CHAPTER IX

ANALYSIS OF STATISTICAL DATA

The general aim of the tests to be considered in this chapter is to examine by means of an analysis of observed data certain hypotheses regarding assumed theoretical probability distributions. (See also beginning of Section 3.1.)

A. Lexis Theory (Sections 1 and 2)

1. *Repeated Equal Alternatives*

1.1. *Groups of Bernoulli trials.* A very common kind of statistical analysis consists of listing the numbers of alternative events occurring in certain groups of observations. Take, for example, the table of deaths registered in the United States during n successive years. Each x_ν is the number of events in a group of m_ν trials, the single trial consisting of the alternative "death or survival" and m_ν being the number of people subject to the observation in the νth year. We may first assume that all m_ν are equal ($m_\nu = m$ in this case) and that there exists one unique value q for the probability of the event "death" in each alternative. We assume that for each $\nu = 1, 2, ..., n$ there exists a probability $p_\nu(x)$ of obtaining an x_ν-value $= x$. Under such conditions, all the distributions $p_\nu(x)$ are equal Bernoulli distributions with the probabilities (for integral x between 0 and m)

$$\binom{m}{x}(1 - q)^{m-x}q^x. \tag{1}$$

We wish to check these assumptions against observations.

The mean and the variance of the distribution (1) have been found to be

$$\alpha = mq, \qquad \sigma^2 = mq(1 - q). \tag{2}$$

If this is introduced into (15) and (18) in Chapter VIII, Section 3.1, we have with

$$a = \frac{1}{n}\sum_{\nu=1}^{n} x_\nu, \qquad s^2 = \frac{1}{n}\sum_{\nu=1}^{n} (x_\nu - a)^2, \tag{3}$$

$$E[a] = mq, \qquad E[s^2] = \frac{n-1}{n}\, mq(1-q). \tag{3'}$$

However, to compare these expectations with the observed values a and s^2, i.e., with the values computed from the $x_1, x_2, \ldots, x_n$ in the table, we would have to know the quantity q. As this is not the case, the following method, first introduced by the economist W. Lexis,[1] may be used to arrive at a conclusion.

The same statistical survey could also be registered in a different way, namely, by listing the nm successive outcomes of the single "trials." This would mean that nm data are given

$$x'_1, x'_2, \ldots, x'_{nm}, \tag{4}$$

each x'-value being either zero (survival) or one (death). The average a' of these observations is the total number of deaths divided by nm:

$$a' = \frac{1}{nm}(x_1' + x_2' + \cdots + x'_{nm}) = \frac{1}{nm}(x_1 + x_2 + \cdots + x_n) = \frac{a}{m}, \tag{4'}$$

and the dispersion s'^2 (since, for $x' = 0$ or 1, $x'^2 = x'$) is

$$s'^2 = \frac{1}{nm}\sum_1^{nm} (x_\nu'^2 - a'^2) = \frac{1}{nm}\sum_1^{nm} (x_\nu' - a'^2) = a'(1-a') = \frac{a}{m}\left(1 - \frac{a}{m}\right). \tag{4''}$$

Each number of the set (4) is the result of a single alternative with the event probability q. Therefore, mean and variance of the corresponding probability distribution are

$$\alpha' = q, \qquad \sigma'^2 = q(1-q). \tag{4'''}$$

Again, we can apply Eq. (18′), Chapter VIII to the set of data (4). The number n has to be replaced by nm and we obtain (see also Chapter VIII, Problem 10).

$$E[s'^2] = E\left[\frac{a}{m}\left(1 - \frac{a}{m}\right)\right] = \frac{nm-1}{nm}\, q(1-q). \tag{5}$$

[1] W. Lexis, *Zur Theorie der Massenerscheinungen in der menschlichen Gesellschaft.* Freiburg i.B., 1887.

The reader should note that in both equations, (3) and (5), E may be taken with respect to the distribution (1). From (3) and (5), the unknown quantity $q(1 - q)$ can be eliminated and this leads to

$$E\left[\frac{n}{n-1} s^2\right] = E\left[\frac{nm}{nm-1} a\left(1 - \frac{a}{m}\right)\right], \tag{6}$$

with a and s^2 of Eq. (3). The relation (6) between two expected values includes no unknown. It enables us to compare the results of a statistical observation with what is to be expected under certain conditions. We compare, accordingly, the observed value of $\frac{n}{n-1} s^2$ with the observed value of $\frac{nm}{mn-1} a\left(1 - \frac{a}{m}\right)$. If n is large we may simply compare s^2 with $ms'^2 = a\left(1 - \frac{a}{m}\right)$.

For an example which the reader can easily duplicate, the first 3000 letters in the Latin text of Ceasar's *Bellum Gallicum* have been divided into $n = 30$ groups of $m = 100$ letters each. In each group the number of a's have been counted. There were six groups with 7 a's, four groups each with 6, 8, and 9 a's, three groups with 11 and 12, two groups with 5, and one group with 3, 10, 13, and 14 each. The total number of "events" was therefore 253, the average 8.43, and the dispersion $s^2 = 77.90 - 71.06 = 6.84$. Hence

$$\frac{n}{n-1} s^2 = 7.08, \qquad \frac{nm}{nm-1} a\left(1 - \frac{a}{m}\right) = 7.72.$$

Whether or not this difference is significant can be judged by computing the variance according to (24), Chapter VIII. The variance formula (24), Chapter VIII includes the fourth moment τ^4 of the Bernoulli distribution (1). This can be found by the method of the generating function (Chapter IV, Section 3.3) if the third and fourth derivatives of the function $(p + qt)^m$ are considered. [See also Eq. (36) and Problem 4 of this chapter.] The result, with $p = 1 - q$ is

$$\tau^4 = mpq[1 + 3(m-2)pq] = \sigma^2\left[1 + 3\sigma^2\left(1 - \frac{2}{m}\right)\right]. \tag{7}$$

If this and $\sigma^4 = m^2p^2q^2$ are introduced into (24), Chapter VIII, the unknown q again appears. Since, however, we want *some estimate only* for $\mathrm{Var}[s^2]$ we may replace q, which according to (3′) is the expectation of a/m, by the observed value of a/m, that is, 0.0843 in our example. This gives $\sigma^2 = 7.72$, then $\tau^4 = 183$ and $\mathrm{Var}[s^2] = 3.97$.

In the same way, the variance of s'^2 can be computed; in (24), Chapter VIII, n has to be replaced by nm and τ^4 by τ'^4, which follows from (7) with $m = 1$, viz., $\tau'^4 = \sigma'^2 - 3\sigma'^4$. This leads to $\tau'^4 = 0.059$ and $\mathrm{Var}[ms'^2] = \mathrm{Var}[a(1 - a/m)] = 0.18$. Since, except for the factors $n/(n - 1)$ and $nm/(nm - 1)$, which for large n are near to 1, the quantities s^2 and ms'^2 have been compared, we now see that the standard deviations which enter the comparison are of the order of magnitude $\sqrt{3.97} \sim \pm 2$ and $\sqrt{0.18} \sim \pm 0.4$. The observed difference between the values 7.08 and 7.72 is thus seen to lie well within the range of the expected inaccuracy.

For large m and n, if the approximation indicated in (24), Chapter VIII is used, one has

$$\mathrm{Var}[s^2] \sim \frac{mpq}{n}(1 + 2mpq), \qquad \mathrm{Var}\left[a\left(1 - \frac{a}{m}\right)\right] \sim \frac{mpq}{n}(1 - 4pq), \tag{8}$$

which, in our example, would supply 4.2 and 0.18, respectively. In general, the variance of ms'^2 will be much smaller than that of s^2, which increases approximately with m^2. In other words: it is more accurate to approximate ms'^2 by mpq than to approximate s^2 by mpq.

1.2. *Lexis ratio.* Sometimes the ratio of the two quantities compared in (6) is studied under the name of the *Lexis ratio* or *Lexis quotient*:

$$L = \frac{n}{n-1}\,\frac{nm-1}{nm}\,\frac{s^2}{a\left(1 - \dfrac{a}{m}\right)}. \tag{9}$$

The fact that numerator and denominator have the same expectation does not imply that L has expectation one.[2] Considering, however, that the variance of the denominator is comparatively small, one can as a first approximation consider this variance zero. One would then have, if (3′), (6) and the last form of (24) and (18′), Chapter VIII are used,

$$L \sim 1 \pm \frac{\sqrt{\mathrm{Var}[s^2]}}{E[s^2]} = 1 \pm \sqrt{\frac{\tau^4 - \sigma^4}{n\sigma^4}}; \tag{10}$$

[2] Indeed, the equality $E[\text{numerator}] = E[\text{denominator}]$ does not imply, in general, that $E\left[\dfrac{\text{numerator}}{\text{denominator}}\right] = 1$. However, in the case of the Lexis number L we do have $E[L] = 1$. This was proved in 1916 by A. A. Tschuprow and in a much simplified way by the same author in *Skand. Aktuarietidskr.* **6** (1918), p. 224. It is assumed that L is defined to be 1 for $a = 0$ and $a = m$.

that is, in our example, 1 ± 0.26 against the observed $L = 7.08/7.72 = 0.92$. If in (10) we use the approximation (8) we obtain

$$\frac{\operatorname{Var}[s^2]}{\{E[s^2]\}^2} \sim \frac{2m^2p^2q^2}{n(mpq)^2} = \frac{2}{n}.$$

In the light of this approach, a set of statistical data is said to have *normal, subnormal,* or *supernormal dispersion* according to whether L approximately equals 1 or is decidedly smaller or greater than 1. The statistic of our example must be considered to have normal dispersion.

The Lexis theory for *the case of equal event probability throughout but varying* $m_1, m_2, \ldots, m_n$ can be developed in the same way. Instead of (1), we now have

$$\binom{m_\nu}{x_\nu}(1-q)^{m_\nu - x_\nu} q^{x_\nu}$$

and

$$\alpha_\nu = m_\nu q, \qquad \sigma_\nu^2 = m_\nu q(1-q).$$

If this is introduced into (15) and (18), Chapter VIII, we have

$$E[a] = \frac{q}{n}\sum_{\nu=1}^{n} m_\nu, \qquad E[s^2] = \frac{n-1}{n^2}q(1-q)\sum_{\nu=1}^{n} m_\nu. \tag{3''}$$

We define, as before, with $\sum_{\nu=1}^{n} m_\nu = N$

$$a' = \frac{1}{N}\sum_{i=1}^{N} x_i', \qquad s'^2 = \frac{1}{N}\sum_{i=1}^{N} x_i'^2 - a'^2 = a'(1-a').$$

Then

$$E[a'] = q, \qquad E[a'(1-a')] = \frac{N-1}{N}(1-q)q. \tag{5'}$$

From (3″) and (5′), the unknown quantity $q(1-q)$ can be eliminated and this leads to

$$E\left[\frac{n^2}{n-1}\frac{s^2}{N}\right] = E\left[\frac{N}{N-1}a'(1-a')\right].$$

The Lexis number is defined accordingly. If n is large, we may compare

$$\frac{ns^2}{N} \qquad \text{with} \qquad a'(1-a').$$

In the next section, we shall discuss the occurrence of non-normal dispersion.

Problem 1. Analyze the statistics of the occurrence of the letter "a" in a text of plain English prose, taking $n = 30$ successive groups of $m = 100$ letters each. Compare in (6) the two expressions in brackets and give an estimate of the respective variances.

Problem 2. Make the same analysis for the occurrence of "zero" in the fifth place of the 3000 entries from 1100 to 4099 in a seven-figure log table, taking 60 groups of 50 entries each. Do the same using a five-figure log table.

Problem 3. Prove that *if* $x_1, x_2, \ldots, x_n$ are not the numbers of events but their frequencies (the former x divided by m), Eq. (6) has to be replaced by

$$E\left[\frac{n}{n-1} s_0^2\right] = E\left[\frac{n}{nm-1} a_0(1-a_0)\right],$$

where a_0 and s_0^2 are average and dispersion of these frequencies. Give also estimates for $\mathrm{Var}[s_0^2]$ and $\mathrm{Var}[a_0(1-a_0)]$ in this case.

Problem 4. The method outlined in Chapter IV, Section 3.3 for computing the moments of a Bernoulli distribution shows that the moment about the mean of order $2k$ is a polynomial of order k with respect to the number m of trials. Derive formula (7) by assuming that $\tau^4 = Am + Bm^2$ and determining A and B through direct evaluation of τ^4 in the cases $m = 1$ and $m = 2$.

2. Non-Equal Alternatives

2.1. *Poisson trials.* If we analyze routine statistics as they appear in statistical yearbooks, etc., from the standpoint of the Lexis quotient, we find in most cases (deaths, births, marriages) excessively supernormal and in some cases (sex rate of newborn children) slightly subnormal dispersions. These facts can be well accounted for by a generalization of the original hypothesis. Instead of assuming that all underlying alternatives have the same basic probability q, we shall now study cases where q varies in some definite manner.

We again consider n groups, each of m trials. First, we take up the case that within each group of m trials the probability q has different values $q_1, q_2, \ldots, q_m$, but that the same values reappear in all groups. (This means that the group of people among whom the inquiry is made is not uniform, but its composition remains stable throughout the inquiry.) Instead of n groups of Bernoulli trials we have now n groups of so-called *Poisson trials.* Applying the formulas of Chapter VIII, Section 3.1, the α_ν and σ_ν^2, $\nu = 1, 2, \ldots, n$, are now the same for all ν; $\alpha_\nu = \alpha$, $\sigma_\nu^2 = \sigma^2$, so that the simpler Eq. (18′) of Chapter VIII can be used, only the relation among α_ν, σ_ν^2, and the q_μ-values have to be stated anew. Each single distribution $p_\nu(x) = p(x)$ represents now a repetition of m independent (but unequal) alternatives, a problem dealt with in Chapter V. Since in the μth alternative the expectation is q_μ and the variance $q_\mu(1 - q_\mu)$, it follows from the addition formulas of mean value and variance that

$$\sum_{\mu=1}^{m} q_\mu = \alpha, \qquad \sum_{\mu=1}^{m} q_\mu(1 - q_\mu) = \sigma^2. \tag{11}$$

This introduced into Eq. (18′), Chapter VIII gives

$$E[s^2] = \frac{n-1}{n} \sum_{\mu=1}^{m} q_\mu(1 - q_\mu). \tag{12}$$

If, on the other hand, the nm alternatives with the results 0 or 1 are listed, the nm values of α' consist of n sets each composed of $q_1, q_2, \ldots, q_m$ and the σ'^2 of n sets each comprising $q_1(1 - q_1), q_2(1 - q_2), \ldots, q_m(1 - q_m)$. Therefore, the general formula (18) of Chapter VIII supplies for the expectation of the dispersion s'^2

$$E[s'^2] = \frac{nm-1}{nm} \frac{1}{nm}\, n \sum_{\mu=1}^{m} q_\mu(1 - q_\mu) + \frac{1}{nm}\, n \sum_{\mu=1}^{m} (q_\mu - q)^2, \tag{13}$$

where q is the arithmetical mean of $q_1, q_2, \ldots, q_m$. It was seen before that s'^2 equals $(a/m)(1 - a/m)$.[1] Thus from (12) and (13)

$$E\Big[\frac{n}{n-1} s^2\Big] - E\Big[\frac{nm}{nm-1} a\Big(1 - \frac{a}{m}\Big)\Big] = -\frac{nm}{nm-1} \sum_{\mu=1}^{m} (q_\mu - q)^2 < 0. \tag{14}$$

Since the expression on the right-hand side in (14) is negative, we have found: *A subnormal dispersion has to be expected if the event probabilities within each group of observations are not uniform, but remain the same in all groups.* This case is often called the *Poisson case.*

This result is confirmed by the statistics of the sex rate of newborns. It seems that the birth rate is a natural characteristic of the race, independent of changing climatic and economic conditions. Accordingly, in countries with a practically homogeneous population, the dispersion should be normal, and it should be subnormal wherever the population is composed of various races.

2.2. *Lexis trials.* We now turn to the other particular case, where the event probability q is supposed to have the same value within each

[1] Here, as well as in the last section, we prefer to write $(a/m)(1 - a/m)$ for s'^2 in order to emphasize that E is used each time in the same sense.

group of m observations but changes from group to group, assuming the values $q_1, q_2, \ldots, q_n$. Here we have

$$\alpha_\nu = mq_\nu, \qquad \sigma_\nu^2 = mq_\nu(1 - q_\nu) \tag{15}$$

and Eq. (18), Chapter VIII supplies

$$E[s^2] = \frac{n-1}{n}\frac{m}{n}\sum_{\nu=1}^{n} q_\nu(1 - q_\nu) + \frac{m^2}{n}\sum_{\nu=1}^{n}(q_\nu - q)^2, \tag{16}$$

where q is the arithmetical mean of $q_1, q_2, \ldots, q_n$. On the other hand, considering the nm alternatives with m sets of mean values and variances, each consisting of $q_1, q_2, \ldots, q_n$ and $q_1(1 - q_1), q_2(1 - q_2), \ldots, q_n(1 - q_n)$, respectively, we obtain

$$E[s'^2] = \frac{nm-1}{nm}\frac{m}{nm}\sum_{\nu=1}^{n} q_\nu(1 - q_\nu) + \frac{1}{nm}\, m\sum_{\nu=1}^{n}(q_\nu - q)^2. \tag{17}$$

Comparing (16) and (17) and writing $(a/m)(1 - a/m)$ for s'^2, it is found that

$$E\left[\frac{n}{n-1}s^2\right] - E\left[\frac{nm}{nm-1}a\left(1 - \frac{a}{m}\right)\right] = \frac{nm^2(m-1)}{(n-1)(nm-1)}\sum_{\nu=1}^{n}(q_\nu - q)^2 > 0. \tag{18}$$

Here the expression to the right is positive and thus: *A supernormal dispersion is to be expected if the event probability is supposed to be uniform in each of the n groups but changing from group to group.* This is often called the *Lexis case*, in contrast to the Bernoulli and the Poisson case.

Note that the absolute value of the difference in (18) is large compared with that in (14) if the dispersion of the q-values is assumed to be of the same order in both cases. If n and m are considered large, the dispersion is multiplied by $nm^2/(nm - 1) \sim m$ in (14) and by $n^2m^2(m - 1)/(n - 1)(nm - 1) \sim m^2$ in (18).

For the discussion of an example of supernormal dispersion, it is often better to use the set of event frequencies[2] x/m instead of the numbers of events x, since this eliminates to a certain extent the influence of a slight variation in the group size m (see also Problem 7). If now a_0, s_0^2 are average and dispersion of the n observed frequencies, we have to replace s^2 by $m^2s_0^2$ and a by ma_0, in (18). This transforms (18) into

$$E\left[\frac{n}{n-1}s_0^2\right] - E\left[\frac{n}{nm-1}a_0(1 - a_0)\right] = \frac{n(m-1)}{(n-1)(nm-1)}\sum_{\nu=1}^{n}(q_\nu - q)^2. \tag{18'}$$

[2] We always use "frequency" in the sense of "relative frequency" (see Chapter I).

The marriage frequency in Germany in the $n = 19$ years from 1877 to 1895 (a period of comparative stability) had values between 7.5 and 8.0 per 1000 capita. The average was $a_0 = 0.00782$ and the dispersion $s_0^2 = 0.000\,000\,029$. At first sight, this seems a very slight variation, and the layman is inclined to be impressed by the apparent uniformity of human behavior. But if we compute the two expressions on the left-hand side of (18′), we find with $m \sim 45 \times 10^6$

$$\frac{n}{n-1} s_0^2 = 30.3 \times 10^{-9} \qquad \text{and} \qquad \frac{n}{nm-1} a_0(1 - a_0) = 0.172 \times 10^{-9},$$

that is, a Lexis quotient $L = 176$. [The variance of $ns_0^2/(n-1)$ estimated from Eq. (8) with $q = 0.00782$ would be about 55×10^{-18}, which would give an inaccuracy of about $\pm\, 7.4 \times 10^{-9}$]. We see that the dispersion is very large and that only (1/176)th of it corresponds to the variation that would be expected under stable conditions (same q in all years). Almost the whole of the observed dispersion may be attributed to a changing q. In fact, marriage probability is extremely sensitive to small changes in economic conditions as they occur from year to year. We see that the set of data under consideration is by no means remarkable for its uniformity; rather, it reflects a considerable variation of external conditions.

We have the same experience with death statistics. As a rule we encounter Lexis quotients of 100 to 200, excluding wartime conditions or other catastrophic events. Here, the change in q-values is due to seasonal variations of weather conditions and also to economic fluctuations.

Equation (18′) can also be used in the following way. If m is very large (as in our example), the coefficient of the sum to the right of (18′) is practically $1/(n-1)$, independent of m. This shows that the same difference is to be expected if the analysis is repeated under the same conditions with another large m (for instance, in two surveys, one made for the entire United States and a second for one significant state). An example of this approach is given in Problem 6.

Finally, we note another interpretation of supernormal dispersion. If in the second term to the left in (18′), "1" is negligible compared to nm, the coefficient of $a_0(1 - a_0)$ becomes $1/m$. If then m is replaced by $m' = m/L$, we reach the conclusion that the observed values of $ns_0^2/(n-1)$ and $a_0(1 - a_0)/m'$ are approximately equal. This means that the observed dispersion can be considered normal if the statistical data are interpreted as the outcome of m/L independent trials instead of m. In other words, in terms of the previous examples, the situation is as if 176 people would act in solidarity with respect to entering marriage, or, in the case of death statistics, as if 100 to 200 people are collectively affected by the general conditions.[3] (Compare our interpretation of the smallpox statistics in Chapter VI, Section 8.3.)

This interdependence is only another form in which the decisive influence of changing (common) external conditions manifests itself upon the statistical data.

2.3. *Asymptotic distribution of the Lexis quotient.* Assume that the number m in Section 1 is large enough that for (1) the Laplace formula may be substituted. We consider the simplest case of equal q's and a single m. Then (1) is asymptotically normal with mean mq and variance mpq. From Eq. (9) we have for the Lexis quotient

$$L = \frac{n}{n-1}\,\frac{nm-1}{nm}\,\frac{s^2}{a\left(1-\frac{a}{m}\right)} \sim \frac{n}{n-1}\,\frac{s^2}{a\left(1-\frac{a}{m}\right)}.$$

If for large m the observed average a is identified with the theoretical mean mq and a/m with q we have

$$(n-1)\,L \sim \frac{ns^2}{mpq}$$

and $ns^2/mpq = ns^2/\sigma^2$ is the quantity denoted by χ^2 in Eq. (71), Chapter VIII which we have found to have the χ^2-distribution with $(n-1)$ degrees of freedom. Hence: *For large m, the distribution of $(n-1)L$ is, in the Bernoulli case, a χ^2-distribution with $(n-1)$ degrees of freedom.*

Here we have replaced a by mq in the denominator of L, but s^2 is computed with the mean a. If we replace a by mq in s^2 too, thus substituting for L the approximation $\dfrac{1}{n-1}\,\dfrac{\sum_{\nu=1}^{n}(x_\nu - mq)^2}{mpq}$, then $(n-1)L$ has the χ^2-distribution with n degrees of freedom.

Problem 5. If the observed data $z_1, z_2, \ldots, z_n$ are the frequencies of events, $z_\nu = x_\nu/m$, and a_0, s_0^2 the average and the dispersion of these frequencies, prove that Eq. (14) is replaced by

$$E\left[\frac{n}{n-1}\,s_0^2\right] - E\left[\frac{n}{nm-1}\,a_0(1-a_0)\right] = -\frac{n}{nm-1}\,\frac{1}{m}\sum_{\mu=1}^{m}(q_\mu - q)^2.$$

Illustrate by an appropriate example.

[3] A different interpretation of non-normal dispersion has been proposed by H. GEIRINGER, *Ecconometrica* **10** (1942), pp. 53-60. It is assumed that the chance variables (4) are not necessarily independent, but in each group of m variables there is interdependence (the same in each group). This leads to an equality, similar to (14) and (18) which accounts for subnormal or for supernormal dispersion, dependent on the sign of a "covariance."

Problem 6. Apply the Lexis method to the following analysis of style. Consider the plain prose texts (stories or essays in magazines) of twenty different authors and count the number of nouns among the first 200 and among the first 400 words. If this yields each time a supernormal dispersion, compare the excess of dispersion in the two cases $n = 200$ and $n = 400$.

Problem 7. In the case of an equal event probability q throughout, but varying group sizes $m_1, m_2, \ldots, m_n$, assume that all m_ν differ only little from an average value m, viz., $m_\nu - m = \Delta_\nu$ and Δ_ν small. With the notation of problem 5 develop the Lexis theory for the frequencies.

B. Student Test and F-Test (Section 3)

3. The Two Tests

3.1. *The problem.* The problem of testing hypotheses considered in this chapter has the following form. In a specific situation, a sequence x_ν, $\nu = 1, 2, \ldots, n$ of observations is made. The assumption or hypothesis concerns the theoretical distribution which may correspond to this situation.[1] For example, in the Lexis theory we tried, in appropriate situations, the hypothesis of nm independent alternatives.

The checking or testing has the following general form: if the "observed result" has a "small" probability under the assumed hypothesis, we reject the hypothesis. The "observed result" is often incorporated into a *test function* $F(x_1, x_2, \ldots, x_n)$ whose probability under the hypothesis we wish to evaluate. The Lexis quotient, L, was such a test function and other examples [Eqs. (21), (26), (30), and (71)] will be discussed.

Sometimes the problem of finding the distribution of F can be solved asymptotically only, or only under certain restrictions imposed by the mathematics of the problem. We may then choose to solve a simpler problem: to determine (rigourously) the mean value and variance of the test function under the hypothesis, and to check whether or not the observed (or "empirical") test function seems to lie in an appropriately small interval around the expected value. (This was the approach of the Lexis theory.) Clearly, terms like "small," "appropriate," etc., need further specification.

All such "tests" are based on the so-called "direct" approach, which is actually a rather indirect procedure: we *assume* theoretical distribu-

[1] In this book the theoretical distribution is a probability distribution (chance). The so-called "sampling from a finite population" is not considered (see e.g. Wilks [30b], Section 8.5, and other parts of this work).

tions, compute expectations, variances, and distributions of test functions and use them to judge the observed results. The more straightforward problem of theoretical statistics is, however, that of inference; we attempt to draw inferences from given observations on the unknown "parameters," "hypotheses," etc., rather than assuming such hypotheses and testing them. The inference problem was introduced in Chapter VII and will be studied further in Chapter X.

We return to the problem of testing and begin with the following important problem. An experiment is repeated n times; it seems reasonable to assume that the observed quantity is subject to a normal distribution with known mean value but unknown variance. One wants to check this assumption. Suppose we are interested in the influence of a specific drug upon insomnia. The drug is given to n patients, and the additional hours of sleep gained by the use of the drug are observed for each patient. Let $x_1, x_2, \ldots, x_n$ be the results. If the drug is, in fact, ineffective, it seems reasonable to *assume that x is normally distributed*, i.e.,

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\alpha)^2}{2\sigma^2}\right), \tag{19}$$

with the mean value $\alpha = 0$ and some unknown σ^2. (Reasons for using a normal probability distribution are discussed in connection with the theory of errors, the method of least squares,[2] etc.) How can we find out whether or not the set of observations consisting of $x_1, x_2, \ldots, x_n$ agrees with this assumption?

If n independent trials are made, the probability density in the n-dimensional collective resulting from the combination will be

$$p_n(x_1, x_2, \ldots, x_n) = p(x_1)p(x_2) \cdots p(x_n) \tag{20}$$

and this can serve to compute the distribution of any function $F(x_1, x_2, \ldots, x_n)$, according to the mixing rule: To find the probability of F falling in some interval x to $x + dx$, one has to integrate $p_n(x_1, x_2, \ldots, x_n)$ over that part of the n-dimensional space which lies between the two surfaces $F(x_1, x_2, \ldots, x_n) = x$, and $F(x_1, x_2, \ldots, x_n) = x + dx$. This integral, divided by dx, is the probability density of F at the point x. But this density will in general depend on σ^2 as well as on α. Our problem is to find a function F whose distribution is independent of σ^2 but reflects the value of α. Such a function was suggested in 1908 by a British

[2] See Chapter VII, Section 5. Elementary considerations go back to Gauss, to his postulate of the arithmetic mean. See an interesting presentation in E. T Whittaker and G. Robinson, *The Calculus of Observations*, p. 215 ff. London, 1932.

statistician, Student.[3] It is known as *Student's test function* or *Student's ratio*, usually denoted by t. Its definition is

$$t = \sqrt{n-1}\,\frac{\bar{x}-\alpha}{s}, \qquad \bar{x} = \frac{1}{n}\sum_{\nu=1}^{n} x_\nu, \qquad s^2 = \frac{1}{n}\sum x_\nu^2 - \bar{x}^2, \tag{21}$$

and α is supposed known. (The $\sqrt{n-1}$ corresponds to the fact that the distribution of s^2 has $n-1$ degrees of freedom.) We write the quotient $(\bar{x}-\alpha)/s = t/\sqrt{n-1}$ in the form

$$\frac{\bar{x}-\alpha}{\sigma/\sqrt{n}} \div \frac{s\cdot\sqrt{n}}{\sigma}.$$

Since we assume x to be normally distributed with mean α and variance σ^2, we know that $u = (\bar{x}-\alpha)\sqrt{n}/\sigma$ is normally distributed with mean zero, variance one: The squared second fraction, $(1/\sigma^2)\sum_{\nu=1}^{n}(x_\nu - \bar{x})^2$, is the χ^2 with $n-1$ degrees of freedom of Eq. (71), Chapter VIII. Since we have found in Chapter VIII (corollary in Section 6.4) that u and χ^2 are independently distributed, we can apply the result of Chapter VIII, Section 7.1 which states that $t = \dfrac{u}{\sqrt{\chi^2/(n-1)}}$ has the Student distribution with $n-1$ degrees of freedom. Hence:

If the $x_1, x_2, \ldots, x_n$ are drawn from a normal population $N(\alpha, \sigma^2)$ and $u = \dfrac{\bar{x}-\alpha}{\sigma/\sqrt{n}}$, $\chi^2 = (1/\sigma^2)\sum_{\nu=1}^{n}(x_\nu - \bar{x})^2$, the quotient

$$t = \frac{u}{\sqrt{\chi^2/(n-1)}} = \sqrt{n-1}\,\frac{\bar{x}-\alpha}{s} \tag{21'}$$

has the distribution

$$p_{n-1}(t) = \frac{1}{\sqrt{\pi(n-1)}}\cdot\frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)}\left(1+\frac{t^2}{n-1}\right)^{-n/2} \tag{22}$$

Since the variate t does not contain σ (while α is assumed known) we can compute t from the sample. We thus obtain a test for the significance of an observed mean $\bar{x}$, given a value α.

When Student's distribution was discovered, it marked an important

[3] STUDENT, "The probable error of a mean." *Biometrika* **6** (1908), p. 1. A more exact derivation was given in 1926 by R. A. Fisher.

step in statistical inference, since it gave the exact distribution of a function of the observations (of a "statistic"), in contrast to the classical results (Chapter VI) which are only asymptotically valid (although under very weak and general assumptions). The fact that certain distributions are correct for any n may have given rise to the unfortunate term of a "small sample theory," meant to distinguish considerations based on exact distributions from asymptotic statements.[4] The answer provided by an exact distribution is correct for any n but in the case of a small sample, it is correspondingly meager; by uncritical or illogical "interpretations" strong but erroneous statements have occasionally been obtained (see section 3.3).

3.2. *Difference of two means.* An important application of Student's distribution is to test assumptions regarding the difference of two means. If two independent samples $x_1, x_2, \ldots, x_{n_1}$, and $y_1, y_2, \ldots, y_{n_2}$ are drawn *from normal populations* with parameters α_1, σ_1^2 and α_2, σ_2^2 and if $\bar{x}_1$, $\bar{x}_2$ denote the respective sample averages, then

$$u = \frac{(\bar{x}_1 - \alpha_1) - (\bar{x}_2 - \alpha_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} \tag{23}$$

is normally distributed with mean 0, variance 1. In fact $\bar{x}_1$ has a normal distribution with α_1, σ_1^2/n_1 and $\bar{x}_2$ one with α_2, σ_2^2/n_2, while $\bar{x}_1 - \bar{x}_2$ has expectation $\alpha_1 - \alpha_2$ and variance $\sigma_1^2/n_1 + \sigma_2^2/n_2$, which proves the statement.

Now assume that $\sigma_1^2 = \sigma_2^2 = \sigma^2$. Then the statistic u becomes

$$u = \frac{(\bar{x}_1 - \alpha_1) - (\bar{x}_2 - \alpha_2)}{\sigma\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}. \tag{23'}$$

Let s_1^2 and s_2^2 be the dispersions of the two samples; then by the corollary just quoted (Chapter VIII, Section 6.4) the statistic

$$\chi^2 = \frac{n_1 s_1^2 + n_2 s_2^2}{\sigma^2} \tag{24}$$

is distributed independently of (23′) in χ^2-distribution with $(n_1 - 1) + (n_2 - 1) = n_1 + n_2 - 2$ degrees of freedom. Hence:

[4] An example of a similarly unfortunate terminology was Bortkiewicz' notation of the Poisson limit law as the "law of small numbers." After this terminology had created some confusion it was abandoned, in favor of the appropriate characterization "rare events."

If two samples $x_1, \ldots, x_{n_1}, y_1, \ldots, y_{n_2}$ *are supposed to be drawn from normal populations* $N(\alpha_1, \sigma^2)$, $N(\alpha_2, \sigma^2)$, *then the ratio*

$$t = \frac{u}{\sqrt{\dfrac{\chi^2}{n_1 + n_2 - 2}}} = \frac{\sqrt{n_1 + n_2 - 2}}{\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}} \frac{(\bar{x}_1 - \alpha_1) - (\bar{x}_2 - \alpha_2)}{\sqrt{n_1 s_1^2 + n_2 s_2^2}} \tag{25}$$

has the Student distribution with $n_1 + n_2 - 2$ *degrees of freedom.* If $n_1 = n_2 = n$, the last expression simplifies to

$$\sqrt{(n-1)}\,\frac{(\bar{x}_1 - \alpha_1) - (\bar{x}_2 - \alpha_2)}{\sqrt{s_1^2 + s_2^2}}.$$

3.3. *An example.* As an example for the application of Student's distribution, we quote the following. A sleeping drug has been tested on $n = 10$ patients. The increase x in length of sleep has been found to range from -1.6 to $+3.7$ hours, with an average $\bar{x} = 0.75$ and dispersion $s^2 = 2.88$. If we make the assumption that the mean efficiency of the drug corresponds to an extension of sleep length of $\alpha = 0.5$ hours, we arrive at the t-value:

$$t = \frac{\sqrt{n-1}\,(\bar{x} - \alpha)}{s} = \frac{(3) \cdot (0.25)}{1.70} = 0.44.$$

We know (Chapter VIII, Section 7.1) that if an α is assumed and the number of degrees of freedom given, the table gives for the symmetric Student distribution the probability of the two "tails," $P(t > x) + P(t < -x)$. In our case, $n = 10$, and we find 70 % probability for $|t| > 0.398$, hence 35 % for $t > 0.398$, and 30 % for $t > 0.543$. Returning to our example, we see that there is a sufficiently large probability for a t as large or larger than 0.44. Also 0.44 is well within the 50 % limits which are ± 0.703. One can conclude that the outcome of this (very modest) statistical investigation is *not in disagreement with the assumption that the ten observed values of x correspond to a normal distribution with mean value* 0.5. The same conclusion would hold, on account of the symmetry of the distribution, if α were assumed to be slightly larger than $\bar{x} = 0.75$, equal to one hour, say.

Unfortunately, the Student test is often applied in an illegitimate way. In discussing the just-quoted experiments, one author concludes: The probability that the drug is effective, that is, the probability of $\alpha \geqslant 0$, is 89 %. He arrives at this conclusion by using definition (21) $\alpha = \bar{x} - ts/\sqrt{n-1}$ and concluding that $\alpha \geqslant 0$ if $t \leqslant \sqrt{n-1}\,(\bar{x}/s)$, where $\sqrt{n-1}\,(\bar{x}/s) = 1.32$ in the present problem. Then he takes from the integrated Eq. (22) or from the tables, with $n - 1 = 9$,

$$P_9(1.32) = \text{Prob}\{t \leqslant 1.32\} = 0.888.$$

There is, of course, not the slightest reason why the probability of $t \leqslant 1.32$, taken from a table with constant α, should be interpreted as a probability of α falling in some region. In fact, the reasoning and its result are equally absurd. In the same way, considering *two* experiments with $x_1 = 0.1$, $x_2 = 0.2$, which gives $t = 3$ (by the above computation) one could draw the conclusion that the probability of the drug being effective is (by Eq. (82), Chapter VIII) $0.5 + (1/\pi)$ (arctan 3) ~ 0.90. Simple common sense should teach us that such an inference from two observations is impossible.

Problem 8. If a sample of 4 is drawn from a normal population, what is the probability of obtaining a t-value between 0.5 and 1.0? Do the same problem for a sample of 10 and compare in both cases the probability with that from the Gaussian, $\alpha = 0$, $\sigma = 1$.

Problem 9. An object of mass production, supposed to have a size of $\frac{1}{4}$ in., is subject to a sampling test. Five measurements supply $x = 0.2510$, 0.2500, 0.2495, 0.2485, 0.2480. Check whether this result fits the assumption of a normal distribution around 0.2500.

Problem 10. A wine manufacturer sent presumably identical samples of wine to two laboratories. Each made five determinations of the alcohol content in percentage as follows: (a) 12, 13.5, 13, 10.5, 12 and (b) 13.5, 14, 11.5, 15, 13. Were the two laboratories measuring the same thing? (Assume normality and a common variance.)

3.4. *The F-Test.* We now consider the following problem. We have two samples of sizes n_1 and n_2, respectively, from normal populations $N(\alpha_1, \sigma_1^2)$, $N(\alpha_2, \sigma_2^2)$; it may seem reasonable to assume that $\sigma_1 = \sigma_2$. One wants to check this assumption. Denote by a_1, a_2 the arithmetical means of the two samples. We know that the two dispersions, $z_1 = \sum_{\nu=1}^{n_1} (x_\nu - a_1)^2/\sigma_1^2$, $z_2 = \sum_{\nu=1}^{n_2} (x_\nu - a_2)^2/\sigma_2^2$, each have a χ^2–distribution with $n_1 - 1$ and $n_2 - 1$ degrees of freedom, respectively. We wish to test the hypothesis $\sigma_1 = \sigma_2$.

It is here convenient to use the notation $s^2 = \sum_{\nu=1}^{n} (x_\nu - a)^2/(n - 1)$ rather than our usual one where the denominator is n. Then

$$z_1 = \frac{(n_1 - 1)s_1^2}{\sigma_1^2}, \qquad z_2 = \frac{(n_2 - 1)s_2^2}{\sigma_2^2}.$$

Under the hypothesis $\sigma_1 = \sigma_2$ we have

$$\frac{z_1/(n_1 - 1)}{z_2/(n_2 - 1)} = \frac{s_1^2}{s_2^2} F. \tag{26}$$

To test the hypothesis $\sigma_1 = \sigma_2$ we take as test function this $F = s_1^2/s_2^2$, which we compute from the data. In Chapter VIII, Section 7.2, Eqs. (86) and (86′), we have computed the distribution of F. We now choose an ϵ and take from the tables a value $w = F_\epsilon$ defined by

$$1 - H(w) = \Pr\{F > w\} = \epsilon, \tag{27}$$

$$H(w) = K \int_0^w F^{(n_1/2)-1} \left(1 + \frac{n_1}{n_2} F\right)^{-(n_1+n_2)/2} dF \tag{27′}$$

and where $\epsilon = 0.05$, for example. If the observed $F = s_1^2/s_2^2$ is greater than $w = F_\epsilon$ we decide to reject the hypothesis. If, for example, $n_1 = 10$, $n_2 = 30$ we find, with $\epsilon = 0.05$, $w = 2.22$. If $\epsilon' = 0.01$ the w' for the same n_1 and n_2 is 3.09; that means that with $n_1 = 10$, $n_2 = 30$ there is at most a chance of 1 %, to obtain an $s_1^2/s_2^2 > 3$ if $\sigma_1 = \sigma_2$ holds true. If $n_1 - 1 = 100$, $n_2 - 1 = 100$, $w = F_\epsilon = 1.39$, in the 5 % case, and 1.60 in the 1 % case.

The main application of the F-test is in the so-called analysis of variance, which is not discussed in this book.

C. The X^2-Test (Sections 4 and 5)

4. Checking a Known Distribution

4.1. *The problem.* A very important and typical problem is the following. The outcome of a collective with known arithmetical probability distribution is observed n times. The set of statistical data then consists of k numbers $n_1, n_2, \ldots, n_k$ with

$$n_1 + n_2 + \cdots + n_k = n, \tag{28}$$

n_i indicating how many times the ith label value has been observed. The probabilities of the k labels, $q_1, q_2, \ldots, q_k$ with

$$q_1 + q_2 + \cdots + q_k = 1 \tag{29}$$

are supposed to be known. For example, assume that a die has been tossed 1200 times and that the points 1, 2, 3, ... appeared $n_1 = 196$,

$n_2 = 188$, $n_3 = 214$, ... times, respectively. The q_i are supposed to equal $\frac{1}{6}$ each. The problem is whether the statistical data confirm or do not confirm the assumption $q_1 = q_2 = \dots = q_6 = \frac{1}{6}$.

The population from which the sample is drawn is here the combination of n equal collectives (n tossing of a die, for example) with the numbers n_1, n_2, ..., n_{k-1} as the results, n_k being determined by Eq. (28); we shall give, in Eq. (44), this $(k - 1)$-dimensional probability distribution.

In the present problem as in previous ones we select first a test function $F(n_1, \dots, n_{k-1})$. The next step is to compute its expectation and its variance and then to compare $E[F] \pm \sqrt{\mathrm{Var}[F]}$ with the observed value of F. If possible, we compute also the distribution of F and use a corresponding table as the basis of the test.

Since in an infinite sequence of trials the ratios n_i/n will approach q_i, a reasonable test function might have the form

$$F(n_1, n_2, \dots, n_{k-1}) = \sum_{i=1}^{k} \lambda_i\left(\frac{n_1}{n} - q_i\right)^2, \tag{30}$$

where λ_1, λ_2, ..., λ_k are appropriately chosen positive constants. Let us compute $E[F]$ for this choice of F. Obviously, we need only the expectation of one single term like $(n_1/n - q_1)^2$; then $E[F]$ will be a linear combination of these k expectations.

As long as we are only interested in a function of n_1, we can "mix" all the other label values 2, 3, ..., k and study the simple alternative "1" or "not 1." In such an alternative with the probabilities q_1 for the event and $(1 - q_1)$ for non-event, the expectation of n_1 is nq_1 and the variance of n_1 or the expectation of $(n_1 - nq_1)^2$ is $nq_1(1 - q_1)$. It follows for any i that

$$E\left[\left(\frac{n_i}{n} - q_i\right)^2\right] = \frac{1}{n^2} E[(n_i - nq_i)^2] = \frac{q_i(1 - q_i)}{n},$$

and thus

$$E[F] = \frac{1}{n}\sum_{i=1}^{k} \lambda_i q_i(1 - q_i). \tag{31}$$

For instance, with $\lambda_1 = \lambda_2 = \cdots = \lambda_k = 1$

$$E[F] = \frac{1}{n}\sum_{i=1}^{k} q_i(1 - q_i) = \frac{1}{n} - \frac{1}{n}\sum_{i=1}^{k} q_i^2 . \tag{32}$$

Instead of taking $\lambda_1 = \lambda_2 = \cdots = \lambda_k = 1$, the British statistician Karl Pearson introduced (1900) another set of λ_i which has certain distinct advantages, namely,

$$\lambda_i = \frac{n}{q_i}, \qquad i = 1, 2, ..., k. \tag{33}$$

If this is introduced into (30), the test function F will be denoted by X^2 and we have

$$\begin{aligned} X^2 = \frac{1}{n}\sum_{i=1}^{k} \frac{(n_i - nq_i)^2}{q_i} &= \frac{1}{n}\sum_{i=1}^{k} \frac{n_i^2}{q_i} - 2\sum_{i=1}^{k} n_i + n\sum_{i=1}^{k} q_i \\ &= \frac{1}{n}\sum_{i=1}^{k} \frac{n_i^2}{q_i} - n. \end{aligned} \tag{34}$$

With λ_i from (33), Eq. (31) becomes

$$E[X^2] = \frac{1}{n}\sum_{i=1}^{k} n(1 - q_i) = \sum_{i=1}^{k} (1 - q_i) = k - 1. \tag{35}$$

The expectation of X^2 is thus seen to be *independent of the specific values of the probabilities* $q_1, q_2, ..., q_k$ *and equal to the number of different label values minus* 1. This simple formula speaks in favor of the choice (33) and more along this line will be found later. Whether Pearson's choice of λ_i has a statistical justification is a different question: it gives great weight to those deviation squares $(n_i - nq_i)^2$ that correspond to small q_i. (See also p. 456). The computation of the variance is more difficult. It necessitates certain formal computations which we shall do in Section 4.3.

4.2. *Preliminary computations.* In this section, we will compute expected values associated with Bernoulli distributions and multinomial distributions which will be needed in the computation of the variance of the present test function (30) and of a related one (the so-called ω^2-test function) to be studied in Sections 6 and 7.

First let us consider the ordinary Bernouilli distribution,

$$p_n(n_1) = \frac{n!}{n_1!(n - n_1)!} q_1^{n_1}(1 - q_1)^{n-n_1}.$$

Let $\delta_1 = n_1 - nq_1$. Then the second, third, and fourth moments about the mean are easily obtained by means of the moment generating function (See Chapter IV, Section 3.3), namely,

$$E[\delta_1^2] = nq_1(1 - q_1), \qquad E[\delta_1^3] = nq_1(1 - q_1)(1 - 2q_1),$$
$$E[\delta_1^4] = 3n^2q_1^2(1 - q_1)^2 + nq_1(1 - q_1)(1 - 6q_1 + 6q_1^2). \tag{36}$$

For mixed moments we have to use the multinomial distribution: where $q_1 + q_2 + q_3 = 1$,

$$p_n(n_1, n_2) = \frac{n!}{n_1!n_2!(n - n_1 - n_2)!} q_1^{n_1}q_2^{n_2}q_3^{n-n_1-n_2} .$$

Let $\delta_1 = n_1 - nq_1$, $\delta_2 = n_2 - nq_2$, $\delta_3 = n_3 - nq_3$. The covariance is equal to $-nq_iq_j$; hence we have

$$E[\delta_1\delta_2] = - nq_1q_2 .$$

More difficult is the computation of $E[\delta_i^2\delta_j^2]$. We know that the first three moments of $p_n(x)$, with respect to the mean, can be found from those of $p_1(x)$ by multiplication by n, and that this is no longer true of the moment of order four. We take as given that [as in the last formula (36)] the fourth moments of $p_n(n_1, n_2)$ are each equal to a constant multiplied by n^2 plus a constant multiplied by n. We may then write

$$E[\delta_1^2\delta_2^2] = An^2 + Bn = -A'n(n - 2) + \tfrac{1}{2}B'n(n - 1), \tag{37}$$

where A' is the required value of $E[\delta_1^2\delta_2^2]$ for $n = 1$ and B' that for $n = 2$. We shall now compute A' and B'.

Take $n = 1$. There are three possibilities:

(a) $n_1 = 1, n_2 = n_3 = 0$; then $\delta_1 = 1 - 1 \cdot q_1 = 1 - q_1$, $\delta_2 = 0 - q_2 = -q_2$,

(b) $n_2 = 1, n_1 = n_3 = 0$; $\delta_1 = -q_1$, $\delta_2 = 1 - q_2$,

(c) $n_1 = n_2 = 0$; $\delta_1 = -q_1$, $\delta_2 = -q_2$.

Multiplying each of these $\delta_1^2\delta_2^2$ by its probability and adding, we obtain, for $n = 1$,

$$E[\delta_1^2\delta_2^2] = (1 - q_1)^2q_2^2q_1 + q_1^2(1 - q_2)^2q_2 + q_1^2q_2^2(1 - q_1 - q_2)$$
$$= q_1q_2(q_1 + q_2 - 3q_1q_2).$$

Next take $n = 2$. There are six different possibilities: $n_1 = 2$, $n_2 = n_3 = 0$; $n_2 = 2$, $n_1 = n_3 = 0$; $n_3 = 2$, $n_1 = n_2 = 0$; $n_1 = n_2 = 1$; $n_1 = n_3 = 1$; $n_2 = n_3 = 1$. The first term is then $(2 - 2q_1)^2 \cdot (0 - 2q_2)^2q_1^2$;

the term corresponding to $n_1 = n_2 = 1$ is $(1 - 2q_1)^2(1 - 2q_2)^2 \cdot 2q_1q_2$, etc. The sum of the six terms is $2q_1q_2$. Hence

$$A' = q_1q_2(q_1 + q_2 - 3q_1q_2), \qquad B' = 2q_1q_2; \tag{38}$$

substituting this into (37) we find

$$\begin{aligned} E[\delta_1^2\delta_2^2] &= n(n-1)q_1q_2 - n(n-2)\cdot q_1q_2(q_1 + q_2 - 3q_1q_2) \\ &= nq_1q_2[1 + (n-2)(1 - q_1 - q_2 + 3q_1q_2)]. \end{aligned} \tag{39}$$

4.3. *Variance of Karl Pearson's X^2.* Now let us return to the problem of Section 4.1 and compute first the variance for the general test function (30):

$$F = \sum_{i=1}^{k} \lambda_i\left(\frac{n_i}{n} - q_i\right)^2 = \frac{1}{n^2}\sum \lambda_i\delta_i^2. \tag{30}$$

We set $F \cdot n^2 = Z^2$, hence

$$Z^2 = \lambda_1\delta_1^2 + \lambda_2\delta_2^2 + \cdots + \lambda_k\delta_k^2, \tag{40}$$

and $Z^2 = n^2X^2$.

To obtain Pearson's X^2 from this Z^2 we shall have to set $\lambda_1 = n/q_i$ as in (33) and to divide the expression Z^2 by n^2. Now

$$\begin{aligned} \operatorname{Var}[Z^2] &= E[Z^4] - (E[Z^2])^2 \\ &= \sum_{i=1}^{k} \lambda_i^2 E[\delta_i^4] + 2\sum_{j<i}^{1\cdots k} \lambda_j\lambda_i E[\delta_i^2\delta_j^2] - \Big[\sum_{i=1}^{k} \lambda_i E[\delta_i^2]\Big]^2. \end{aligned} \tag{41}$$

Here, $E[\delta_i^2]$, $E[\delta_i^4]$ and $E[\delta_i^2\delta_j^2]$ are given by (36) and (39). Substituting these equations into (41), we obtain

$$\begin{aligned} \operatorname{Var}[Z^2] &= 2n^2\Big[\sum_{i=1}^{k} \lambda_i^2 q_i^2(1 - q_i)^2 + 2\sum_{j<i}^{1\cdots k} \lambda_i\lambda_j q_i^2 q_j^2\Big] \\ &+ n\Big[\sum_{i=1}^{k} \lambda_i^2 q_i(1 - q_i)(1 - 6q_i + 6q_i^2) - 2\sum_{j<i}^{1\cdots k} \lambda_i\lambda_j q_i q_j(6q_iq_j - 2q_i - 2q_j + 1)\Big]. \end{aligned} \tag{42}$$

Dividing the right-hand side of Eq. (42) by n^4, substituting $\lambda_i = n/q_i$ from Pearson's assumption (33), and using

$$\sum_{i=1}^{k} q_i = 1, \qquad 2\sum_{i<j}^{1\cdots k} q_iq_j = 1 - \sum_{i=1}^{k} q_i^2,$$

we obtain for Pearson's X^2, given by Eq. (34), the following variance:

$$\operatorname{Var}[X^2] = 2(k-1) + \frac{1}{n}\left[\sum_{i=1}^{k} \frac{1}{q_i} - k^2 - 2k + 2\right]. \tag{43}$$

Equations (35) and (43) enable one to form a quick judgement, valid for any n, as to whether an observed value of X^2 is in the interval $E(X^2] \pm (\operatorname{Var}[X^2])^{1/2}$ or not.

4.4. *Asymptotic distribution of* X^2. We shall now proceed to find the distribution of X^2, that is, $\operatorname{Prob}\{a \leqslant X^2 \leqslant b\}$ for an arbitrary interval $[a, b]$. This problem is solved asymptotically only.

To compute the distribution of X^2 we need the multinomial probability $p(n_1, n_2, \ldots, n_{k-1})$ of getting a set of labels $n_1, n_2, \ldots, n_{k-1}$ in n trials; then the probabilities p of all those sets $n_1, n_2, \ldots, n_{k-1}$ which fulfill the condition

$$a \leqslant X^2 \leqslant b$$

have to be summed and this sum is the probability of X^2 falling in the interval a to b. We have seen in Chapter IV that the probability $p(n_1, n_2, \ldots, n_{k-1})$ is given by the multinomial formula (with $n_1 + \cdots + n_k = n$)

$$p(n_1, n_2, \ldots, n_{k-1}) = \frac{n!}{n_1! n_2! \cdots n_k!}\, q_1^{n_1} q_2^{n_2} \cdots q_k^{n_k}, \tag{44}$$

which reduces to the binomial distribution, if k equals 2.

We now consider large n and use the Stirling formula $n! \sim n^n e^{-n}\sqrt{2\pi n}$ for all factorials appearing in (44). From (44) we obtain

$$\log p = n \log n - \sum_{i=1}^{k} n_i \log n_i - n + \sum_{i=1}^{k} n_i + \sum_{i=1}^{k} n_i \log q_i + R, \tag{45}$$

where R denotes the terms of lower order of magnitude in n or n_i. In the place of n_i we introduce

$$u_i = \frac{n_i - nq_i}{\sqrt{nq_i}}, \qquad \text{or} \qquad n_i = nq_i\left(1 + \frac{u_i}{\sqrt{nq_i}}\right), \tag{46}$$

which transform the X^2 of (34) into[1]

$$X^2 = u_1^2 + u_2^2 + \cdots + u_k^2. \tag{47}$$

[1] In the metric defined by (46) the u_i are ordinary rectangular coordinates.

It will be understood that while n and n_i go to ∞, the u_i remain bounded. Using the expansion $\log(1 + x) = x - x^2/2 + \cdots$, we have with definition (46),

$$\sum_{i=1}^{k} n_i \log n_i = \sum_{i=1}^{k} n_i \log n + \sum_{i=1}^{k} n_i \log q_i$$

$$+ \sum_{i=1}^{k} nq_i \left(1 + \frac{u_i}{\sqrt{nq_i}}\right) \cdot \frac{u_i}{\sqrt{nq_i}} \left(1 - \frac{u_i}{2\sqrt{nq_i}} + \cdots\right)$$

$$= n \log n + \sum_{i=1}^{k} n_i \log q_i + \sum_{i=1}^{k} u_i \sqrt{nq_i} + \sum_{i=1}^{k} \tfrac{1}{2} u_i^2 , \tag{45'}$$

if terms which have $\sqrt{n}$ in the denominator are omitted.

Definition (45) in connection with $\Sigma_{i=1}^k n_i = n$ and $\Sigma_{i=1}^k q_i = 1$ shows that $\Sigma_{i=1}^k u_i\sqrt{nq_i} = 0$. Therefore, if (45′) is introduced into (45), the right-hand side of (45) reduces to $- \frac{1}{2} \Sigma_{i=1}^k u_i^2 + R = - \frac{1}{2} X^2 + R$. On the other hand, neglecting all terms that vanish for $n \to \infty$, we have

$$R \sim - \frac{k-1}{2} \log 2\pi - \frac{k-1}{2} \log n - \tfrac{1}{2} \sum_{i=1}^{k} \log q_i ,$$

which is independent of the n_i. Thus the asymptotic expression for p becomes

$$p(n_1 , n_2 , \ldots, n_{k-1}) \sim \text{const. } e^{-X^2/2} = \frac{e^{-X^2/2}}{\sqrt{(2n\pi)^{k-1} q_1 q_2 \cdots q_k}} . \tag{48}$$

Since X^2 is a quadric in the variables $u_1 , u_2 , \ldots, u_{k-1}$, the result is: *As the n_i increase indefinitely in the above described way, the polynomial distribution* (44) *tends toward the normal distribution* (48); *it has a constant value for all sets $n_1 , n_2 , \ldots, n_{k-1}$ to which one and the same value of X^2 belongs.*

For the last step[2], the computation of the distribution of X^2, let us start with the case $k = 3$. Here we have only two independent variables,

$$X^2 = u_1^2 + u_2^2 + u_3^2 = u_1^2 + u_2^2 + \left(\frac{u_1\sqrt{q_1} + u_2\sqrt{q_2}}{\sqrt{q_3}}\right)^2 .$$

In a u_1, u_2-coordinate system, $X^2 = \text{const} = c^2$ is represented by an ellipse. All the ellipses corresponding to different c-values are similar

[2] From here on the integration problem is the same as in Chapter VIII, Section 6.5, but we proceed now in a more geometric way.

and their axes are proportional to c. The probability of X^2 lying in the interval c^2 to $(c + dc)^2$ is given by the sum of all p-values that correspond to points u_1, u_2 within the annular ring shown in Fig. 32. All these p are

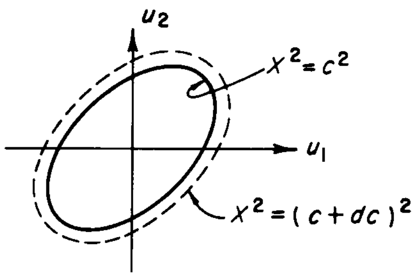

FIG. 32.

equal (except for infinitesimal differences) since they belong to essentially the same X^2-value. We have, therefore, to multiply $Ce^{-c^2/2}$ by the number of points in the ring. Since two consecutive values of n_1 or n_2 differ by one unit, two consecutive u_1 differ, according to (45) by $1/\sqrt{nq_1}$ and two consecutive u_2 by $1/\sqrt{nq_2}$. The points which belong to integers n_i, thus, form a rectangular lattice in the u_1, u_2-plane, the meshes of which have dimensions proportional to $1/\sqrt{n}$. As n increases, the number of meshes and of lattice points in a given region become larger and larger, and are proportional to the area of the annular region in Fig. 32. This area, again, is proportional, for $k - 1 = 2$, to dc and to the length of the elliptical curve, hence to $c\,dc$. Thus, for $k = 3$ (writing here "prob" for the densities)

$$\text{prob}\{c^2 \leqslant X^2 \leqslant (c + dc)^2\} = \text{const} \cdot e^{-c^2/2} c\,dc. \tag{49}$$

This holds for $k = 3$ only. For general k the ellipse becomes a $(k - 1)$-dimensional ellipsoid and the differential $c\,dc$ in Eq. (49) is replaced by $c^{k-2}\,dc$. Hence

$$\text{prob}\{c^2 \leqslant X^2 \leqslant (c + dc)^2\} = \text{const} \cdot e^{-c^2/2} c^{k-2}\,dc\,.$$

Finally, we write $x = c^2$, $c = \sqrt{x}$, $dc = dx/2\sqrt{x}$, $c^{k-2}\,dc = \frac{1}{2}(\sqrt{x})^{k-3}\,dx$, and arrive at

$$\text{prob}\{x \leqslant X^2 \leqslant x + dx\} = \tfrac{1}{2}\,\text{const} \cdot e^{-x/2} x^{(k-3)/2}\,dx. \tag{50}$$

The factor multiplied by dx on the right-hand side is the probability density $p_k(x)$ of X^2, which we can write as

$$p_k(x) = Ce^{-x/2} \cdot \left(\frac{x}{2}\right)^{(k-3)/2}. \tag{51}$$

Hence, for any a and b:

$$\Pr\{a \leqslant X^2 \leqslant b\} = C \int_a^b e^{-x/2} \left(\frac{x}{2}\right)^{(k-3)/2} dx. \tag{52}$$

In order to find the constant C, we choose $a = 0$, $b = \infty$, thus covering the whole range of possible X^2-values and obtain

$$1 = C \int_0^\infty e^{-x/2} \left(\frac{x}{2}\right)^{(k-3)/2} dx = 2C\Gamma\left(\frac{k-1}{2}\right), \qquad C = \frac{1}{2\Gamma\left(\frac{k-1}{2}\right)}.$$

The asymptotic probability density of X^2, is, independent of the original probabilities $q_1, q_2, \ldots, q_k$, equal to[3]

$$p_{k-1}(x) = \frac{1}{\Gamma\left(\frac{k-1}{2}\right) 2^{(k-1)/2}} x^{(k-3)/2} e^{-x/2}. \tag{53}$$

This is the density of the chi-square distribution with $k - 1$ degrees of freedom [Chapter VIII, Eq. (65)]. The formula for the corresponding c.d.f. $P_m(x)$ ($m = k - 1$ = number of degrees of freedom) is in Chapter VIII, Eqs. (64) and (67).

Since mean and variance of the chi-square distribution (53) are $(k - 1)$ and $(2k - 2)$ we may write, for large n:

$$X^2 \sim k - 1 \pm \sqrt{2k - 2}.$$

Note that $(k - 1)$ was found in (35) to be the $E[X^2]$ for any n whatsoever, while $(2k - 2)$ is the variance only for infinite n. As can be seen from (43), the complete expression includes in addition a term in $1/n$.

4.5. *Applications.* If a die, supposed to be unbiased, is thrown sufficiently often, the limits of the expectation for the quantity

$$X^2 = \frac{6}{n} \sum_{i=1}^{6} \left(n_i - \frac{n}{6}\right)^2 = \frac{6}{n} \sum_{i=1}^{6} n_i^2 - n$$

are $5 \pm \sqrt{10}$, and this may be checked with the observed value of X^2.

[3] The result is due to K. Pearson, "On the criterion that a given system of deviations is such that it can be reasonably supposed to have arisen from random sampling." *Phil. Mag.* **50** (1900), p. 157.

Or, using the second Eq. (67), Chapter VIII, we find the probability of $X^2 \leqslant x$ as

$$P_5(x) = 2\Phi(\sqrt{x}) - 1 - \sqrt{\frac{2x}{\pi}}\, e^{-x/2}\left(1 + \frac{x}{3}\right),$$

that is, 0.793 for $x = 5 + 0.674 \cdot \sqrt{10} = 7.1$ and 0.285 for $x = 2.9$. There is therefore for X^2 a probability of 0.508 to lie between the probable limits 2.9 and 7.1.

Let us consider an application of the X^2-method to an experiment with the Galton board (Chapter VI, Section 4.2). We consider $n = 1000$ balls; the board has 60 rows and correspondingly $k = 61$ cells at the bottom, where the balls land. Take the simplest assumption that each ball after hitting a nail goes to the left or right with probabilities $p = q = \frac{1}{2}$; then the q_i of Eq. (34) are

$$q_i = \binom{60}{i}\left(\frac{1}{2}\right)^{60}, \qquad i = 0, 1, ..., 60.$$

The expectation of X^2 equals $k - 1 = 60$, and the first term of the variance equals $2k - 2 = 120$. If, however, we compute from (43) the exact value of the variance, it turns out to be of the order of 10^{16} notwithstanding the large value of $n = 1000$. The reason lies in the small values of the extreme q_i's. Pearson's definition of X^2 gives here disproportionate weight to these results—the number of balls near the two ends, left and right—since we divide each $(n_i - nq_i)^2$ by nq_i. Hence the computed value of X^2 depends almost entirely on those results in the end cells and we obtain no information on the more central part of the results, which is actually of greater interest. If we wish to apply the X^2-method to this problem it seems reasonable to mix the results in the end cells; mixing, for example, the cells from $i = 0$ to $i = 15$ and those from $i = 45$ to $i = 60$, we have $k' = 31$, and the smallest q-value equals $q_0 + q_1 + ... + q_{15} \sim 5 \cdot 10^{-5}$. The variance is then of the order of magnitude 100 and application of the method becomes reasonable. This also makes the asymptotic results apply better to finite samples. On the other hand there is much arbitrariness in such a mixing procedure.

Problem 11. Prove that for $k = 2$ formula (44) reduces to the Laplace-Bernoulli formula (35) of Chapter VI.

Problem 12. Prove that the value of the constant in the asymptotic expression (48) is in agreement with the rule expressed in Eqs. (118) and (119) of Chapter VIII. Note that if X^2 is expressed in terms of n_1,

$n_2, \ldots, n_{k-1}$, the coefficients of the quadric in these $(k-1)$ variables are $a_{ii} = (1/nq_i) + (1/nq_k)$, $a_{ij} = 1/nq_k$ for all $i \neq j$.

Problem 13. The first $n = 608$ figures of the decimal development of π consist of 60 0's, 62 1's, 67 2's, 68 3's, 64 4's, 56 5's, 62 6's, 44 7's, 58 8's, 67 9's. Analyze these data with respect to the assumption that each of the ten digits 0 to 9 has the same probability.[4]

Problem 14. A pair of dice has been thrown $n = 300$ times. The points 2 to 12 (in arithmetical order) appeared 10, 10, 21, 32, 44, 45, 43, 31, 30, 23, 11 times. Check this result with respect to the assumption that the dice are unbiased.

5. X^2-Test if Certain Parameters of the Theoretical Distribution Are Estimated from the Sample

5.1. *Estimates.* We consider a situation commonly encountered in the X^2-test: the $q_1, q_2, \ldots, q_k$ are unknown or they are dependent on some unknown parameters $\theta_1, \theta_2, \ldots, \theta_r$, where the mathematical form of the dependence of the q_1 on the θ_ρ is known.[1] For example, the $q_1, \ldots, q_k$ may be Poisson probabilities whose parameter λ we do not know. We then test the hypothesis that the observed sample has been drawn from a Poisson distribution with λ estimated from the data.

If the true values of the parameters were known, then from (34) we would have to consider

$$X^2 = \sum_{i=1}^{k} \frac{[n_i - nq_i(\theta_1, \ldots, \theta_r)]^2}{nq_i(\theta_1, \ldots, \theta_r)} . \tag{54}$$

In order to be able to apply the test, one has to "estimate" the unknown values of the $\theta_1, \ldots, \theta_r$ from the sample. The q_i are then no longer "given" constants and the considerations and results of Section 4 must be re-examined.

Let us insert a remark regarding the systematics of our subject. As has been said in Section 3.1, in probability theory and in statistics two types of problems appear. In type A, probabilities are given and

[4] Regarding the meaning of the term "probability" in such a problem see a paper by H. Geiringer, quoted in footnote 9, p. 48.

[1] The customary notation which we adopt here is not consistent with our usual distinction between Greek and Latin letters. One should denote an estimate of θ by t, i.e., $t(x_1, x_2, \ldots, x_n)$ or $t(n_1/n, n_2/n, \ldots, n_k/n)$ and use θ for the theoretical value. Then the various estimates would be t', or $\bar{t}$ and the true value simply θ. However, by using $\tilde{\theta}, \theta', \ldots$ we show *which* parameter we are estimating.

observations are made. We want to know the distributions of certain functions of the observations; or, in a more restricted way, we investigate certain properties of these distributions, for example their mean value and variance, or asymptotic properties (laws of large numbers, asymptotic distributions, etc.). From the point of view of probability theory this is the "direct" problem. We were concerned with it in Chapters V and VI and—after some preparations in Chapter VIII—again in the present Chapter IX, and we shall return to it in Chapters XI and XII.

In problem B, the inference problem, the probabilities are considered unknown. We are interested in the distributions of certain functions of these unknown probabilities or in parameters characterizing them; or, we may be satisfied to establish certain properties of these distributions (laws of large numbers, limit distributions, etc.). From the point of view of probability theory, this is the "inverse" problem, while for the statistician it appears as the "direct" approach. We were concerned with it in Chapter VII and it is the subject of Chapter X.

It would, however, be artificial to keep strictly to this classification under all circumstances. The present section deals with estimation and according to the above classification it should belong to Chapter X, Sections 6 and 7, which treat this subject. In particular, in Chapter X, Section 6 we consider the likelihood estimate which will be introduced and discussed in the present section. On the other hand the present section is very closely related to the preceding Section 4. In Section 5.2 of this chapter, we return again to problem A (the main problem of Chapter IX) when we compute the asymptotic distribution of the likelihood estimate, considered as a function of the observations. In Chapter X, we study estimates and in particular likelihood estimates in the framework of Problem B, but Sections 6.3 and 6.4 of Chapter X follow up our present section since it seems natural, after all, to discuss the distribution of an estimate right after the estimate has been established.

Returning now to the X^2-test based on (54), we have to choose a *method of estimation*. A natural method is the *minimum X^2-method* which consists in choosing θ so that the right-hand side of (54) becomes minimum. Equating to zero the r partial derivatives of (54) with respect to the θ_ρ, $\rho = 1, 2, \ldots, r$, we obtain the equations

$$-\frac{1}{2}\frac{\partial X^2}{\partial \theta_\rho} = \sum_{i=1}^{k}\left[\frac{n_i - nq_i}{q_i} + \frac{(n_i - nq_i)^2}{2nq_i^2}\right]\frac{\partial q_i}{\partial \theta_\rho} = 0. \tag{54'}$$

[2] They have been investigated by J. Neyman and E. S. Pearson, "On the use and interpretation of certain test criteria for purposes of statistical inference." *Biometrica* **20A** (1928), pp. 175 and 263.

These equations are, however, quite complicated and in general they are not used.[2]

Much simplification results if we consider the minimum of a modified X^2, which we call X_c^2, where the q_i in the denominator of (54) are taken as constants, c_i. We speak of a X_c^2-*method* or *modified* X^2-*method.* We introduce

$$X_c^2 = \sum_{i=1}^{k} \frac{[n_i - nq_i(\theta_1, ..., \theta_r)]^2}{nc_i}. \tag{55}$$

(We may for example, replace nq_i by n_i.) Differentiation of (55) gives then equations much simpler than (54′):

$$\sum_{i=1}^{k} \frac{n_i - nq_i(\theta)}{c_i} \frac{\partial q_i(\theta)}{\partial \theta_\rho} = 0, \qquad \rho = 1, 2, ..., r. \tag{56}$$

(We write θ as an abbreviation for $\theta_1, \theta_2, ..., \theta_r$.)

A third procedure is to use an estimate derived from the *maximum likelihood method.* (This method will be treated systematically in the next chapter.) We know that

$$\frac{n!}{n_1! n_2! \cdots n_k!} q_1^{n_1} q_2^{n_2} \cdots q_k^{n_k}, \qquad q_i = q_i(\theta_1, \theta_2, ..., \theta_r)$$

is the probability of the joint results $n_1, n_2, ..., n_k$. The product

$$L = q_1^{n_1} q_2^{n_2} \cdots q_k^{n_k} \tag{57}$$

is called the *likelihood* [see Chapter X, Eq. (103)]. One seeks an extremum of L or rather of $\log L$. This leads to the equation

$$\frac{\partial \log L}{\partial \theta_\rho} = \sum_{i=1}^{k} \frac{n_i}{q_i(\theta)} \frac{\partial q_i}{\partial \theta_\rho} = 0, \qquad \rho = 1, 2, ..., r. \tag{58}$$

(Here one does not assume constant q_i in the denominator.) Since, on account of $\Sigma\, q_i = 1$, we have $\Sigma_{i=1}^{k}\, \partial q_i / \partial \theta_\rho = 0$, Eqs. (58) may be written as

$$\sum_{i=1}^{k} \frac{[n_i - nq_i(\theta)]}{q_i(\theta)} \frac{\partial q_i(\theta)}{\partial \theta_\rho} = 0, \qquad \rho = 1, ..., r. \tag{58′}$$

The solutions θ' of (56) and the solutions $\tilde{\theta}$ of (58′) are the *modified* X^2-*estimates* and the *maximum-likelihood estimates,* respectively. We write $X^2(\theta') = X_c^2$, $X^2(\tilde{\theta}) = \tilde{X}^2$. Equations (56) are simpler than (58′).

In either case, we have first to show that the system of r equations with r unknowns has a solution; then this solution has to be found; finally, we have to find the asymptotic distribution of X_c^2 or of $\tilde{X}^2$. A treatment of the problem based on Eqs. (58′) is in Cramér's textbook [4], pp. 425-434. He shows that under certain regularity conditions the limit distribution of $\tilde{X}^2$ is a χ^2-distribution with $k - r - 1$ degrees of freedom. We shall consider both (56) and (58′), omitting occasionally some details, and shall follow the elegant presentation of van der Waerden [28], pp. 189-209.

The structure of the proof is not changed if we assume for simplification that the q_i depend *linearly* on the θ. The results can then be generalized to non-linear differentiable functions by use of Taylor's formula; here we shall examine the linear case only.

Consider first *the X_c^2-method.* Writing x_i and x_i' for n_i and $nq_i(\theta)$ we see that

$$X_c^2 = \sum_{i=1}^{k} \frac{(x_i - x_i')^2}{nc_i} \tag{55'}$$

can be regarded as the square of a weighted distance between the point P with coordinates $x_i = n_i$ (which is given) and the point P' with coordinates $x_i' = nq_i(\theta)$. The equations

$$x_i' = nq_i(\theta), \qquad i = 1, 2, ..., k, \tag{59}$$

define a linear r-dimensional space G and $X_c^2 =$ minimum states that the point P should be as close as possible to P'. Hence (see Fig. 33)

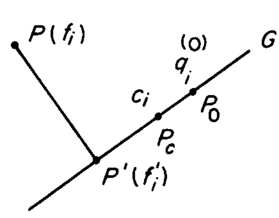

FIG. 33.

P' is the intersection with G of the line through P normal to G. *The problem of minimizing X_c^2 is a problem of least squares.* Consequently, the corresponding system of linear equations is positive definite and has one solution.

Let us set up this system of linear equations. Since we have assumed that the $q_i(\theta)$ are linear functions of θ_ρ we have, with $q_{i\rho} = \partial q_i/\partial \theta_\rho$

$$q_i(\theta) = q_i(0) + \sum_\rho q_{i\rho}\theta_\rho ,$$

where the $q_{i\rho}$ are constants. One can choose the origin in the θ-space such that for $\theta = 0$, q_i equals c_i, the chosen values in (55′). Substituting these $q_i(0) = c_i$ into (56) we obtain the equations, for the θ_ρ,

$$\begin{aligned} \sum_{\rho=1}^{r} h_{\alpha\rho}\theta_\rho &= \sum_{i=1}^{k} \frac{[x_i - nc_i]q_{i\alpha}}{c_i} \\ h_{\alpha\rho} &= n \sum_{i=1}^{k} \frac{q_{i\alpha}q_{i\rho}}{c_i} . \end{aligned} \tag{60}$$

The $h_{\alpha\rho}$ have indeed the form of the coefficients in a system of least-square equations.

Let us see how a change of the arbitrarily chosen c_i influences the point P'. Consider as coordinates of point P not the x_i but the frequencies $x_i/n = f_i$ and put $f_i' = x_i'/n = q_i(\theta)$. We assume that both the point P and the point P_c with coordinates c_i are at a distance of the order $\epsilon = n^{-(1/2)}$ from the point P_0 with the coordinates q_i^0, the unknown true values; all q_i are supposed to be bounded from below: $q_i \geqslant \delta > 0$. We define a "distance"

$$r^2 = \frac{1}{n} X_c^2 = \sum_{i=1}^{k} \frac{(f_i - f_i')^2}{c_i} .$$

If then c_i is replaced by some t_i in the ϵ-neighborhood of q_i^0 the differences $c_i - t_i$ are of the order ϵ. The point P and the space G remain unchanged and the metric defined by r^2 changes at most by values of the order ϵ. The direction of the normal PP' may change by an angle of the order ϵ; since the length of PP' is also of order ϵ the change of the coordinates of P' is of the order $\epsilon^2 = n^{-1}$.

Before proceeding to the next step, the study of the distribution of X_c^2 and of the solution θ' of Eqs. (56) *we consider the likelihood equations* (58). The space S of the q_i defined by the conditions

$$q_i \geqslant 0, \qquad \sum_{i=1}^{k} q_i = 1$$

is closed and bounded and the part of G which lies in S is likewise closed and bounded. The likelihood function (57) is continuous and has thus a maximum on G. This maximum cannot lie on the boundary since there, $L = 0$. Therefore, a solution of Eqs. (58) exists.

The computation of the $\tilde{\theta}$ may be done by iteration.[3] In the denominator of (58) we replace the $\tilde{\theta}_\rho$ by some $\theta_\rho^{(0)}$ [for example $q_i(\theta^{(0)}) = c_i = n_i/n$]. We obtain the equations of the form (56) which are least-square equations just as in the case of $X_c{}^2$ and therefore can be solved; they give values $\theta_\rho^{(1)}$ which may be introduced into the denominator of (58) and a new solution $\theta_\rho^{(2)}$ is obtained, etc. The procedure converges and is independent of the choice of the $\theta_\rho^{(0)}$.

Indeed, consider two different initial values $q_i^{(0)}$ and $t_i^{(0)}$ whose difference is of the order $\epsilon = n^{-1/2}$; then the $q_i^{(1)}$, $t_i^{(1)}$ differ in the order ϵ^2 only, and so on. Now assume that $t_i^{(0)}$ is a solution of the likelihood equations; then $t_i^{(0)} = t_i^{(1)} = t_i^{(2)} = \cdots$ and the sequence $q_i^{(0)}$, $q_i^{(1)}$, ... converges toward $t_i^{(0)}$ independent of the value $q_i^{(0)}$ as long as this value lies in an ϵ-neighborhood of $t_i^{(0)}$. The iterations $q_i^{(\nu)}$, $\nu = 0, 1, \ldots$ converge geometrically toward the solution $\tilde{q}_i$, such that $q_i^{(\nu)} - \tilde{q}_i$ is of the order $\epsilon^{\nu+1}$, if $q_i^{(1)}$ differs from $q_i^{(0)}$ in the order ϵ. It then follows that all $q_i^{(\nu)}$, $\nu > 1$ differ from $q_i^{(1)}$ in the order $\epsilon^2 = n^{-1}$ only. Hence, if the $q_i^{(0)}$-choice is fairly good, *one* iteration (giving $q_i^{(1)}$) may be sufficient.

We have seen before that if in the $X_c{}^2$-procedure we change the initial value c_i to another value in an ϵ-neighborhood, the change of the coordinates of P' is of the order ϵ^2 only. On the other hand, the solution of the $\tilde{X}^2$-equations is nothing but the solution of a succession of least-square equations. If the difference between $q_i^{(1)}$ and $q_i^{(0)}$ is of the order ϵ we have just seen that the difference between the $\tilde{q}_i$ and the $q_i^{(1)}$ is of the order of ϵ^2; since the coordinates of P' for initial values $q_i^{(1)}$ differ from those for initial values $q_i^{(0)}$ in the order of ϵ^2 only, we see that *the estimates* $q_i' = q_i(\theta')$ *and* $\tilde{q}_i = q_i(\tilde{\theta})$ *which correspond to Eqs.* (56) *and* (58), *respectively, differ in the order* n^{-1} *only; therefore the* $X_c{}^2$ *and* $\tilde{X}^2$ *have asymptotically the same distributions.*

5.2. *Asymptotic distributions of* $X_c{}^2$, $\tilde{X}^2$, $\boldsymbol{\theta}'$ *and* $\tilde{\boldsymbol{\theta}}$. We wish now to find *the probability* $P(x)$ that $X_c{}^2 \leqslant x$. We found in Chapter VIII, Section 6.3 that the distribution of k independently and normally $N(\alpha, \sigma^2)$ distributed squares, when both α and σ^2 were given, has the χ^2-distribution with k degrees of freedom. We have seen in Section 4 of the present chapter that when *one* relation exists among the n_i (or the u_i), $i = 1, 2, \ldots, k$, the asymptotic distribution of X^2 is a χ^2-distri-

[3] We do not pretend that this is always the most practical method for actual computation. The iteration method is here of theoretical interest since it shows the relation between the present problem and that of minimum $X_c{}^2$.

bution with $k - 1$ degrees of freedom; similarly, we saw in Chapter VIII, Section 6.4 that the distribution of s^2, the sum of k squares with the mean "estimated from the observations", was a χ^2-distribution with $k - 1$ degrees of freedom. From all this we are led to expect that in the present case, where r parameters are estimated, an analogous reduction of the degrees of freedom will take place.

Consider now with $\tilde{q}_i = q_i(\tilde{\theta})$ (where $\tilde{\theta}$ stands for $\tilde{\theta}_1, \ldots, \tilde{\theta}_r$) the expression

$$\tilde{X}^2 = \sum_{i=1}^{k} \frac{(x_i - n\tilde{q}_i)^2}{n\tilde{q}_i} . \tag{61}$$

If in the denominator, we replace the $\tilde{q}_i$ by some c_i, in particular by the (unknown) true $q_i = q_i^0$, we obtain the X_c^2 of (55′). We have seen in the preceding section that the minimum X_c^2-estimates q_i' differ from the $\tilde{q}_i$ in the order of magnitude n^{-1} only; and this difference will not influence the asymptotic behavior. It will, therefore, be sufficient to establish the asymptotic distribution of X_c^2 and the result will then apply to $\tilde{X}^2$ as well.

As in Section 4, Eq. (46) we introduce new variables u_i such that asymptotically

$$P(x) = \Pr\{X_c^2 \leqslant x\} = \frac{1}{(2\pi)^{(k-1)/2}} \cdot \int \underset{X_c^2 \leqslant x}{\cdots} \int e^{-\frac{1}{2}(u_1^2 + \ldots + u_k^2)}\, d\omega_{k-1} .$$

By an orthogonal transformation, taking the hyperplane $u_1 + \cdots + u_k$ as the kth coordinate plane, we obtain

$$P(x) = \frac{1}{(2\pi)^{(k-1)/2}} \cdot \int \underset{X_c^2 \leqslant x}{\cdots} \int e^{-\frac{1}{2}(u_1^2 + \ldots + u_{k-1}^2)}\, du_1 \cdots du_{k-1} .$$

We then make another orthogonal transformation such that r of the new axes which we denote by $y_1, y_2, \ldots, y_r$ lie in the space G of Eqs. (59) and the other ones are normal to G. Denote the new coordinates of the points P and P', by y_i and y_i', $i = 1, 2, \ldots, k - 1$. Since P' lies in G all coordinates of P' except the first r coordinates must be zero: $y_{r+1}' = 0$, $y_{r+2}' = 0, \ldots, y_{k-1}' = 0$ and X_c^2 becomes

$$X_c^2 = (y_1 - y_1')^2 + \cdots + (y_r - y_r')^2 + y_{r+1}^2 + \cdots + y_{k-1}^2 . \tag{62}$$

Here, only the $y_1', \ldots, y_r'$ contain the parameters θ. Thus, X_c^2 becomes a minimum if θ is such that $y_1 - y_1' = 0, \ldots, y_r - y_r' = 0$. Then, we obtain in these coordinates, the decisive result

$$X_c^2 = y_{r+1}^2 + y_{r+2}^2 + \cdots + y_{k-1}^2 . \tag{62′}$$

Hence (compare Chapter VIII, Section 6.5), the integration limits, given by $X_c^2 \leqslant x$, relate to $y_{r+1}, \ldots, y_{k-1}$ only. Therefore, the integration with respect to the first r variables can be carried out and influences the constant in front of the integral sign only. We have $u_1^2 + \cdots + u_{k-1}^2 = y_{r+1}^2 + \cdots + y_{k-1}^2$ and obtain

$$P(x) = \frac{1}{(\sqrt{2\pi})^{k-r-1}} \int \cdots \int e^{-\frac{1}{2}(u^2_{r+1}+\cdots+u^2_{k-1})}\, du_{r+1} \cdots du_{k-1},$$

with the integration over the region $y_{r+1}^2 + \cdots + y_{k-1}^2 \leqslant x$. The final step which consists of the same computation as in Chapter VIII, Section 6.5, leads to

$$P(x) = \frac{1}{2^\lambda \Gamma(\lambda)} \int_0^x y^{\lambda-1}\, e^{-\frac{1}{2}y}\, dy, \qquad \lambda = \frac{k-1-r}{2}. \tag{63}$$

$P(x)$ is a χ^2-distribution with $(k-1-r)$ degrees of freedom. This result holds also for the distribution of $\tilde{X}^2$, the maximum likelihood estimate.

Let us now investigate the asymptotic distribution of θ' and of $\tilde{\theta}$. In our theoretical investigation we may again assume that the c_i in (55′) and in (60) are the (unknown) true values q_i^0, and, with $q_i^0 = q_i(\theta^0)$, that θ^0, the true value, equals zero: $\theta_\rho^0 = 0$, $\rho = 1, 2, \ldots, r$. Equations (60) for the θ' have then the form

$$\sum_{\rho=1}^{r} h_{\alpha\rho}\theta_\rho' = \sum_{i=1}^{k} \frac{[x_i - nq_i^0]q_{i\alpha}}{q_i^0} \tag{60′}$$

$$h_{\alpha\rho} = \sum_{i=1}^{k} \frac{nq_{i\alpha}q_{i\rho}}{q_i^0}. \tag{60″}$$

Let us compute the expected value of the solutions θ_ρ'. Since the expected values of $x_i - nq_i^0$ are zero it is seen that $E[\theta_\rho'] = 0$, hence $E[\theta_\rho'] = \theta_\rho^0$; *the θ_ρ' are "unbiased" estimates* (see Chapter X, Section 7.1). These unbiased estimates θ_ρ' are linear "statistical functions", since by (60) the θ_ρ' are linear functions of the relative frequencies.

We know that the $x_i - nq_i^0$ are of the order of magnitude of $\sqrt{n}$ in probability and since the $h_{\alpha\rho}$ are of the order n, the θ_ρ' are of the order $n^{-1/2}$ in probability. For the sake of simplicity consider *one* parameter θ, i.e., $r = 1$; this θ' is the parameter of the point P' (see p. 460), the projection of the observed point P onto the straight line G. We wish to find the asymptotic probability that $\theta' \leqslant tn^{-1/2}$. Denote by P_t the point whose θ-value equals $tn^{-1/2}$; through P_t we draw a hyperplane H_t normal to G; the point P must then lie in the half space to the left of H_t.

Next, the rectangular coordinates u_i are orthogonally transformed so that the u_1-axis coincides with G; the point P_t has then the coordinate $u_1 = at$ and the above-mentioned half-space has the equation $u_1 \leqslant at$.

We have to determine the probability of all the lattice points in this half space. In the same way as in Section 4 we approximate the sum by an integral. The computation is similar to that in Section 4 except that the region of integration (which was there $X^2 \leqslant x$) is now that part of the hyperplane H_t which lies in the half space $u_1 \leqslant at$. The integration over the $u_2, u_3, \ldots, u_k$ can be carried out and there remains the integral

$$\frac{1}{\sqrt{2\pi}} \int^{at} e^{\frac{1}{2}u_1^2} \, du_1 = \Phi(at) . \tag{64}$$

The distribution of θ' is asymptotically normal. The mean is the true value θ^0 ($= 0$).

One has still to find the a in the last integral. The result is $a^2 = h_{11}/n$ [h_{11} given by (60″)]. Now $\Phi(at)$ is the probability of $\theta' \leqslant t/\sqrt{n}$; then with $t' = t/\sqrt{n}$ and $a' = a\sqrt{n}$ we have $at = a't'$ and $\Phi(at) = \Phi(a't')$ is the probability of $\theta' \leqslant t'$; hence *a'^2 is the asymptotic variance σ^2* of θ' and we obtain, with definition (60″),

$$\frac{1}{\sigma^2} = h_{11} . \tag{64'}$$

Equation (64) gives the asymptotic variance, σ^2. By asymptotic variance we mean *the variance of the asymptotic distribution of θ'*. (This is not necessarily the same as the asymptotic value of the variance.) Analogous formulas hold in the case of r parameters. Actually, under our assumption of a *linear* space G, hence constant $h_{\alpha\rho}$, Eq. (64′) *for the variance of θ' holds not only asymptotically but for all n.*

To find the asymptotic distribution of $\tilde{\theta}$, we use θ' as an approximation: $\tilde{\theta} = \theta' + \eta$ where $\theta'\sqrt{n}$ is normally distributed with mean value zero and variance independent of n, and $\eta\sqrt{n}$ converges in probability toward zero. It can then be shown without difficulty[4] that *$\tilde{\theta}\sqrt{n}$ is also asymptotically normally distributed with the same mean value and the same variance as $\theta'\sqrt{n}$.* However, $\tilde{\theta}$, unlike θ', is not a *linear* function of the relative frequencies; it is a differentiable statistical function whose asymptotic distribution is the same as that of its linear approximation, i.e., the same as that of θ'. The asymptotic mean value and variance of $\tilde{\theta}$, i.e., *the mean value and variance*[5] *of the asymptotic distribution of $\tilde{\theta}$*, are however not valid for *any n* as in the case of θ'.

5.3. *Applications.* (1) *Two rare events.* Suppose rare events have been observed n_1 times during the time t_1 and n_2 times during t_2. If the observed frequencies n_1/t_1 and n_2/t_2 are different we would like to

---

[4] See the lemma used in Chapter VII, Section 6.2 and Chapter XII, Section 5.2.

[5] A direct study of asymptotic mean value and variance of $\tilde{\theta}$ is in Chapter X, Section 6.4.

know whether this difference may be due to chance i.e., whether there is a single true average $\theta_1 = \theta_2$, or whether $\theta_1 \neq \theta_2$. The expected values of n_1 and n_2, respectively, are $\lambda_1 = \theta_1 t_1$, $\lambda_2 = \theta_2 t_2$. In order to examine the assumption $\theta_1 = \theta_2 = \theta$ we write

$$\lambda_1 = \theta t_1, \qquad \lambda_2 = \theta t_2$$

with unknown θ. The probability of the joint occurrence of the two events is

$$\frac{\lambda_1^{n_1}}{n_1!} e^{-\lambda_1} \frac{\lambda_2^{n_2}}{n_2!} e^{-\lambda_2} = \frac{1}{n_1! n_2!} (\theta t_1)^{n_1} (\theta t_2)^{n_2} e^{-\theta(t_1+t_2)} .$$

The logarithm of $L = \theta^{n_1+n_2} e^{-\theta(t_1+t_2)}$ is $(n_1 + n_2) \log \theta - (t_1 + t_2)\theta$ and

$$\frac{\partial \log L}{\partial \theta} = \frac{n_1 + n_2}{\theta} - (t_1 + t_2);$$

the maximum of L is at $\tilde{\theta} = (n_1 + n_2)/(t_1 + t_2)$.

With this $\tilde{\theta}$ we form, writing X^2 rather than $\tilde{X}^2$:

$$X^2 = \frac{(n_1 - \tilde{\theta} t_1)^2}{\tilde{\theta} t_1} = \frac{(n_2 - \tilde{\theta} t_2)^2}{\tilde{\theta} t_2} . \tag{65}$$

In (65) two independently observed values n_1 and n_2 occur; hence if we knew θ the number of degrees of freedom would be $m = 2$; since we have estimated θ the number of degrees of freedom is $m = 1$. The X^2-test may now be applied.

(2) *Comparison of several probabilities.* We consider a problem with k attributes a_1, $i = 1, 2, \ldots, k$, corresponding unknown probabilities $q_1, q_2, \ldots, q_k$, $\sum_{i=1}^{k} q_i = 1$. We make n_1 observations and obtain n_{1i} times the attribute a_i. Then, we observe the same attributes under different circumstances n_2 times and find that a_i has appeared n_{2i} times; in a last series of n_h observations a_i appeared n_{hi} times, $i = 1, 2, \ldots, k$. We wish to check whether the $q_1, q_2, \ldots, q_k$ have remained the same.

We arrange the hk observed numbers n_{ji}, $j = 1, 2, \ldots, h$, $i = 1, 2, \ldots, k$ in a rectangular scheme

					Sum
	n_{11}	n_{12}	$\cdots$	n_{1k}	n_1
	n_{21}	n_{22}	$\cdots$	n_{2k}	n_2
	.	.		.	.
	.	.		.	.
	.	.		.	.
	n_{h1}	n_{h2}	$\cdots$	n_{hk}	n_h
Sum:	m_1	m_2	$\cdots$	m_k	n

Assuming that the probabilities q_k are constant for the h series of trials we estimate them by the maximum likelihood method, where

$$L = q_1^{n_{11}+\cdots+n_{h1}} q_2^{n_{12}+\cdots+n_{h2}} \cdots q_r^{n_{1k}+\cdots+n_{hk}}$$

$$= q_1^{m_1} q_2^{m_2} \cdots q_k^{m_k} = q_1^{m_1} q_2^{m_2} \cdots (1 - q_1 - \cdots - q_{k-1})^{m_k};$$

$$\log L = m_1 \log q_1 + m_2 \log q_2 + \cdots + m_k \log (1 - q_1 - \cdots - q_{k-1})$$

$$\frac{\partial \log L}{\partial q_i} = \frac{m_i}{q_i} - \frac{m_k}{q_k} = 0,$$

and with $m_k/q_k = \rho$ we see that $m_i/q_i = \rho$ for all $i = 1, 2, \ldots, k$. Adding these equations in the form $m_i = \rho q_i$ we find $\rho = n$ and therefore the obvious estimates $\tilde{q}_i = m_i/n$. Accordingly, we form

$$X^2 = \sum_{j=1}^{h} \frac{(n_{j1} - \tilde{q}_1 m_j)^2}{\tilde{q}_1 m_j} + \sum_{j=1}^{h} \frac{(n_{j2} - \tilde{q}_2 m_j)^2}{\tilde{q}_2 m_j} + \cdots + \sum_{j=1}^{h} \frac{(n_{jk} - \tilde{q}_k m_j)^2}{\tilde{q}_k m_j}. \quad (66)$$

Let us find the number of degrees of freedom. Among the hk observed number n_{ji} there are h relations $n_j = \sum_{i=1}^{k} n_{ji}$. In addition we estimated $k - 1$ probabilities $\tilde{q}_1, \tilde{q}_2, \ldots, \tilde{q}_{k-1}$. Hence the number of degrees of freedom is $hk - h - (k - 1) = (h - 1)(k - 1)$.

These are two from among the many simple applications of the X^2-test. A survey has been given by Cochran[6]. We consider now a case where the estimation of the parameter is more difficult.

(3) *Comparison with a Poisson distribution of unknown parameter.* Suppose that we wish to test the hypothesis that a given sample of n results $x_1, \ldots, x_n$ is drawn from *some* Poisson distribution. We arrange the x_i into k groups, pooling data, if necessary. As always we denote by n_i the number of results equal to a_i or briefly to i, and compare each n_i with nq_i where $q_i = (\lambda^i/i!)e^{-\lambda}$, for $i = 0, 1, \ldots, k - 1$. We assume in the example *one* "pooled" result only. The pooled probability $q_k + q_{k+1} + \cdots$ is compared with n_k/n. Using Eqs. (58) and $dq_i/d\lambda = q_{i-1} - q_i$, we obtain

$$\sum_{i=0}^{k-1} \frac{n_i}{q_i} (q_{i-1} - q_i) + \frac{n_k}{\sum_{i=k}^{\infty} q_i} \sum_{i=k}^{\infty} (q_{i-1} - q_i) = 0. \quad (67)$$

[6] W. G. Cochran, "The χ^2-test." *Ann. Math. Statist.* **23**, p. 328. See further applications in various textbooks, e.g. Cramér [4], pp. 434-450 and our quotations, pp. 458, 460.

Equation (67) has a solution $\tilde{\lambda}$ which we should like to know. As $q_{i-1} = (i/\lambda)q_i$, we obtain from (67)

$$\frac{1}{\lambda}\sum_{i=0}^{k-1} in_i - \sum_{i=0}^{k-1} n_i + n_k \frac{\sum_{i=k}^{\infty} q_i\left(\frac{i}{\lambda} - 1\right)}{\sum_{i=k}^{\infty} q_i} = \frac{1}{\lambda}\sum_{i=0}^{k-1} in_i - \sum_{i=0}^{k} n_i + \frac{1}{\lambda} n_k \frac{\sum_{i=k}^{\infty} iq_i}{\sum_{i=k}^{\infty} q_i} = 0,$$

or

$$\lambda = \frac{1}{n}\left[\sum_{i=0}^{k-1} in_i + n_k \frac{\sum_{i=k}^{\infty} iq_i}{\sum_{i=k}^{\infty} q_i}\right]. \tag{68}$$

Equation (68) is merely Eq. (67) rewritten. The fraction to the right equals $k + (\sum_{\nu=0}^{\infty} \nu q_{k+\nu})/(\sum_{\nu=0}^{\infty} q_{k+\nu})$. With $1/n \sum_{i=1}^{k} in_i = \bar{x}$, Eq. (68) becomes

$$\lambda = \bar{x} + \frac{n_k}{n} \frac{\sum_{\nu=0}^{\infty} \nu q_{k+\nu}}{\sum_{\nu=0}^{\infty} q_{k+\nu}}. \tag{68'}$$

which is still an identity in λ. If we neglect the last term in (68′) we obtain $\hat{\lambda} = \bar{x}$.

We note that $\bar{x}$ is the maximum likelihood estimate $\hat{\lambda}$ for ungrouped data. Indeed, if we wish to determine the value of λ which maximizes

$$L = \frac{\lambda^{x_1}}{x_1!} e^{-\lambda} \cdots \frac{\lambda^{x_n}}{x_n!} e^{-\lambda} = C\lambda^{x_1+\cdots+x_n} e^{-n\lambda}$$

or $\log L = \log C + (x_1 + x_2 + \cdots + x_n) \log \lambda - n\lambda$ we have

$$\frac{d \log L}{d\lambda} = (x_1 + \cdots + x_n)\frac{1}{\lambda} - n, \qquad \hat{\lambda} = (x_1 + \cdots + x_n)/n = \bar{x}.$$

We shall see on p. 469 that the asymptotic distribution of the $\hat{X}^2$ computed with $\hat{\lambda} = \bar{x}$ is not exactly Helmert's distribution with $k - 1 - r = k - 2$ degrees of freedom.

To find $\tilde{\lambda}$ we may consider $\bar{x}$ as a first approximation $\lambda^{(1)}$ and use an iteration method. If in the last fraction in (68′) we consider the terms for $\nu = 0$ and $\nu = 1$ only, we replace the fraction by $q_{k+1}/(q_k + q_{k+1}) = \lambda/(\lambda + k + 1)$ and $\lambda^{(2)} = \lambda^{(1)} + \frac{n_k}{n} \cdot \frac{\lambda^{(1)}}{\lambda^{(1)} + k + 1}$, will be a second approximation to $\tilde{\lambda}$. The numerical data will show whether $\lambda^{(2)}$ and $\lambda^{(1)}$ differ within the limits of the accuracy of the computation and it is a problem of numerical analysis to find $\tilde{\lambda}$.[7]

[7] Regarding the problem of finding $\tilde{\lambda}$, we quote also a paper by C. G. KHATRI, "A method for estimating approximately the parameters of a certain class of distribution from grouped observations." *Ann. Inst. Statist. Math.* **14** (1962), pp. 57-62. It relates specifically to Poisson and Gaussian distributions.

Comparing $\lambda^{(1)} = \bar{x}$ with $\tilde{\lambda}$, we see that $\tilde{\lambda} > \lambda^{(1)}$ and from a certain i on $q_i(\tilde{\lambda}) > q_i(\lambda^{(1)})$; the X^2 computed with $\lambda^{(1)}$ will be too large and we may reject an assumption which might be acceptable on the basis of $\tilde{\lambda}$.

The following rather obvious remark may be added. Suppose that instead of using pooled probabilities $q_k + q_{k+1} + \cdots$, we compare each observed n_i with nq_i, $i = 0, 1, ..., k$, where $q_i = (1/i!)\lambda^i e^{-\lambda}$. The sum of the q_i will then not be unity. If we compute a maximum likelihood estimate by considering $\log L = n_0 \log q_0 + n_1 \log q_1 + \cdots + n_k \log q_k$, then $d(\log L)/d\lambda = 0$ leads to the root $\lambda = \bar{x}$. But the procedure is not justified since now $\sum_{i=0}^{k} q_i < 1$, while $\sum_i q_i = 1$ is one of the basic assumptions of the X^2-test.

(4) We add the following interesting result (which is not limited to a Poisson distribution). We have found that the $\tilde{X}^2$ as well as the X_c^2 have asymptotically Helmert's distribution with $k - r - 1$ degrees of freedom if r parameters are estimated from the sample. If the k results designate k classes (cells) agreed upon in the organization of the material, and the original observations x_i are available, we may want to estimate the parameters from the x_i without using cells i.e., to use a likelihood estimate $\hat{\theta}$ based on all x_i, $i = 1, 2, ..., n$. Also $\hat{\theta}$ is often easier to find than $\tilde{\theta}$ or θ'.

Chernoff and Lehman[8] have studied the distribution of X^2 computed with this (these) $\hat{\theta}$, which we denote by $\hat{X}^2$. Their result[9] is that *the asymptotic distribution of $\hat{X}^2$ is that of*

$$\sum_{i=1}^{k-r-1} y_i^2 + \sum_{i=k-r}^{k-1} c_i y_i^2, \tag{69}$$

where all y_i are independently and normally distributed $N(0, 1)$ *and the* c_i (which are determined by a characteristic-value problem) *are between* 0 *and* 1.

If we compare this result with that for $\tilde{X}^2$ [see Eq. (62′)] and that for K. Pearson's original X^2 where $r = 0$, we see that the distribution of $\hat{X}^2$ is in a way "in between" those two. If we use $\hat{X}^2$ in connection with the rejection rule for $m = k - r - 1$ degrees of freedom we might reject a hypothesis which would be acceptable on the basis of $\tilde{X}^2$. (This is the same conclusion we have reached before in application (3) in an elementary way.) It is obvious that the excess in rejecting will be

[8] H. Chernoff and E. L. Lehman, "The use of the maximum likelihood estimates in χ^2-tests for goodness of fit." *Ann. Math. Statist.* **25** (1954), pp. 579-586.

[9] This result applies to the estimate $\lambda^{(1)} = \bar{x}$ considered on p. 468.

more serious if k is small. The authors show by numerical examples that in the case of a Poisson distribution where $\hat{\lambda} = \bar{x}$, this excess will not be serious. In the case of the Gauss distribution (with unknown α and σ^2) the situation is less favorable.

(5) Consider[10] the first 2035 digits of (π); by that we mean the first 2035 digits after the period of the decimal expansion of π. From these we form $n = 1017$ *non-overlapping* pairs beginning with 14, 15, ... [These pairs form at the same time the first 1017 "digits" of (π) for the basis 100.] We compare with an assumed uniform distribution, where $q_i = 1/100$ for each of the possible 100 pairs. Here $k = 100$ and nothing has been estimated from the observations. Computation gives $X^2 = 100.31$. Expression (43) for the variance of X^2 reduces for a uniform distribution to $\text{Var}[X^2] = 2(k - 1)(1 - (1/n))$. In our case $E[X^2] = 99$, $\text{Var}[X^2] \sim 14$, hence the result is in line with our hypothesis.

If we consider in the same way for the first 2010 digits of (e) the $n = 1005$ pairs beginning with 71, 82, ... we have expectation and variance as before. The computed value of X^2 is 67.14. This is a remarkably small value considering that $99 - \sqrt{198} \sim 85$; the fit is unusually close.[11]

An author considered testing by overlapping pairs, taking for (π) the pairs 14, 41, 15, 59, ..., and accordingly for (e): 71, 18, 82, ... The X^2 computed with $q_i = 1/100$ were evaluated by means of the χ^2-distribution with $k^2 - k = 90$ degrees of freedom, since "there exist $k = 10$ relations between the observed frequencies." This is a wrong procedure: in fact, the original X^2-distribution (Section 4) is established under the assumption that the observed frequencies of the k attributes are results of n "drawings from an urn," containing k labels in a distribution $q_1, q_2, \ldots, q_k$, $\sum_{i=1}^{k} q_i = 1$. In the case of overlapping pairs, however, the elements depend on each other as in a Markov chain. (Make the place selection: "retain only those pairs which follow the digit 4," then for the 10 pairs 40, 41, 42, ..., the limit frequency is $\frac{1}{10}$, for the other 90 pairs it is zero; of course $10 \cdot \frac{1}{10} + 90.0 = 1$ is the same as $100 \cdot \frac{1}{100}$ but the distributions are quite different.) The limit distribution of the X^2 computed from overlapping pairs is *not* the usual χ^2-distribution, i. e., Helmert's distribution. The mean value of the new distribution is actually 99 (and not 90) its variance is larger than

[10] See H. Geiringer, "On the statistical investigation of transcendental numbers." *In Studies in Mathematics and Mechanics*, presented to R. v. Mises, New York, 1954, pp. 310-322, and literature quoted there.

[11] It cannot be concluded that this is a significant feature of the number e since the computation for $n = 3000, 4000$, etc., leads again to values close to expectation.

$2(k - 1)(1 - (1/n))$. The student should be careful not to assume that every sum of squares has, automatically the χ^2-distribution.[12]

D. The ω^2-Tests (Sections 6 and 7)

6. von Mises' Definition

6.1. *Definition. Historical comments.* If the theoretical distribution which we used as the basis of comparison in the X^2-test is a *continuous distribution* $P(x)$, then this test becomes rather arbitrary: one divides the range (a, b) of $P(x)$ into k parts a, x_1; x_1, x_2; x_2, x_3; ...; x_{k-1}, b, and one takes $q_i = P(x_i) - P(x_{i-1})$, $i = 1, 2, ..., k$, $x_0 = a$, $x_k = b$. Apart from the arbitrariness of this procedure, we see that for Pearson's X^2 both the expectation and the number of degrees of freedom tend toward infinity as the number k of parts tends to infinity. We have also seen that in the case of small q_i-values the variance of Pearson's X^2 increases very strongly, while by pooling values, the procedure becomes still more arbitrary; furthermore, computations may become involved.

If we take the weights λ_i [of Eq. (30)] as equal constants, the above discussed difficulty connected with small q-values disappears. Consider, however, the example shown in Fig. 34 where the theoretical distribution P is uniform and arithmetical; I and II are two observed distributions which lead to the same $X_I^2 = X_{II}^2$ (also for Pearson's definition of the λ_i) although I deviates from P systematically while II deviates very little from P.

A test which avoids these disadvantages and which applies without any arbitrariness to both discrete and continuous distributions is obtained if the theoretical distribution function $P(x)$ is compared with the observed distribution, or repartition $S_n(x)$. The $S_n(x)$ is always a step function, $P(x)$ only if it is a discrete distribution. We introduce

$$\epsilon(x) = n[S_n(x) - P(x)], \tag{70}$$

where $S_n(x)$ is defined in the case of n observed values $x_1, x_2, ..., x_n$ as

$$\begin{aligned} &S_n(x) = \frac{\nu}{n}, \qquad x_\nu \leqslant x < x_{\nu+1}, \qquad \nu = 1, 2, ..., n - 1 \\ &S_n(x) = 0, \qquad x < x_1, \qquad S_n(x) = 1, \qquad x \geqslant x_n, \end{aligned} \tag{70'}$$

[12] Relevant to this problem is an article by P. Billingsley, "Statistical methods in Markov chains." *Ann. Math. Statist.* 32 (1961), pp. 12-40, where additional literature is also given.

and

$$\omega_n^2 = \int \lambda(x)\epsilon^2(x)\, dx, \tag{71}$$

where $\lambda(x)$ is a weight function. We shall take it as a constant, since even with this simplest choice ω_n^2 is a comprehensive measure of the deviation.

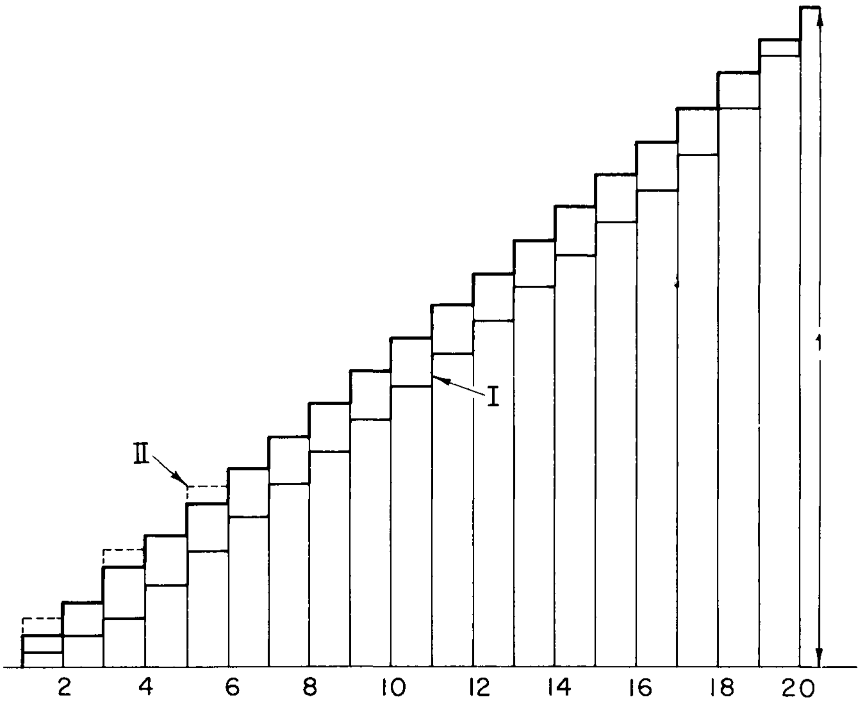

FIG. 34. Two different distributions compared with uniform distribution give same X^2.

The use of ω_n^2 as a test function, defined by (70) and (71) i.e., the ω^2-*criterium*, or ω^2-*test*, was proposed in 1931 by v. Mises,[1] who studied it in detail. The idea of comparing the theoretical and empirical distribution functions was advanced by Cramér (1928)[2]; however, he proposed no test of goodness of fit of the type (71)

[1] v. Mises [21], pp. 316-335.

[2] H. CRAMER, "On the composition of elementary errors," second paper. *Skand. Aktuarietidskr.* (1928), pp. 171-180. (His first paper of the same title has nothing to do with our problem.) Cramér discusses in [4] briefly a particular case of Smirnov's test of 1936; Smirnov, in turn, calls his test function "a slight modification of v. Mises' well-known expression."

or of Pearson's X^2, but considered a *sequence of expressions* I_i, $i = 1, 2, \ldots$, for which he computed the expected values: $P(x)$ is expanded in what is a rearrangement of the Bruns-Charlier series[3] (Chapter III, Section 3):

$$P(x) = \Phi(x) + P_1(\Phi) + P_2(\Phi) + \cdots, \qquad P_1(\Phi) = a_3\Phi^{(3)}(x),$$
$$P_2(\Phi) = a_4\Phi^{(4)}(x) + a_6\Phi^{(6)}(x),$$

etc., where Φ is the Gaussian, $\Phi^{(\nu)}$ its νth derivative. Then *the differences*

$$\Delta_1(x) = S_n(x) - \Phi(x), \qquad \Delta_i(x) = \Delta_{i-1}(x) - P_{i-1}(\Phi), \qquad i = 2, 3, \ldots, \tag{a}$$

are formed and the integrals

$$I_i = \int [\Delta_i(x)]^2\, dx, \qquad i = 1, 2, 3, \ldots, \tag{b}$$

are taken as test functions. The expectations $E[I_i]$ have been computed by Cramér for $i = 1, 2, 3$; for $i = 3$, these are already very long expressions. The variances are not considered.

Smirnov[4] has modified the definition of ω_n^2 by considering $P(x)$ as the independent variable; in this way the test becomes "distribution free." He computed the limit distribution and this distribution has been tabulated. We shall report these and related investigations in the next section.

v. Mises proposed to choose $\lambda(x)$ in (71) so as to have *the expected value of ω_n^2 always equal to one.* Let us compute this expected value. If n is given, then for a fixed $x = \xi$, $P(\xi)$ is a constant probability; $S_n(\xi) = \nu/n$ can take on one of the $n + 1$ values $0, 1/n, 2/n, \ldots, (n-1)/n, 1$. Then $\Pr\{S_n(x) = \nu/n\}$ is the probability that in n trials, each with the two possibilities $x > \xi$ or $x \leqslant \xi$, the result $x \leqslant \xi$ happened ν times. Hence we have a Bernoulli problem with $q = P(\xi)$, $p = 1 - q = 1 - P(\xi)$, $E[\nu] = nP(\xi)$, $E[\epsilon] = 0$, $E[\epsilon^2] = E[(\nu - nP)^2] = nP(\xi)(1 - P(\xi))$, and since this holds for any ξ, we have

$$E[\epsilon^2] = nP(x)(1 - P(x)), \tag{72}$$

We now choose a (constant) λ such that

$$E[\omega_n^2] = \int \lambda E[\epsilon^2(x)]\, dx = 1; \tag{73}$$

using (72) this gives

$$\frac{1}{\lambda} = n \int P(x)(1 - P(x))\, dx. \tag{74}$$

Our ω_n^2 is given by (71) *with λ defined by* (74).

---

[3] See F. Y. EDGEWORTH, "The law of error." *Proc. Cambridge Phil. Soc.* **20** (1905), pp. 36-141.

[4] N. V. S. SMIRNOFF, "Sur la distribution de ω^2 (criterium de M. R. v. Mises)." *Com. Rend. Acad. Sci.* (*Paris*) **202** (1936), pp. 449-452, and a paper of the same title in Russian *Recueil Math.* **2** (1937), pp. 973-993, which gives the proofs.

If, for example $P(x)$ is a uniform distribution between 0 and c:

$$P(x) = \frac{x}{c}, \qquad \text{for} \quad 0 \leqslant x \leqslant c, \tag{75}$$

and equal to 0 and 1 for $x \leqslant 0$ and $x \geqslant c$, respectively, then

$$\frac{1}{\lambda} = n \int_0^c \frac{x}{c}\left(1 - \frac{x}{c}\right) dx = \frac{nc}{6}, \qquad \text{or} \qquad \lambda = \frac{6}{nc}. \tag{75'}$$

Let us rewrite definition (71) for the sake of easier computation. If the observed values are $x_1, x_2, \ldots, x_n$ (each written as often as it appears) the expression for ω_n^2 is, according to (70′)

$$\begin{aligned} \frac{1}{\lambda n^2}\, \omega_n^2 &= \int [S_n(x) - P(x)]^2\, dx \\ &= \int^{x_1} P^2\, dx + \sum_{\nu=1}^{n-1} \int_{x_\nu}^{x_{\nu+1}} \left[\frac{\nu}{n} - P(x)\right]^2 dx + \int_{x_n} (1 - P(x))^2\, dx. \end{aligned} \tag{76}$$

We introduce, for the moment, $Q(x) = \int_0^x P\, dx$. Then expanding the squares, which are of the form [constant $-$ $P(x)]^2$, we obtain by a brief computation

$$\frac{1}{\lambda n^2}\, \omega_n^2 = -\frac{1}{n^2} \sum_{\nu=1}^{n} (2\nu - 1)x_\nu + \frac{2}{n} \sum_{\nu=1}^{n} Q(x_\nu) + \int^0 P^2\, dx + \int_0 (1 - P)^2\, dx$$

and finally

$$\frac{1}{n^2}\, \omega_n^2 = \lambda \left[\int^0 P^2\, dx + \int_0 (1 - P)^2\, dx + \frac{2}{n} \sum_{\nu=1}^{n} \int_0^{x_\nu} P\, dx - \frac{1}{n^2} \sum_{\nu=1}^{n} (2\nu - 1)x_\nu\right], \tag{76'}$$

with the λ from (74). This is the desired formula. Hence, to compute ω_n^2 we need the indefinite integral $Q(x)$ and the definite integrals of P^2 and $(1 - P)^2$.

As a first example we compute ω_n^2 for the uniform distribution (75). There

$$Q(x) = \int_0^x P\, dx = \frac{x^2}{2c}, \qquad \int^0 P^2\, dx = 0, \qquad \int_0 (1 - P)^2\, dx = \frac{c}{3}.$$

Substituting into (76′) and rearranging we obtain

$$\omega_n^2 = \frac{1}{2n} + 6 \sum_{\nu=1}^{n} \left(\frac{x_\nu}{c} - \frac{2\nu - 1}{2n}\right)^2, \qquad \text{where} \quad E[\omega_n^2] = 1. \tag{77}$$

6.2. Variance of ω_n^2. The computation of the variance of ω_n^2 is analogous to that of Pearson's X^2. With definitions (71) and (74) we have [for general $\lambda(x)$]

$$\mathrm{Var}[\omega_n^2] = E[\omega_n^4] - \{E[\omega_n^2]\}^2 = \iint \lambda(x)\lambda(y)E[\epsilon^2(x)\epsilon^2(y)]\,dx\,dy - 1$$

$$= \iint\limits_{x=y} \lambda(x)\lambda(y)E[\epsilon^2(x)\epsilon^2(y)]\,dx\,dy + 2\iint\limits_{x} \lambda(x)\lambda(y)\,E[\epsilon^2(x)\epsilon^2(y)]\,dy\,dx - 1.$$

Here the factor "2" in the second integral is due to the symmetry. Since the first integral is zero, we obtain

$$\mathrm{Var}[\omega_n^2] = 2\iint\limits_{x} \lambda(x)\lambda(y)E[\epsilon^2(x)\epsilon^2(y)]\,dy\,dx - 1. \tag{78}$$

In order to find

$$\iint\limits_{x} E[\epsilon^2(x)\epsilon^2(y)]\,dy\,dx$$

we proceed as in Section 4.2 in the computation of $E(\delta_1^2\delta_2^2]$. The x-axis is divided into 3 parts: $-\infty, x$; x, y; y, ∞. The probability that a result is in the first segment is $P(x)$, and, analogously it is $P(y) - P(x)$ and $1 - P(y)$ for the other two segments. The corresponding frequencies are $nS_n(x) = n_1$, $n[S_n(y) - S_n(x)] = n - n_1 - n_2$ and $n - nS_n(y) = n_2$, respectively. We put

$$P(x) = q_1, \qquad 1 - P(y) = q_2, \qquad P(y) - P(x) = 1 - q_1 - q_2 = q_3 \tag{79}$$

and have then

$$\delta_1 = n_1 - nq_1 = n[S_n(x) - P(x)] = \epsilon(x)$$
$$\delta_2 = n_2 - nq_2 = n[1 - S_n(y) - 1 + P(y)] = n[P(y) - S_n(y)] = -\epsilon(y).$$

As in Section 4.2 we have to compute $E[\delta_1^2\delta_2^2]$ with respect to the multinomial distribution $p_n(n_1, n_2)$. We shall also need $E[\delta_1\delta_2]$. From Section 4.2, we have $E[\delta_1\delta_2] = E[-\epsilon(x)\epsilon(y)] = -nq_1q_2$. Hence

$$E[\epsilon(x)\epsilon(y)] = nq_1q_2 = nP(x)(1 - P(y)), \qquad \text{if} \quad x \leqslant y$$

and, similarly we have

$$E[\epsilon(x)\epsilon(y)] = nP(y)(1 - P(x)), \qquad \text{if} \quad y \leqslant x.$$

The expectation of $\epsilon^2(x)\epsilon^2(y)$ follows from Eq. (39) if we substitute the q_1, q_2 of (79) and use integration instead of summation. Thus, $\mathrm{Var}[\omega_n^2]$ is given by (78) with

$$E[\epsilon(x)^2\epsilon(y)^2] = nq_1q_2[1 + (n-2)(1 - q_1 - q_2 + 3q_1q_2)] \tag{80}$$

[from (39)], and the q_i have to be replaced from (79).

We compute the variance of the uniform distribution. Here we have to substitute from (79) $q_1 = x/c$, $q_2 = 1 - y/c$, $q_3 = (y-x)/c$, and have for the first term

$$\int_0^c \int_x^c \frac{x}{c}\left(1 - \frac{y}{c}\right) dy\, dx = \frac{1}{c^2}\int_0^c x\left[\int_x^c (c-y)\, dy\right] dx$$
$$= \frac{1}{c^2}\int_0^c x\frac{(c-x)^2}{2}\, dx = \frac{c^2}{24}.$$

The other terms follow in the same way and we obtain from (78), using $\lambda = 6/nc$

$$\mathrm{Var}[\omega_n^2] = 2\lambda^2\left[n^2\left(\frac{c^2}{24} - \frac{c^2}{60}\right) - n\left(\frac{c^2}{24} - \frac{c^2}{30}\right)\right] - 1 = \frac{4}{5} - \frac{3}{5n}. \tag{80'}$$

We have therefore the result:

If the observations $x_1 \leqslant x_2 \leqslant \cdots \leqslant x_n$ are compared with a uniform distribution on the segment $x = 0$ to $x = c$, one computes expression (77) *and compares it with*

$$E[\omega_n^2] \pm \sqrt{\mathrm{Var}[\omega_n^2]} = 1 \pm \sqrt{\frac{4}{5} - \frac{3}{5n}}. \tag{80''}$$

6.3. ω_n^2 *for normal distribution.* As a next important application *we take for $P(x)$ the normal distribution*

$$P(x) = \Phi(x) = \frac{1}{\sqrt{2\pi}}\int^x e^{-x^2/2}\, dx. \tag{81}$$

Here, too, all computations can be carried out explicitly.

Remember that with $\Phi' = \phi$, $\phi' = -x\phi$, $\phi^2 = \phi(x\sqrt{2})/\sqrt{2\pi}$ and we obtain by means of integration by parts

$$\int^x \Phi\, dx = x\Phi + \phi, \qquad \int^x \Phi^2\, dx = x\Phi^2 + 2\phi\Phi - \frac{1}{\sqrt{\pi}}\Phi(x\sqrt{2}). \tag{82}$$

which may be verified by differentiation. Subtracting the second of these from the first we have

$$\int \Phi(1-\Phi)\,dx = \frac{1}{\sqrt{\pi}}, \qquad \text{hence} \qquad \lambda = \frac{\sqrt{\pi}}{n}. \tag{83}$$

Next we have from (82), with A any constant

$$\int^x (A-\Phi)^2\,dx = x(A-\Phi)^2 - 2\phi(A-\Phi) - \frac{1}{\sqrt{\pi}}\,\Phi(x\sqrt{2}) + \text{constant}. \tag{82'}$$

As always, the value of $\omega_n{}^2$ is a sum of integrals of the form (82′) since $S_n(x) = \nu/n$ in $(x_\nu, x_{\nu+1})$. The result of the straightforward computation is, with (83) used,

$$\omega_n{}^2 = 2\sqrt{\pi}\sum_{\nu=1}^{n}\left[\phi(x_\nu) + x_\nu\left(\Phi(x_\nu) - \frac{2\nu-1}{2n}\right)\right] - n. \tag{84}$$

This follows also from (76′) (see Problem 15).

In the computation of the variance all integrals can be carried out. We use (80) and compute (as in Section 4.2) for $n=1$ and for $n=2$. *For* $n=2$, the right-hand side of (80) reduces to $2q_1q_2$, where $q_1 = \Phi(x)$, $q_2 = 1-\Phi(y)$. Now

$$\int_x [1-\Phi(y)]\,dy = -x(1-\Phi)+\phi \quad \text{and} \quad \int \Phi\,dx \int_x [1-\Phi(y)]\,dy = \tfrac{1}{2}, \tag{85}$$

since the integral from $-\infty$ to $+\infty$ of the odd function $x\Phi(1-\Phi)$ is zero. Hence with $\lambda^2 = \pi/4$ [see (83)]

$$E[\omega_2{}^4] = \frac{\pi}{4}\cdot 2 = \frac{\pi}{2}, \qquad \operatorname{Var}[\omega_2{}^2] = \frac{\pi}{2} - 1. \tag{86}$$

To find the variance *for* $n=1$, we take from (84) the value of

$$\omega_1{}^2 = 2\sqrt{\pi}[\phi(x) + x(\Phi(x) - \tfrac{1}{2})] - 1; \quad \text{then} \quad \omega_1{}^2 + 1 = 2\sqrt{\pi}(x\Phi + \phi - \tfrac{1}{2}x),$$

and

$$\frac{(\omega_1{}^2+1)^2}{4\pi} = x^2\Phi^2 + \phi^2 + \frac{x^2}{4} + 2x\Phi\phi - x^2\Phi - x\phi. \tag{87}$$

We compute the expected values of these six terms, where each time

$$E[f(x)] = \int f(x)\phi(x)\,dx.$$

Using formulas (82) and partial integrations we find in turn

$$\begin{aligned} &E[\phi] = \frac{1}{2\sqrt{\pi}}, && E[(\phi^2] = \frac{1}{2}\frac{1}{\pi\sqrt{3}}, && E[x\Phi] = \frac{1}{2\sqrt{\pi}}, \\ &E[x\phi\Phi] = \frac{1}{4\pi\sqrt{3}}, && E[x^2\Phi] = \frac{1}{2}, && E[x^2\Phi^2] = \frac{1}{3} + \frac{1}{2\pi\sqrt{3}}. \end{aligned} \tag{88}$$

For example

$$E[x^2\Phi^2] = \int x^2\Phi^2\phi\, dx = -\int x\Phi^2\, d\phi = \int \phi(\Phi^2 + 2x\Phi\phi)\, dx = \frac{1}{3} + \frac{1}{2\pi\sqrt{3}}.$$

Substituting these values into (87) we obtain

$$\frac{E[(\omega_1^2+1)^2]}{\pi} = 4\left(\frac{1}{3} + \frac{1}{2\pi\sqrt{3}}\right) + \frac{2}{\pi\sqrt{3}} + 1 + \frac{2}{\pi\sqrt{3}} - 2 - 0 = \frac{1}{3} + \frac{2\sqrt{3}}{\pi}.$$

$$\begin{aligned} E[(\omega_1^2+1)^2] &= E[\omega_1^4] + 2\cdot 1 + 1 = \pi\left(\frac{1}{3} + \frac{2\sqrt{3}}{\pi}\right) = \frac{\pi}{3} + 2\sqrt{3}, \\ \operatorname{Var}[\omega_1^2] &= E[\omega_1^4] - 1 = \frac{\pi}{3} + 2\sqrt{3} - 4. \end{aligned} \tag{89}$$

Now, if $\operatorname{Var}[\omega_n^2] = \sigma_n^2 = A + B/n$, then $A + B = \sigma_1^2$, $2A + B = 2\sigma_2^2$, and from (86) and (89)

$$\begin{aligned} A &= 2\sigma_2^2 - \sigma_1^2 = \pi - 2 - \frac{\pi}{3} - 2\sqrt{3} + 4 = \frac{2\pi}{3} + 2 - 2\sqrt{3} \\ B &= 2\sigma_1^2 - 2\sigma_2^2 = \frac{2\pi}{3} + 4\sqrt{3} - 8 - \pi + 2 = 4\sqrt{3} - 6 - \frac{\pi}{3}, \\ \operatorname{Var}(\omega_n^2) &= \left(\frac{2\pi}{3} + 2 - 2\sqrt{3}\right) - \frac{1}{n}\left(\frac{\pi}{3} + 6 - 4\sqrt{3}\right) \\ &= 0.6303 - \frac{0.1190}{n}. \end{aligned} \tag{90}$$

To compare[5] *the results* $x_1 \leqslant x_2 \leqslant \cdots \leqslant x_n$ *with a Gaussian distribution* Φ *one computes* ω_n^2 *from* (84) *and compares it with*

$$1 \pm \sqrt{0.63 - \frac{0.12}{n}}.$$

As in our two examples (uniform distribution and normal distribution) the computations may be carried out for other given $P(x)$.

6.4. *The* ω^2*-test for a discrete distribution.* We show, finally, the application of the ω^2-test *to the case of a discrete* $P(x)$. Let $q_1, q_2, \ldots, q_k$ be the probabilities corresponding to $a_1, a_2, \ldots, a_k$ (which are in arithmetical order). Each a_i has happened n_i times, $\sum_{i=1}^k n_i = n$. We introduce the cumulative distributions

$$\begin{aligned} q_1 + q_2 + \cdots + q_i &= Q_i, \quad i = 1, 2, \cdots, k; \quad Q_k = 1, \\ n_1 + n_2 + \cdots + n_i &= N_i, \quad i = 1, 2, \cdots, k; \quad N_k = n. \end{aligned} \tag{91}$$

[5] See v. Mises [21], p. 330.

Then

$$P(x) = Q_i\,, \qquad S_n(x) = \frac{N_i}{n} \qquad \text{for} \quad a_i \leqslant x \leqslant a_{i+1} \tag{92}$$

$$\epsilon(x) = N_i - nQ_i \qquad \text{for} \quad a_i \leqslant x \leqslant a_{i+1} \tag{92'}$$

We assume now that *the a_i are equidistant.* As before, we take $\lambda =$ constant, and such that

$$\frac{1}{\lambda} = n \sum_{i=1}^{k} Q_i(1 - Q_i), \tag{93}$$

so as to have $E[\omega_n{}^2] = 1$. Then with this λ,

$$\omega_n{}^2 = \lambda \sum_{i=1}^{k} (N_i - nQ_i)^2. \tag{94}$$

The variance of $\omega_n{}^2$ is given by Eq. (42) where, with the notation of this formula, q_j is replaced by Q_j and q_i by $1 - Q_i$. We consider now an application where the q_i are estimated from the observations.

Charlier[6] reports measurements of the cranial index of 22,505 Swedish soldiers. The observed values have been divided by him into 13 classes, as follows:

Class:	−6	−5	−4	−3	−2	−1	0	1	2	3	4	5	6
Frequency:	12	87	510	1952	4346	6039	5050	2822	1172	377	94	31	13

and the material is compared with a Gaussian distribution. Figure 35 shows the stepline of the frequencies and the Φ-line. From the observations, a and s^2 have been found which we simply identify[7] with α and σ^2, thus

$$\alpha = -0.7211, \qquad \sigma^2 = 2.3864, \qquad \sigma = 1.5448.$$

From the normal distribution we derive an arithmetical distribution by considering the given class intervals, $(-\infty, -5.5)$, $(-5.5, -4.5)$ $\cdots$ $(-0.5, +0.5)$, ..., $(5.5, +\infty)$. Let $x_i = (i - \alpha)/\sigma$, $i = -5.5, -4.5, \ldots,$

[6] C. V. L. Charlier, *Math. Statist.* Lund (1920), p. 76.

[7] That means we estimate the unknown parameters from the observations in the simplest way. (This is a problem similar to that discussed in Section 5 for the X^2-distribution; Darling [footnote 13], p. 489 has examined it for the ω^2-distribution.) Since n is here very large this method of estimation is adequate. We are not going to use here the distribution of ω^2.

+5.5. The interval Δ between two consecutive x-values is therefore $\Delta = 1/\sigma = 0.64733$.

The constant λ is from (93): $1/\lambda = n \sum_{i=1}^{k} \Phi_i(1 - \Phi_i)$, which we can replace with sufficient accuracy by

$$\frac{1}{\lambda} = n \frac{1}{\Delta} \int \Phi(1 - \Phi)\, dx = n \cdot \sigma \cdot \frac{1}{\sqrt{\pi}},$$

where (83) has been used. The computation is seen in the accompanying table.

Class	x	$\Phi(x)$	nQ_i	N_i	$(nQ_i - N_i)^2$
	$-\infty$	0			
−6			23	12	121
	−3.0935	0.0010			
−5			162	99	3669
	−2.4462	0.0072			
−4			811	609	40804
	−1.79888	0.03605			
−3			2807	2561	60516
	−1.5155	0.1247			
−2			6911	6907	16
	−0.5042	0.3071			
−1			12533	12946	170569
	+0.1431	0.5569			
0			17674	17996	103684
	0.79045	0.7853			
1			20811	20818	49
	1.4378	0.9247			
2			22087	21990	9409
	2.0851	0.9814			
3			22434	22367	4489
	2.7324	0.9968			
4			22496	22461	1225
	3.3798	0.9996			
5			22504	22492	144
	4.0271	0.9999			
6			22505	22505	0
	∞	1			
					3.95×10^5

We find

$$\omega_n^2 = \frac{3.95 \times 10^5}{22{,}505} \cdot \frac{\sqrt{\pi}}{\sigma} = 20.1.$$

The variance of ω_n^2 can be taken with sufficient accuracy from (90); the expected value of ω_n^2 is 1; we see that there is *no* agreement with the normal distribution. In fact the differences $N_i - nQ_i$ show a considerable skewness. The deviations to the left are much larger than those to the right. The graph (Fig. 35) gives a different (erroneous) impression of fit, since, in contrast to the computation, it does not reflect the high value of n.

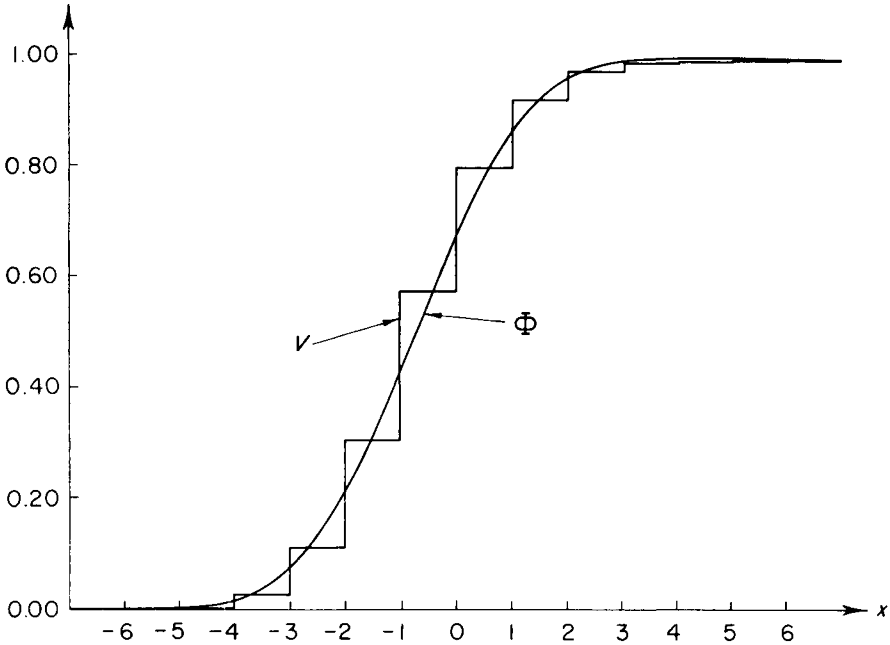

FIG. 35. Normal distribution fitted to curve of cranial index.

Problem 15. Derive the result (84) from the general formula (76′).

Problem 16. Do the last problem in the text by the X^2-method and compare with the result obtained by the ω^2-method.

Problem 17. In a table of random numbers from a uniform distribution on the interval $(-0.5, +99.5)$ the first 10 measurements are 79, 7, 7, 30, 61, 83, 0, 14, 94, 69. Compare with a uniform distribution [Eqs. (77) and (80″)].

Problem 18. Prove that in the case of a uniform arithmetical distribution, $q_1 = q_2 = \cdots = q_k = 1/k$,

$$Q_i = \frac{i}{k}, \qquad E[\omega_n^2] = \frac{k^2 - 1}{6k} \cdot \lambda n.$$

Problem 19. Prove, in preparation for Problem 20,

$$\sum_{i=1}^{k} i = \frac{k(k+1)}{2}, \qquad \sum_{i=1}^{k} i^2 = \frac{k(k+1)(2k+1)}{6},$$

$$\sum_{i=1}^{k} i^3 = \frac{k^2(k+1)^2}{4}, \qquad \sum_{i=1}^{k} i^4 = \frac{k(k+1)(2k+1)}{30}(3k^2+3k-1)$$

$$\sum_{i=1}^{k} i^2(k-i)^2 = \frac{k(k^4-1)}{30}.$$

Problem 20 (difficult). Prove that the variance of ω_n^2 for a uniform arithmetical distribution equals

$$\mathrm{Var}[\omega_n^2] = \frac{4}{5}\frac{2k^2+7}{2k^2-2} - \frac{3}{5n}\frac{k^2+6}{k^2-1}.$$

Problem 21. In the first 50 figures of the decimal expansion of π, the figures 0, 1, ..., 9 appeared with frequencies $n_0 = n_1 = n_4 = n_7 = 5$, $n_2 = n_8 = 8$, $n_3 = n_5 = n_6 = 4$, $n_9 = 2$. Compare with a uniform arithmetical distribution using the results of the preceding problem. Study the same material by means of Pearson's X^2.

7. Smirnov's ω^2-Test

7.1. *Definition.* We turn now to Smirnov's definition. He sets

$$\omega_n^2 = n \int g[P(x)][S_n(x) - P(x)]^2 \, dP(x), \tag{95}$$

which amounts to using $P(x)$, *which is assumed continuous, monotonically increasing*, as the independent variable; $g(t)$ is a positive weight function in $0 \leqslant t \leqslant 1$ and such that $t(1-t)\,g(t)$ has there a continuous derivative. We introduce the variable $y = P(x)$, with the inverse $x = \psi(y)$

$$y = P(x), \qquad x = \psi(y). \tag{96}$$

The observations $x_1 \leqslant x_2 \leqslant \cdots \leqslant x_n$ are transformed into

$$y_\nu = P(x_\nu), \qquad \text{where} \qquad y_1 \leqslant y_2 \leqslant \cdots \leqslant y_n. \tag{97}$$

They have the distribution $T_n(y)$, y in (0, 1)

$$S_n[\psi(y)] = T_n(y), \qquad T_n[P(x)] = S_n(x). \tag{97'}$$

and we obtain

$$\omega_n{}^2 = n \int_0^1 g(y)[T_n(y) - y]^2 \, dy. \tag{95'}$$

The $\omega_n{}^2$ *of* (95) *is transformed, therefore, into a particular* $\omega_n{}^2$*—that of a uniform*[1] *theoretical distribution.* Thus we see that Smirnov's $\omega_n{}^2$ does not depend on the given $P(x)$: it is *distribution free.*

For constant g, we may, for the sake of comparison with v. Mises' $\omega_n{}^2$, choose this constant so that *the expected value of* $\omega_n{}^2$ *will be one.* Using our result (74′), viz $\lambda = 6/n$, we obtain then

$$\omega_n{}^2 = 6n \int [S_n(x) - P(x)]^2 \, dP(x) = 6n \int [T_n(y) - y]^2 \, dy. \tag{95''}$$

If one compares (95) with general $g(x)$ and (71) with general $\lambda(x)$ one finds that there is no formal difference between them. With $\lambda[\psi(y)]\,\psi'(y) = g(y)$, Eq. (71) is transformed into (95′), or by setting $g[P(x)]P'(x) = \lambda(x)$, (95) becomes (71). Actually, however, there are differences: Smirnov's definition is distribution free, a definite advantage in the computation of the distribution of ω^2. On the other hand, it introduces automatically a weight equal to $dP/dx = p(x)$, which *multiplies* the square of the deviations. (In Pearson's X^2 of Section 4, the deviations are *divided* by $p(x)$—there q_i .) If we pass to the second form of (95″) we merely transfer the influence of $P(x)$ onto $T_n(y)$, since we obtain in $T_n(y)$ a distorted representation of the observations. For example, for a $p(x)$ similar to $\phi(x)$ the deviations in the middle are then given great weight and those at the ends are neglected, which is perhaps less desirable than the opposite bias in Pearson's X^2-test. Of course, one can counterbalance the weight $p(x)$ by a suitable choice of g; but if this g is adapted to the $P(x)$ under consideration the $\omega_n{}^2$ would then again depend on $P(x)$—as in v. Mises' definition. Another point is that Smirnov's $\omega_n{}^2$ presupposes a continuous monotonically increasing $P(x)$ while v. Mises' definition holds for any $P(x)$. Let us now study Smirnov's test.

We transform (95″) for actual computation. We need $\omega_n{}^2$ for $P(x) = x$, in (0, 1), hence, we use our result (77). We substitute the y_ν of (97) for the x_ν of (77) and transform finally back to the x_ν . We thus immediately obtain the simple result

$$\omega_n{}^2{}_{\text{(Smirnov)}} = \frac{1}{2n} + 6 \sum_{\nu=1}^{n} \left[P(x_\nu) - \frac{2\nu - 1}{2n}\right]^2.$$

[1] We could also transform it into an $\omega_n{}^2$ with some other theoretical distribution; $P(x) = \Phi(x)$, for example.

Consider the variance of ω_n^2. In all the considerations of Section 6, mean value and variance of ω_n^2 depend on $P(x)$ and on $\lambda(x)$ but not on $S_n(x)$; this is clear since they are expected values of $S_n(x)$ and of functions of $S_n(x)$. Therefore, on account of (95′) the variance of Smirnov's ω_n^2 *will be the same for any* $P(x)$, *dependent, however, on* g. For v. Mises' ω_n^2 with $P(x) = x$ *the variance is given by our result* (80′) *and this result is then valid for Smirnov's* ω_n^2 *with any* P. (To obtain these results our computations of Section 6.2 were needed.) We review:

If in (95), $g =$ constant, then for a continuous distribution $P(x)$ we have, with $g = 6$,

$$\omega_n^2 = 6n \int [S_n(x) - P(x)]^2 \, dP(x) = \frac{1}{2n} + 6 \sum_{\nu=1}^{n} \left[P(x_\nu) - \frac{2\nu - 1}{2n}\right]^2$$
$$E[\omega_n^2] = 1, \qquad \operatorname{Var}[\omega_n^2] = \frac{4}{5} - \frac{3}{5n}. \tag{98}$$

7.2. *Limit distribution of* ω_n^2. *The result.* Smirnov's main achievement is *the determination of the limit distribution* $H(z)$ *of* ω_n^2. It is, of course, distribution free: This $H(z)$ (see footnote 4, p. 473) *is given by the series*

$$H(z) = 1 - \frac{1}{\pi} \sum_{n=1}^{\infty} \int_{\lambda_{2n-1}}^{\lambda_{2n}} \frac{e^{-(\lambda/2)z}}{\sqrt{-D(\lambda)}} \frac{d\lambda}{\lambda}, \tag{99}$$

where $D(\lambda)$ *is the Fredholm determinant of the kernel*

$$\begin{aligned} K(x, y) &= \sqrt{g(x)g(y)}\,x(1 - y), \qquad x \leqslant y \\ &= \sqrt{g(x)g(y)}\,y(1 - x), \qquad y \leqslant x, \end{aligned} \tag{100}$$

and the λ_n *are the eigenvalues of* $K(x, y)$, that is, the eigenvalues of the integral equation

$$f(y) - \lambda \int_0^1 K(x, y) f(x) \, dx = 0. \tag{101}$$

If $g = 1$ the $D(\lambda)$ of (100) is well known. We have then

$$g = 1 : K = \begin{matrix} x(1 - y), & x \leqslant y \\ y(1 - x), & x \geqslant y \end{matrix}, \qquad D(\lambda) = \frac{\sin \sqrt{\lambda}}{\sqrt{\lambda}}, \quad \lambda_n = n^2\pi^2. \tag{102}$$

and (99) is accordingly simplified.

Another form of this main result is that the *characteristic function $h(u)$ of $H(z)$ equals $1/\sqrt{D(2ui)}$, where $D(\lambda)$ is the Fredholm determinant of $K(x, y)$. This* can also be written[2]

$$h(u) = \prod_{\nu=1}^{\infty} \left(1 - \frac{2ui}{\lambda_\nu}\right)^{-1/2}. \tag{103}$$

Smirnov's original derivation of (99) has been replaced by a simpler one due to Kac and Siegert,[3] Doob[4], and Anderson and Darling.[5] We sketch the proof.

7.3. *Derivation.* We have so far considered the random variable $\epsilon(x) = n[S_n(x) - P(x)]$, whose mean value is zero, $E[\epsilon(x)] = 0$, and whose variances are $E[\epsilon(x)^2] = nP(x)(1 - P(x))$, $E[\epsilon(x)\epsilon(y)] = nP(x)(1 - P(y))$ for $x \leqslant y$ and $nP(y)[1 - P(x)]$ for $x \geqslant y$. Now we consider in particular $P(x) = x$ and $T_n(x)$ [see (97′)] instead of $S_n(x)$ and put

$$\delta_n(x) = \sqrt{n}[T_n(x) - x], \qquad 0 \leqslant x \leqslant 1, \tag{104}$$

for which

$$E[\delta_n(x)] = 0, \qquad E[\delta_n(x)\delta_n(y)] = \begin{matrix} x(1 - y), & x \leqslant y \\ y(1 - x), & x \geqslant y \end{matrix}. \tag{105}$$

From the k-dimensional central limit theorem it follows that for any fixed k, and $n \to \infty$, the joint distribution of any k components $\delta_n(x_i)$, $i = 1, 2, \ldots, k$ is asymptotically normal. A one-parameter family of random variables u_t, such that the joint distribution of any finite set of u_t's is Gaussian (multivariate normal), is called a *Gaussian process.*[6]

---

[2] Remember that $(1 - 2ui/\lambda)^{-\frac{1}{2}}$ is the characteristic function of Helmert's distribution for $m = 1$ degrees of freedom. Hence, $H(z)$ has the form of the distribution of an infinite sum of independent χ^2-variates, each with $m = 1$.

[3] M. Kac and A. J. F. Siegert, "An explicit representation of a stationary Gaussian process." *Ann. Math. Statist.* **18** (1947), pp. 438-442.

[4] J. L. Doob, "Heuristic approach to the Kolmogorov Smirnov theorems." *Ann. Math. Statist.* **20** (1949), pp. 393-403. The paper relates to a different problem, but the idea applies to our case.

[5] T. W. Anderson and D. A. Darling, "Asymptotic theory of certain 'Goodness of Fit' criteria based on stochastic processes." *Ann. Math. Statist.* **23** (1952), pp. 193-212. Anderson and Darling make the surprising statement that "the principal innovation in (their) paper is the incorporation of a weight function." Both v. Mises and Smirnov *had* used weight functions. Also Smirnov's distribution is computed by him for general $g(x)$.

[6] See Doob [6], p. 71.

Hence the $\delta(x) = \lim_{n\to\infty} \delta_n(x)$ form a Gaussian process, $0 \leqslant x \leqslant 1$ with

$$E[\delta(x)] = 0, \qquad E[\delta(x)\delta(y)] = k(x,y) = \begin{matrix} x(1-y), & x \leqslant y \\ y(1-x), & y \leqslant x \end{matrix} . \tag{105'}$$

Also $\delta(0) = 0$ with probability one. Doob suggests heuristically that in *calculating asymptotic distributions of $\delta_n(x)$—processes, when $n \to \infty$ one may replace the $\delta_n(x)$-process by the $\delta(x)$-process.*

We wish to find the asymptotic distribution $\lim_{n\to\infty} H_n(z)$, where

$$\begin{aligned} H_n(z) &= \text{Prob}\{\omega_n^2 \leqslant z\} \\ &= \text{Prob}\left\{\int_0^1 \delta_n(x)^2 g(x)\, dx \leqslant z\right\} . \end{aligned} \tag{106}$$

Following the heuristic approach (footnote 4, p. 485), we consider instead of $H_n(z)$

$$H(z) = \text{Prob}\left\{\int_0^1 \delta(x)^2 g(x)\, dx \leqslant z\right\} , \tag{107}$$

and in calculating the asymptotic distribution of the $\delta_n(x)$-process we replace the $\delta_n(x)$-process by the $\delta(x)$-process. This amounts here to assuming that $\lim_{n\to\infty} H_n(z) = H(z)$. This can indeed be justified in our case.[7] Following Smirnov's lead we seek then the characteristic function of $H(z)$, i.e., the expectation with respect to $H(z)$ of $e^{iu\omega^2}$, where $\omega^2 = \int_0^1 \delta(x)^2 g(x)\, dx$.

With $\delta(x)\sqrt{g(x)}$ instead of $\delta(x)$ the $k(x,y)$ on the right-hand side of (105′) is replaced by (100). This kernel is symmetric and positive definite. The corresponding integral equation (101) defines a sequence $\lambda_1, \lambda_2, \ldots,$ of positive eigenvalues and a corresponding complete orthogonal system of eigenfunctions $f_1, f_2, \ldots$. For $K(x,y)$ the well-known bilinear formula holds[8]:

$$K(x,y) = \sum_{j=1}^{\infty} \frac{1}{\lambda_j} f_j(x) f_j(y), \tag{108}$$

[7] M. D. Donsker, "Justification and extension of Doob's heuristic approach" etc. *Ann. Math. Statist.* **23** (1952), pp. 277-281.

[8] See any book dealing with integral equations, for example Ph. Frank and R. v. Mises, *Differentialgleichungen der Physik*, Vol. I, p. 531, 1930. The basic theory is due to Hilbert and E. Schmidt. The convergence of the bilinear formula has been proved in 1923 by A. Hammerstein, under the condition that $K(x,y)$ satisfies a Lipschitz condition.

where the series is absolutely and uniformly convergent. Following Kac and Siegert, we introduce a system $G_1, G_2, \ldots$, of independent normally distributed random variables each having mean 0 and variance 1, and consider the series

$$\sum_{j=1}^{\infty} \frac{1}{\sqrt{\lambda_j}} f_j(x) \cdot G_j . \tag{109}$$

Since for each x the sum of the squares of the coefficients of the G_j converges, on account of

$$\sum_{j=1}^{\infty} \left(\frac{f_j(x)}{\sqrt{\lambda_j}} \right)^2 = \sum_{j=1}^{\infty} \frac{f_j^2(x)}{\lambda_j} = K(x, x),$$

we may conclude that (109) converges "in the mean,"[9] for each x, toward a random variable $z(x)$.[10] We write therefore

$$z(x) = \sum_{j=1}^{\infty} \frac{1}{\sqrt{\lambda_j}} f_j(x) G_j , \tag{109'}$$

and shall show that $z(x)$ is a Gaussian process in (0, 1) with "correlation function" $K(x, y)$. In fact, $E[z(x)] = 0$, since this holds for the G_j. Also $E[G_j^2] = 1$ and we find

$$E[z(x)z(y)] = E\Big[\sum_{j=1}^{\infty} \frac{1}{\lambda_j} f_j(x) f_j(y) G_j^2\Big] = \sum_{j=1}^{\infty} \frac{1}{\lambda_j} f_j(x) f_j(y) = K(x, y). \tag{110}$$

Therefore $z(x)$ is the same stochastic process as $\sqrt{g(x)}\delta(x)$, and

$$\int_0^1 z(x)^2 \, dx = \int_0^1 g(x)\delta(x)^2 \, dx = \omega^2. \tag{111}$$

Next, by Parseval's formula[11] which applies here since the kernel is positive definite and therefore the system of eigenfunctions complete,

$$\int_0^1 z(x)^2 \, dx = \sum_{j=1}^{\infty} \frac{1}{\lambda_j} G_j^2 , \tag{112}$$

[9] See for example Frank and v. Mises, *loc. cit.*, p. 374

[10] Cf. also A. N. Kolmogorov, "Ueber die Summen durch den Zufall bestimmter unabhangiger Grossen." *Math. Ann.*, **99** (1928), pp. 309-319, see p. 314.

[11] See for example Frank and v. Mises, *loc. cit.*, pp. 375 and 376.

and therefore

$$\omega^2 = \sum_{j=1}^{\infty} \frac{1}{\lambda_j} G_j^2 . \tag{112'}$$

We proceed now to compute the characteristic function $h(u)$ of the random variable ω^2. Using (112),

$$\begin{aligned} h(u) &= E[e^{iu\omega^2}] = E\left[\exp\left(iu \sum_{j=1}^{\infty} \frac{1}{\lambda_j} G_j^2\right)\right] \\ &= \prod_{j=1}^{\infty} E\left[\exp\left(\frac{iu}{\lambda_j} G_j^2\right)\right] = \prod_{j=1}^{\infty} \left(1 - \frac{2ui}{\lambda_j}\right)^{-1/2}. \end{aligned} \tag{113}$$

The last equality follows from the fact that G_j^2 has a χ^2-distribution with one degree of freedom (see footnote 2, p. 485) and its characteristic function is $(1 - 2ui/\lambda_j)^{-\frac{1}{2}}$. Hence, we have recovered Eq. (103). Finally, from the well-known product formula $D(\lambda) = \Pi_{j=1}^{\infty} (1 - \lambda/\lambda_j)$, and considering the last expression in (113) *we obtain Smirnov's above quoted result*:

$$h(u) = \frac{1}{\sqrt{D(2ui)}} . \tag{103}$$

To find $H(z)$, this formula must be inverted. As in Chapter V, Section 6.2, one starts with

$$H(z) - H(0) = \frac{1}{2\pi} \int h(u) \frac{1 - e^{-iuz}}{iu} du, \tag{114}$$

which has then to be evaluated. This has been done by Smirnov and leads to the result (99).

It is well known that the integral equation (101) is equivalent to a "first boundary value problem" of an ordinary second order differential equation. In our case, it is particularly simple, namely,

$$y'' + \lambda g(x)y = 0, \qquad y(0) = y(1) = 0, \tag{115}$$

corresponding to the problem of the vibrating string with variable mass $g(x)$ (see e.g. Frank and v. Mises, *loc. cit.*, p. 503).

7.4. *Comments.* (a) *Special cases. Tabulation.* (1) Consider the weight function $g = 1$. In this case a solution of (115) with $y(0) = 0$ is $y = \sin \sqrt{\lambda}x$ and this satisfies $y(1) = 0$ if $\sin \sqrt{\lambda} = 0$ or $\lambda_j = j^2\pi^2$,

$f_j = \sin j\pi x$, and $D(\lambda) = (\sin \sqrt{\lambda})/\sqrt{\lambda}$. For this particular problem, where

$$h(u) = \left(\frac{\sqrt{2iu}}{\sin\sqrt{2iu}}\right)^{1/2}, \tag{116}$$

Anderson and Darling have worked out the inversion (in a way which differs from that of Smirnov) and have obtained $H(z)$ in the form of a very well converging series. The authors have tabulated $H(z)$ (see p. 203 of footnote 5, p. 485). Our Table VIII gives $H(z)$ for the ω_n^2 of Eq. (98) which is normed so as to have the expected value one:

$$H(z) = \lim_{n\to\infty} \text{Prob}\left\{6n \int_0^1 [T_n(y) - y]^2\, dy \leqslant z\right\}.$$

The same authors consider the weight function

$$g(x) = [x(1 - x)]^{-1}.$$

The kernel $K(x, y) = \sqrt{x(1 - y)/y(1 - x)}$, $x \leqslant y$, etc., is known. They worked out the characteristic function and computed some significant values of $H(z)$.[12]

(b) *The ω^2-test in the case of unknown parameters.* As in the X^2-test (see Section 5) the problem arises that $P(x)$ may not be completely given; only its form may be known while some parameters have to be estimated. The problem has been studied by Darling.[13] It would lead too far to discuss this problem here.

(c) *Some general results of v. Mises.* In 1947 v. Mises took up the problem of the limiting distribution of statistics like ω_n^2 from a general point of view, considering ω_n^2 as well as Pearson's X^2 as examples of "statistical functions of second class" which were defined by him. Roughly speaking these "statistical functions" are functions of $S_n(x)$, for which the *linear* approximation vanishes. He proved[14] that *for functions of this type, the limit law is in general, of the type of Smirnov's result* with some corresponding kernel: *The characteristic function $h(u)$ of the limit distribution always equals $1/\sqrt{D(\lambda)}$, $\lambda = 2ui$, where $D(\lambda)$ is the Fredholm determinant of the kernel.* From $h(u)$ the distribution can in principle, be found.

[12] T. W. Anderson and D. A. Darling, "A test of goodness of fit." *J. Amer. Statist. Assn.* **49** (1954), pp. 765-769.

[13] D. A. Darling, *Ann. Math. Statist.* **26** (1955), pp. 1-20.

[14] R. v. Mises, "On the asymptotic distribution of differentiable statistical functions." *Ann. Math. Statist.* **18** (1947), pp. 309-348; "Théorie et applications des fonctions statistiques." *Rend. Mat. e appl.* Ser. V. **11** (1952), pp. 1-37.

For v. Mises' ω_n^2 (Sect. 6) the kernel is

$$\begin{aligned} K(x, y) &= \sqrt{\lambda(x)\lambda(y)}[P(x) - P(x)P(y)], \qquad x \leqslant y \\ &= \sqrt{\lambda(x)\lambda(y)}[P(y) - P(x)P(y)], \qquad y \leqslant x. \end{aligned} \tag{117}$$

In the case of Pearson's X^2, $D(\lambda)$ reduces to the determinant of a quadratic form in k variables. For any particular $P(x)$, the characteristic function of the limit distribution of v. Mises' ω_n^2 is therefore determined.

The reader will notice that there was nothing in our proof of Smirnov's result which would not work for v. Mises' ω_n^2, with kernel (117), and the $\delta_n(x)$ of (104) replaced by $\sqrt{n}[S_n(x) - P(x)]$. We obtain in this way a proof of v. Mises' result regarding his ω_n^2 (found by him in a completely different way). This is not surprising since, formally, there is no difference between the definitions of v. Mises and of Smirnov (see Sect. 7.1).

E. Deviation Tests (Section 8)

8. On the Kolmogorov-Smirnov Tests

8.1. *Difference between empirical and theoretical distributions.* The problem of the deviation between an empirical and a theoretical distribution has been approached by Kolmogorov[1] in a way different from Smirnov's and v. Mises' ω_n^2-methods. He considers *the maximum of all deviations* $| S_n(x) - P(x) |$, where $P(x)$ is continuous:

$$D_n = \sup_{-\infty < x < +\infty} | P(x) - S_n(x) | . \tag{118}$$

The one-sided deviations

$$D_n^+ = \sup_{-\infty < x < \infty} [P(x) - S_n(x)], \tag{119}$$

and

$$D_n^- = \sup_{-\infty < x < \infty} [S_n(x) - P(x)], \tag{120}$$

are also considered. The D_n, D_n^+, D_n^- are random variables and we are interested in their distribution. It is obvious that a monotonic transformation of x will not change the differences. As before we may use $[y - T_n(y)]$ instead of $P(x) - S_n(x)$. Or we may from the beginning take x for $P(x)$. We write $S_n(x)$ rather than $T_n(y)$.

---

[1] A. N. Kolmogorov, "Determinazione empirica di una legge di distribuzione." *Giorn. Ist. Ital. Attuari* 4 (1933), pp. 83-91.

Considering first D_n^+ we write

$$D_n^+ = \sup_{0 \leqslant \lambda \leqslant 1} [\lambda - S_n(\lambda)]. \tag{119'}$$

The distribution of D_n^+ has been determined, for any n, by Birnbaum and Tingey.[12] *The result is*

$$Q_n = \Pr\{D_n^+ > \epsilon\} = \sum_{\nu=0}^{[n(1-\epsilon)]} \binom{n}{\nu} \epsilon \left(\epsilon + \frac{\nu}{n}\right)^{\nu-1} \left(1 - \epsilon - \frac{\nu}{n}\right)^{n-\nu}. \tag{121}$$

Here $[n(1 - \epsilon)]$ denotes the greatest integer contained in $n(1 - \epsilon)$. Previous to this result is Smirnov's asymptotic distribution[3] of D_n^+, which reads

$$Q_n \sim e^{-2n\epsilon^2}. \tag{122}$$

This is a one-sided test to be applied in the usual way. The hypothesis that P is the true d.f. is rejected if the greatest observed difference $P - S_n$ is greater than ϵ, where ϵ is determined from (121) or (122), as follows: a certain β is choosen (to be identified with Q_n), for example $\beta = 5\,\%$ or $10\,\%$, or $1\,\%$; then we find from (122), $\epsilon = \sqrt{(-\log \beta/2n}$ and, for finite n, the ϵ corresponding to $Q_n = \beta$ is tabulated, from (121). For example for $n = 10$, $\beta = 5\%$, we find $\epsilon = 0.37$ by the correct formula, and $\epsilon = 0.39$ by the asymptotic one; for $n = 50$, $\beta = 5\,\%$, $\epsilon = 0.17$ by both formulas (see Table IX).

If we replace x by $1 - x$, the differences change sign only and we obtain a one-sided test in the other direction

$$D_n^- = \sup_{0 \leqslant \lambda \leqslant 1} [S_n(\lambda) - \lambda]. \tag{120'}$$

The probability that $D_n^- > \epsilon$ is again given by (121) and (122). Note that (122) with $\epsilon = \lambda/\sqrt{n}$ is equivalent to

$$\lim_{n \to \infty} \Pr\{\sqrt{n}\, D_n^+ > \lambda\} = \lim_{n \to \infty} \Pr\{\sqrt{n}\, D_n^- > \lambda\} = e^{-2\lambda^2} \tag{122'}$$

A two-sided test is Kolmogorov's D_n-test. For the same ϵ and n as

[2] Z. W. BIRNBAUM and F. H. TINGEY, "One-sided confidence contours for distribution functions." *Ann. Math. Statist.* **22** (1951), pp. 592-596.

[3] N. V. SMIRNOV, "Sur les escarts de la courbe de distribution empirique." (Russian, French summary) *Mat. Sb.* **48** (1939), pp. 3-26. Here the author gives also simplified proofs of Kolmogorov's result (123).

before, the error bound is now obviously $\leqslant 2\beta$. For large n Kolmogorov found the limit distribution $\mathscr{L}(z) = \lim_{n\to\infty} \mathscr{L}_n(z)$ of $\sqrt{n}\, D_n$, namely,

$$\lim_{n\to\infty} \Pr\{\sqrt{n}\, D_n > \lambda\} = 2\sum_{r=1}^{\infty} (-1)^{r-1} e^{-2r^2\lambda^2} = 1 - \mathscr{L}(\lambda), \qquad (123)$$

or equivalently

$$\lim_{n\to\infty} \Pr\{D_n > \epsilon\} \sim 2\sum_{r=1}^{\infty} (-1)^{r-1} e^{-2r^2 n\epsilon^2}. \qquad (123')$$

The assumption that a certain $P(x)$ is the d.f. of an observed $S_n(x)$ is then rejected if the observed maximum of $| P - S_n |$ is greater than ϵ where ϵ follows from tables of $\mathscr{L}(z)$.[4] (See Tables X and XI.)

The series (123′) converges rapidly and as an approximation one may use its first term only, namely $2e^{-2n\epsilon^2}$, which is twice the right-hand side of (122); this corresponds to the above-mentioned rule of having 2β as maximum error probability.

8.2. *Two-sample test.* A related problem studied by Smirnov[5] is the classical *problem of two samples*, which has been investigated by various authors and various methods. Smirnov considers the maximum difference of two observed sample distributions which are assumed to have the same theoretical d.f. $P(x)$. Let $x_1, x_2, \ldots, x_n; y_1, y_2, \ldots, y_m$ be the two samples, and we assume that all observations are mutually independent; one notices certain differences between the x and the y and wants to test whether they might be due to chance so that the two samples *can be ascribed* to the same $P(x)$. This is a very general situation since nothing is assumed regarding $P(x)$, except continuity.

The problem is mathematically similar to Kolmogorov's where we compared $S_n(x)$ with $P(x)$; now we compare two empirical distributions $S_n(x)$ and $R_m(x)$ and put

$$D_{mn} = \sup_{-\infty<x<\infty} | S_n(x) - R_m(x) |,$$

[4] The asymptotic values have been tabulated by Smirnov, 1939, and amplified in N. V. Smirnov, "Table for estimating the goodness of fit of empirical distributions." *Ann. Math. Statist.* **19** (1948), pp. 279-281. Exact values based on Kolmogorov's original work have been tabulated by Z. W. Birnbaum, "Numerical tabulation of the distribution of Kolmogorov's statistic." *J. Amer. Statist. Assn.* **47** (1952), p. 431 ff.

[5] N. V. Smirnov, "On the estimation of the discrepancy between empirical curves of distribution for two independent samples." *Bull. Math. Univ. Moscow* **2** (1939), pp. 3-14.

or

$$D_{mn} = \sup_{0 \leqslant \lambda \leqslant 1} \left| \frac{\nu_n(\lambda)}{n} - \frac{\nu_m'(\lambda)}{m} \right|, \tag{124}$$

where $\nu_n(\lambda)$ is the number of x's which are $\leqslant \lambda$, and $\nu_m'(\lambda)$ is the corresponding number of y's. If we introduce $N = mn/(m + n)$, the asymptotic distribution of $\sqrt{N}\, D_{m,n}$ is the same as that of $\sqrt{n}\, D_n$. More precisely, Smirnov has proved: *If* $m \to \infty$, *in a way that* $m/n \to a$ (*a* constant), *then, with* $N = mn/(m + n)$

$$\lim_{n \to \infty} Pr\, \{\sqrt{N}\, D_{mn} > \lambda\} = 1 - \mathscr{L}(\lambda), \tag{125}$$

where $\mathscr{L}(\lambda)$ *is the same as in* (123).

The application is similar as before. The hypothesis is rejected if the observed difference between the two samples is greater than a certain bound. Take, for example, $\beta = 0.08$ we look for the value of λ in table X such that $\mathscr{L}(\lambda) = 0.92$ and find $\lambda = 1.27$; let then, for example $n = 30$, $m = 40$, then $N^{\frac{1}{2}} = \sqrt{mn/(m + n)} = 4.14$ and $1.27/4.14 = 0.31$; if the observed maximum $D_{m,n}$ surpasses 0.31 (which corresponds to the previous ϵ) we reject the hypothesis; the analogous value, if $\beta = 0.05$, is $1.36/4.14 = 0.33$. Again one may consider the first term as an approximation to the complete series.

These theorems can be proved in a similar way as the result regarding the distribution of ω_n^2 (see the papers quoted in Sect. 7 of Kac, Siegert, Doob, Anderson, Darling). The new proofs are much simpler than the original proofs of Smirnov and Kolmogorov. Smirnov derived his and Kolmogorov's results from general deep results regarding the number of intersections of $S_n(x)$ and $P(x) \pm \epsilon/\sqrt{n}$, and of $S_n(x)$ and $R_m(x) \pm \epsilon/\sqrt{N}$ respectively. A useful survey article by Darling[1] shows the various ramifications of the problems.

Problem 22. From a table of $\mathscr{L}(z)$ we find $z = 1.22$, $\mathscr{L}(z) = .898104$; $z = 1.23$, $\mathscr{L}(z) = .902972$. For various $n = 5, 10, 15, 20, \ldots 40, 50, \ldots 100$, construct a table with error probability $\beta = 10\,\%$ for the two sided Kolmogorov test.

[1] A. Darling, "The Kolmogorov-Smirnov, Cramér-v. Mises tests." *Ann. Math. Stat.* *28* (1957), pp. 823-838.