $\Sigma u^2 u = -\Sigma u^3 = -3\Sigma uuu$.

Similarly

$$\mu_4 = \frac{n_1(N_1-n_1)}{(N_1-1)(N_1-2)(N_1-3)}\{(N_1^2 - 6n_1N_1 + 6n_1^2 + N_1)\,m_4 + 3(n_1-1)N_1(N_1-n_1-1)\,m_2^2\}$$

and $$\frac{\mu_4}{\mu_2^2} - 3 = \frac{1}{n_1}\left\{\left(\frac{m_4}{m_2^2} - 3\right)\frac{1-6k+6k^2}{1-k} - 6k\right\}, \quad \ldots\ldots\ldots(26)$$

if $\frac{3}{N_1}$ is neglected.

If the analysis is continued further, it is found that

$$\frac{\mu_5}{\mu_2^{\frac{5}{2}}} - 10\frac{\mu_3}{\mu_2^{\frac{3}{2}}} = \frac{1}{n_1^{\frac{3}{2}}}\left\{\left(\frac{m_5}{m_2^{\frac{5}{2}}} - 10\frac{m_3}{m_2^{\frac{3}{2}}}\right)f_1(k) + 10\frac{m_3}{m_2^{\frac{3}{2}}}f_2(k)\right\} \quad \ldots(27)$$

and

$$\frac{\mu_6}{\mu_2^3} - 15 - 15\left(\frac{\mu_4}{\mu_2^2} - 3\right) - 10\frac{\mu_3^2}{\mu_2^3}$$
$$= \frac{1}{n_1^2}\left\{\left(\frac{m_6}{m_2^3} - 15\right)f_3(k) - 15\left(\frac{m_4}{m_2^2} - 3\right)f_4(k) - 10\frac{m_3^2}{m_2^2}f_5(k) - f_6(k)\right\}, \quad \ldots\ldots(28)$$

if $\frac{4}{N_1}$ and $\frac{5}{N_1}$ are neglected, where $f_1(k), f_2(k) \ldots f_6(k)$ are functions of k, finite unless $k = 1$.

As in the general analysis (p. 26), let $\frac{m_r}{m_2^{\frac{r}{2}}}$ be finite in the universe, and neglect terms involving $\frac{1}{n_1}$.

We have from formulae (24) to (28)

$$\sigma^2 = \mu_2 = n_1 m_2 (1-k),$$
$$\kappa_1 = \frac{\mu_3}{\sigma^3} = \frac{1}{n_1^{\frac{1}{2}}}\cdot\frac{1-2k}{(1-k)^{\frac{1}{2}}}\cdot\frac{m_3}{m_2^{\frac{3}{2}}},$$
$$\kappa_2 = \frac{\mu_4}{\mu_2^2} = 3,$$
$$\kappa_3 = \frac{\mu_5}{\sigma^5} = 10\kappa,$$
$$\kappa_4 = \frac{\mu_6}{\sigma^6} = 15 - 10\kappa^2 = 15, \text{ since } \kappa^2 \text{ is of order } \frac{1}{n_1}.$$

Hence, up to the 6th moment at least, the frequency of the sum of n_1 things taken at random without replacement from a universe con-

taining $n_1 \div k$ things (in which the distance of the bulk of the observations from the centre does not exceed a small multiple of the standard deviation) is given by the equation

$$y = \frac{1}{\sigma\sqrt{(2\pi)}} \cdot e^{-\frac{x^2}{2\sigma^2}} \left\{1 - \frac{\kappa_1}{2}\left(\frac{x}{\sigma} - \frac{x^3}{\sigma^3}\right)\right\}, \quad \text{.....(29)}$$

correct to terms involving $\frac{1}{\sqrt{n_1}}$, but neglecting terms involving $\frac{1}{n_1}$.

The frequency curve of the average of n_1 things is obtained by writing

$$\sigma_a = \frac{\sigma}{n_1} = \sqrt{\left\{\frac{m_2\,(1-k)}{n_1}\right\}} \quad \text{.....(30)}$$

for σ in this equation.

III. Stratified Selection. *Variables*

Suppose a universe containing N things to be divided into d strata containing $N_1, N_2 \ldots N_d$ things respectively.

Select at random, but without replacement, $n_1 = kN_1$ things from the first stratum, $n_2 = kN_2$ from the second ... $n_d = kN_d$ from the last.

Write

$$n = n_1 + n_2 + \ldots + n_d = k\,(N_1 + N_2 + \ldots + N_d) = kN.$$

In any, the tth, stratum let the average be written $\bar{x}_t$ and the moments about the average ${}_tm_2$, ${}_tm_3$ ....

Let the results of the selection be to give totals $n_1\bar{x}_1 + {}_1u_s$, $n_2\bar{x}_2 + {}_2u_s$ ... for the strata, and $n\bar{x} + u_s$ for the universe, where $\bar{x}$ is the average for the universe.

Then

$$n\bar{x} = kN\bar{x} = k\,(N_1\bar{x}_1 + N_2\bar{x}_2 + \ldots) = n_1\bar{x}_1 + n_2\bar{x}_2 + \ldots + n_d\bar{x}_d,$$

and

$$n\bar{x} + u_s = (n_1\bar{x}_1 + {}_1u_s) + (n_2\bar{x}_2 + {}_2u_s) + \ldots,$$

so that

$$u_s = {}_1u_s + {}_2u_s + \ldots + {}_tu_s + \ldots + {}_du_s.$$

As shown in the preceding section (formulae (24), (25), (26)) the moments for ${}_tu_s$ are given by

$$\sigma_t{}^2 = {}_t\mu_2 = n_t\,(1-k) \cdot {}_tm_2, \quad {}_t\mu_3 = n_t\,(1-k)\,(1-2k) \cdot {}_tm_3$$

and

$${}_t\mu_4 - 3 \cdot {}_t\mu_2{}^2 = n_t\,(1-k)\,\{({}_tm_4 - 3 \cdot {}_tm_2{}^2)\,(1 - 6k + 6k^2) - {}_tm_2{}^2\,6k\,(1-k)\} \text{ etc.}$$

Apply the methods of the general section (pp. 24–7), the quantities ${}_1u_s$, ${}_2u_s$ ... being independent of each other. Then

$$1 + \alpha M_1 + \frac{\alpha^2}{2} M_2 + \frac{\alpha^3}{3!} M_3 + \ldots = e^{\frac{\alpha^2}{2}\Sigma_t\mu_2 + \frac{\alpha^3}{3!}S^3 \cdot C_3 + \frac{\alpha^4}{4!}S^4 \cdot C_4 + \frac{\alpha^5}{5!}S^5 \cdot C_5 + \ldots},$$

where M_1, M_2 ... are the moments of the frequency group of u_s.

Equate coefficients of α, α^2, α^3.

$$M_1 = 0,$$

$$S^2 = M_2 = \Sigma_t\mu_2 = (1-k)\,\Sigma n_t \cdot {}_tm_2,$$

$$M_3 = \Sigma_t\mu_3 = (1-k)\,(1-2k)\,\Sigma n_t \cdot {}_tm_3.$$

Take S as of unit order.

Let ${}_1m_2, {}_2m_2 \ldots {}_dm_2$ be of the same order of magnitude. Then this order is $\frac{1}{n}$, from the equation just written for S^2.

Let $\frac{{}_tm_r}{({}_tm_2)^r}$ be finite for all values of t and of r.

Then

$$C_3 = (1 - 2k)\ \Sigma\left\{n_t \cdot \frac{{}_tm_3}{({}_tm_2)^{\frac{3}{2}}} \cdot ({}_tm_2)^{\frac{3}{2}} \div (1-k)^{\frac{1}{2}}\ (\Sigma n_t \cdot {}_tm_2)^{\frac{3}{2}}\right\},$$

and is of order $n^{-\frac{1}{2}}$; and

$$C_4 = \Sigma\left[n_t\left\{\left(\frac{{}_tm_4}{({}_tm_2)^2} - 3\right)(1 - 6k + 6k^2) - 6k(1-k)\right\}({}_tm_2)^2 \div (1-k)\,(\Sigma n_t \cdot {}_tm_2)^{\frac{3}{2}}\right],$$

and is of order n^{-1}.

Similarly C_5, C_6 are of orders $n^{-\frac{3}{2}}$, n^{-2} respectively.

Neglect terms of order n^{-1} or lower.

Then, as on pp. 25–6, M_2, M_3 ... are the moments of the curve

$$y = \frac{1}{S\sqrt{(2\pi)}} \cdot e^{-\frac{x^2}{2S^2}}\left\{1 - \frac{K}{2}\left(\frac{x}{S} - \frac{x^3}{3S^3}\right)\right\}, \quad \ldots\ldots\ldots\ldots(31)$$

where $S^2 = (1-k)\ \Sigma n_t \cdot {}_tm_2$, and $KS^3 = (1-k)(1-2k)\ \Sigma n_t \cdot {}_tm_3$.

The values of S and K can be given in another form as follows.

Write $\bar{x}_1 = \bar{x} + v_1 \ldots \bar{x}_t = \bar{x} + v_t$, where $\bar{x}$ is the average in the universe, and $\bar{x}_t$ the average in the tth district.

Then $n\bar{x} = \Sigma n_t \bar{x}_t$, $\Sigma n_t v_t = 0$, and $ns_1{}^2 = \Sigma n_t\,({}_tm_2 + v_t{}^2)$, where s_1 is the standard deviation in the universe.

Write $$n\sigma_v{}^2 = \Sigma n_t v_t{}^2 = n_1(\bar{x}_1 - \bar{x})^2 + n_2(\bar{x}_2 - \bar{x})^2 + \ldots, \quad \ldots\ldots(32)$$

so that σ_v is a measure of the scattering of the averages in the districts about the average in the universe.

Then $$S^2 = (1-k)\ \Sigma n_t \cdot {}_tm_2 = n\,(1-k)\,(s_1{}^2 - \sigma_v{}^2). \quad \ldots\ldots\ldots(33)$$

If no attention had been paid to stratification we should have been dealing with one restricted universe and have obtained from formula (24) after the proper change of symbols $S^2 = n\,(1-k)\,s_1{}^2$.

Stratification therefore improves the precision of the sample.

By a similar procedure we obtain

$$M_3 = n\,(1-k)\,(1-2k)\,(M_3{}' - 3V_{12} - {}_3\mu_v),$$

where $M_3{}'$ is the third moment in the universe, $nV_{12} = \Sigma n_t v_t \cdot {}_tm_2$, and $n \cdot {}_3\mu_v = \Sigma n_t v_t{}^3$.

For the average

$${}_aS_{\bar{d}} = S \div n = \frac{1}{\sqrt{n}}\sqrt{\{(1-k)\,(s_1{}^2 - \sigma_v{}^2)\}}^*. \quad \ldots\ldots\ldots(34)$$

* s_1 is replaced by σ on p. 19.

IV. Application to Attributes

Now consider a universe in which PN persons (or things) possess a certain attribute, and QN do not. $P + Q = 1$.

Select a number of persons from the universe, and make a total by counting each that possesses the attribute as *one*, and each that does not as *zero*.

We may then consider the universe as a frequency group in which QN things are at 0 and PN at 1.

The average is $$\frac{QN \times 0 + PN \times 1}{(P + Q)\,N} = P.$$

With reference to the average, the QN are at $-P$ and the PN at $1 - P = Q$.

The moments about the average are

$$\sigma_0^2 = {}_0\mu_2 = \{QN \times (-P)^2 + PN \times Q^2\} \div (P + Q)\,N = PQ,$$
$${}_0\mu_3 = -P^3Q + PQ^3 = PQ\,(Q^2 - P^2) = PQ\,(Q - P),$$
$${}_0\mu_4 = P^4Q + PQ^4 = PQ\,(1 - 3PQ),$$
$${}_0\mu_5 = -P^5Q + PQ^5 = PQ\,(Q - P)\,(1 - 2PQ),$$
$${}_0\mu_6 = P^6Q + PQ^6 = PQ\,(1 - 5PQ + 5P^2Q^2),$$
$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$

$${}_0\kappa = {}_0\mu_3 \div \sigma_0^3 = (Q - P) \div \sqrt{(PQ)},$$

$${}_0\kappa_2 - 3 = \frac{{}_0\mu_4}{\sigma_0^4} - 3 = \frac{1 - 6PQ}{PQ},$$

$${}_0\kappa_3 - 10\,{}_0\kappa = \frac{{}_0\mu_5}{\sigma_0^5} - \frac{10\,{}_0\mu_3}{\sigma_0^3} = \frac{(Q - P)\,(1 - 12PQ)}{(PQ)^{\frac{3}{2}}},$$

$${}_0\kappa_4 - 15 + 15\,({}_0\kappa_2 - 3) - 10\,{}_0\kappa^2 = (1 - 30PQ + 120P^2Q^2) \div P^2Q^2,$$

where $${}_0\kappa_4 \cdot \sigma_0^6 = {}_0\mu_6,$$

and generally $${}_0\kappa_r = \{(-P)^{r-1} + Q^{r-1}\} \div (PQ)^{\frac{r}{2}-1}.$$

If P is finite and n large we have all the conditions necessary for the general case discussed in Section I (pp. 24–6); for each $a = 1$, each $\sigma_t = PQ$, and each ${}_t\kappa_r$ equals the value of ${}_0\kappa_r$ just written and is finite, if P is finite*.

Write s for S in that notation. Then

$$s^2 = M_2 = \Sigma PQ = nPQ.$$ The σ's are then of the order $sn^{-\frac{1}{2}}$,
$$M_3 = \Sigma PQ\,(Q - P) = nPQ\,(Q - P),$$
$$\kappa = \frac{M_3}{s^3} = \frac{Q - P}{\sqrt{(nPQ)}}.$$

Hence formula (20), which now gives the frequency of x when $Pn + x$ are found in an *unrestricted* selection of n things, becomes

$$y = \frac{1}{\sqrt{(2PQn\pi)}} \cdot e^{-\frac{x^2}{2PQn}} \cdot \left\{1 - \frac{Q - P}{2\sqrt{(PQn)}}\left(\frac{x}{\sqrt{(PQN)}} - \frac{x^3}{3\,(PQN)^{\frac{3}{2}}}\right)\right\}, \quad (35)$$

terms in $\frac{1}{n}$ being neglected.

* See Section V below for the case where P is small.

This formula is usually obtained by considering the limit of the expression $P^{nP+x} \cdot Q^{nQ-x} \cdot {}_nC_{nP+x}$ when n is increased, P being finite.

In a *restricted* universe, where $n = kN$ are taken without replacement from N, the frequency of x is obtained from formula (29). Correct to the 6th moment at least

$$y = \frac{1}{\sigma\sqrt{(2\pi)}} \cdot e^{-\frac{x^2}{2\sigma^2}} \left\{1 - \frac{\kappa_1}{2}\left(\frac{x}{\sigma} - \frac{x^3}{3\sigma^3}\right)\right\}, \quad \ldots\ldots\ldots\ldots(36)$$

when $\frac{1}{n}$, $\frac{1}{N} \ldots \frac{5}{N}$ are neglected, where

$$\sigma^2 = n\,(1-k)\,PQ, \quad \kappa_1\sigma^3 = n\,(1-k)\,(1-2k)\,PQ\,(Q-P).$$

In a stratified selection we have formula (31) with

$$S^2 = (1-k)\,\Sigma\,(n_t P_t Q_t), \quad \kappa S^3 = (1-k)\,(1-2k)\,\Sigma\,\{n_t P_t Q_t\,(Q_t - P_t)\}, \quad (37)$$

where P_t is the proportion with, Q_t without, the attribute in the tth district, which contains N_t persons, from which $n_t = kN_t$ are selected.

Write
$$P_t = P + v_t,$$
$$n\sigma_v^2 = \Sigma n_t v_t^2 = \Sigma n_t\,(P_t - P)^2,$$
$$nV_{12} = \Sigma n_t v_t \cdot {}_t m_2 = \Sigma\,\{n_t v_t\,(P + v_t)\,(Q - v_t)\}$$
$$= \Sigma n_t\,(Q - P)\,v_t^2 - \Sigma n_t v_t^3, \text{ since } \Sigma n_t v_t = 0,$$
$$n \cdot {}_3\mu_v = \Sigma n_t v_t^3.$$

Then as in formulae (32), (33), (34) if σ_d^2 and $\kappa_d\sigma_d^3$ are the second and third moments of the frequency of x,

$$\sigma_d^2 = n\,(1-k)\,(s_1^2 - \sigma_v^2) = n\,(1-k)\,(PQ - \sigma_v^2), \quad \ldots\ldots\ldots(38)$$
$$\kappa_d\sigma_d^3 = n\,(1-k)\,(1-2k)\,\{PQ\,(Q-P) - 3\,(Q-P)\,\sigma_v^2 + 2 \cdot {}_3\mu_v\}.$$

The precision is improved by stratification. If we take an extreme case where all the persons with the attribute were concentrated in the first district,

$$P_1 n_1 = Pn, \; v_1 = P_1 - P, \; v_2 = v_3 = \ldots = v_c = -P,$$
$$n\sigma_v^2 = n_1\,(P_1 - P)^2 + (n - n_1)\,P^2 = nP\,(P_1 - P) = nP\,(Q - Q_1),$$

and
$$\sigma_d^2 = n\,(1-k)\,PQ_1 = nP\,(1-k)\left(1 - P\,\frac{N}{N_1}\right).$$

It is found that $\kappa_d\sigma_d^3 = n\,(1-k)\,(1-2k)\,PQ_1\,(Q_1 - P_1)$.

In all the cases discussed in this section the frequency of x when $Pn + x$ appear in the selection has been given.

Now let p be the proportion found in the selection, so that $pn = Pn + x$, and $p = P + z$, where $x = nz$.

For the frequency curve of z, write $\sigma_p = \frac{\sigma}{n}$ in formula (36). The κ in the equations is unchanged. We have

$$y = \frac{1}{\sigma_p\sqrt{(2\pi)}} \cdot e^{-\frac{z^2}{2\sigma_p^2}} \left\{1 - \frac{\kappa_1}{2}\left(\frac{z}{\sigma_p} - \frac{z^3}{3\sigma_p^3}\right)\right\}, \quad \ldots\ldots\ldots(39)$$

where $\sigma_p = \sqrt{\left(\frac{(1-k)\,PQ}{n}\right)}$, and $\kappa_1 = \frac{(1-2k)\,(Q-P)}{\sqrt{\{n\,(1-k)\,PQ\}}}$, k being $\frac{n}{N}$.

If the universe is unrestricted we have merely to put $k = 0$ in this formula.

For a stratified universe, we write ${}_a\sigma_d$ for σ_p, where

$$ {}_a\sigma_d = \sqrt{\left\{\frac{(1-k)\,(PQ - \sigma_v^2)}{n}\right\}}. \quad \ldots\ldots\ldots\ldots\ldots\ldots(40) $$

V. Law of Small Numbers

If the sample is stratified, it is indifferent to the argument how the P_t's are distributed; for since in formula (37) some of the terms such as $P_t n_t$ must be comparable with Pn, the aggregate of the terms is of order n when P is finite, and the coefficients of α^2, α^3 ... in the exponent on p. 25 are of order 1, $n^{-\frac{1}{2}}$, n^{-1} ... even though many of the P_t's in the *strata* are small or even zero.

But if P in the *universe* is so small that Pn is no longer large, the whole argument breaks down*, and we must start again from the beginning.

Let P be small, but $Pn = w$ be finite.

Let N be the number in the universe, $n = kN$, k finite. Write $M = N - n$.

[To elucidate the order in which terms are neglected suppose that $N = 10{,}000$, $n = 1000$, $P = \cdot 02$. Then $k = \cdot 1$, $w = 20$, $PN = 200$, $M = 9000$.]

Write E_x for the chance that $Pn + x = r$ will be found with the attribute in question in the sample.

Then out of ${}_NC_n$ possible selections ${}_{PN}C_{Pn+x} \times {}_{QN}C_{Qn-x}$ contain exactly $Pn + x$ cases.

Therefore

$$ E_x = \frac{{}_{PN}C_{Pn+x} \times {}_{QN}C_{Qn-x}}{{}_NC_n} = \frac{(PN)!\,(QN)!\,n!\,M!}{(Pn+x)!\,(PM-x)!\,(Qn-x)!\,(QM+x)!\,N!} $$

$$ = \frac{n!}{r!\,(n-r)!} \cdot \frac{M!}{(PM-x)!\,(QM+x)!} \cdot \frac{(PN)!\,(QN)!}{N!}. $$

Apply Stirling's formula, viz.

$$ m! = m^{m+\frac{1}{2}} \cdot e^{-m} \cdot \sqrt{(2\pi)}, $$

in which $\frac{1}{12m}$ is neglected in comparison with 1 in the index, to all the factorials except $r!$ Then $\frac{1}{12PN}$, $\frac{1}{12\,(n-r)}$ and smaller fractions are neglected.

* If we use the limit of $P^{nP+x} \cdot Q^{nQ-x} \cdot {}_nC_{nP+x}$ for our analysis, we find that we have to assume that $\frac{1}{Pn}$ is negligible. In the method of Section I, $M_2 = M_3$, and C_4, C_5 ... are not in descending order of magnitude.

After a little reduction we have

$$E_x = \frac{e^{-r}n^r}{r!\left(1-\frac{r}{n}\right)^n}\cdot\left(1-\frac{r}{n}\right)^{r-\frac{1}{2}}\cdot\frac{N^{\frac{1}{2}}}{M^{\frac{1}{2}}}\cdot P^r\cdot Q^{n-r}$$

$$\times\left(1-\frac{x}{PM}\right)^{-PM+x-\frac{1}{2}}\cdot\left(1+\frac{x}{QM}\right)^{-QM-x-\frac{1}{2}}.$$

Here $\left(1-\frac{r}{n}\right)^n = e^{-r}$, if $\frac{1}{n}$ is neglected, since r is comparable with w and is finite.

Also $\left(1-\frac{r}{n}\right)^{r-\frac{1}{2}}Q^{-r} = \left(Q-\frac{x}{n}\right)^{r-\frac{1}{2}}Q^{-r} = Q^{-\frac{1}{2}}\left(1-\frac{x}{nQ}\right)^{r-\frac{1}{2}}$

$$= \left(1-\frac{w}{n}\right)^{-\frac{1}{2}}\left(1-\frac{x}{nQ}\right)^{w+x-\frac{1}{2}} = 1,$$

since $\frac{wx}{n} = Px$ is negligible.

Again $N^{\frac{1}{2}} \div M^{\frac{1}{2}} = (1-k)^{-\frac{1}{2}}, \quad Q^n = e^{-w},$

and $n^r P^r = w^r.$

Therefore

$$E_x = \frac{e^{-w}\cdot w^r}{r!}(1-k)^{-\frac{1}{2}}\cdot\left(1-\frac{x}{PM}\right)^{-PM+x-\frac{1}{2}}\cdot\left(1+\frac{x}{QM}\right)^{-QM-x-\frac{1}{2}}, \qquad \text{.....(41)}$$

when terms containing $\frac{1}{n}$ are neglected.

It is only possible to proceed further if we now regard $\frac{1}{PN} = \frac{k}{w}$, $= \frac{1}{200}$ in the illustration, as small.

Write $PN = W.$

Then $PQM = PN\,(1-k)\,Q = W\,(1-k)$, neglecting $\frac{w}{n}$.

Write $L = \log\left\{\left(1-\frac{x}{PM}\right)^{-PM+x-\frac{1}{2}}\cdot\left(1+\frac{x}{QM}\right)^{-QM-x-\frac{1}{2}}\right\}$

$$= (PM-x+\tfrac{1}{2})\left(\frac{x}{PM}+\frac{x^2}{2P^2M^2}+\ldots\right)-(QM+x+\tfrac{1}{2})\left(\frac{x}{QM}-\frac{x^2}{2Q^2M^2}+\ldots\right)$$

$$= -\frac{x^2}{2PQM}+\frac{x\,(Q-P)}{2PQM}-\frac{x^3\,(Q-P)}{6P^2Q^2M^2}+\ldots.$$

The standard deviation of x is $\sqrt{\{PQn\,(1-k)\}} = \sqrt{\{w\,(1-k)\}}$, neglecting $\frac{w}{2n}$.

Write $x^2 = \tau^2\,.\,w\,(1-k)$, so that τ is finite.

$$L = -\frac{\tau^2 k}{2}+\frac{\tau\sqrt{w}}{2W\sqrt{(1-k)}}-\frac{\tau^3 w^{\frac{3}{2}}}{6W^2\,(1-k)}-\ldots.$$

Now neglect $\frac{1}{W}$. Then

$$L = -\frac{\tau^2 k}{2} = -\frac{x^2 k}{2w(1-k)} = -\frac{x^2}{2PM},$$

and from (41) $$E_x = \frac{e^{-w} \cdot w^r}{r!}(1-k)^{-\frac{1}{2}} \cdot e^{-\frac{x^2}{2PM}}, \quad \ldots\ldots\ldots\ldots\ldots(42)$$

approximately, when $\frac{1}{PN}$ is neglected.

If the universe is unrestricted, $k = 0$, and this reduces to the well-known form

$$E_x = \frac{e^{-w} \cdot w^r}{r!}, \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots(43)$$

as may be obtained directly from the limit of ${}_nC_r \cdot P^r \cdot Q^{n-r}$.

The following table shows comparisons between the results obtained by applying the general law and the law of small numbers, distinguishing between a restricted and an unrestricted universe.

	(1) $E(x)$			(2) $E(x)$		
x	r	(a)	(b)	r	(c)	(d)
−10	0	·000	·00005	1, 2 or 3	·000	000
−9	1	·000	·0005	4, 5 ,, 6	·000	·000
−8	2	·003	·002	7, 8 ,, 9	·003	·003
−7	3	·008	·008	10, 11 ,, 12	·028	·030
−6	4	·019	·019	13, 14 ,, 15	·109	·113
−5	5	·037	·038	16, 17 ,, 18	·230	·231
−4	6	·062	·063	19, 20 ,, 21	·280	·276
−3	7	·089	·090	22, 23 ,, 24	·208	·205
−2	8	·112	·113	25, 26 ,, 27	·100	·100
−1	9	·126	·125	28, 29 ,, 30	·033	·033
0	10	·127	·125	31, 32 ,, 33	·008	·008
1	11	·115	·113	34, 35 ,, 36	·001	·001
2	12	·095	·095	37, 38 ,, 39	·000	000
3	13	·072	·073			
4	14	·051	·052			
5	15	·034	·035			
6	16	·022	·022			
7	17	·013	·013			
8	18	·007	·007			
9	19	·004	·004			
10	20	·002	·002			

Take (1) $n = 1000$, $P = \cdot 01$, N indefinitely large. $w = 10$.
(2) $n = 1000$, $P = \cdot 02$, $N = 10{,}000$. $k = \cdot 1$, $w = 20$.
Apply to (1) the formulae

$$E_x = \frac{1}{s\sqrt{(2\pi)}} \cdot e^{-\frac{x^2}{2s^2}}\left\{1 - \frac{Q-P}{2s}\left(\frac{x}{s} - \frac{x^3}{3s^3}\right)\right\}, \quad \ldots\ldots\ldots(a)$$

where $s^2 = PQn$, and

$$E_x = \frac{e^{-w}}{r!} w^r, \text{ where } r = 10 + x. \quad \ldots\ldots\ldots\ldots\ldots(b)$$

Apply to (2) the formulae

$$E_x = \frac{1}{\sigma\sqrt{(2\pi)}} \cdot e^{-\frac{x^2}{2\sigma^2}} \left\{1 - \frac{Q-P}{2\sigma}\left(1 - \frac{2n}{N}\right)\left(\frac{x}{\sigma} - \frac{x^3}{3\sigma^3}\right)\right\}, \quad \ldots\ldots(c)$$

where $\sigma^2 = PQn\left(1 - \frac{n}{N}\right)$, and

$$E_x = \frac{e^{-w}}{r!} \cdot w^r (1-k)^{-\frac{1}{2}} \cdot e^{-\frac{x^2 k}{2w(1-k)}}, \quad \ldots\ldots\ldots\ldots(d)$$

where $r = 20 + x$.

It is noticeable that even when w is as small as 10 the differences between the results of (a) and (b) are insignificant.

An abbreviated table of the values of $\frac{e^{-w} \cdot w^r}{r!}$ is given for convenience of reference. A more complete statement will be found in *Tables for Biometricians*, p. 113.

Values of $\frac{e^{-w} \cdot w^r}{r!}$.

	w															
r	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
0	·368	·135	·050	·018	·007	·002	·001	·000	·000	—	—	—	—	—	—	—
1	·368	·271	·149	·073	·034	·015	·006	·003	·001	—	—	—	—	—	—	—
2	·184	·271	·224	·147	·084	·045	·022	·011	·005	·002	·001	—	—	—	—	—
3	·061	·180	·224	·195	·140	·089	·052	·029	·015	·008	·004	·002	·001	—	—	—
4	·015	·090	·168	·195	·175	·134	·091	·057	·034	·019	·010	·005	·003	·001	·001	—
5	·003	·036	·101	·156	·175	·161	·128	·092	·061	·038	·022	·013	·007	·004	·002	·001
6*	—	·012	·050	·104	·146	·161	·149	·122	·091	·063	·041	·025	·015	·009	·005	·003
7	—	·003	·022	·060	·104	·138	·149	·140	·117	·090	·065	·044	·028	·017	·010	·006
8	—	·001	·008	·030	·065	·103	·130	·140	·132	·113	·089	·066	·046	·030	·019	·012
9	—	—	·003	·013	·036	·069	·101	·124	·132	·125	·109	·087	·066	·047	·032	·021
10	—	—	·001	·005	·018	·041	·071	·099	·119	·125	·119	·105	·086	·066	·049	·034
11	—	—	—	·002	·008	·023	·045	·072	·097	·114	·119	·114	·101	·084	·066	·050
12	—	—	—	—	·003	·011	·026	·048	·073	·095	·109	·114	·110	·098	·083	·066
13	—	—	—	—	·001	·005	·014	·030	·050	·073	·093	·106	·110	·106	·096	·081
14	—	—	—	—	—	·002	·007	·017	·032	·052	·073	·090	·102	·106	·102	·093
15	—	—	—	—	—	·001	·003	·009	·019	·035	·053	·072	·088	·099	·102	·099
16	—	—	—	—	—	—	·001	·005	·011	·022	·037	·054	·072	·087	·096	·099
17	—	—	—	—	—	—	—	·002	·006	·013	·024	·038	·055	·071	·085	·093
18	—	—	—	—	—	—	—	·001	·003	·007	·015	·026	·040	·055	·071	·083
19	—	—	—	—	—	—	—	—	·001	·004	·008	·016	·027	·041	·056	·070
20	—	—	—	—	—	—	—	—	·001	·002	·005	·010	·018	·029	·042	·056
21	—	—	—	—	—	—	—	—	—	·001	·002	·006	·011	·019	·030	·043
22	—	—	—	—	—	—	—	—	—	—	·001	·003	·006	·012	·020	·031
23	—	—	—	—	—	—	—	—	—	—	·001	·002	·004	·007	·013	·022
24	—	—	—	—	—	—	—	—	—	—	—	·001	·002	·004	·008	·014
25	—	—	—	—	—	—	—	—	—	—	—	—	·001	·002	·005	·009
26	—	—	—	—	—	—	—	—	—	—	—	—	—	·001	·003	·006
27	—	—	—	—	—	—	—	—	—	—	—	—	—	·001	·002	·003
28	—	—	—	—	—	—	—	—	—	—	—	—	—	—	·001	·002
29	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	·001*

* In each case the total of the quantities for greater values of r than those computed is less than ·001.

It is worth while to examine specially the chance of missing altogether any instance of an attribute, and to test how far this chance is diminished by stratification.

In an unstratified sample the chance is Q^n, which is less than e^{-w}, where $w = Pn$. This is only perceptible when w is a small number, say less than 5.

In a stratified sample we find, using the notation of p. 11, that the chance of selecting none is

$$C = (1 - P_1)^{n_1} . (1 - P_2)^{n_2} \ldots (1 - P_d)^{n_d}.$$

$$\log C = n_1 \log (1 - P_1) + n_2 \log (1 - P_2) + \ldots$$

$$= - (n_1 P_1 + n_2 P_2 + \ldots) - \tfrac{1}{2} (n_1 P_1^2 + n_2 P_2^2 + \ldots)$$

$$= - nP - \frac{n}{2} (P^2 + \sigma_v^2), \text{ when } n_1 P_1^3 \text{ etc. are neglected.}$$

But $$\log Q^n = n \log Q = n \log (1 - P) = - nP - \frac{n}{2} P^2,$$

when nP^3 etc. are neglected.

Therefore $$C = Q^n . e^{-\frac{1}{2} n \sigma_v^2}.$$

If, however, there is much concentration, we ought not to neglect $n_1 P_1^3$ etc. For example, if all the persons with the attribute are concentrated in the first district, $P_1 n_1 = Pn$, and $Q_2 = Q_3 = \ldots = Q_c = 1$.

Then $$C \div Q^n = \left(1 - \frac{w}{r_1}\right)^{n_1} \div \left(1 - \frac{w}{n}\right)^n = 1 - \frac{w^2}{2n_1}, \quad \ldots\ldots\ldots(44)$$

if P_1^2 and $P_1 w$ are neglected.

The reduction of Q^n is small. In the last formula, if $n = 10{,}000$, $n_1 = 100$ and $P_1 = \cdot 04$, then $w = 4$ and $P = \cdot 004$.

$Q^n = \cdot 0182$. $C = \cdot 0169 = Q^n \times \cdot 923$.

VI. Distribution of Alternative Attributes

In a population containing N persons, $P_1 N$ are in one class (have a certain attribute), $P_2 N$ in an alternative class and so on to $p_c N$, so that

$$P_1 + P_2 + \ldots + P_c = 1.$$

n are selected (without replacement) from the whole population, and there are found in the respective classes

$$p_1 n = P_1 n + x_1, p_2 n = P_2 n + x_2 \ldots p_c n = P_c n + x_c.$$

Then $$p_1 + p_2 + \ldots + p_c = 1,$$

$$x_1 + x_2 + \ldots + x_c = 0.$$

The chance, E_x, that this selection should be made is the product of the separate chances that $p_1 n$ should be found out of $P_1 N$, $p_2 n$ out of $P_2 N$ ... divided by the whole number of possible selections of n out of N.

Therefore

$$E_x = ({}_{P_1 N}C_{p_1 n} \times {}_{P_2 N}C_{p_2 n} \times \ldots \times {}_{P_c N}C_{p_c n}) \div {}_N C_n$$

$$= \frac{(P_1 N)!}{(P_1 n + x_1)! (P_1 M - x_1)!} \times \frac{(P_2 N)!}{(P_2 n + x_2)! (P_2 M - x_2)!} \times \ldots \times \frac{n! M!}{N!},$$

where $M = N - n$.

Write E_0 for the value of E_x when $0 = x_1 = x_2 = \ldots = x_c$.

Apply Stirling's formula to all the factorials, neglecting such terms as $\frac{1}{12P_1n}$ in the index, and take logarithms. Then

$$\log (E_x \div E_0)$$

$$= -(P_1n + x_1 + \tfrac{1}{2}) \log\left(1 + \frac{x_1}{P_1n}\right) - (P_2n + x_2 + \tfrac{1}{2}) \log\left(1 + \frac{x_2}{P_2n}\right) - \ldots$$

$$- (P_1M - x_1 + \tfrac{1}{2}) \log\left(1 - \frac{x_1}{P_1M}\right) - (P_2M - x_2 + \tfrac{1}{2}) \log\left(1 - \frac{x_2}{P_2M}\right) - \ldots.$$

The standard deviation of x_1, when only the first class is considered, is

$$\sqrt{\left\{P_1 (1 - P_1)\, n \left(1 - \frac{n}{N}\right)\right\}},$$

and hence $x_1, x_2 \ldots$ are of the order $\sqrt{n}$, if $P_1, P_2 \ldots$ are all finite.

Expand, collect terms, and neglect those of order $\frac{1}{n}$ or lower. Then

$$\log (E_x \div E_0) = -\frac{N}{2M}\left(\frac{x_1^2}{P_1n} + \frac{x_2^2}{P_2n} + \ldots\right)$$

$$- \tfrac{1}{2}\left(1 - \frac{2n}{N}\right)\frac{N}{M}\left[\left(\frac{x_1}{P_1n} - \frac{x_1^3N}{3P_1^2n^2M}\right) + \left(\frac{x_2}{P_2n} - \frac{x_2^3N}{3P_2^2n^2M}\right) + \ldots\right].$$

Write $$x_1^2 = \frac{M}{N} \cdot P_1nz_1^2,\ x_2^2 = \frac{M}{N} \cdot P_2nz_2^2 \ldots,$$

and $$\frac{n}{N} = k, \text{ so that } M = N(1 - k).$$

By Stirling's formula it is found that

$$E_0 = \{2\pi n (1 - k)\}^{-\frac{c-1}{2}} \cdot (P_1 \cdot P_2 \ldots P_c)^{-\frac{1}{2}}.$$

Therefore

$$E_x = E_0 \cdot e^{-\frac{1}{2}(z_1^2 + z_2^2 + \ldots) - \frac{1}{2}(1-2k)N^{\frac{1}{2}}(Mn)^{-\frac{1}{2}}[P_1^{-\frac{1}{2}}(z_1 - \frac{1}{3}z_1^3) + P_2^{-\frac{1}{2}}(z_2 - \frac{1}{3}z_2^3) + \ldots]}$$

$$= \{2\pi n (1 - k)\}^{-\frac{c-1}{2}} \cdot (P_1 \cdot P_2 \ldots P_c)^{-\frac{1}{2}} \cdot e^{-\frac{1}{2}\Sigma z^2}$$

$$\times [1 - \tfrac{1}{2}(1 - 2k)(1 - k)^{-\frac{1}{2}} n^{-\frac{1}{2}} \Sigma \{P^{-\frac{1}{2}}(z - \tfrac{1}{3}z^3)\}], \quad \ldots(45)$$

since in the expansion of the second part of the index terms involving $\frac{1}{n}$ may be neglected.

[Formula (39), p. 32, is a particular case of this when $c = 2$, $x_1 = x - x_2$, $P_1 = P$, $P_2 = Q$.]

The formula can be used in full, but it is easier to discuss its use when n is so large that $n^{-\frac{1}{2}}$ is negligible.

In this case write $m_1 = P_1n$, $m_2 = P_2n \ldots$, so that $m_1, m_2 \ldots$ are the numbers that would be obtained in classes 1, 2 ... if the proportions in the sample were exactly the same as in the universe. Then

$$x_1^2 = (1 - k)\, m_1z_1^2 \text{ etc.}$$

and $$E_x = E_0 \cdot e^{-\frac{1}{2(1-k)}\left(\frac{x_1^2}{m_1} + \frac{x_2^2}{m_2} + \ldots + \frac{x_c^2}{m_c}\right)}.$$

Write $$\chi^2 = \frac{1}{1-k}\left(\frac{x_1^2}{m_1} + \frac{x_2^2}{m_2} + \ldots + \frac{x_c^2}{m_c}\right)$$
$$= \frac{n}{1-k}\sum_{t=1}^{t=c}\frac{(P_t - p_t)^2}{p_t}.$$

Then $$E_x = E_0 \cdot e^{-\frac{1}{2}\chi^2}. \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(46)$$

Complexes of errors have the same probability if they result in the same value of χ^2.

A table has been computed (*Biometrika*, vol. I, pp. 155 *et seq.*) which shows the probability that given values of χ^2 will be exceeded. From it the rough generalization can be made that it is rather more likely than not that χ^2 will exceed the value $c - 2$, where $c\ (> 2)$ is the number of classes, and that the odds are more than 20 to 1 against χ^2 exceeding $2c$. (Bowley, *Elements of Statistics*, 4th ed., p. 431.)

VII. The Inverse Problem

So far we have considered the frequency of errors in sampling from a known universe, and the results have been expressed in terms of the data from the universe. In practice our data must be drawn from the sample, and the first step is to transform the principal formulae in this sense.

A. *Adjustment of the formulae*

One attribute.

Use formula (36). Write $\sigma'^2 = pqn\,(1-k)$, while $\sigma^2 = PQn\,(1-k)$; $pn = Pn + x$, $qn = Qn - x$. Then

$$\frac{1}{\sigma^2} = \frac{n}{1-k}\cdot\frac{1}{(pn-x)(qn+x)} = \frac{1}{pqn\,(1-k)}\cdot\left(1-\frac{x}{pn}\right)^{-1}\cdot\left(1+\frac{x}{qn}\right)^{-1}$$
$$= \frac{1}{\sigma'^2}\left\{1 + \frac{x\,(q-p)}{pqn}\right\}, \text{ if } \frac{x^2}{p^2n^2} \text{ is neglected;}$$

and $$\frac{1}{\sigma} = \frac{1}{\sigma'}\left\{1 + \frac{x\,(q-p)}{2pqn}\right\}.$$

Hence $$e^{-\frac{x^2}{2\sigma^2}} = e^{-\frac{x^2}{2\sigma'^2}}\cdot e^{-\frac{x^3(q-p)}{2\sigma'^2 pqn}} = e^{-\frac{x^2}{2\sigma'^2}}\left\{1 - \frac{x^3\,(q-p)}{2\sigma'^2 pqn}\right\},$$

since terms in $\frac{1}{n}$ are neglected. Again

$$\frac{\kappa_1}{2}\left(\frac{x}{\sigma} - \frac{x^3}{3\sigma^3}\right) = \tfrac{1}{2}(1-2k)(Q-P)\left(\frac{x}{\sigma^2} - \frac{x^3}{3\sigma^4}\right)$$
$$= \tfrac{1}{2}(1-2k)(Q-P)\left[\frac{x}{\sigma'^2}\left\{1 + \frac{x\,(q-p)}{pqn}\right\} - \frac{x^3}{3\sigma'^4}\left\{1 + \frac{2x\,(q-p)}{pqn}\right\}\right]$$
$$= (1 - \tfrac{1}{2}k)(q-p)\left(\frac{x}{\sigma'^2} - \frac{x^3}{3\sigma'^4}\right),$$

since terms in $\frac{1}{n}$ are neglected.

Hence, neglecting terms in $\frac{1}{n}$, we have

$$E_x = \frac{1}{\sigma' \sqrt{(2\pi)}} \cdot e^{-\frac{x^2}{2\sigma'^2}} \left\{1 + \frac{x\,(q-p)}{2pqn}\right\} \left\{1 - \frac{x^3\,(q-p)}{2\sigma'^2 pqn}\right\}$$
$$\times \left\{1 - (q-p)\,(1 - \tfrac{1}{2}k)\left(\frac{x}{\sigma'^2} - \frac{x^3}{3\sigma'^4}\right)\right\}$$
$$= \frac{1}{\sigma' \sqrt{(2\pi)}} \cdot e^{-\frac{x^2}{2\sigma'^2}} \left\{1 - (q-p)\,(1 - \tfrac{1}{2}k)\,\frac{x^3}{3\sigma'^4}\right\}, \quad \text{.....(47)}$$

since the coefficient of x is $\frac{q-p}{2pq\,(N-n)}$ and is negligible.

This result may also be obtained from the form

$$E_x = \frac{{}_{PN}C_{pn} \times {}_{QN}C_{qn}}{{}_{N}C_{n}} = {}_{pN+x\frac{N}{n}}C_{pn} \times {}_{qN-x\frac{N}{n}}C_{qn} \div {}_{N}C_{n},$$

by writing $\frac{x}{n} = x'$, replacing factorials by Stirling's formula, taking logarithms, expanding and collecting terms.

In a stratified sample we cannot get an explicit result, if we retain terms of order $\frac{1}{\sqrt{n}}$; but if $\frac{1}{\sqrt{n}}$ is negligible, the chance that pn would be found in the aggregate, if the proportion in the universe was $p - \frac{x}{n}$, is readily shown to be

$$E_x = \frac{1}{\sigma_d' \sqrt{(2\pi)}} \cdot e^{-\frac{x^2}{2\sigma_d'^2}}, \quad \text{.....(48)}$$

where $\sigma_d'^2 = (pq - \sigma_v'^2)\,n\,(1-k)$, and

$$n\sigma_v'^2 = n_1\,(p_1 - p)^2 + n_2\,(p_2 - p)^2 + \ldots.$$

Distribution of attributes.

Replace $P_1, P_2 \ldots$ in formula (45) by $p_1 + \frac{x_1}{n}$, $p_2 + \frac{x_2}{n} \ldots$ and neglect terms of order $\frac{1}{n}$. Then

$$(P_1 \cdot P_2 \ldots)^{-\frac{1}{2}} = (p_1 \cdot p_2 \ldots)^{-\frac{1}{2}} \left\{1 + \frac{1}{2n}\left(\frac{x_1}{p_1} + \frac{x_2}{p_2} + \ldots\right)\right\},$$

$$\Sigma z^2 = \frac{N}{M}\left(\frac{{x_1}^2}{p_1 n + x_1} + \frac{{x_2}^2}{p_2 n + x_2} + \ldots\right)$$
$$= \frac{N}{M}\left\{\frac{{x_1}^2}{p_1 n} + \frac{{x_2}^2}{p_2 n} + \ldots - \frac{1}{n^2}\left(\frac{{x_1}^3}{{p_1}^3} + \frac{{x_2}^3}{{p_2}^3} + \ldots\right)\right\},$$

$$\tfrac{1}{2}\,(1-2k)\,(1-k)^{-\frac{1}{2}}\,n^{-\frac{1}{2}}\,\Sigma\,\{P^{-\frac{1}{2}}\,(z - \tfrac{1}{3}z^3)\}$$
$$= (\tfrac{1}{2} - k)\,(1-k)^{-1}\,\Sigma\left(\frac{x}{pn} - \frac{x^3}{3\,(1-k)\,p^2 n^2}\right),$$

$$E_x = \{2\pi n\,(1-k)\}^{-\frac{c-1}{2}}\,(p_1 . p_2 \ldots)^{-\frac{1}{2}}\, e^{-\frac{1}{2(1-k)}\Sigma\frac{x^2}{pn}}\left\{1 - \frac{(2-k)}{6\,(1-k)^2 n^2}\,\Sigma\frac{x^3}{p^2}\right\}$$

$$= Ce^{-\frac{1}{2}\chi^2}\left\{1 - \frac{(2-k)\,n}{6\,(1-k)^2}\,\Sigma\frac{(P_t - p_t)^3}{p_t{}^2}\right\}, \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(49)$$

where $$\chi^2 = \frac{n}{1-k}\,\Sigma\frac{(P_t - p_t)^2}{p_t}.$$

If $\frac{1}{\sqrt{n}}$ is neglected, we have

$$E_x = Ce^{-\frac{1}{2}\chi^2}, \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(50)$$

and for given observations of $p_1, p_2 \ldots$ the universes can be classified according to the probabilities that they would yield these values.

Magnitude of an average.

If a universe contains N magnitudes, whose average is $\bar{x}$, standard deviation σ and moments about the average are $\mu_2, \mu_3 \ldots$, the chance that a random sample of n things will yield an average $\bar{u}_1 = \bar{x} + x$ is given by formula (29).

Given the sample only, however, we do not know the values of μ_2 or μ_3, which are there involved.

Let μ_2', μ_4' be the observed second and fourth moments in the sample about the observed average, and σ' the standard deviation.

The chances that various values of the differences $\mu_2' - \mu_2$ should be found in the sample are given by a frequency curve which approximates to normality when n is increased with standard deviation $\sqrt{\left(\frac{\mu_4 - \mu_2{}^2}{n}\right)}$*, which may be written $\sqrt{\left(\frac{\mu_4' - \mu_2'{}^2}{n}\right)}$ when $\frac{1}{n}$ is neglected.

Write $b_2 = \mu_4' \div \mu_2'{}^2$, and $\mu_2 = \mu_2' + \frac{d}{\sqrt{n}}$. Then d is a finite quantity comparable with $\mu_2'\sqrt{(b_2 - 1)}$.

When this value is written for μ_2 in the expression for E_x, a term is introduced which contains $\frac{d}{\sqrt{n}}$, which is of the same order as κ_1. We therefore can only carry our approximation definitely as far as the first term†.

In the case, then, where n is so large that $\frac{1}{\sqrt{n}}$ is negligible, we have

$$E_x = \frac{1}{\sigma'\sqrt{(2\pi)}} \cdot e^{-\frac{x^2}{2\sigma'^2}}, \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(51)$$

* *Biometrika*, vol. II, part III, p. 280. Bowley, *Elements of Statistics*, 4th ed. p. 417. Tshuprov in *Biometrika*, vol. XIII, p. 295, when $\frac{1}{N^3}$ is neglected and the original frequency curves are identical.

† See further on this question p. 45 below.

where $\sigma'^2 = \mu_2' \frac{1-k}{n}$ and differs from σ only by terms now supposed negligible.

In a stratified sample we may similarly replace the unknown constants in the universe by the constants computed from the sample (averages and standard deviations of the strata) when $\frac{1}{\sqrt{n}}$ is negligible.

B. *Inference from sample to universe*

The preceding formulae express the chances that the samples would be found from given universes, whose constants are computed from the sample. We have now to complete the problem by expressing the chances that the universes had these constants, a procedure which involves certain hypotheses.

One attribute.

Let $F(P)$ be the *à priori* chance that a universe should contain a proportion P having the attribute in question. The necessary hypothesis is that $F(P)$ should be continuous, and that its derived functions should be finite (so that their product with $\frac{1}{n}$ is negligible) in the neighbourhood of $P = p$, where p is the proportion found in the sample.

Write $f(p)$ for the chance that, when P is the proportion in the universe, p should be found in the sample.

Then the *à priori* combined chance that the universe should contain PN and the sample pn is $F(P) \times f(p)$.

We may write $pn = P_1N - x_1$, $pn = P_2N - x_2$ ....

For a number of pairs of values, (P_1, x_1), (P_2, x_2) ..., such chances are in the ratios $F(P_1) \times f_1(p) : F(P_2) \times f_2(p) : \ldots$, where $f_1(p), f_2(p)$ ... are the chances that p shall be found in the sample, when P_1, P_2 ... are the proportions in the universe.

One of the events resulting in p has by hypothesis taken place, and therefore the sum of these chances is unity.

Hence the *à posteriori* chance that, p being found in the sample, P was the proportion in the universe is

$$\frac{F(P) \times f(p)}{\Sigma \{F(P_t) \times f_t(p)\}},$$

the summation being extended over all values from $P = 0$ to $P = 1$.

Write Q_x for the chance that, p being found, $p - \frac{x}{n}$ was the proportion in the universe, and replace summation by integration.

Then, since $x = pn$ if $P = 0$, and $x = -qn$ if $P = 1$,

$$Q_x = \frac{F\left(p - \frac{x}{n}\right) . E_x}{\int_{pn}^{-qn} F\left(p - \frac{x}{n}\right) . E_x . dx},$$

where E_x has the value in formula (47).

The chance that P was within the limits $p \pm \frac{x}{n}$ is then

$$C_x = \int_{-x}^{x} Q_x . dx = \frac{\int_{-x}^{x} F\left(p - \frac{x}{n}\right) . E_x . dx}{\int_{-qn}^{pn} F\left(p - \frac{x}{n}\right) . E_x . dx} .$$

Write $x = z\sigma'$, where $\sigma'^2 = pqn\,(1 - k)$, $k = \frac{n}{N}$, and write

$$l = \frac{q - p}{6\sigma'}\,(2 - k).$$

Then $$E_x . dx = \frac{1}{\sqrt{(2\pi)}} . e^{-\frac{1}{2}z^2}\,(1 - lz^3) . dz.$$

σ' is of order $\sqrt{n}$, z is finite, x is of the order $\frac{1}{\sqrt{n}}$ and so is l. Also

$$F\left(p - \frac{x}{n}\right) = F\left(p - z\frac{\sigma'}{n}\right) = F\,(p) - z\frac{\sigma'}{n} . F'\,(p) + \tfrac{1}{2} . \frac{z^2\sigma'^2}{n^2} . F''\,(p) + \ldots$$

$$= F\,(p) - z\frac{\sigma'}{n} . F'\,(p), \text{ if terms of the order } \frac{1}{n} \text{ are neglected.}$$

Therefore

$$\int_{-x}^{x} F\left(p - \frac{x}{n}\right) . E_x . dx = \int_{-z}^{z} \left\{F\,(p) - z\frac{\sigma'}{n} . F'\,(p)\right\} . (1 - lz^3) . \frac{1}{\sqrt{(2\pi)}} . e^{-\frac{1}{2}z^2} . dz$$

$$= \int_{-z}^{z} \frac{1}{\sqrt{(2\pi)}} . e^{-\frac{1}{2}z^2} . dz \times F\,(p),$$

since terms in z and z^3 disappear, and $\frac{l\sigma' z^4}{n}$ is of the order $\frac{1}{n}$.

For the denominator of C_x it can be shown that pn and $-qn$ are replaceable by $\pm\,\infty$ when $\frac{1}{n}$ is neglected, so that the denominator becomes

$$F\,(p) \times \int_{-\infty}^{\infty} \frac{1}{\sqrt{(2\pi)}} . e^{-\frac{1}{2}z^2} . dz = F\,(p).$$

Therefore

$$C_x = \int_{-z}^{z} \frac{1}{\sqrt{(2\pi)}} . e^{-\frac{1}{2}z^2} . dz = \int_{-x}^{x} \frac{1}{\sigma'\sqrt{2\pi}} . e^{-\frac{x^2}{2\sigma'^2}} . dx, \ldots\ldots(52)$$

where $\sigma'^2 = pqn\left(1 - \frac{n}{N}\right)$. C_x is independent of $F\,(p)$.

This is the chance that the proportion in the universe was within the limits $p \pm \frac{x}{n}$, when p is found in a sample of n things drawn at random from N, subject to the hypotheses enumerated*.

The same argument applies to a stratified sample, but in that case we must have n sufficiently large to allow us to neglect $\frac{1}{\sqrt{n}}$ throughout (p. 40).

Distribution of attributes.

We can argue on the same lines, using formula (49), that it is very unlikely that the universe was such that

$$\frac{nN}{N-n}\left\{\frac{(P_1-p_1)^2}{p_1}+\frac{(P_2-p_2)^2}{p_2}+\ldots\right\}$$

exceeded $2c$ (see p. 39), where c is the number of classes, if we assume continuity of the *à priori* chances of the values of $P_1, P_2 \ldots$.

Average of variables.

The discussion proceeds on similar lines*.

Let $F(\bar{x})$ be the *à priori* chance that the average in the universe is $\bar{x}$, and $f(\bar{u}_1)$ the chance that the average in the sample will then be $\bar{u}_1$. Suppose that $F(\bar{x})$ is continuous and that its derived functions in the neighbourhood of $\bar{x} = \bar{u}_1$ are finite.

Write $\bar{u}_1 = \bar{x} + x$.

The chance that, given $\bar{u}_1$, the average in the universe was $\bar{x}$ is

$$Q_x = \frac{F(\bar{u}_1 - x) \,.\, E_x}{\int_b^a F(\bar{u}_1 - x) \,.\, E_x \,.\, dx},$$

where E_x is given by formula (51), and a and b are the greatest and least possible values of x. In this formula $\frac{1}{\sqrt{n}}$ is neglected.

The chance that the average in the universe was within the limits $\bar{u}_1 \pm x$ is

$$C_x = \int_{-x}^{x} Q_x \,.\, dx = \frac{\int_{-x}^{x} F(\bar{u}_1 - x) \,.\, E_x \,.\, dx}{\int_a^b F(\bar{u}_1 - x) \,.\, E_x \,.\, dx}.$$

Write $x = \tau\sigma'$. $\sigma'^2 = \mu_2'\left(\frac{1}{n} - \frac{1}{N}\right)$, μ_2' being the second moment found from the sample. τ is finite, σ' of the order $\frac{1}{\sqrt{n}}$, the standard deviation of the frequency group exhibited by the sample being considered to be finite.

* For this proof see *Metron*, vol. II, No. 3, "The precision of measurements estimated by samples."

The numerator of C_x

$$= \int_{-\tau}^{\tau} \{F(\bar{u}_1) - \tau\sigma' F'(\bar{u}_1) + \tfrac{1}{2}\tau^2\sigma'^2 F''(\bar{u}_1) \ldots\} \cdot \frac{1}{\sqrt{(2\pi)}} \cdot e^{-\frac{1}{2}\tau^2} \cdot d\tau$$

$$= F(\bar{u}_1) \cdot \int_{-\tau}^{\tau} \frac{1}{\sqrt{(2\pi)}} \cdot e^{-\frac{1}{2}\tau^2} \cdot d\tau, \text{ when } \frac{1}{\sqrt{n}} \text{ is neglected.}$$

The denominator of C_x is

$$F(\bar{u}_1) \cdot \int_a^b \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\tau^2} \cdot d\tau = F(\bar{u}_1),$$

since the whole chance of a value of τ is unity.

Therefore $$C_x = \int_{-x}^{x} \frac{1}{\sigma' \sqrt{(2\pi)}} \cdot e^{-\frac{x^2}{2\sigma'^2}} \cdot dx. \quad \ldots\ldots\ldots\ldots\ldots\ldots(53)$$

Here we have neglected $\frac{1}{\sqrt{n}}$, but we can retain it with a slight modification.

Let us go back to formula (29) and consider σ in the universe as known. We can then keep the term involving κ_1 which is of the order $\frac{1}{\sqrt{n}}$ throughout the analysis just given, till it disappears in integration. In formula (53) we then have σ instead of σ', correct to terms involving $\frac{1}{\sqrt{n}}$.

But σ differs from σ' (p. 41) in accordance with a frequency curve whose standard deviation is $\sqrt{\frac{\mu_4' - \mu_2'^2}{4n\mu_2'}}$, since the standard deviation for μ_2 is $\sqrt{\frac{\mu_4' - \mu_2'^2}{n}}$ and $\sigma^2 = \mu_2$. Here μ_4' and μ_2' can be found from the sample.

In using formula (53) enlarge σ by say three times this standard deviation, and we can safely say that C_x is less than the expression so found, even though $\frac{1}{\sqrt{n}}$ is not neglected.

II. PURPOSIVE SELECTION

Summary

INTRODUCTION

THE problems presented by purposive selection differ in emphasis, rather than in kind, from those already discussed when the selection is random. In both methods we are concerned with the proportion, with the average, or with the distribution of some quantity or attribute. In both methods the two fundamental factors in the measurement of the precision of the observations are the *dispersion* from their mean of the proportions or averages through the "universe" under consideration, and the *number of entries* (in random selection the number of individuals, in purposive that of districts) that are included in the sample. In each method the precision may be increased by *stratification*. The essential difference is that in purposive selection the unit is an aggregate, such as a whole district*, and the sample is an aggregate of these aggregates, while in random selection the unit is a person or thing, which may or may not possess an attribute, or with which some measurable quantity is associated. It results that we are concerned with *weighted*, instead of with unweighted averages. Further the fact that the selection is purposive very generally involves intentional dependence on *correlation*, the correlation between the quantity sought and one or more known quantities. Consequently the most important additional investigation in this section relates to the question

* Or, as M. March suggests, an "establishment."

how far the precision of the measurements is increased by correlation, and how best an enquiry can be arranged to maximize this precision.

As before we first deal with averages and proportions in which one attribute is involved, and secondly with the distribution of a population within grades on a scale of measurements.

1. Averages and Single Proportions

Notation

The country, population, or "universe" under investigation consists of N districts.

The population, area, or other fundamental quantity in any, the sth, district is a_s units, that of the universe A units; so that

$$A = \sum_1^N a_s. \quad \text{.....(1)}$$

It is required to find P, the proportion of the A units that have a certain attribute, or X the average of some variable connected with every unit.

If p_s is the proportion, or x_s the average, in the sth district,

$$AP = \sum_1^N a_s p_s, \text{ or } AX = \sum_1^N a_s x_s. \quad \text{.....(2)}$$

Regard the N values of the p's, or of the x's, as frequency groups, whose (unweighted) means are $\bar{p}$ or $\bar{x}$, and standard deviations are σ_p or σ_x.

Let there be one or more allied measurements, whose magnitude is already known in every district. In the sth district write $u_s, v_s, w_s \ldots$ for the magnitudes of these "controls," and let their magnitudes in the universe be $U, V, W \ldots$; so that

$$AU = \sum_1^N a_s u_s,\ AV = \sum_1^N a_s v_s,\ AW = \sum_1^N a_s w_s \ldots. \quad \text{.....(3)}$$

Regard the N values of the u's, v's, w's ... as frequency groups whose (unweighted) averages are $\bar{u}, \bar{v}, \bar{w} \ldots$ and standard deviations $\sigma_u, \sigma_v, \sigma_w \ldots$.

Write $r_{1u}, r_{1v}, r_{1w} \ldots$ for the correlation coefficients between x (or p) and $u, v, w \ldots$, and $\rho_{uv}, \rho_{uw}, \rho_{vw} \ldots$ for those between $u, v, w \ldots$, so that

$$r_{1u} = \frac{\text{Mean } (p_s - \bar{p})\,(u_s - \bar{u})}{\sigma_p \sigma_u} \text{ or } \frac{\text{Mean } (x_s - \bar{x})\,(u_s - \bar{u})}{\sigma_x \sigma_u} \ldots,$$

$$\rho_{uv} = \frac{\text{Mean } (u_s - \bar{u})\,(v_s - \bar{v})}{\sigma_u \sigma_v} \ldots. \quad \text{.....(4)}$$

Assume that the partial regression equation connecting p (or x) with $u, v, w \ldots$ is rectilinear with sufficient approximation, and write it in the form

$$p - \bar{p} \text{ (or } x - \bar{x}) = G_u (u - \bar{u}) + G_v (v - \bar{v}) + G_w (w - \bar{w}) + \ldots. \quad \text{...(5)}$$

The values of $G_u, G_v, G_w \ldots$ are known in terms of the standard

deviations and correlation coefficients (see pp. 50–1 below). E.g. if U is the only control, $G_u = r_{1u} \cdot \frac{\sigma_p}{\sigma_u}$.

Write e_s for the error resulting from calculating p_s or x_s from the regression equation, so that

$$e_s = p_s - \bar{p} \text{ (or } x_s - \bar{x}) - G_u (u_s - \bar{u}) - G_v (v_s - \bar{v}) - G_w (w_s - \bar{w}) \ldots . \quad (6)$$

A number, n, of districts is selected in such a way that the average for each control is the same in the aggregate of them as in the universe, so that

$$U \cdot \sum_1^n a_s = \sum_1^n a_s u_s, \quad V \cdot \sum_1^n a_s = \sum_1^n a_s v_s, \quad W \cdot \sum_1^n a_s = \sum_1^n a_s w_s \ldots . \quad \ldots\ldots(7)$$

Write P_n or X_n for the value of the unknown as computed from the selected districts (P or X being the true value), so that

$$P_n = \frac{\sum_1^n a_s p_s}{\sum_1^n a_s} \text{ or } X_n = \frac{\sum_1^n a_s x_s}{\sum_1^n a_s} . \quad \ldots\ldots\ldots\ldots\ldots\ldots(8)$$

The problem is to measure the precision of P_n or of X_n.

The subsequent analysis is written in terms of X; for P it would be exactly similar.

We have from (8) and (6)

$$X_n = \frac{1}{\sum_1^n a_s} \cdot \sum_1^n a_s \{e_s + \bar{x} + G_u (u_s - \bar{u}) + G_v (v_s - \bar{v}) + G_w (w_s - \bar{w}) + \ldots\}$$

$$= \frac{\sum a_s e_s}{\sum a_s} + \bar{x} + G_u (U - \bar{u}) + G_v (V - \bar{v}) + G_w (W - \bar{w}) + \ldots \text{ from (7).}$$

Therefore $$X = X_n - K - \frac{\sum_1^n a_s e_s}{\sum_1^n a_s}, \quad \ldots\ldots\ldots\ldots\ldots\ldots(9)$$

where $$K = -(X - \bar{x}) + G_u (U - \bar{u}) + G_v (V - \bar{v}) + G_w (W - \bar{w}) + \ldots$$
$$\ldots\ldots\ldots\ldots\ldots\ldots(10)$$

K, which depends on the differences between the weighted and the unweighted averages of the quantities, is small unless there is considerable correlation between the sizes of the districts and the variables; in any case all the terms can either be computed exactly from the data, or (if they involve x) given approximate values from the sample. K appears as a slight, but probably perceptible, correction to the value of X_n found in the sample.

We have now to determine the nature of the error $\frac{\sum_1^n a_s e_s}{\sum_1^n a_s}$.

Write $n\bar{a} = \sum_1^n a_s$, and $n\sigma_a^2 = \sum_1^n (a_s - \bar{a})^2 = \sum_1^n a_s^2 - n\bar{a}^2$. $\ldots\ldots(11)$

Then $$n\bar{a}\,(X_n - K - X) = \sum_1^n a_s e_s.$$

Let σ_e be the standard deviation of e_s, and equally of $e_1, e_2 \ldots e_n$, it being assumed that these are uncorrelated.

Write σ_n for the standard deviation of the error made in writing $X_n - K$ instead of X.

Then $$n^2\bar{a}^2\sigma_n^2 = \sum_1^n a_s^2 . \sigma_e^2 = \text{(from (11))}\ n\sigma_e^2\,(\bar{a}^2 + \sigma_a^2).$$

Therefore $$\sigma_n^2 = \sigma_e^2 . \frac{1}{n}\left(1 + \frac{\sigma_a^2}{\bar{a}^2}\right). \quad \ldots\ldots\ldots\ldots\ldots\ldots(12)$$

Now it is known* that $\sigma_e^2 = \dfrac{R}{R_t} . \sigma_x^2$, where t is the number of controls, and

$$R = \begin{vmatrix} 1 & r_{1u} & r_{1v} & r_{1w} \ldots \\ r_{1u} & 1 & \rho_{uv} & \rho_{uw} \ldots \\ r_{1v} & \rho_{uv} & 1 & \rho_{vw} \ldots \\ r_{1w} & \rho_{uw} & \rho_{vw} & 1 \ \ldots \\ \vdots & \vdots & \vdots & \vdots \ \ldots \end{vmatrix}, \text{ and } R_t = \begin{vmatrix} 1 & \rho_{uv} & \rho_{uw} \ldots \\ \rho_{uv} & 1 & \rho_{vw} \ldots \\ \rho_{uw} & \rho_{vw} & 1 \ \ldots \\ \vdots & \vdots & \vdots \ \ldots \end{vmatrix},$$

where the r's must be estimated from the sample, while the ρ's can be calculated from the universe.

Also σ_x must be estimated from the sample, and is subject to a standard deviation $\dfrac{\sigma_x}{\sqrt{2n}}$.

Hence finally from (12) we find that the standard deviation of the error in estimating X (or P) from $X_n - K$ (or $P_n - K$) is

$$\frac{\sigma_x}{\sqrt{n}} \cdot \sqrt{\left(1 + \frac{\sigma_a^2}{\bar{a}^2}\right)} \cdot \sqrt{\frac{R}{R_t}} \text{ or } \frac{\sigma_p}{\sqrt{n}} \cdot \sqrt{\left(1 + \frac{\sigma_a^2}{\bar{a}^2}\right)} \cdot \sqrt{\frac{R}{R_t}}. \quad (13)$$

The advantage obtained by the use of the controls depends solely on the value of $\sqrt{\dfrac{R}{R_t}}$, being greatest when this is least.

If $t = 0$, $\dfrac{R}{R_t} = 1$.

If $t = 1$, $\dfrac{R}{R_t} = 1 - r_{1u}^2$.

If $t = 2$, $\dfrac{R}{R_t} = \dfrac{1 + 2r_{1u}r_{1v}\rho_{uv} - r_{1u}^2 - r_{1v}^2 - \rho_{uv}^2}{1 - \rho_{uv}^2}$.

We may examine these values further by considering special cases.

Suppose $r_{1u} = r_{1v} = r_{1w} = \ldots = r$, and $\rho_{uv} = \rho_{uw} = \rho_{vw} = \ldots = \rho$.

Then it can be shown, from the elementary properties of determinants, that $R_t = \{\rho\,(t-1) + 1\}\,(1-\rho)^{t-1}$ and $R = R_t - tr^2\,(1-\rho)^{t-1}$, so that

$$\frac{R}{R_t} = 1 - \frac{tr^2}{(t-1)\,\rho + 1}.$$

* See pp. 50–1 below.

For a given r the expression is greatest when ρ is least. If in fact there is no correlation between the controls and therefore ρ is zero, we have $\frac{R}{R_t} = 1 - tr^2$, or more generally $\frac{R}{R_t} = 1 - r_{1u}^2 - r_{1v}^2 - r_{1w}^2 - \ldots$.

If two of the controls are perfectly correlated, they have the same effect as if one were omitted. If they were all perfectly correlated, we should have $\frac{R}{R_t} = 1 - r^2$, as in the case of one control.

In the same case of equal r's and equal ρ's, $\frac{R}{R_t}$ diminishes as t increases, as is *à priori* obvious, but does not become less than $1 - \frac{r^2}{\rho}$.

If r_0 is the greatest of the r's, and ρ_0 the least of the ρ's, then $\frac{R}{R_t}$ is greater than $1 - \frac{r_0^2}{\rho_0}$.

Thus the advantage of increasing the number of controls is in ordinary cases quite small. E.g. if $r = \frac{2}{3}$, $\rho = \frac{1}{2}$, which are values that might easily arise in fact, then $\sqrt{\frac{R}{R_t}}$ for $t = 1, 2, 3, \ldots 10 \ldots \infty$ is ·745, ·638, ·577 ... ·438 ... ·333.

In fact, however, there are necessary relationships between the r's and the ρ's, e.g. if $\rho = 0$, $1 \nless tr^2$, and probably even so rapid a fall would not be obtained.

It is clear that the standard deviation of the error of the result is in ordinary cases dominated by the value of σ_x (or σ_y) and by n the number of observations, rather than by the controls exercised in purposive selection.

Note on the regression formula

Write $z, y_1, y_2 \ldots$ for $x - \bar{x}, u - \bar{u}, v - \bar{v} \ldots$.

Let the regression equation be

$$z = b_1 y_1 + b_2 y_2 + \ldots,$$

it being assumed that it should be satisfied by $z = 0 = y_1 = y_2 = \ldots$.

Obtain values of $b_1, b_2 \ldots$ by minimizing

$$f = \Sigma\,(-z + b_1 y_1 + b_2 y_2 + \ldots)^2.$$

Then $$0 = \tfrac{1}{2}\frac{\partial f}{\partial b_1} = N\,(-r_{1z}\sigma_1\sigma_z + b_1\sigma_1^2 + b_2 r_{12}\sigma_1\sigma_2 + \ldots).$$

Therefore $$-r_{1z}\sigma_z + b_1\sigma_1 + b_2\sigma_2 r_{12} + \ldots = 0.$$

Similarly $$-r_{2z}\sigma_z + b_1\sigma_1 r_{12} + b_2\sigma_2 + \ldots = 0,$$

$$-r_{3z}\sigma_z + b_1\sigma_1 r_{13} + b_2\sigma_2 r_{23} + \ldots = 0.$$

Also $$f = N\,(\sigma_z^2 - b_1\sigma_1\sigma_z r_{1z} - \ldots) + \frac{N}{2}\cdot\frac{\partial f}{\partial b_1} + \ldots.$$

Therefore $$\sigma_z - b_1\sigma_1 r_{1z} - b_2\sigma_2 r_{2z} - \ldots = \frac{f}{N\sigma_z}.$$

These equations are all satisfied by

$$\frac{-\sigma_z}{R_{11}} = \frac{b_1\sigma_1}{R_{12}} = \frac{b_2\sigma_2}{R_{13}} = \ldots, \text{ and } f = N\sigma_z^2 \cdot \frac{R}{R_{11}},$$

where $R = \begin{vmatrix} 1 & r_{1z} & r_{2z} & r_{3z} \ldots \\ r_{1z} & 1 & r_{12} & r_{13} \ldots \\ r_{2z} & r_{12} & 1 & r_{23} \ldots \\ r_{3z} & r_{13} & r_{23} & 1 \ldots \\ \vdots & \vdots & \vdots & \vdots \ldots \end{vmatrix}$ and R_{1k} is the co-factor obtained by striking out the first row and the kth column for $k = 1, 2, 3 \ldots$.

Hence the mean square error $= \frac{f}{N} = \sigma_z^2 \cdot \frac{R}{R_t}$ in the notation above (p. 49), where

$$r_{1z} = r_{1u}, r_{2z} = r_{1v} \ldots, r_{12} = r_{uv}, r_{13} = r_{uw} \ldots.$$

The effect of successive controls may be seen from the following equation (based on Yule, *Introduction to Theory of Statistics*, ed. 1922, p. 237 (10)):

$$\frac{R}{R_t} = \sqrt{(1 - r_{1u}^2)} \cdot \sqrt{(1 - r^2_{1v \cdot c})} \cdot \sqrt{(1 - r^2_{1w \cdot c})} \ldots \text{ to } t \text{ factors},$$

where $r_{1v \cdot c}$ is the correlation coefficient (in Mr Yule's terminology) between x and v when u is constant and $w \ldots$ are not introduced; $r_{1w \cdot c}$ is that between x and w when u and v are constant and terms subsequent to w are not introduced; and so on.

Two experiments may be given to illustrate the relative importance of the quantities involved.

1. *Required* the number of males occupied in transport by road, in England and Wales (excluding London), in 1911.

Control. Proportion of rural to whole population.

Districts. The 61 Administrative Counties (including County Boroughs).

The counties selected were Kent, Bucks, Norfolk, Peterborough, Lincoln (Lindsey), Derby, Yorks (West Riding), Stafford, Warwick, Worcester, Brecknock, and Flint; 12 in all. Except that the proportion, rural to all, is approximately the same (·245) in the aggregate as in all England and Wales (excluding London), viz. ·250, the selection followed no rule.

$U = \cdot 250$. $n = 12$.

$\bar{a}$, average number of persons in the 12 counties, $= 748{,}000$.

$\sigma_a = 803{,}000$.

P the proportion of all persons occupied in transport by road in England and Wales (excluding London).

P_n the proportion of all persons occupied in transport by road found in the 12 counties.

$r_{1u} = r_{up}$, computed from the selection, $= -\cdot47$.
σ_p, computed from the selection, $= \cdot0028$.
The quantity K (p. 48 (10)) $= -\cdot0007$ approx.
$P_n = \cdot0110$.
Forecast: $P_n - K = \cdot0117$.

$$\sigma_n = \sigma_p \times \frac{1}{\sqrt{n}} \times \sqrt{\left(1 + \frac{\sigma_a^2}{\bar{a}^2}\right)} \times \sqrt{1 - r_{up}^2}$$
$$= \cdot0028 \times \cdot288 \times 1\cdot44 \times \cdot883 = \cdot00116 \times \cdot883 = \cdot00102.$$

Forecast: $\cdot0117 \pm \cdot0010$.
Fact, computed from the census for the whole country: $\cdot0115$.

In the same case 12 counties were taken purely at random, and the forecast was $\cdot0097 \pm \cdot0010$.

Again 12 counties were taken in a certain geographical order, and the forecast was $\cdot0105 \pm \cdot0010$.

2. From an official report the wage-rates in 1912 of compositors, masons, and engineering labourers were extracted for 47 towns, and the problem was set to find the average wage-rates of ironmoulders in these towns by purposive selection of 12.

Here no weights were used, so that $1 = a_1 = a_2 = \ldots$; $\sigma_a = 0$. $n = 12$.

The towns were selected so that the averages of the three occupations used as controls were approximately the same in them as in the 47 towns together.

Compositors: $U = 33\cdot49$, $\sigma_u = 2\cdot326$, $\bar{u} = 33\cdot50$. Shillings per week.
Masons: $V = 9\cdot120$, $\sigma_v = \cdot529$, $\bar{v} = 9\cdot2$. Pence per hour.
Labourers: $W = 20\cdot33$, $\sigma_w = 1\cdot464$, $\bar{w} = 20\cdot33$. Shillings per week.
Ironmoulders: $\bar{x} = 38\cdot17$, $\sigma_x = 2\cdot01$. Shillings per week.
$\rho_{uv} = \cdot712$, $\rho_{uw} = \cdot176$, $\rho_{vw} = \cdot477$.
$r_{xu} = \cdot54$, $r_{xv} = \cdot38$, $r_{xw} = \cdot002$.
For all three controls, $R = \cdot247$, $R_3 = \cdot354$. $\sqrt{(R \div R_3)} = \cdot835$.
For u and v only, $R = \cdot349$, $R_2 = \cdot493$. $\sqrt{(R \div R_2)} = \cdot841$.
For u only, $R = \cdot7084$, $R_1 = 1$. $\sqrt{(R \div R_1)} = \cdot842$.

Therefore $\sigma_n = \sigma_e \times \frac{1}{\sqrt{12}} = \frac{1}{\sqrt{12}} \sigma_x \times \cdot835 = \frac{1}{\sqrt{12}} \times 2\cdot01 \times \cdot835 = \cdot484$,

or $\cdot488$ for u only.
$K = -\cdot0316$.
Forecast: $\bar{x} - K \pm \sigma_n = 38\cdot21 \pm \cdot484$.
Fact: $39\cdot12$.

The difference between the forecast and the fact is 1·9 times the standard deviation of the error, which is greater than would be anticipated. But not much dependence can be placed on a sample based on only 12 districts, since the errors of the terms involving x are considerable.

Stratification

It is not easy to distinguish the advantages of the method of stratification, in which the universe is regarded as consisting of divisions, in the districts within each of which the variable in question is confined within a narrow grade and where one district is selected from each division, from the general method of purposive selection. But an analysis of a simple case will throw some light on the question.

Suppose that we are investigating the average value of a quantity X by the help of a control U. Let there be $N = k \times \nu$ districts with equal populations in the country. Arrange them in order of ascending magnitude of U in k equal divisions, the resulting quantities being as follows, where the values connected with U are known and those connected with X are unknown:

	1st Division		2nd Division		.....	kth Division	
	U	X	U	X		U	X
Averages	$\bar{u}+d_1$	$\bar{x}+\delta_1$	$\bar{u}+d_2$	$\bar{x}+\delta_2$	.....	$\bar{u}+d_k$	$\bar{x}+\delta_k$
District values	$\bar{u}+d_1+{}_1u_1$	$\bar{x}+\delta_1+{}_1x_1$	$\bar{u}+d_2+{}_2u_1$	$\bar{x}+\delta_2+{}_2x_1$	.....	$\bar{u}+d_k+{}_ku_1$	$\bar{x}+\delta_k+{}_kx_1$
	$\bar{u}+d_1+{}_1u_2$	$\bar{x}+\delta_1+{}_1x_2$	$\bar{u}+d_2+{}_2u_2$	$\bar{x}+\delta_2+{}_2x_2$	.....	$\bar{u}+d_k+{}_ku_2$	$\bar{x}+\delta_k+{}_kx_2$
	⋮	⋮	⋯	⋮		⋮	⋮
	$\bar{u}+d_1+{}_1u_\nu$	$\bar{x}+\delta_1+{}_1x_\nu$	$\bar{u}+d_2+{}_2u_\nu$	$\bar{x}+\delta_2+{}_2x_\nu$	.....	$\bar{u}+d_k+{}_ku_\nu$	$\bar{x}+\delta_k+{}_kx_\nu$
Standard deviations	${}_1\sigma_u$	${}_1\sigma_x$	${}_2\sigma_u$	${}_2\sigma_x$	.....	${}_k\sigma_u$	${}_k\sigma_x$
Correlation coefficient	r_1		r_2		.....	r_k	

General averages, $\bar{u}$ and $\bar{x}$; so that

$$\sum_{t=1}^{t=k} d_t = 0 = \sum_{t=1}^{t=k} \delta_t;\ \sum_{s=1}^{s=\nu} {}_tu_s = 0 = \sum_{s=1}^{s=\nu} {}_tx_s, \text{ for } t = 1, 2 \ldots k.$$

General standard deviations and correlation coefficient: σ_u, σ_x, r_{ux}.

Then $$N\sigma_u{}^2 = \sum_{t=1}^{t=k}\sum_{s=1}^{s=\nu} (d_t + {}_su_t)^2 = \sum_{t=1}^{t=k} ({}_t\sigma_u{}^2 + d_t{}^2),$$

and $\sigma_u{}^2 = \sigma_u{}'^2 + \sigma_d{}^2$, where $k\sigma_u{}'^2 = {}_1\sigma_u{}^2 + {}_2\sigma_u{}^2 + \ldots + {}_k\sigma_u{}^2$,

$$k\sigma_d{}^2 = d_1{}^2 + d_2{}^2 + \ldots + d_k{}^2.$$

Similarly $\sigma_x{}^2 = \sigma_x{}'^2 + \sigma_\delta{}^2$, where $k\sigma_x{}'^2 = {}_1\sigma_x{}^2 + {}_2\sigma_x{}^2 + \ldots + {}_k\sigma_x{}^2$,

$$k\sigma_\delta{}^2 = \delta_1{}^2 + \delta_2{}^2 + \ldots + \delta_k{}^2.$$

Also $$Nr_{ux}\sigma_u\sigma_x = \sum_{t=1}^{t=k}\sum_{s=1}^{s=\nu} (d_t + {}_su_t)(\delta_t + {}_sx_t) = \sum_{t=1}^{t=k} \nu\,(d_t\delta_t + r_t \cdot {}_t\sigma_u \cdot {}_t\sigma_x),$$

$$r_{ux}\sigma_u\sigma_x = \frac{1}{k}\Sigma d_t\delta_t + \frac{1}{k}\Sigma r_t \cdot {}_t\sigma_u \cdot {}_t\sigma_x.$$

In the special case where the divisions are similar, and

$${}_1\sigma_u = {}_2\sigma_u = \ldots = {}_k\sigma_u = \sigma_u',\ {}_1\sigma_x = {}_2\sigma_x = \ldots = {}_k\sigma_x = \sigma_x',$$

and $$r_1 = r_2 = \ldots = r_k = r',$$

we should have $$r_{ux}\sigma_u\sigma_x = r_{d\delta} \cdot \sigma_d\sigma_\delta + r'\sigma_u'\sigma_x'. \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\text{(A)}$$

This equation shows that, other things being unchanged, the smaller is σ_u', the larger is r'; that is when the divisions are nearly homogeneous within themselves in respect to U, the greater is the correlation between

U and X within the divisions. But it is clear from the equation that the relation between the general correlation of U and X in the country and the correlations within the divisions is complex, depending *inter alia* on the correlation of the division averages.

Now select one district from each division, choosing (as is generally possible) one whose U is practically equal to the average for the division. In the selected districts measure the X.

Write the values found $\bar{u} + d_1, \bar{x} + \delta_1 + {}_1x_m$,
$\bar{u} + d_2, \bar{x} + \delta_2 + {}_2x_m$,
$\vdots$
$\bar{u} + d_k, \bar{x} + \delta_k + {}_kx_m$.

Let $\bar{x}_k$ be the resulting estimate for X. Then

$$\bar{x}_k = \frac{1}{k}\sum_{t=1}^{t=k}(\bar{x} + \delta_t + {}_tx_m) = \bar{x} + \frac{1}{k}\Sigma_t x_m, \text{ since } \Sigma\delta_t = 0.$$

Now the value of X expected from the correlation in the tth district is

$$\bar{x} + \delta_t + r_t\frac{{}_t\sigma_x}{{}_t\sigma_u} \times 0 = \bar{x} + \delta_t,$$

so that ${}_tx_m$ is the difference between expectation and fact. The standard deviation of ${}_tx_m$ is therefore ${}_t\sigma_x \cdot \sqrt{1 - r_t^2}$.

Hence the standard deviation of the error in taking $\bar{x}_k$ for $\bar{x}$ is that derived from such standard deviations as ${}_t\sigma_x \cdot \sqrt{1 - r_t^2}$, and equals

$$\sqrt{\left\{\frac{1}{k^2}\sum_{t=1}^{t=k}{}_t\sigma_x^2\,(1 - r_t^2)\right\}} = \sigma_e, \text{ say.}$$

To reduce this further make the assumptions that led to equation (A). In this case

$$\sigma_e^2 = (1 - r'^2)\cdot\frac{1}{k}\sigma_x'^2 = \frac{\sigma_x^2}{k}(1 - r'^2)\left(1 - \frac{\sigma_\delta^2}{\sigma_x^2}\right). \quad \text{.....(B)}$$

We can compare this result with that from unstratified control as follows.

Write $$\delta_t = r_{ux}\cdot\frac{\sigma_x}{\sigma_u}\cdot d_t + h_t,$$

where h_t is the error in estimating the X average of the tth division from the regression equation connecting X with U.

Write $$k\sigma_h^2 = \sum_1^k h_t^2,$$

and since no significant correlation is to be expected between h and d, take *mean* $h_t d_t = 0$.

Then squaring the equation for δ_t, we have

$$\sigma_\delta^2 = r_{ux}^2\frac{\sigma_x^2}{\sigma_u^2}\sigma_d^2 + \sigma_h^2;$$

and multiplying the same equation by d_t, we obtain

$$r_{d\delta}\sigma_d\sigma_\delta = r_{ux}\cdot\frac{\sigma_x}{\sigma_u}\cdot\sigma_d^2.$$

Eliminate $r_{d\delta}$ with the help of equation (A); then

$$r'\sigma_u'\sigma_x' = r_{ux}\frac{\sigma_x}{\sigma_u}(\sigma_u^2 - \sigma_d^2) = r_{ux}\frac{\sigma_x}{\sigma_u}\sigma_u'^2.$$

Therefore $$r'^2\sigma_x'^2 = r_{ux}^2 \cdot \frac{\sigma_x^2}{\sigma_u^2} \cdot \sigma_u'^2.$$

From (B),

$$\sigma_e^2 = \frac{1}{k}\{\sigma_x'^2 - r'^2\sigma_x'^2\} = \frac{1}{k}\left\{\sigma_x^2 - \sigma_\delta^2 - r_{ux}^2 \cdot \frac{\sigma_x^2}{\sigma_u^2}\sigma_u'^2\right\}$$

$$= \frac{1}{k}\left\{\sigma_x^2 - r_{ux}^2 \cdot \frac{\sigma_x^2}{\sigma_u^2}\sigma_d^2 - \sigma_h^2 - r_{ux}^2 \cdot \frac{\sigma_x^2}{\sigma_u^2}(\sigma_u^2 - \sigma_d^2)\right\}$$

$$= \frac{\sigma_x^2}{k}\left(1 - r_{ux}^2 - \frac{\sigma_h^2}{\sigma_x^2}\right).$$

Whereas, if we took the control without stratification, we should have

$$\sigma_e^2 = \frac{\sigma_x^2}{k}(1 - r_{ux}^2).$$

Now there is no reason to expect any large value of $\frac{\sigma_h^2}{\sigma_x^2}$, which may be written $\eta^2 - r_{ux}^2$, where η is akin to the "correlation ratio," when each district is regarded as an array; so that

$$\sigma_e^2 = \frac{\sigma_x^2}{k}(1 - \eta^2).$$

The advantage obtained by stratification therefore, though it exists, may be expected to be slight. It depends on the non-rectilinearity of the regression between the control and the quantity sought in the divisions.

[*Note.* If in each division all the U's were the same, every r_t would be zero, and we should have exactly the correlation ratio in the usual sense; and in fact in this case the result is obtainable immediately.]

This method was tried on the statistics of wages discussed above. For the control 45 towns were arranged in order of compositors' wages and the list was divided into 9 groups from top to bottom each containing 5 towns. $N = 45$, $k = 9$, $\nu = 5$. In each group a town was selected in which the wage was approximately the average in that group, and the ironmoulders' wage was written down for that town. The average of these wages was 39·85*s*. In all the towns together the average was 39·1*s*. In the former method when 12 towns were taken the average was 38·2*s*. Thus the new method based on only 5 towns gives a somewhat closer result.

It was found that r' was approximately zero and $\sigma_\delta = 1{\cdot}3$. σ_x, not obtainable from the sample, was in fact 2·01. Thus

$$\sigma_e = \frac{2{\cdot}01}{\sqrt{\cdot 5}} \cdot \sqrt{1 - 0} \cdot \sqrt{1 - \left(\frac{1{\cdot}3}{2{\cdot}01}\right)^2} = {\cdot}68,$$

which is just less than the difference between the true average (39·1) and the sample average (39·85).

There is an evident difficulty in computing σ_e, in that we have no means of determining σ_x, unless we take also a random sample for that purpose.

2. Distribution in Grades

The problem being to assign the proportions in various age-groups, in grades of income or in some other classification, districts are selected which each satisfy certain controlling conditions, and the proportions found in their aggregate form the required estimate.

As regards the proportion in any one grade by itself, the problem is that discussed earlier in this section. The new circumstance is that we are considering several grades together. There is the governing condition that the aggregate of the proportions expressed as percentages is 100, and there may be other controls, e.g. that the average income is known. Unless the number of grades is quite small, it is shown below that such conditions add little to the accuracy. A more important consideration may be that there is an approximation to a law of distribution, such as Pareto's income formula, or other correlations between the proportions.

The additional security obtained by increasing the number of districts, by stratification, and by correlated controls, is similar in kind and extent to that already discussed for single averages or proportions.

We will use the following notation; and for simplicity take the case when the districts have equal populations.

	Proportions				
Grades	In whole country	In N districts			
1	p_1	$p_1 + {}_1x_1$	$p_1 + {}_2x_1$	...	$p_1 + {}_Nx_1$
2	p_2	$p_2 + {}_1x_2$	$p_2 + {}_2x_2$	...	$p_2 + {}_Nx_2$
...	...	...	...	...	...
m	p_m	$p_m + {}_1x_m$	$p_m + {}_2x_m$	...	$p_m + {}_Nx_m$

Then

$$\sum_{t=1}^{t=m} p_t = 1; \quad \sum_{s=1}^{s=N} {}_sx_t = 0, \text{ for } t = 1, 2 \ldots m, \text{ and } \sum_{t=1}^{t=m} {}_sx_t = 0, \text{ for } s = 1, 2 \ldots N.$$

Write $$N\sigma_t^2 = \sum_{s=1}^{s=N} {}_sx_t^2, \text{ for } t = 1, 2 \ldots m.$$

Without controls, for a single grade the standard deviation for p_t as estimated from the average of N districts taken at random is $\frac{\sigma_t}{\sqrt{N}}$.

Suppose that ${}_1x_t, {}_2x_t \ldots {}_Nx_t$ are normally distributed for each value of t.

Write $$\chi^2 = \frac{{}_1x_1^2}{\sigma_1^2} + \frac{{}_1x_2^2}{\sigma_2^2} + \ldots + \frac{{}_1x_m^2}{\sigma_m^2}.$$

Then, if ${}_1x_1, {}_1x_2 \ldots$ were completely independent of each other, the chance that they would be found in a single district would be $Ce^{-\frac{1}{2}\chi^2}$, where C is a constant.

Any given value of χ^2 corresponds to a complex of errors, and may be taken as a measure of that complex (cf. p. 39 above, where a different form is used for the complex).

The chance that so great a value as ${\chi_1}^2$ will be found is

$$\int_{\chi_1}^{\infty} e^{-\frac{1}{2}\chi^2} \cdot \chi^{m-1} \cdot d\chi \div \int_0^{\infty} e^{-\frac{1}{2}\chi^2} \cdot \chi^{m-1} \cdot d\chi = P_1{}^*.$$

The necessary linear relation between the x's, viz.

$$\sum_{t=1}^{t=m} {}_1x_t = 0,$$

allows the elimination of one x, reduces χ^2 to a quadratic expression in one variable less, and reduces the index of the expression for P_1 by unity, so that we may begin by writing the index as $m - 2$.

If the average in the district is known to equal the average in the country, we have a further relation such as

$$a_1 \cdot {}_1x_1 + a_2 \cdot {}_1x_2 + \ldots = 0,$$

where $a_1, a_2 \ldots$ are the scale readings of the centres or averages of the grades. The index is then to be taken as $m - 3$. The table annexed indicates how far such a relation reduces the chance of extreme deviations. Thus if $m = 10$ and $\chi^2 = 8$, if the average were not known $P = \cdot534$, i.e. the chance of obtaining so great a χ^2 is $\cdot534$. If the average were known we should regard m as 9, and the chance is only $\cdot433$.

This would be the case if we depended on only one district. If we merged n districts, as a first and rough approximation we could write

$$\chi^2 = \frac{({}_1x_1 + {}_2x_1 + \ldots + {}_nx_1)^2}{n{\sigma_1}^2} + \frac{({}_1x_2 + {}_2x_2 + \ldots {}_nx_2)^2}{n{\sigma_2}^2} + \ldots.$$

[The mathematics of this have not been verified.]

Whether this is the form or not, we have the combined effect of the increased precision that arises from averaging, and of the virtual reduction of the number of grades that comes from the controls.

Values of $P = \int_{\chi}^{\infty} e^{-\frac{1}{2}\chi^2} \cdot \chi^{m-2} \cdot d\chi \div \int_0^{\infty} e^{-\frac{1}{2}\chi^2} \cdot \chi^{m-2} \cdot d\chi.$

	m								
χ^2	2	3	4	5	6	7	8	9	10
1	·317	·607	·801	·910	·963	·986	·995	·998	·999
2	·157	·368	·572	·736	·849	·920	·960	·981	·991
3	·083	·223	·392	·558	·700	·809	·885	·934	·964
4	·046	·135	·261	·406	·549	·677	·780	·857	·911
5	·025	·082	·172	·287	·416	·544	·660	·758	·834
6	·014	·050	·112	·199	·306	·423	·540	·647	·740
7	·008	·030	·072	·136	·221	·321	·429	·537	·637
8	·005	·018	·046	·092	·156	·238	·333	·433	·534
9	·003	·011	·029	·061	·109	·174	·253	·342	·437
10	·002	·007	·019	·040	·075	·125	·189	·265	·350

* Obtained for a different purpose by Professor Karl Pearson (*Philosophical Magazine*, 1900, vol. I, pp. 157 *et seq.*). The index is considered as giving an m-dimensional ellipsoid, and P_1 is the ratio of the volume of $z = Ce^{-\frac{1}{2}\chi^2}$ from χ_1 to ∞ to the volume from 0 to ∞. See Bowley, *Elements of Statistics*, 4th ed. pp. 426 *et seq.*

A different method of approaching the problem is to consider the reduction of the standard deviation in one grade only, owing to the fact that

$$\sum_{t=1}^{t=m} {}_1x_t = 0.$$

Take only one district, and suppress the prefix.

The chance of finding x_1, x_2 ..., subject only to the condition

$$x_1 + x_2 + \ldots + x_m = 0$$

and assuming that there are no other correlations, is

$$Ce^{-\frac{1}{2}\left(\frac{x_1^2}{\sigma_1^2}+\frac{x_2^2}{\sigma_2^2}+\ldots+\frac{x^2_{m-1}}{\sigma^2_{m-1}}+\frac{(x_1+x_2+\ldots+x_{m-1})^2}{\sigma_m^2}\right)}.$$

Take $x_1 \ldots x_{m-2}$ as fixed, and integrate between extreme limits for x_{m-1}. The result is the chance of finding x_1, $x_2 \ldots x_{m-2}$ whatever the value of x_{m-1}. Similarly integrate away x_{m-2}, $x_{m-3} \ldots x_2$ in succession.

The whole chance of finding x_1, whatever the other values, is

$$Ce^{-\frac{1}{2}\cdot\frac{\Delta_m}{\Delta_{m-1}}\cdot x_1^2},$$

where $\Delta_m =$

$$\begin{vmatrix} \frac{1}{\sigma_1^2}+\frac{1}{\sigma_m^2} & \frac{1}{\sigma_m^2} & \frac{1}{\sigma_m^2} & \ldots \\ \frac{1}{\sigma_m^2} & \frac{1}{\sigma_2^2}+\frac{1}{\sigma_m^2} & \frac{1}{\sigma_m^2} & \ldots \\ \frac{1}{\sigma_m^2} & \frac{1}{\sigma_m^2} & \frac{1}{\sigma_3^2}+\frac{1}{\sigma_m^2} & \ldots \\ \vdots & \vdots & \vdots & \ldots \end{vmatrix} = \frac{1}{\sigma_m^{2m-2}} \begin{vmatrix} \frac{\sigma_m^2}{\sigma_1^2}+1 & 1 & 1 & \ldots \\ 1 & \frac{\sigma_m^2}{\sigma_2^2}+1 & 1 & \ldots \\ 1 & 1 & \frac{\sigma_m^2}{\sigma_3^2}+1 & \ldots \\ \vdots & \vdots & \vdots & \ldots \end{vmatrix}$$

$(\overline{m-1}$ rows and columns).

In the particular case when $\sigma_1 = \sigma_2 = \ldots = \sigma_m$, $\Delta_m = \frac{m}{\sigma_1^{2m-2}}$, and $\frac{\Delta_m}{\Delta_{m-1}} = \frac{m}{m-1}\cdot\frac{1}{\sigma_1^2}$.

The standard deviation is only reduced in the ratio $\sqrt{m} : \sqrt{\overline{m-1}}$, and instead of σ_1 is approximately $\sigma_1\left(1-\frac{1}{2m}\right)$.

If the average is also given we have

$$a_1x_1 + a_2x_2 + \ldots + a_mx_n = 0,$$

where a_1, a_2 ... are approximately known.

Take the simplified case of equal grades, and write

$$a_m - a_{m-1} = a_{m-1} - a_{m-2} = \ldots = a_2 - a_1 = g.$$

Then $$a_1x_1 + (a_1+g)\,x_2 + \ldots + (a_1+\overline{m-1}g)\,x_m = 0$$

and $$x_1 + x_2 + \ldots + x_m = 0.$$

Therefore $$x_m = (m-2)\,x_1 + (m-3)\,x_2 + \ldots + x_{m-2}$$

and $$-x_{m-1} = (m-1)\,x_1 + (m-2)\,x_2 + \ldots + 2x_{m-2}.$$

Further assume that $\sigma_1 = \sigma_2 = \ldots = \sigma_m$; then by eliminating x_m and x_{m-1} we find that the chance of obtaining $x_1, x_2 \ldots x_{m-2}$ is

$$Ce^{-\frac{1}{2\sigma_1^2}\{x_1^2+x_2^2+\ldots+x^2_{m-2}+(\overline{m-1}x_1+\overline{m-2}x_2+\ldots+2x_{m-2})^2+(\overline{m-2}x_1+\overline{m-3}x_2+\ldots+x_{m-2})^2\}}.$$

The result of integrating away $x_2, x_3 \ldots x_{m-2}$ is to give that the whole chance of obtaining x_1 is

$$Ce^{-\frac{1}{2\sigma_1^2}\cdot\frac{D_m}{D_{m-1}}\cdot x_1^2},$$

where (after a troublesome reduction) it is found that

$$D_m = \tfrac{1}{12}(m^4 - m^2),$$

$$\sqrt{\frac{D_{m-1}}{D_m}} = \sqrt{\frac{(m-1)(m-2)}{m(m+1)}} = \text{approx. } 1 - \frac{2}{m}.$$

Instead then of σ_1 for the standard deviation of x_1, we have approximately $\sigma_1\left(1 - \frac{2}{m}\right)$.

On the other hand by taking n districts independently without controls we have $\frac{\sigma_1}{\sqrt{n}}$.

It is evidently more important to increase the number of independent districts than to increase the number of controls, if there is no correlation or law connecting the proportions within each district.

Correlation between the proportions in the grades

It is doubtful whether such correlations are generally present in the problems to which this method is applied, beyond those arising from the fact that excess in one proportion must be balanced by defects in others, or from equality of averages, with which we have already dealt. In M. Jensen's problems (Méthods permettant de réaliser une économie de travail dans la statistique), for example:—In the division of land among certain crops we need not expect much correlation between the areas under wheat and barley. In the distribution by income there is no necessary correlation between the proportions in the grades, and similarly in the distribution by ages. But there may be such correlations. A deficit of wheat may be compensated by an excess of barley. In such a city as London we find a deficiency both of the very young and of the very old. The number of young children is presumably generally correlated with the number of potential mothers.

The maximum effect of correlation between one pair of grades is to reduce virtually the number of grades by 1, for with perfect correlation between the first pair if we were given ${}_1x_1$ we should know ${}_1x_2$. The effect of this can be studied from the table of values of P given above, p. 57.

The effect on χ^2 can be tested theoretically. Instead of writing

$$Ce^{-\frac{1}{2}\sum_{i=1}^{i=m}\frac{{}_1x_i^2}{\sigma_i^2}},$$

we should write

$$Ce^{-\frac{1}{2R}\left(\frac{{}_1x_1^2}{\sigma_1^2}\cdot R_{11}+\ldots+2\frac{{}_1x_1\cdot{}_1r_2}{\sigma_1\sigma_2}R_{12}+\ldots\right)},$$

where R, R_{11} ... are the determinants written on p. 51 above. Here, of course, ${}_1x_1+{}_1x_2+\ldots{}_1x_m=0$.

There is no evident simplification, and it is not even obviously necessary that a reduction of the aggregate of the errors, however measured, is probable if there is correlation. An accidental excess in one grade will (if the correlation is positive) be found with an excess in another grade, so that for example all the lower grades might show an excess and all the higher ones a defect; while if there was no correlation the distribution of excesses and defects would be sporadic, and by widening the grades we should increase the precision. Of course, if we had many districts we could test the correlation, but the process would be laborious.

We have thus two opposite tendencies from correlation between grades. Suppose for example that in an age distribution in 16 grades of five years each, there is strong correlation between the numbers in the groups 0 to 5 and 5 to 10 years, we tend to have 15 distinct grades only and so far a measurement of greater precision; but the errors in the remaining grades tend to be greater than if those in the first two could neutralize one another.

We may test this as follows. The chance of an aggregate of errors $x_1, x_2 \ldots x_m$ is $Ce^{-\frac{1}{2}\chi^2}$, where if there is no correlation

$$\chi_1^2=\frac{x_1^2}{\sigma_1^2}+\frac{x_2^2}{\sigma_2^2}+\frac{x_3^2}{\sigma_3^2}+\ldots+\frac{x_m^2}{\sigma_m^2},\text{ where } x_1+x_2+\ldots+x_m=0,$$

while if the first two (only) are correlated,

$$\chi_2^2=\frac{1}{1-r^2}\left(\frac{x_1^2}{\sigma_1^2}+\frac{x_2^2}{\sigma_2^2}-\frac{2rx_1x_2}{\sigma_1\sigma_2}\right)+\frac{x_3^2}{\sigma_3^2}+\ldots+\frac{x_m^2}{\sigma_m^2}.$$

For a given x_1, the value expected for x_2 is

$$r\frac{\sigma_2}{\sigma_1}x_1\pm\sigma_2\sqrt{1-r^2}.$$

Write

$$\frac{x_2}{\sigma_2}=r\frac{x_1}{\sigma_1}+\lambda\sqrt{1-r^2}.$$

Then

$$\chi_2^2-\chi_1^2=r^2\left\{\lambda-\frac{x_1}{r\sigma_1}(1+\sqrt{1-r^2})\right\}\left\{\lambda+\frac{x_2}{r\sigma_1}(1-\sqrt{1-r^2})\right\},$$

as may be shown after some reduction.

Here λ, x_1 and r may be positive or negative. λ and $\frac{x_1}{\sigma_1}$ are both likely to be between 2 and -2. λ and x_1 are equally likely to be of the same or of opposite signs.

If, for example,

$$\lambda = \frac{x_1}{\sigma_1} = 1 \text{ and } r = \tfrac{1}{2}, \quad \chi_2^2 - \chi_1^2 = -\frac{\sqrt{3}}{2},$$

the chance of so great a complex is increased.

Again, if

$$\lambda = \frac{x_1}{\sigma_1} = 1 \text{ and } r = -\tfrac{1}{2}, \quad \chi_2^2 - \chi_1^2 = \frac{\sqrt{3}}{2},$$

and the chance is diminished.

But if $-\lambda = \frac{x_1}{\sigma_1} = 1$, the results are interchanged.

Correlation then appears on the average to give no increase in precision.

Finally, if there is some law of distribution to which the observations in a district may be expected to approximate closely, such as the normal law of error, Makeham's law, or Pareto's law, the constants which determine such a formula may be determined with accuracy for each district. Suppose that two such constants are involved, we should get their values from two grades in each district, and from their variation in the group of districts selected estimate their value and precision for the country as a whole. This is an extreme case of correlation within the districts. The index in the formula for P would be considerably reduced, and the chance of finding any assigned complex of errors would also be reduced. Since, however, we cannot in general expect the existence of any such law, these are mainly theoretic considerations.

III. TEST BY SUB-SAMPLES, AND GENERAL CONTROLS

(RANDOM AND PURPOSIVE SELECTION)

If the number of persons, or the number of districts, in our aggregate selection is considerable, we can with a little care divide it into four sub-samples each satisfying the governing conditions. The consilience or difference of the results evidently affords some guidance as to their precision. But it is important to notice that if there is any concealed bias throughout the selection, the method of sub-division affords no help in detecting it. In the case of random selection, if P_1, P_2, P_3, P_4 are the proportions of persons possessing a certain attribute in four sub-samples, and if σ is the computed standard deviation of the error of P, the proportion in their aggregate, then $\sigma \times \frac{\sqrt{5}}{2}$ (or $\sigma \times \sqrt{1+\frac{1}{m}}$ if there are m sub-samples) affords a measure of the difference to be expected between P and P_1 (P_2, P_3 or P_4). If the P's are markedly closer together than this value, we may have overestimated σ, or they may all err in the same direction. If on the other hand the P's are more dispersed, we have underestimated σ either by ignorance of some factor involved or by breaking some rule in sampling.

We may be able to make general controls by calculating from our sample some quantity whose magnitude in the whole population is known (and one that has not already been used as a control if the selection is purposive). If this estimate differs from the known magnitude by more than twice (say) the standard deviation computed, then there is evidence that some rule has been broken, either by taking a biased sample, or by faulty definition, or from erroneous information. We can thus sometimes ascertain that our procedure has been faulty. But on the other hand if we get agreement between expectation and fact in respect of some known magnitude, it is still quite possible that there should be errors in information or in method of collection in respect of the quantities which cannot be verified.