CHAPTER VIII

MORE ON DISTRIBUTIONS

A. Sample Distribution and Statistical Parameters (Sections 1-3)

1. Repartition

1.1. *The problem.* In Chapter II we learned that any one-dimensional distribution can be completely represented by the so-called cumulative distribution function (c.d.f.) which indicates, for each value of x, the proportion of the distributed quantity lying at points with abscissas $\leqslant x$. Such a function $P(x)$ is monotonic, not decreasing, starting with the value 0 and ending with the value 1. If $P(x) = \int^x p(x)\, dx$, we customarily speak of a continuous distribution and call the derivative $p(x)$ the density at the point x. If $P(x)$ is discontinuous, jumping at discrete points and remaining constant between two consecutive jumps, we have the case of a discrete distribution with lumped, concentrated values of the distributed quantity. A distribution may also be partly discrete, partly continuous. All these terms were originally borrowed from the field of mechanics with mass or weight distributions. They have been applied in the first part of this book to distributions of probabilities (or chances) which appeared in both forms, as continuous and as discrete probabilities.

Now, in this second part, which deals with theoretical statistics based on probability theory, the concept of distribution will be applied concurrently to two different subjects. Let n real numbers $X_1, X_2, \ldots, X_n$ (not necessarily all different) be given. If x is any real quantity, there must exist an integer n_x indicating how many of the numbers $X_1, X_2, \ldots, X_n$ are either smaller than or equal to x. The quotient n_x/n is a function of x with the following properties: it is never decreasing; its value is 0 for sufficiently small x (x smaller than all given X_i) and equal to 1 for sufficiently large x (x greater than all X_i). Moreover, it is constant in each interval that includes none of the X_i and it has a jump equal to $1/n$ (or a multiple of it) at each point whose abscissa x coincides with one (or several) of the given quantities. We denote this function n_x/n by $S_n(x)$ and call it the (cumulative) *frequency distribution* or the *repartition* of the n quantities $X_1, X_2, ..., X_n$. Obviously $S_n(x)$ is a cumulative distribution function. It is a particular case of the discontinuous type, its distinction being that all jumps are integral multiples of $1/n$. We also use the term *sample distribution* $S_n(x)$.

The repartition $S_n(x)$ does not depend on the ordering of the given X_i. Besides, $S_n(x)$ does not change if the number n is, say, doubled (tripled, etc.) and each of the original X_i appears twice (three times, etc.) in the new set.

In some cases the set of n numbers X_i may be given by direct enumeration. For example, the number of deaths occurring in a certain country during a year may be given for a succession of 20 years. Here, in general, the 20 numbers will all be different. In other cases, however, the same number might occur very often. Assume for instance that we throw a die 100 times and note each time the outcome of the throw. Since each time one of the six numbers 1, 2, ..., 6 must appear, it is obvious that repetitions must occur. In order to compute the repartition $S_n(x)$, it is only necessary to know how many times each of the six quantities appears. In a general way, instead of the explicit enumeration $X_1, X_2, ..., X_n$, it is sufficient to give the possible attributes $x_1, x_2, ..., x_k$ and the corresponding frequencies, i.e., to give the data in the form

$$\begin{array}{l} n_1 \text{ times the value } x_1 \\ n_2 \text{ times the value } x_2 \\ \vdots \qquad\qquad\qquad\quad \vdots \\ n_k \text{ times the value } x_k\,, \end{array}$$

where, clearly, $n_1 + n_2 + \cdots + n_k = n$. If the $x_1, x_2, ..., x_k$ are in arithmetical order and x_κ is the largest among them not surpassing x, we have $S_n(x) = (n_1 + n_2 + \cdots + n_\kappa)/n$.

In many statistical problems, we shall have to deal simultaneously with a repartition and a corresponding probability distribution. Take, for instance, the above-mentioned case of repeated throwings of a die. For any definite number n of trials, we have a repartition of the n results expressed by the c.d.f. $S_n(x)$ with jumps of magnitude n_1/n, n_2/n, ..., n_6/n at the points $x = 1, 2, ..., 6$. At the same time there exists a set of six probabilities $p(1), p(2), ..., p(6)$ which in turn determine the c.d.f. $P(x)$ which has, at the same abscissas 1, 2, ..., 6, jumps of magnitude $p(1), p(2), ..., p(6)$, respectively. The two c.d.f.'s $S_n(x)$ and $P(x)$ then describe the sample distribution and the probability distribution respectively. The n results $X_1, X_2, \ldots, X_n$ are the results of n independent trials each subject to the same distribution $P(x)$.

1.2. *Sample mean and sample variance.* What was said in Chapters I, II, and III on distributions in general and on discrete distributions in particular applies now to sample distributions also; e.g., the *sample mean* or *mean* or *average* can be defined as the Stieltjes integral $\int x\, dS_n(x)$ and each of the following forms is an equivalent definition:

$$a = \int x\, dS_n(x) = \frac{1}{n}(X_1 + X_2 + \cdots + X_n) = \frac{1}{n}(n_1 x_1 + n_2 x_2 + \cdots + n_k x_k). \tag{1}$$

The notations $\bar{X}$ or $\bar{x}$ for the sample mean are often used also.

In a general way, if $f(x)$ is any continuous function of x, one will have

$$\begin{aligned} \int f(x)\, dS_n(x) &= \frac{1}{n}[f(X_1) + f(X_2) + \cdots + f(X_n)] \\ &= \frac{1}{n}[n_1 f(x_1) + n_2 f(x_2) + \cdots + n_k f(x_k)]. \end{aligned} \tag{2}$$

In particular the *sample variance*, which is sometimes also called the *dispersion* of the sample, can be written in various forms:

$$\begin{aligned} s^2 &= \int (x-a)^2\, dS_n(x) = \int x^2\, dS_n(x) - a^2 \\ &= \frac{1}{n}\sum_{\nu=1}^{n}(X_\nu - a)^2 = \frac{1}{n}\sum_{\kappa=1}^{k} n_\kappa (x_\kappa - a)^2 \\ &= \frac{1}{n}\sum_{\kappa=1}^{k} n_\kappa x_\kappa{}^2 - \frac{1}{n^2}\left(\sum_{\kappa=1}^{k} n_\kappa x_\kappa\right)^2. \end{aligned} \tag{3}$$

The positive square root of s^2 is called the *standard deviation.*

As a general principle of notation, we state: whenever a probability (chance) distribution and a corresponding sample distribution occur in the same problem, we shall use Latin letters for functions of the sample distribution; such functions are often termed *statistics.* We shall use the corresponding Greek letters for the corresponding functions of the probability distribution and denote them often as *parameters.* Take, for instance, sample mean a and sample variance s^2 as compared to mean value α and variance σ^2:

$$\begin{aligned} a &= \int x\, dS_n(x), & \alpha &= \int x\, dP(x), \\ s^2 &= \int (x-a)^2\, dS_n(x), & \sigma^2 &= \int (x-\alpha)^2\, dP(x). \end{aligned}$$

Sometimes one also refers to a, s^2 as the "empirical" and to α, σ^2 as the "theoretical" values of mean and variance, etc. In the discussion of properties of distributions, when it is not specified whether the sample or probability distribution are in question, we shall often use the letter $V(x)$ to denote the c.d.f. Thus $\int x\, dV(x)$, $\int (x - a)^2\, dV(x)$, etc., will be mean value, variance, etc., in general.

Note on numerical computation. In computing the mean and variance of a repartition, the method of the "provisory mean" or "working origin" which makes use of the shift-of-origin rule (Chapter III) is often useful. If c is some constant and a the mean value, one can easily see that

$$a = \int (x - c)\, dS_n(x) + c \int dS_n(x) = \int (x - c)\, dS_n(x) + c\,, \tag{4}$$

$$\begin{aligned} s^2 = \int (x - c)^2\, dS_n(x) - (a - c)^2 &= \int (x - c)^2\, dS_n(x) \\ &- \left[\int (x - c)\, dS_n(x)\right]^2 . \end{aligned} \tag{4'}$$

If, for instance, the given X are

$$27.49, \quad 26.91, \quad 27.15, \quad 27.44, \quad \ldots,$$

one might choose $c = 27.00$ so as to have values of $X - c$

$$0.49, \quad -0.09, \quad 0.15, \quad 0.44, \quad \ldots .$$

Here, only two-digit numbers have to be squared in order to obtain s^2. If the original definition were used, one would have to compute the squares of the $X - a$ where a should be computed to at least three or four decimal figures, which would in turn make $X - a$ a four-digit number. Further techniques are concerned with the *grouping* of data. Suppose the data are such that one does not distort the pertinent information by grouping them. One introduces *class intervals* or *cells* and assumes, in general, that all observations belonging to a class interval are concentrated at the midpoint of that cell. This influences the values of both a and s^2, which are then a' and s'^2. Consideration of the error thus committed leads to the often used *Sheppard's correction* for s'^2.

Problem 1. In the years 1927–1941, the numbers of suicides in New York State were as follows:

1856, 1932, 2135, 2345, 2503, 2711, 2357,
2216, 2180, 2088, 2122, 2253, 2207, 2266, 1844.

Compute the mean and the dispersion.

Problem 2. Compute the mean and the dispersion for the sequence of last digits of the four-digit telephone numbers appearing in column 2 of a typical page of the local telephone directory.

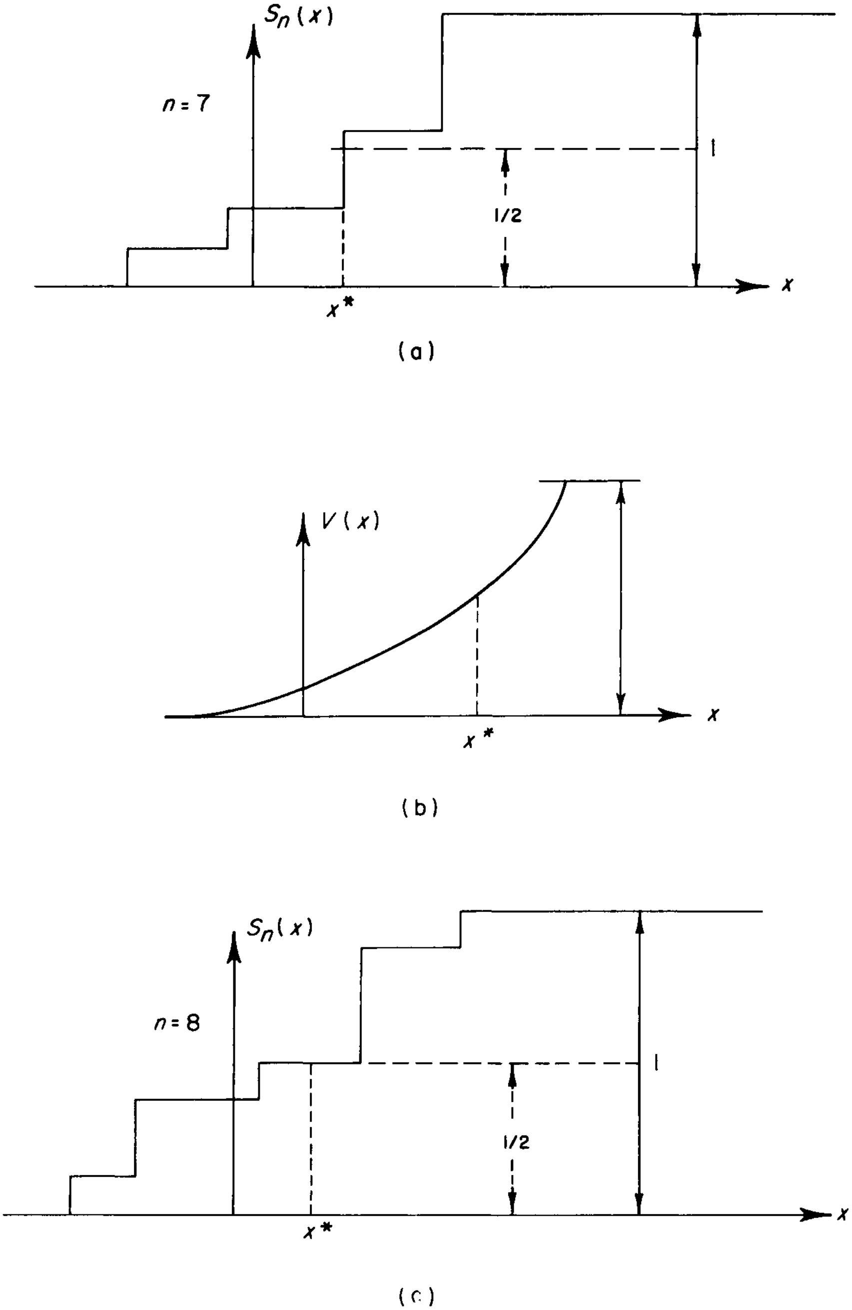

FIG. 22. Median for discrete cases with odd and even n and for continuous case.

2. Some Statistical Parameters[1]

2.1. *Quantiles.* In addition to mean value and variance, various kinds of statistical parameters can be used to characterize a distribution $V(x)$. The term parameter here means a quantity which is uniquely determined by a distribution $V(x)$. On a higher level of mathematical analysis, one might consider such a parameter as a function of $V(x)$, regarding $V(x)$ as a "variable." Such a function will be defined, in general, only for a certain "domain" of $V(x)$, e.g., for all repartitions or for all continuous $V(x)$ whose derivatives (densities) vanish at $x = \pm\infty$ to a certain order. The variance may be defined, for example, for all discontinuous distributions with a finite number of jumps and for those continuous distributions whose derivatives vanish as the power $|x|^{-3-\epsilon}$, $\epsilon > 0$ [so as to ensure convergence of $\int x^2\, dV(x)$].

A comprehensive class of statistical parameters consists of the values $\int f(x)\, dV(x)$ where $f(x)$ is a continuous function of x. In particular, if $f(x)$ is of the form $(x - c)^m$, we obtain the *moments* of order m. (See Chapter III, Section 4.) A more detailed study of moments will follow in Sections 4 and 5 of the present chapter.

A statistical parameter of a quite different type is the so-called *median*, x^*. (See Chapter III, Section 1.1.) It is primarily defined for repartitions with an uneven $n = 2m - 1$. If the $X_1, X_2, \ldots, X_n$ are arranged in arithmetical order, then the mth among the X-values is called the median. In the case of an even $n = 2m$, the term median is often used to denote *the arithmetical mean of the* mth *and the* $(m + 1)$th *of the* X *when arranged in arithmetical order* (see Fig. 22c).

It seems perhaps more logical to denote as median any number x^* for which

$$V(x^*) = \tfrac{1}{2}\,. \tag{5}$$

We see that every $V(x)$ has at least one median. (This does not hold for the mean.) In Figs. 22a and 22b, there is exactly one median. If the $V(x)$-line and the line $y = \frac{1}{2}$ have a whole horizontal piece in common, as in Fig. 22c [or in a continuous $V(x)$-line with a horizontal piece], the median is indeterminate.

We know that the second moment of a distribution reaches its minimum if taken about the mean, $\int (x - \bar{x})^2\, dV(x) = \text{minimum}$. Similarly, *the first absolute moment of a distribution reaches its minimum if taken about the median*:

$$E[|x - x^*|] = \int |x - x^*|\, dV(x) = \text{minimum}. \tag{6}$$

[1] The term "statistical parameters" is used (in this section) for either statistics or parameters, in line with our use of $V(x)$.

In fact, let x_0 be any quantity; we set $\int_{-\infty}^{x_0} = \int_{-\infty}^{x^*} + \int_{x^*}^{x_0}$, $\int_{x_0}^{\infty} = \int_{x_0}^{x^*} + \int_{x^*}^{\infty}$, and $x - x_0 = x - x^* + x^* - x_0$, and obtain

$$E[|\, x - x_0 \,|] = E[|\, x - x^* \,|] + 2\int_{x_0}^{x^*} (x - x_0)\, dV + (x^* - x_0)\left[\int_{x^*}^{\infty} dV - \int_{-\infty}^{x^*} dV\right].$$

The last term on the right-hand side is zero by definition of x^*. The second term $\int_{x_0}^{x^*} (x - x_0)\, dV$ is $\geqslant 0$ no matter whether x^* is to the right or left of x_0. The minimum of $E[|\, x - x_0 \,|]$ is therefore reached if this term vanishes, i.e., for $x_0 = x^*$.

In the indeterminate case we obtain the same value of the absolute moment for *any* median. Hence the indeterminary remains. Consider the arithmetical distribution $p(1) = 2/16$, $p(2) = p(3) = 3/16$, $p(4) = 0$, $p(5) = p(6) = 4/16$; all values $3 \leqslant x < 5$ are medians and for all these values $3 + y$, $0 \leqslant y < 2$, we obtain $E[|x - 3 - y|] = 27/16$. If we agree that in an arithmetical case the median should be one of the actual discontinuity points, X_i, we obtain as a unique answer the discontinuity point to the left, here $x^* = 3$.

Similarly, the so-called *lower* and *upper quartiles* of a repartition are introduced. Arrange the n X-values in arithmetical order; compute $n/4$ and $3n/4$. If either one is an integer m, then X_m is the respective quartile. If $n/4$ is not an integer and m', $m``$ are the two neighboring integers, the arithmetical mean of $X_{m'}$ and $X_{m``}$ may be designated as the lower quartile. In analogy to Eq. (5), the quartiles x' and x'' are defined as the solutions of

$$V(x') = \tfrac{1}{4}, \qquad V(x'') = \tfrac{3}{4}. \tag{5'}$$

The explanations given for the median apply here with some obvious modifications.

In a quite analogous way, so-called *percentage points* are sometimes used to describe the repartition of a large number of statistical data. They correspond, for example, to the solutions of

$$V(x_1) = 0.1, \quad V(x_2) = 0.2, \ldots, V(x_9) = 0.9, \tag{5''}$$

whereby x_5 coincides with the median. The median, quartiles, deciles, and percentiles are called *quantiles*.

Another statistical parameter sometimes mentioned in the literature is the so-called *mode*. Suppose the X are given in the form

$$n_1 \text{ times } x_1, \quad n_2 \text{ times } x_2, \ldots, n_k \text{ times } x_k,$$

then the x_κ with the greatest n_κ, if it is unique, is the mode. It is the most frequent of the x-values. If the greatest n_κ appears several times (e.g., if all n_κ equal 1), the mode is not defined. In the case of a continuous distribution with c.d.f. $V(x)$, one might take for the mode the point where the density dV/dx reaches a maximum, provided it is unique. A distribution with a unique mode is called *unimodal.*

2.2. *Skewness and kurtosis. Disparity measure.* Two parameters derived from moments, the *skewness* and the *kurtosis*, were introduced in Chapter III, p. 129. The quantities S, skewness, and the *excess*, closely connected with the kurtosis have been introduced by Charlier,[2] in the form

$$S = \frac{m_3}{2s^3}, \qquad E = \frac{1}{8}\left(\frac{m_4}{s^4} - 3\right),$$

where m_3, m_4 denote moments around the mean of the distribution $V(x)$. These measures differ slightly from the ones used by Pearson and given in Chapter III, which do not contain the factors $\frac{1}{2}$ and $\frac{1}{8}$. Charlier also introduced a measure dependent on the fifth moment, viz.,

$$T = -\frac{1}{120}\left(\frac{m_5}{s^5} - 10\,\frac{m_3}{s^3}\right).$$

The standard deviations or standard errors (see pp. 370 and 383) of these measures were first computed by Charlier under the hypothesis that the distribution is approximately normal. They are *at most* of the order $n^{-1/2}$.

Lindeberg[3] has introduced different measures for skewness and excess.[4] If $X_1, X_2, \ldots, X_n$ are the observed values and if r denotes the number of these X_i which are greater than the empirical mean a, the Lindeberg skewness is

$$S_L = 100\frac{r}{n} - 50.$$

Further, if R denotes the number of those X_i which fall between the limits $a \pm \frac{1}{2}\, s$ the quantity

$$E_L = 100\frac{R}{n} - \frac{100}{\sqrt{2\pi}}\int_{-1/2}^{1/2} e^{-t^2/2}\,dt = 100\frac{R}{n} - 38.29$$

---

[2] CHARLIER, *Arkiv mat.* **9** (1914), pp. 12–13.

[3] J. W. LINDEBERG, "Über die Begriffe Schiefe und Exzess in der mathematischen Statistik." *Skand. Aktuarietidskr.* (1925), p. 106.

[4] See also H. CRAMÉR, "On the composition of elementary errors," II. *Skand. Aktuarietidskr.* (1928), pp. 141–180, see p. 163.

has been proposed by Lindeberg as a measure of excess. The standard errors of these quantities are $30/\sqrt{n}$ and $42/\sqrt{n}$, respectively. The numerical calculations are simple.

Some other parameters, in particular those that apply to distributions in more than one dimension (Chapter XI), will be introduced later. Here is an example of a different type of parameter which proves useful for certain purposes.

Assume that the yearly income of a great number of people is tabulated. The X are supposed to be non-negative. One is interested in having a measure of the "uniformity" or "non-uniformity" of income. If all X are equal, i.e., all people have the same income, the uniformity is complete. The opposite extreme would be that one person has all the income and all others have nothing, i.e., all X except one are zero. Thus we are looking for a quantity depending on a distribution $V(x)$ in such a way that it takes extreme values in the two cases mentioned and intermediate values in all other cases. A possible choice for a *disparity measure* is

$$d = \frac{\int_0^\infty V(1 - V)\,dx}{\int_0^\infty (1 - V)\,dx}. \tag{7}$$

In fact, this d is non-negative since both V and $1 - V$ are non-negative and cannot exceed the value 1 since the factor V in the numerator is $\leqslant 1$ and $1 - V$ is non-negative. Thus, the extremes of d are 0 and 1. The expression d is independent of the unit of the abscissas. If all people have the same income $\bar{x}$, the corresponding c.d.f. is

$$\begin{aligned} V(x) = S_n(x) &= 0 \qquad \text{for} \quad x \leqslant \bar{x} \\ &= 1 \qquad \text{for} \quad x > \bar{x}, \end{aligned}$$

i.e., for any x either V or $1 - V$ are zero and, therefore, in this case, and only in this case, $d = 0$. If, on the other hand, one person has the income x_1 and all other $n - 1$ persons have zero, the c.d.f. is

$$\begin{aligned} V(x) = S_n(x) &= \frac{n-1}{n} \qquad \text{for} \quad x \leqslant x_1 \\ &= 1 \qquad\quad\ \ \text{for} \quad x > x_1 \end{aligned}$$

and (7) gives

$$d = \frac{\frac{1}{n}\left(1 - \frac{1}{n}\right)x_1}{\frac{1}{n}x_1} = 1 - \frac{1}{n},$$

that is, d approaches one for large n.

The denominator in (7) is seen to equal the mean value a:

$$\int_0 (1 - V)\,dx = (1 - V)x\Big]_0^\infty - \int_0 x\,d(1 - V) = \int_0 x\,dV, \tag{8}$$

and therefore Eq. (7) can also be written

$$d = \frac{\int_0 V(1 - V)\,dx}{a} = 1 - \frac{\int_0 (1 - V)^2\,dx}{a}. \tag{7'}$$

For another form of d, applying in the case where $V(x)$ is replaced by $S_n(x)$, see Problem 6. The disparity coefficient (7) was introduced by Gini.[5]

Problem 3. Give the median, quartiles, and deciles for the statistics of Problem 1.

Problem 4. What are the median and the quartiles in the case of the normal (Gaussian) distribution?

Problem 5. Find mean and median for the distribution with density proportional to x^3e^{-x}, $x \geqslant 0$.

Problem 6. Prove that if the X are arranged in arithmetical order, i.e., $X_1 \leqslant X_2 \leqslant X_3 \leqslant \cdots \leqslant X_n$, the disparity defined in (7) equals

$$d = \frac{1}{n^2 a}\sum_{\nu=1}^{n} (2\nu - 1 - n)X_\nu$$

and if n_κ out of the n equal x_κ with $x_1 < x_2 < x_3 < \cdots < x_k$, then

$$d = \frac{1}{n^2 a}\sum_{\kappa=1}^{k} n_\kappa[2(n_1 + n_2 + \cdots + n_{\kappa-1}) + n_\kappa - n]x_\kappa .$$

Problem 7. The per capita income (dollars) in the six New England states were in 1929: Connecticut 918, Maine 566, Massachusetts 897, New Hampshire 652, Rhode Island 851, Vermont 601, and in 1944: Connecticut 1509, Maine 1059, Massachusetts 1299, New Hampshire 882, Rhode Island 1279 and Vermont 981. Compute the disparity coefficients for both years.

3. Expectations and Variances of Sample Mean and Sample Variance

3.1. *Expectations of average and dispersion.* We take now the first step in the probabilistic analysis of observed statistical data.

---

[5] C. Gini, *Ist. Veneto Sci., Lettere Arti* **73**, II, pp. 1203–1248. Gini has studied this and similar measures also in papers published in *Metron*.

Let

$$x_1, \quad x_2, \quad x_3, \ldots, x_n \tag{9}$$

be a set of statistical data,[1] e.g., the numbers of deaths or of marriages that occurred in the United States in n consecutive years, or the sex ratio of newborn children (male births/total number of births) observed in some country during n months, etc. We compute the average a and the dispersion s^2:

$$a = \frac{1}{n}\sum_{\nu=1}^{n} x_\nu, \qquad s^2 = \frac{1}{n}\sum_{\nu=1}^{n}(x_\nu - a)^2 = \frac{1}{n}\sum_{\nu=1}^{n} x_\nu{}^2 - a^2. \tag{10}$$

In order to analyze the set of data from the standpoint of probability theory, we make the assumption that each $x_\nu(\nu = 1, 2, \ldots, n)$ is the outcome of a certain "game of chance" or, quite generally, the result of some procedure to which the rules of probability calculus apply. For example, in the case of the sex ratio of births, we may suppose that at each single birth the sex of the newborn is the result of a gamble with two label values, "male" and "female," and that x_ν is the observed relative frequency of male births in an m_ν-times repeated alternative with an event probability q_ν. In general, it is assumed that for each $\nu = 1, 2, \ldots, n$ there exists a certain probability distribution with the d.f. $P_\nu(x)$ indicating the probability of obtaining an x_ν-value smaller than or equal to x. Accordingly, $P(x_1, \ldots, x_n)$ is the joint distribution function of $x_1, x_2, \ldots, x_n$, with marginal distribution $P_\nu(x_\nu)$, $P_{\nu\mu}(x_\nu, x_\mu)$, etc. Then, with respect to the expectation of any function $F(x_1, x_2, \ldots, x_n)$ or of functions $F(x_\nu)$, or of functions $F(x_\nu, x_\mu)$, the formulas of Chapter III, Section 8.8 hold. For example,

$$E[F(x_\nu, x_\mu)] = \iint F(x_\nu, x_\mu)\, dP_{\nu\mu}(x_\nu, x_\mu). \tag{11}$$

With $F(x_1, \ldots, x_n)$, $G(x_1, \ldots, x_n)$ two c.d. functions we have, as in Chapter III,

$$E[aF + bG] = aE[F] + bE[G], \tag{12}$$

with a and b constants.

An often plausible assumption will be that the n "games" leading to $x_1, x_2, \ldots, x_n$ can be considered independent. Then the multiplication rule applies: the combined probability of getting a set of values $x_1, x_2, \ldots, x_n$ will be subject to a distribution function with the element

[1] We use now $x_1, \ldots, x_n$ rather than $X_1, \ldots, X_n$ as on p. 368. There we had in mind the distinction between the n direct observations $X_1, \ldots, X_n$ and their arrangement in a frequency table where the label x_i, $i = 1, 2, \ldots, k$ appeared n_i times.

$dP_1(x_1)\, dP_2(x_2) \cdots dP_n(x_n)$. The above results hold with obvious modifications. For example, $dP_\nu(x_\nu)\, dP_\mu(x_\mu)$ take the place of $dP_{\nu\mu}(x_\nu, x_\mu)$.

We proceed to compute the expectations and the variances of the two quantities a and s^2, average and dispersion, defined in (10), as functions of the variables $x_1, x_2, ..., x_n$. We first compute the expectations and consider as known the mean value and the variance of x_ν, $\nu = 1, 2, ..., n$, namely,

$$\alpha_\nu = \int x\, dP_\nu(x), \qquad \sigma_\nu^2 = \int (x - \alpha_\nu)^2\, dP_\nu(x) = \int x^2\, dP_\nu(x) - \alpha_\nu^2, \tag{13}$$

where $P_\nu(x)$ is the marginal distribution of $P(x_1, x_2, ..., x_n)$. Then, with E denoting the expectation with respect to $P(x_1, x_2, ..., x_n)$, and with $F = x_\nu$ or x_ν^2, we have

$$E[x_\nu] = \int x\, dP_\nu(x) = \alpha_\nu, \qquad E[x_\nu^2] = \int x^2\, dP_\nu(x) = \sigma_\nu^2 + \alpha_\nu^2. \tag{14}$$

Now, applying (12) we find

$$E[a] = \frac{1}{n}\sum_{\nu=1}^{n} E[x_\nu] = \frac{1}{n}(\alpha_1 + \alpha_2 + \cdots + \alpha_n). \tag{15}$$

In words, *the expectation of the average of* $x_1, x_2, ..., x_n$ *is the average of the mean values* α_ν. In the case of independence, $P(x_1, ..., x_n) = P_1(x_1) \cdots P_n(x_n)$; $P_\nu(x)$ is, of course, the individual distribution of x_ν.

The second definition (10) leads to

$$\begin{aligned} E[s^2] &= \frac{1}{n}\sum_{\nu=1}^{n} E[x_\nu^2] - E[a^2] \\ &= \frac{1}{n}\sum_{\nu=1}^{n} (\sigma_\nu^2 + \alpha_\nu^2) - \frac{1}{n^2} E[(x_1 + x_2 + \cdots + x_n)^2]. \end{aligned}$$

Here we assume independence, $P(x_1, ..., x_n) = P_1(x_1) \cdots P_n(x_n)$. Let us compute $E[x_\kappa x_i]$. We have

$$\iint x_\kappa x_i\, dP_\kappa(x_\kappa)\, dP_i(x_i) = \alpha_\kappa \alpha_i, \qquad i \neq \kappa. \tag{16}$$

Observing that

$$(x_1 + x_2 + \cdots + x_n)^2 = \sum_{\nu=1} x_\nu^2 + 2\sum_{i<\kappa}^{1\ldots n} x_i x_\kappa$$

we have

$$E[(x_1 + x_2 + \cdots + x_n)^2] = \sum_{\nu=1}^{n} (\sigma_\nu^2 + \alpha_\nu^2) + 2 \sum_{i<\kappa}^{1\ldots n} \alpha_i\alpha_\kappa = \sum_{\nu=1}^{n} \sigma_\nu^2 + \Big(\sum_{\nu=1}^{n} \alpha_\nu\Big)^2 . \tag{14'}$$

If we introduce this into the equation for $E[s^2]$, we obtain

$$E[s^2] = \Big(\frac{1}{n} - \frac{1}{n^2}\Big) \sum_{\nu=1}^{n} \sigma_\nu^2 + \frac{1}{n}\sum_{\nu=1}^{n} \alpha_\nu^2 - \Big(\frac{1}{n}\sum_{\nu=1}^{n} \alpha_\nu\Big)^2 .$$

By the shift of origin rule, the last two terms combined are equal to the dispersion of the n quantities α_ν, namely,

$$\frac{1}{n}\sum_{\nu=1}^{n} \alpha_\nu^2 - \alpha^2 = \frac{1}{n}\sum_{\nu=1}^{n} (\alpha_\nu - \alpha)^2 \qquad \text{with} \qquad \alpha = \frac{1}{n}\sum_{\nu=1}^{n} \alpha_\nu . \tag{17}$$

Thus the final expression for $E[s^2]$ is, independence assumed,

$$E[s^2] = \frac{n-1}{n}\frac{1}{n}\sum_{\nu=1}^{n} \sigma_\nu^2 + \frac{1}{n}\sum_{\nu=1}^{n} (\alpha_\nu - \alpha)^2 . \tag{18}$$

The expectation of the dispersion s^2 of $x_1, x_2, \ldots, x_n$ equals $(n-1)/n$ times the average of the variances σ_ν^2 plus the dispersion of the mean values α_ν of the individual distributions $P_\nu(x)$.

If all $P_\nu(x)$ are equal (or have at least the same α and σ^2), Eq. (18) becomes

$$E[s^2] = \frac{n-1}{n}\sigma^2 , \tag{18'}$$

and, in this case, it follows from (15) that $E[a] = \alpha$. If a sample of n is taken from one and the same distribution, the expectation of s^2 is not equal to the variance of the probability distribution, but is smaller by the ratio $(n-1)/n$. (In fact, if we had $n = 1$, there would be only one sample value; the dispersion is necessarily zero, and so is its expectation.) Equation (18′) can also be written as

$$\sigma^2 = E\left[\frac{n}{n-1}s^2\right] = E\left[\frac{1}{n-1}\sum_{\nu=1}^{n} (x_\nu - a)^2\right] . \tag{19}$$

Some statisticians draw from this formula the conclusion that one should define the dispersion of n quantities as the sum of the n squared deviations divided by $(n - 1)$ instead of n. While it is true that this would simplify Eq. (18′), it would make the factor $n/(n - 1)$ reappear in the second term to the right of Eq. (18) and, likewise, in many other places. We see no reason for modifying our definition.

We have computed so far the expectations of a and of s^2. Expectations of higher sample moments can be found in a similar way (see Problem 13, p. 384) and will be considered as needed.

3.2. *Variances of average and dispersion.* If the observed values of a and of s^2 are close to their computed expectations, we shall feel that the hypothesis about the data $x_1, x_2, ..., x_n$, namely, the assumption regarding the probability distribution $P(x_1, x_2, ..., x_n)$, may have been justified. But what does "close to" mean? How is the deviation of the observed value from the expected one to be appraised? A partial answer to these questions is supplied when, in addition to the expectations $E[a]$ and $E[s^2]$, the variances of these quantities are also computed. We shall first carry out the computation and shall then discuss its use.

We first assume the x_ν to be independent. Then

$$\mathrm{Var}[a] = \mathrm{Var}\Big[\frac{1}{n}\Big(\sum_{\nu=1}^{n} x_\nu\Big)\Big] = \frac{1}{n^2}\mathrm{Var}\Big[\sum_{\nu=1}^{n} x_\nu\Big] = \frac{1}{n^2}\sum_{\nu=1}^{n}\mathrm{Var}[x_\nu] .$$

Hence

$$\mathrm{Var}[a] = \frac{1}{n^2}\sum_{\nu=1}^{n}\sigma_\nu^2 , \tag{20}$$

which gives σ^2/n in the case of equal $P_\nu(x)$.

Next we consider the case of a general joint distribution $P(x_1, x_2, ..., x_n)$:

$$\mathrm{Var}[a] = E[(a - E[a])^2] ,$$

$$\begin{aligned} &E[(x_1 + x_2 + \cdots + x_n - \alpha_1 - \alpha_2 - \cdots - \alpha_n)^2] \\ &\quad = E\Big[\sum_{\nu=1}^{n}(x_\nu - \alpha_\nu)^2 + 2\sum_{\nu<\mu}^{1\ldots n}(x_\nu - \alpha_\nu)(x_\mu - \alpha_\mu)\Big] \\ &\quad = \sum_{\nu=1}^{n}\sigma_\nu^2 + 2E\Big[\sum_{\nu<\mu}(x_\nu - \alpha_\nu)(x_\mu - \alpha_\mu)\Big] \\ &\quad = \sum_{\nu=1}^{n}\sigma_\nu^2 + 2\sum_{\nu<\mu} M_{\nu\mu} , \end{aligned}$$

where $M_{\nu\mu}$ is the second moment with respect to the mean of $P_{\nu\mu}(x_\nu, x_\mu)$ (also called the *covariance* of the random variables x_μ and x_ν). Hence

$$\mathrm{Var}[a] = \frac{1}{n^2}\sum_{\nu=1}^{n} \sigma_\nu{}^2 + \frac{2}{n^2}\sum_{\nu<\mu}^{1\ldots n} M_{\nu\mu} . \tag{20'}$$

If the variables x_μ, x_ν are pairwise independent, then $M_{\nu\mu} = 0$ and (20′) reduces to (20).

The variance of s^2 shall be evaluated here only under the restriction to equal and independent distributions $P_\nu(x)$. In addition to α and σ^2, we need the moment of fourth order with respect to the mean, which will be called τ^4:

$$\int (x-\alpha)^4\, dP(x) = \tau^4 . \tag{21}$$

Using for brevity the notation $x_\nu' = x_\nu - \alpha$, we state

$$E[x_\nu'] = 0, \qquad E[x_\nu'^2] = \sigma^2, \qquad E[x_\nu'^4] = \tau^4,$$
$$E[x_\nu'^2 x_\mu'^2] = \sigma^4, \qquad \text{for} \quad \nu \neq \mu . \tag{22}$$

Besides, the expectation vanishes for all products which contain at least one variable to the first power, e.g., $E[x_1'x_2'] = 0$, $E[x_1'x_2'^2] = 0$, etc. on account of the independence and since $E[x_\nu'] = 0$. Now

$$\mathrm{Var}[s^2] = E[s^4] - \{E[s^2]\}^2 . \tag{23}$$

Here s^2 can be written if the shift-of-origin rule is used:

$$s^2 = \frac{1}{n}\sum_{\nu=1}^{n} x_\nu'^2 - \frac{1}{n^2}\Big(\sum_{\nu=1}^{n} x_\nu'\Big)^2 = \frac{n-1}{n^2}\sum_{\nu=1}^{n} x_\nu'^2 - \frac{2}{n^2}\sum_{\nu<\mu}^{1\ldots n} x_\nu' x_\mu' . \tag{10'}$$

In s^4, the cross product of the two sums in (10′) includes only terms like $x_i'x_\kappa'x_\nu'^2$ or like $x_\nu'x_\mu'^3$, the expectations of all of which vanish. Thus,

$$E[s^4] = \frac{(n-1)^2}{n^4}\Big\{\sum_{\nu=1}^{n} E[x_\nu'^4] + 2\sum_{\nu<\mu}^{1\ldots n} E[x_\nu'^2 x_\mu'^2]\Big\} + \frac{4}{n^4}\sum_{\nu<\mu}^{1\ldots n} E[x_\nu'^2 x_\mu'^2] .$$

Using (22) we find, since the number of terms $\nu < \mu$ is $\frac{1}{2}n(n-1)$,

$$E[s^4] = \frac{(n-1)^2}{n^3}\tau^4 + \Big[\frac{(n-1)^3}{n^3} + 2\,\frac{n-1}{n^3}\Big]\sigma^4 .$$

If this is combined with (23) and (18′) and the coefficient of σ^4 appropriately reduced, the result is

$$\mathrm{Var}[s^2] = \frac{n-1}{n^3}[(n-1)\tau^4 - (n-3)\sigma^4] \sim \frac{1}{n}(\tau^4 - \sigma^4), \tag{24}$$

the last expression being equivalent to the second for large n.[2]

If for any function F both the expected value and the variance are known, the Tchebycheff inequality permits a conclusion about the probability of F: let λ be any constant greater than 1; then the probability that F falls in the interval from $E[F] - \lambda\sqrt{\mathrm{Var}[F]}$ to $E[F] + \lambda\sqrt{\mathrm{Var}[F]}$ is at least $1 - (1/\lambda)^2$. This gives 50% probability for $\lambda = \sqrt{2} = 1.41$. More can be said if the assumption appears justified that the distribution of F is, at least approximately, normal. According to the tables, a 50% probability holds then for the interval of half-length $0.674\sqrt{\mathrm{Var}[F]}$, i.e., the factor $\sqrt{2} = 1.41$ is reduced to 0.674. It is usual to indicate expectation and variance of a variable F briefly by the symbol $A \pm B$ where A is the expectation of F and B the square root of its variance (=standard deviation). Thus, *in the case of n equal distributions with mean α and variance σ^2*, one would write, if n is large,

$$a = \alpha \pm \frac{\sigma}{\sqrt{n}}, \qquad s^2 = \frac{n-1}{n}\sigma^2 \pm \sqrt{\frac{\tau^4 - \sigma^4}{n}}. \tag{25}$$

It is seen that for both a and s^2 the standard deviation goes to zero as n increases. In connection with the Tchebycheff inequality, this means: *If a sample of n is drawn from a population, the probability that the sample average a and the sample dispersion s^2 lie in any close neighborhood of their expected values α and $(n-1)\sigma^2/n$, respectively, tends toward unity as n increases.* The first part of this statement was found in Chapter V.[3]

The square root of $\mathrm{Var}[F]$ is called *standard error* (or standard deviation) of F. Thus $\sigma/\sqrt{n}$ is the standard error of a, and $\sqrt{(\tau^4 - \sigma^4)/n}$ is the standard error of s^2. In analogy to Eq. (25) we write

$$F = E[F] \pm \sqrt{\mathrm{Var}[F]}, \tag{25′}$$

where $\sqrt{\mathrm{Var}[F]}$ is the standard error or standard deviation of F.

[2] A computation of $\mathrm{Var}[s^2]$ in the case of different distributions $V_\nu(x)$ is given, for example, in v. Mises [21], pp. 280–287.

[3] Using Tchebycheff's inequality we need the existence of the variance of a and s^2, respectively. Khintchine's theorem (Chapter V, Section 3) gives the result without assuming existence of the variance.

Problem 8. Give the expressions for expectation and variance of a and s^2 in the case that all x_ν are subject to the same normal distribution with mean α and variance σ^2.

Problem 9. Compute, under the assumption that all α_ν and all $\sigma_\nu{}^2$ are equal, the expectation and variance of a quadric

$$Q = \tfrac{1}{2}\sum_{\nu=1}^{n} A_\nu x_\nu{}^2 + \sum_{i<\kappa}^{1\ldots n} B_{i\kappa} x_i x_\kappa \, .$$

Problem 10. Find, under the same assumption, expectation and variance of $a(1 - a/m)$ where m is a constant, either by direct computation or by using the result of Problem 9.

Problem 11. Compute $E[s^2]$ if the assumption of independence is abandoned.

Problem 12. Compute $\mathrm{Var}[s^2]$ if the restriction to equal distributions $P_\nu(x)$ is abandoned; use the notations $\rho_\nu{}^3 = E[x_\nu'^3]$, $\tau_\nu{}^4 = E[x_\nu'^4]$, where the x_ν' are defined as in (22).

Problem 13. Assume all $P_\nu(x) = P(x)$ and find $E[m_k]$, where

$$m_k = \frac{1}{n}\sum_{\nu=1}^{n} (x_\nu - a)^k \, ;$$

introduce the moments $\mu_i = \int (x - \alpha)^i \, dP(x)$ up to the order k.

B. Moments. Inequalities (Sections 4 and 5)

4. Determining a Distribution by Its First (2m — 1) Moments

4.1. *Lemma.* Knowledge of mean and variance of a distribution provided us with a first crude orientation regarding $V(x)$. The idea prompts itself that by giving the sequence of moments, M_0, M_1, M_2, ... up to a certain order k, the distribution $V(x)$ will be described with increasing accuracy as k increases. In this problem, it is immaterial whether the origin or any point $x = c$ is taken as reference point for the moments. In fact, if the moments with respect to the origin are denoted[1] by M_k,

$$M_k = \int x^k \, dV(x), \tag{26}$$

[1] In Chapter III, Section 3, M_k denoted the kth moment about the mean value.

one finds immediately that

$$M_k' = \int (x-c)^k \, dV(x) = M_k - \binom{k}{1} cM_{k-1} + \binom{k}{2} c^2 M_{k-2} + \cdots + (-1)^k c^k M_0 , \tag{27}$$

and vice versa

$$M_k = \int [(x-c)+c]^k \, dV(x) = M_k' + \binom{k}{1} cM'_{k-1} + \binom{k}{2} c^2 M'_{k-2} + \cdots + c^k M_0' , \tag{28}$$

where, of course, $M_0 = M_0' = 1$. This shows: *Giving the first k moments with respect to any one point is equivalent to giving the same moments with respect to any other point.*

The main questions that arise in connection with the problem of determining a distribution by its first moments are: (1) What are the common properties of distributions that have the same first moments? (2) To what extent can a sequence of real numbers be chosen arbitrarily if the numbers are to be the first moments of some (at least one) distribution? Both questions (and some others) have been answered (1874) by P. L. Tchebycheff. In what follows, only the case of odd k will be discussed, that is, the case where the number of given moments, including M_0, is even. (See a remark at the beginning of Section 4.5.)

Let us first assume that two distributions $V_1(x)$ and $V_2(x)$ exist which have the same moments $M_0, M_1, M_2, \ldots, M_{2m-1}$. Then

$$\int x^\kappa \, dV_1(x) = \int x^\kappa \, dV_2(x) \qquad \text{or} \qquad \int x^\kappa \, d[V_1(x) - V_2(x)] = 0$$
$$\text{for} \quad \kappa = 0, 1, \ldots, 2m-1. \tag{29}$$

If the $V(x)$ are step functions with a finite number of steps, no question of existence arises for the integrals in (29) since these integrals are then only sums of a finite number of terms. If $V(x)$ has a density, we assume that this density dV/dx vanishes at infinity as $x^{-(2m+\epsilon)}$ with $\epsilon > 0$. In this case $V(x)$ at $x = -\infty$ and $1 - V(x)$ at $x = +\infty$ will vanish as $x^{-(2m-1+\epsilon)}$ and moments up to the order $2m-1$ exist. Similarly, if, in the case of countably many steps, $p(x)$ vanishes at infinity like $x^{-(2m+\epsilon)}$, the series $\sum x^{2m-1}p(x)$ converges.[2]

Multiplying each Eq. (29) by an arbitrary constant and summing, we find

$$\int Q_{2m-1}(x) \, d[V_1(x) - V_2(x)] = 0, \tag{30}$$

[2] Remember that these conditions for the behavior of $V(x)$ at infinity are sufficient but not necessary for the convergence of the integrals or series.

where $Q_{2m-1}(x)$ is an arbitrary polynomial of order $2m - 1$. If we integrate (30) by parts and call Q_{2m-2} the derivative of Q_{2m-1}, we have

$$Q_{2m-1}(x)[V_1(x) - V_2(x)]\Big|_{-\infty}^{+\infty} - \int (V_1 - V_2) Q_{2m-2}\, dx = 0 .$$

The first expression is zero since $V_1 - V_2$ vanishes at both limits as $x^{-(2m-1+\epsilon)}$. Thus it is found that

$$\int (V_1 - V_2) Q_{2m-2}\, dx = 0 , \tag{31}$$

where Q_{2m-2} is an *arbitrary polynomial* of degree $(2m - 2)$ or less.

Suppose now that the difference $V_1 - V_2$ is not everywhere zero, but changes its sign at n points of intersection $x = a_1$, $x = a_2$, ..., $x = a_n$ where n is smaller than $k = 2m - 1$ (Fig. 23). [If $y = V_1 - V_2$ vanishes

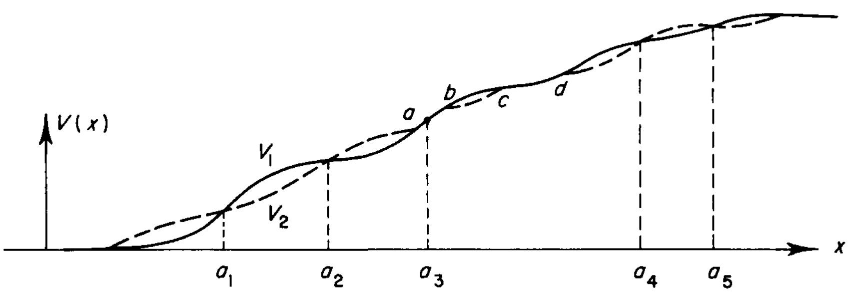

FIG. 23. Intersections of two distribution functions.

along a common arc of the two curves and y is positive (negative) before, negative (positive) after that common arc (see arc ab in Fig. 23), then we take an arbitrary point on that common arc as the "point of intersection" (in Fig. 23 the point with abscissa a_3). If, however, $y = 0$ along an arc like cd in Fig. 23 and has the same sign before and after cd, then there is no "intersection" on this arc.] We can take for Q_{2m-2} the product $(x - a_1)(x - a_2) \cdots (x - a_n)$. This polynomial changes its sign each time the sign of $V_1 - V_2$ changes and only then. Therefore, the product $(V_1 - V_2) Q_{2m-2}$ has the same sign everywhere, unless it is zero. But an integral cannot vanish if the integrand always has the same sign (except where it vanishes). Thus the assumption $n < k$ is contradictory and the result reached is:

LEMMA. *If two distributions have the same moments up to the* kth *order, the graphs of the corresponding c.d.f.'s* $V_1(x)$ *and* $V_2(x)$ *must have at least* k *points of intersection.*

4.2. *First theorem.* A step function with m steps is determined by $2m$ constants: the m abscissas $a_1, a_2, ..., a_m$ of the steps and the m heights $A_1, A_2, ..., A_m$ of the jumps. If $2m$ moments are given, it is, therefore, reasonable to presume that an m-step function can be found that has these moments. Assume that two m-step functions exist which both have these moments. In Fig. 24 the graphs of two functions each

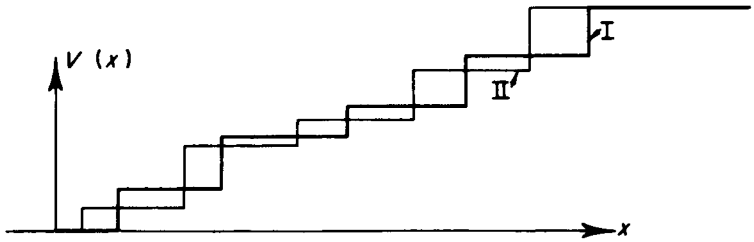

FIG. 24. Maximum number of intersections of two step lines.

with $m = 5$ steps and as many intersections as possible are shown; it is seen that the number of possible intersections (points at which the sign of $V_1 - V_2$ changes) cannot be greater than 8—in general, not greater than $2m - 2$. On the other hand, if the two distributions have the same moments up to the order $k = 2m - 1$, they should have at least $(2m - 1)$ intersections. To avoid the contradiction, we must conclude: *There cannot exist more than one m-step function having a given set of $2m$ moments $M_0, M_1, M_2, ..., M_{2m-1}$.*

Now, assume that for a given set $M_0, M_1, ..., M_{2m-1}$ the m-step function has been found. (Under what conditions it exists will be seen later.) The graph of this $V_1(x)$ consists of m vertical and $(m - 1)$ horizontal line segments (in addition to the two infinite horizontal lines at the levels 0 and 1). The graph of any other $V(x)$ with the same moments $M_0, ..., M_{2m-1}$ must have at least $(2m - 1)$ intersections with the graph of $V_1(x)$. Since only one intersection in the above-defined sense is possible on each of the $(2m - 1)$ horizontal and vertical segments, the graph of $V(x)$ must cross all steps as shown in Fig. 25; that is, $V(x)$ must intersect the m-step line determined by the given $2m$ moments once on each of its vertical segments and once on each of its horizontal segments. (Again, we cannot exclude the case that V_1 and V_2 have a segment ab in common. Then, in line with our definition of intersection, there is on ab one point of intersection.) This leads to the important theorem which gives an answer to the first of the above-formulated questions:

If $2m$ quantities $M_0, M_1, ..., M_{2m-1}$ are such that an m-step function exists with these moments, then all other distributions $V(x)$ with the same *moments are restricted in such a way that, at m definite abscissas* $a_1, a_2, ..., a_m$, *the value of* $V(x)$ *falls in a definite interval of length* $A_1, A_2, ..., A_m$, *respectively.*

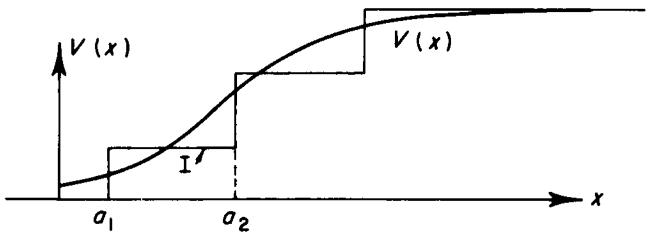

FIG. 25. Graph of arbitrary $V(x)$ must intersect each step.

4.3. *Determination of a step function by moments.* Turning to the second question, we consider the integral

$$F = \int (y_0 + y_1 x + y_2 x^2 + \cdots + y_{m-1} x^{m-1})^2 \, dV(x) = \sum_{i,\kappa=0}^{m-1} y_i y_\kappa M_{i+\kappa}$$
$$= y_0^2 M_0 + y_1^2 M_2 + \cdots + y_{m-1}^2 M_{2m-2} + 2y_0 y_1 M_1 + \cdots + 2y_{m-2} y_{m-1} M_{2m-3}, \tag{32}$$

where $y_0, y_1, ..., y_{m-1}$ are arbitrary and real (independent of x). It is evident that F cannot be negative. In addition, if $V(x)$ is increasing (not constant) at more than $(m - 1)$ points, F will never be zero except when all y vanish. In fact the integrand in (32) is the square of a polynomial in x of degree $(m - 1)$ and therefore can vanish at not more than $(m - 1)$ points. With respect to $y_0, y_1, ..., y_{m-1}$ the function F is a *quadric* (=homogeneous symmetric function of second order). Its matrix is

$$\begin{vmatrix} M_0 & M_1 & M_2 & \cdots & M_{m-1} \\ M_1 & M_2 & M_3 & \cdots & M_m \\ M_2 & M_3 & M_4 & \cdots & M_{m+1} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ M_{m-1} & M_m & M_{m+1} & \cdots & M_{2m-2} \end{vmatrix}, \tag{32'}$$

and its determinant may be denoted by D_{m-1}. Hence

$$D_0 = M_0, \qquad D_1 = \begin{vmatrix} M_0 & M_1 \\ M_1 & M_2 \end{vmatrix}, \qquad D_2 = \begin{vmatrix} M_0 & M_1 & M_2 \\ M_1 & M_2 & M_3 \\ M_2 & M_3 & M_4 \end{vmatrix}, \ldots. \tag{33}$$

If a quadric assumes only positive values (except when all variables are zero) it is called *positive definite*. It is a well-known theorem that the determinant of the matrix of a positive definite form and the determinants of its successive "squares from the upper left-hand corner" must also be positive. In other words, *the first $2m - 1$ moments of a distribution whose $V(x)$ increases at more than $(m - 1)$ points must satisfy the following m conditions*:

$$M_0 > 0, \quad \begin{vmatrix} M_0 & M_1 \\ M_1 & M_2 \end{vmatrix} > 0, \quad \begin{vmatrix} M_0 & M_1 & M_2 \\ M_1 & M_2 & M_3 \\ M_2 & M_3 & M_4 \end{vmatrix} > 0, \ldots, \begin{vmatrix} M_0 & M_1, \ldots, M_{m-1} \\ M_1 & M_2, \ldots, M_m \\ \vdots & \\ M_{m-1} & M_m, \ldots, M_{2m-2} \end{vmatrix} > 0 \tag{33'}$$

or: $D_0 > 0$, $D_1 > 0$, $D_2 > 0$, ..., $D_{m-1} > 0$.

Now, we are going to solve the following problem. We assume known that M_0, M_1, ..., M_{2m-1} are moments of some distribution $V(x)$ and fulfill the inequalities (33'); *we want to find the m-step distribution* $V_1(x)$—there can be only one—*that has these moments*. The abscissas and the magnitudes of the m steps will again be called a_1, a_2, ..., a_m and A_1, A_2, ..., A_m. Then, the following equations hold:

$$\int x^k \, dV_1(x) = \sum_{\mu=1}^{m} a_\mu{}^k A_\mu = \int x^k \, dV(x), \qquad k = 0, 1, \ldots, (2m - 1) \tag{34}$$

or, if we combine these equations multiplied by some constants,

$$\int Q_{2m-1}(x) \, dV_1(x) = \sum_{\mu=1}^{m} Q_{2m-1}(a_\mu) A_\mu = \int Q_{2m-1}(x) \, dV(x), \tag{34'}$$

where $Q_{2m-1}(x)$ is any polynomial whose order is not higher than $(2m - 1)$. If we introduce the polynomial of order m,

$$\omega(x) = (x - a_1)(x - a_2) \cdots (x - a_m) = x^m + \sum_{i=0}^{m-1} c_i x^i, \tag{35}$$

we can use for $Q_{2m-1}(x)$ any one of the expressions

$$\omega(x), \quad x\omega(x), \quad x^2\omega(x), \ldots, x^{m-1}\omega(x). \tag{36}$$

Since each of these expressions vanishes if one of the a_μ is introduced for x, it is seen that with these choices for $Q_{2m-1}(x)$, the middle term in

(34′) vanishes. Consequently the third term (and the first term) must also vanish and this means that, with $\lambda = 0, 1, 2, ..., m - 1$

$$0 = \int x^\lambda \omega(x)\, dV(x) = \int \Big[x^{m+\lambda} + \sum_{i=0}^{m-1} c_i x^{i+\lambda}\Big]\, dV(x)$$

$$= M_{m+\lambda} + \sum_{i=0}^{m-1} c_i M_{i+\lambda}\,. \tag{37}$$

Writing this explicitly for $\lambda = 0, 1, 2, ..., m - 1$, we have

$$\begin{aligned}
c_0M_0 \quad &+ c_1M_1 + \cdots + c_{m-1}M_{m-1} = -M_m \\
c_0M_1 \quad &+ c_1M_2 + \cdots + c_{m-1}M_m \quad = -M_{m+1} \\
&\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\
c_0M_{m-1} &+ c_1M_m + \cdots + c_{m-1}M_{2m-2} = -M_{2m-1}\,.
\end{aligned} \tag{38}$$

This is a system of m linear equations for the m unknowns $c_0, c_1, ..., c_{m-1}$. The determinant of the system is just the last of the determinants (33′), which, by hypothesis, is not zero but positive; hence Eqs. (38) have a solution. Assume the values of $c_0, c_1, ..., c_{m-1}$ have been found; then

$$f(x) = x^m + c_{m-1}x^{m-1} + c_{m-2}x^{m-2} + \cdots + c_1x + c_0 \tag{39}$$

is a known polynomial. If its roots are all real numbers and different from each other, they represent, according to (35), the abscissas $a_1, a_2, ..., a_m$ of the steps of $V_1(x)$.

It can easily be shown that $f(x)$ changes its sign m times. The bracket in the middle term of (37) equals $x^\lambda f(x)$. Since (37) is correct for $\lambda = 0, 1, 2, ..., m - 1$, it follows—in the same way as above—that

$$\int Q_{m-1}(x) f(x)\, dV(x) = 0, \tag{40}$$

where $Q_{m-1}(x)$ is any polynomial whose order is not higher than $(m - 1)$. Assume that $f(x)$ changes its sign only $n < m$ times, at $x = b_1, b_2, ..., b_n$. Then we would take for $Q_{m-1}(x)$ the polynomial $(x - b_1)(x - b_2) \cdots (x - b_n)$ so that $Q_{m-1}(x) f(x)$ has no change of sign. In this case, (40) could not be true; that means that the assumption $n < m$ is contradictory. The result is that the roots of (39) *are m* real values, all different from each other. They are the abscissas $a_1 < a_2 < \cdots < a_m$. The m zeros $a_1, ..., a_m$ must be situated in the smallest interval a, b which contains all points of increase of $V(x)$.

Once the a_μ have been found, the A_μ can be computed from the m equations

$$\sum_{\mu=1}^{m} A_\mu = M_0 = 1, \qquad \sum_{\mu=1}^{m} a_\mu A_\mu = M_1, \qquad \sum_{\mu=1}^{m} a_\mu{}^2 A_\mu = M_2, \ldots,$$

$$\sum_{\mu=1}^{m} a_\mu^{m-1} A_\mu = M_{m-1}, \tag{41}$$

or briefly

$$L_i = M_i - \sum_{\mu=1}^{m} a_\mu{}^i A_\mu = 0, \qquad i = 0, 1, 2, \ldots, m-1. \tag{41'}$$

The determinant of this system of equations is

$$\begin{vmatrix} 1 & 1 & & 1 \\ a_1 & a_2 & & a_m \\ a_1{}^2 & a_2{}^2 & & a_m{}^2 \\ \cdots & \cdots & \cdots & \cdots \\ a_1^{m-1} & a_2^{m-1} & & a_m^{m-1} \end{vmatrix}. \tag{42}$$

This determinant is known to be different from zero if all a_μ are different. Consequently, Eqs. (41) can be solved for $A_1, A_2, \ldots, A_m$. The A_μ, so computed, together with the previously computed a_μ, fulfill all the equations (34) or (34').

It remains to be shown that the A_μ found from (41) are positive. With the a_μ and A_μ, which we have found we compute a step line which at the abscissas a_μ has steps A_μ. It follows from (34) that this step line and $V(x)$ have the first $2m$ moments in common. Hence $V(x)$ must intersect the step line in $2m - 1$ points, one on each horizontal and one on each vertical segment; this, however, could not be possible if any A_μ were negative, i.e., if the step line had a downward step (see Fig. 26).

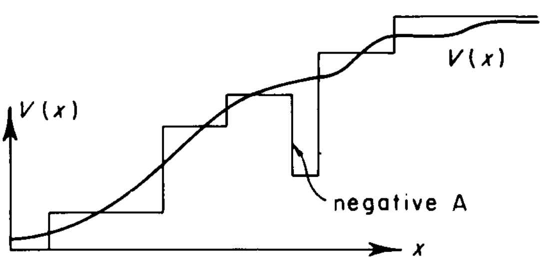

FIG. 26. Steps must be positive.

The result is: *If* M_0, M_1, ..., M_{2m-1} *are the moments of a distribution* $V(x)$, *which increases at m points at least, then there exists one and only one m-step distribution* $V_1(x)$ *which has these moments. The abscissas* a_μ *of the steps,* $\mu = 1, 2, ..., m$ *are the m real roots of Eq.* (39) *where the* c_i *are the solutions of* (38); *the heights* A_μ *of the m steps follow from Eqs.* (41). *The m steps are in the interior of the smallest interval which contains all points of increase of* $V(x)$. Either $V \equiv V_1$, or V crosses each step of V_1.

It may also be verified that the m Eqs. (38) can be written in the form [see definition (41′)]

$$\begin{aligned}
c_0L_0 \quad &+ c_1L_1 + \cdots + L_m \quad = 0\\
c_0L_1 \quad &+ c_1L_2 + \cdots + L_{m+1} = 0\\
&\vdots\\
c_0L_{m-1} &+ c_1L_m + \cdots + L_{2m-1} = 0\,.
\end{aligned}$$

If these are solved, one after the other, starting with $L_0 = L_1 = \cdots = L_{m-1} = 0$, we obtain

$$L_m = 0, \quad L_{m+1} = 0, ..., L_{2m-1} = 0,$$

n addition to (41′).

4.4. *Examples.* The cases $m = 1$, $m = 2$, and $m = 3$ may be briefly discussed. In the first case M_0 and M_1 only are known; $M_0 = 1$, $M_1 = a$. The one-step function with these moments has its step at $x = a$, and the magnitude of the jump is obviously 1. The theorem of p. 387 states that $V(x)$ must cross the vertical segment (Fig. 27a), in

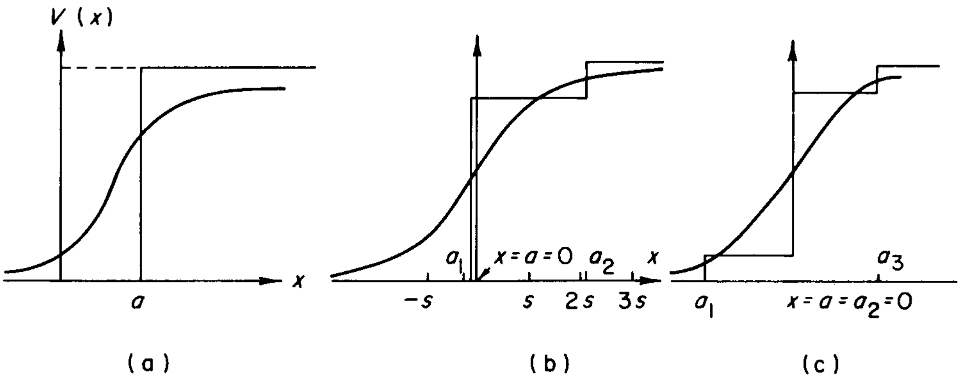

FIG. 27. Determination of m-step function by $2m$ moments, $m = 1, 2, 3$.

other words, that some "mass" must lie to the left as well as to the right of $x = a$.

In the case $m = 2$, we choose the origin to coincide with the mean value so that $M_0 = 1$, $M_1 = 0$, $M_2 = s^2$. The third moment which also has to be given may be written in terms of the "skewness" r (see p. 375), $M_3 = rs^3$. The two-step distribution with these moments is found by solving Eqs. (38) which now read

$$\begin{aligned} c_0M_0 + c_1M_1 &= -M_2 \\ c_0M_1 + c_1M_2 &= -M_3 \end{aligned} \qquad \text{i.e.,} \qquad \begin{aligned} c_0 &= -M_2 = -s^2 \\ c_1 &= -M_3/M_2 = -rs\,. \end{aligned}$$

The polynomial (39) and its roots are then

$$-s^2 - rsx + x^2 = 0, \qquad x = \frac{s}{2}(r \pm \sqrt{4 + r^2})$$

and with

$$a_1 = \frac{s}{2}(r - \sqrt{4 + r^2}), \qquad a_2 = \frac{s}{2}(r + \sqrt{4 + r^2})\,,$$

Eqs. (41) supply

$$A_1 + A_2 = 1, \qquad a_1A_1 + a_2A_2 = 0;$$

$$A_1 = \frac{1}{2}\left(1 + \frac{r}{\sqrt{4 + r^2}}\right), \qquad A_2 = \frac{1}{2}\left(1 - \frac{r}{\sqrt{4 + r^2}}\right).$$

The graph (for $r = 2$) is shown in Fig. 27b. Any $V(x)$ with skewness $r = 2$ is bound to cross the two vertical segments seen in the figure. In other words: to the left of a_1 no V can be greater than A_1; to the right of a_2 it cannot be less than A_1. For $r = 2$, $a_1 = -0.14s$, $a_2 = 2.14s$; hence any $V(x)$ must fulfill the conditions

$$0 < V(a_1) < 0.85, \qquad 0.85 < V(a_2) < 1.$$

In the case $m = 3$, we restrict ourselves to discussing distributions for which $M_1 = M_3 = M_5 = 0$. The decisive quantities are now the moments $M_2 = s^2$ and $M_4 = s^4(K + 3)$ where K is the so-called kurtosis of Section 2.2 and of Chapter III. We first study the conditions to which $M_0, M_1, ..., M_4$ are subject according to (33). The first is always satisfied and the second is satisfied, since $M_1 = 0$. The third gives with $M_0 = 1$, $M_1 = 0$, $M_2 = s^2$, $M_3 = rs^3$, $M_4 = s^4(K + 3)$:

$$\begin{vmatrix} 1 & 0 & s^2 \\ 0 & s^2 & rs^3 \\ s^2 & rs^3 & s^4(K+3) \end{vmatrix} = s^6(K + 2 - r^2) > 0, \qquad \text{or} \qquad K + 2 > r^2\,.$$

Hence, for any distribution with more than two steps $r^2 < K + 2$.

In the case $r = 0$, this means $K + 2 > 0$. The three equations (38) for c_0, c_1, c_2 are then (using $M_5 = 0$)

$$\begin{aligned} c_0 \quad + s^2 c_2 \qquad &= 0 \\ c_1 s^2 \qquad\qquad &= -(K+3)s^4 \\ c_0 s^2 + (K+3)c_2 s^4 &= 0, \end{aligned}$$

or

$$c_0 = 0, \qquad c_1 = -(K+3)s^2, \qquad c_2 = 0.$$

The polynomial (39) of third order and its roots are

$$-(K+3)s^2 x + x^3 = 0; \qquad a_1 = -s\sqrt{K+3}, \quad a_2 = 0, \quad a_3 = s\sqrt{K+3}\,.$$

For the jumps A_1, A_2, A_3 we have from (41)

$$\begin{aligned} A_1 \quad + A_2 + A_3 &= 1 \qquad & A_1 = A_3 &= \frac{1}{2(K+3)} \\ a_1 A_1 \qquad\quad a_3 A_3 &= 0 & & \\ a_1^2 A_1 \qquad\quad a_3^2 A_3 &= s^2 & A_2 &= 1 - \frac{1}{K+3}\,. \end{aligned}$$

We have seen above that $K + 3 > 1$ so that in fact A_1, A_2, A_3 are positive quantities. In Fig. 27c the graph of the three-step function is shown for $K = 1$. Any distribution with this value of kurtosis and $M_1 = M_3 = M_5 = 0$ must pass through the three vertical segments shown in the graph. This is the only correct way of judging the significance of K. Customary statements about the general behavior of a distribution with given kurtosis are unfounded, though they may apply to "most commonly occurring" distributions.

4.5. *Comments.* So far only the case of an *even* number of given moments (including M_0) has been discussed. If $M_0, M_1, ..., M_{2m}$ are given, an analysis similar to the one given here shows that an $(m + 1)$-step distribution with these moments can be found for which the location of one of the steps can be arbitrarily chosen (within certain limits).[3]

All this covers only the elementary part of what is called the *problem of moments.* The aim of the problem of moments is to answer questions like this: Under what conditions and in what way does an *infinite* sequence of real quantities $M_0, M_1, M_2, \ldots$ determine a distribution. The first fundamental investigation along these general lines, preceded by the work of Tchebycheff, Markov, and others, is by Stieltjes.[4] In the paper quoted in Chapter II as the source of the

[3] R. v. Mises, "Bestimmung einer Verteilung durch ihre ersten Momente." *Skand. Aktuarietidskr.* (1937), pp. 220–243 (see p. 237). This case is connected with the Stieltjes problem (see below).

[4] T. J. Stifltjes, "Recherches sur les fractions continues." *Ann. Fac. Sci. Toulouse* **8**, **9** (1894, 1895).

Stieltjes' integral, he also sets up and solves the problem of moments for the interval $(0, \infty)$; i.e., he asks for a distribution $V(x)$ equal to zero for $x \leqslant 0$ and satisfying the infinitely many equations (integral equations)

$$\int_0^{\infty} x^n \, dV(x) = M_n \,, \qquad n = 0, 1, 2, \cdots .$$

The problem most important for us is the one defined and solved by Hamburger[5] who considers the problem of moments for the interval $(-\infty, +\infty)$ where

$$\int_{-\infty}^{+\infty} x^n \, dV(x) = M_n \,, \qquad n = 0, 1, 2, \ldots \tag{26}$$

and obtains basic results which we shall briefly review. In 1923 F. Hausdorff answered the problem of moments for a finite interval.

With the definition (33) Hamburger considers the sequence D_0, D_1, D_2, ... formed by means of a sequence $M_0 > 0$, M_1, M_2, ... of real numbers. Then the necessary and sufficient condition for the existence of at least one d.f. $V(x)$ which satisfies the infinitely many equations (26) is that the sequence D_0, D_1, D_2, ... has one of the following properties:

(A) $D_0 > 0, \quad D_1 > 0, \quad D_2 > 0, \ldots, D_{n-1} > 0, \quad D_n = D_{n+1} = \cdots = 0.$

(B) All the D_n are positive.

In case (A) there always exists one single solution $V(x)$, namely, a discrete distribution with n steps. In case (B) the problem is either determined, i.e., it admits one solution or indeterminate, i.e., it admits more than one solution $V(x)$. In case (B) no $V(x)$ is a discrete distribution with n steps. For the indeterminate case Hamburger[6] has given a method to indicate *all* solutions $V(x)$. One of them is a discrete distribution with denumerably many steps. We repeat that the necessary and sufficient condition for $n = 2m$ numbers to be the moments (26) of an m-step function is that they satisfy conditions (A).

Among the many criteria which have been devised in order to decide whether in the case (B) the problem is determinate or indeterminate, we mention the following. If the M_0, M_1, M_2, ... are such that all determinants (33′) are positive, there always exists at least one distribution with these moments; the distribution is uniquely determined by these

---

[5] H. Hamburger, "Ueber eine Erweiterung des Stieltjes'schen Momentenproblems." *Math. Ann.* **81**, pp. 235–319; **82**, pp. 120–169; **83**, pp. 168–187.

[6] H. Hamburger, "Jacobi matrices and undetermined moment problem." *Amer. J. Math.* **66** (1944).

moments if $\sqrt[n]{|M_n|}/n$ remains bounded as n increases indefinitely.[7] The problem of moments is presented in a monograph of Shohat and Tamarkin.[8]

Problem 14. What relation between mean value a, variance s^2, and skewness r must be fulfilled if it is required that all values of x [for which $V(x)$ is increasing] are non-negative?

Problem 15. Find the three-step distribution with given moments: $M_0 = 1$, $M_1 = 0$, $M_2 = s^2$, $M_3 = rs^3$, $M_4 = (K + 3)\,s^4$, $M_5 = r_1 s^5$, in particular with the assumption $K = 10$, $r = 2$, $r_1 = 8$.

Problem 16. Using the notation of Problem 15 and in addition $K_1 = (M_6/s^6) - 15$, evaluate and discuss condition (33) of order four. Prove that for a distribution with exactly two steps $r_1 = r(K + 4) = r(r^2 + 2)$.

5. Some Inequalities

5.1. *Inequalities based on Schwarz's inequality.* In many statistical problems, it is useful to have estimates for quantities like moments, mean values, or other expectations, i.e., for quantities that depend on distributions not completely defined. Such estimates are mathematically expressed by certain inequalities. Two groups of inequalities of this kind have already been discussed. One group is represented by the Tchebycheff inequality, Chapter III, which can be written

$$\Pr\{|x - a| \leqslant X\} \geqslant 1 - \frac{s^2}{X^2}. \tag{43}$$

It can also be proved that if $M_{|m|}$ is the absolute moment of mth order with respect to the mean value of a probability distribution, then [see Chapter III, Section 3.1, Eq. (10″)]

$$\Pr\{|x - a| \leqslant X\} \geqslant 1 - \frac{M_{|m|}}{X^m}. \tag{44}$$

A second group is supplied by the sequence of inequalities (33′) in the preceding section.

[7] G. Pólya, *Math. Z.* **8** (1920), pp. 171–181.

[8] J. A. Shohat and J. D. Tamarkin, *The Problem of Moments* (Math. Surveys, Vol. 1). New York, 1943.

We obtain other inequalities starting from the *Schwarz inequality* (Chapter III, Section 2.3) which reads

$$\left[\int f(x)g(x)\,dV(x)\right]^2 \leqslant \left[\int f^2(x)\,dV(x)\right]\left[\int g^2(x)\,dV(x)\right]. \tag{45}$$

A particular case of (45) is

$$\left(\sum_1^n f_\nu g_\nu p_\nu\right)^2 \leqslant \left(\sum_1^n f_\nu^2 p_\nu\right)\left(\sum_1^n g_\nu^2 p_\nu\right). \tag{46}$$

If the f_ν and g_ν are considered as the components of two vectors in n dimensions and all $p_\nu = 1$, (46) states that the scalar product of two vectors never exceeds the product of their lengths. For $n = 3$, (46) states then that $\cos^2 \alpha \leqslant 1$ where α is the angle between the two vectors. For $n = 2$, (46) reduces to

$$\frac{f_2^2 g_1^2 + f_1^2 g_2^2}{2} \geqslant f_1 g_1 f_2 g_2\,, \qquad \text{or} \qquad \tfrac{1}{2}(A^2 + B^2) \geqslant AB\,. \tag{47}$$

The above is true on account of $(f_1 g_2 \pm f_2 g_1)^2 \geqslant 0$.

Returning to the general case, Eq. (45) can be read as a statement on expectations as in Chapter III, namely,

$$(E[f(x)g(x)])^2 \leqslant E[f^2(x)] \cdot E[g^2(x)]. \tag{48}$$

Choose now, for example,

$$f(x) = |\,x\,|^\alpha\,, \qquad g(x) = |\,x\,|^\beta$$

so as to have $E[f^2(x)] = M_{2\alpha}$, moment of order 2α; $E[f(x)\,g(x)] = M_{|\alpha+\beta|}$ absolute moment of order $\alpha + \beta$, etc.; then (45) leads to

$$M_{|\alpha+\beta|} \leqslant \sqrt{M_{2\alpha} M_{2\beta}}\,, \qquad \text{or} \qquad \log M_{|\alpha+\beta|} \leqslant \tfrac{1}{2}(\log M_{2\alpha} + \log M_{2\beta}). \tag{49}$$

Consider a fixed $V(x)$ that increases in $0 \leqslant x < \infty$ only. Then, there is no difference between absolute moments and moments. We call *line of moments the function of* ν *defined by* $M_\nu = \int x^\nu\,dV(x)$. Then (49) expresses the fact that *the curve representing* $\log M_x$ *as a function of* $x(>0)$ *is convex* (*from below*) (*Fig.* 28), since at the abscissa $(\alpha + \beta)$, i.e., in the middle of the chord AB, the curve is below the chord, whatever α and β are.

The curve $\log M_x$ versus x passes through the origin, since for

$x = 0$, $M_0 = 1$, $\log M_0 = 0$. This fact, in connection with the downward convexity, shows that

$$\frac{\log M_\beta}{\beta} \geqslant \frac{\log M_\alpha}{\alpha} \qquad \text{for} \quad \beta \geqslant \alpha > 0, \tag{50}$$

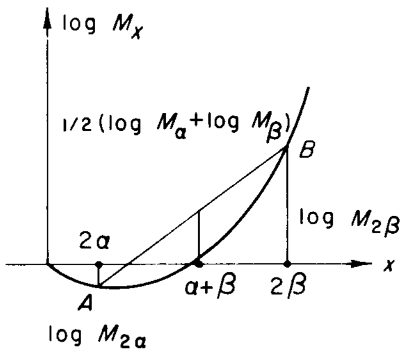

FIG. 28. Curve of $\log M_x$ is convex downward for $x > 0$.

i.e., the function $\log M_x/x$ for $x > 0$ increases monotonically. (If a point P moves along a downwardly convex curve passing through O, the vector OP rotates all the time in a positive, i.e., counterclockwise, sense.) In other terms, if the xth root of M_x is called a_x, one has

$$a_\beta \geqslant a_\alpha \qquad \text{for} \quad \beta \geqslant \alpha > 0\,, \tag{51}$$

e.g., $M_4{}^2 \geqslant M_2{}^4$. It can also be seen that if $\alpha \neq \beta$ the equality sign in (49) or (50) can only hold if $V(x)$ has one single jump. In this degenerate case, M_x equals a^x where a is the abscissa of the jump (and the mean value), and a_x equals a.

5.2. *Inequalities using known a and s^2.* We ask now what we can conclude regarding moments of $V(x)$ if only $M_0 = 1$, $M_1 = a$, $M_2 = a^2 + s^2$ are known. As before, we assume $x \geqslant 0$. The convexity of the $\log M_r$ curve supplies inequalities for the moments of any positive order. In Fig. 29 the three points with abscissas 0, 1, 2, and ordinates 0, $\log a$, $\log (a^2 + s^2)$ are marked A, B, C. The two straight lines AB and BC have the equations

$$y = x \log a \qquad \text{and} \qquad y = \log a + (x-1)\log\Big(a + \frac{s^2}{a}\Big)\,. \tag{52}$$

If x lies between 0 and 1, the corresponding point of the curve must lie below AB and above BC (excluding the degenerate case $s = 0$), that is,

$$\log M_x < x \log a, \qquad \log M_x > \log\left[a\left(a + \frac{s^2}{a}\right)^{x-1}\right], \qquad \text{for} \quad 0 < x < 1.$$

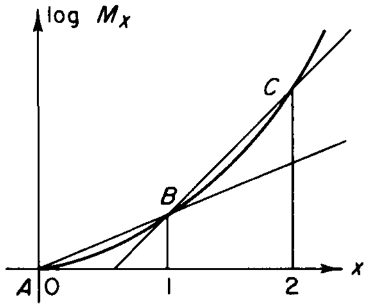

FIG. 29. A property of log M_x.

On the other hand, for an x in the interval 1 to 2, the curve is below BC and above AB; that is,

$$\log M_x > x \log a, \qquad \log M_x < \log\left[a\left(a + \frac{s^2}{a}\right)^{x-1}\right] \qquad \text{for} \quad 1 < x < 2,$$

and, finally, one has

$$\log M_x > \log\left[a\left(a + \frac{s^2}{a}\right)^{x-1}\right] \qquad \text{for} \quad x > 2.$$

These inequalities[1] can also be written as

$$\begin{aligned} a^x\left(1 + \frac{s^2}{a^2}\right)^{x-1} < M_x < a^x \qquad & \text{if} \quad 0 < x < 1 \\ a^x < M_x < a^x\left(1 + \frac{s^2}{a^2}\right)^{x-1} \qquad & \text{if} \quad 1 < x < 2 \\ M_x > a^x\left(1 + \frac{s^2}{a^2}\right)^{x-1} \qquad & \text{if} \quad x > 2 \end{aligned} \tag{53}$$

Note that we obtain from (53)

$$M_x > a^x \quad \text{for} \quad x > 1, \qquad M_x < a^x \quad \text{for} \quad 0 < x < 1. \tag{53'}$$

[1] We remember that our inequalities are derived under the assumption $V(0) = 0$. They remain correct in the general case $[V(-\infty) = 0]$ if M_ν is the *absolute* moment, the expected values of $|x|^\nu$.

From (53) we have in the cases $x = \frac{1}{2}$, $x = \frac{3}{2}$, $x = 3$

$$\sqrt{a}\left(1 + \frac{s^2}{a^2}\right)^{-1/2} < E[\sqrt{x}] < \sqrt{a}$$

$$\sqrt{a^3} < E[\sqrt{x^3}] < \sqrt{a^3}\left(1 + \frac{s^2}{a^2}\right)^{1/2} \qquad (54)$$

$$E[x^3] > a^3\left(1 + \frac{s^2}{a^2}\right)^2 .$$

The first of these, the estimate of $E[\sqrt{x}]$ is often used. It may be written as

$$\frac{a\sqrt{a}}{\sqrt{a^2 + s^2}} < E[\sqrt{x}] < \sqrt{a} ,$$

or if s/a is small,

$$\left(1 - \frac{s^2}{2a^2}\right)\sqrt{a} < E[\sqrt{x}] < \sqrt{a} .$$

If s/a is small, the estimates in the first two cases (54) are very close.

5.3. *Expectation of product and quotient.* A further type of inequality that can be derived from (48) concerns the expectations of the *product and the quotient of two arbitrary functions* $f(x)$ and $g(x)$. Assume that the two mean values a_1, a_2 and the two variances s_1^2, s_2^2 are known, viz.,

$$a_1 = \int f(x)\, dV(x), \qquad a_2 = \int g(x)\, dV(x),$$

$$s_1^2 = \int [f(x) - a_1]^2\, dV(x), \qquad s_2^2 = \int [g(x) - a_2]^2\, dV(x).$$

Let us apply Eq. (48) to $f_0 = f - a_1$, $g_0 = g - a_2$. The left-hand side of (48) gives $E[f_0 g_0] = E[(f - a_1)(g - a_2)] = E[fg] - a_1 a_2$, and $E[f_0^2] = \mathrm{Var}\,[f]$, $E[g_0^2] = \mathrm{Var}\,[g]$. We obtain

$$| E[fg] - E[f]E[g] | \leqslant s_1 s_2 = \sqrt{\mathrm{Var}[f]\,\mathrm{Var}[g]} . \qquad (55)$$

The difference between the expectation of a product and the product of the expectations does not exceed, in absolute value, the geometrical mean of the variances. The equality sign holds if $g = f$ or, more generally, if there is a linear relation between f and g.

Let us now assume that the mean value a_2 of $g(x)$ is different from 0 and that the quotient f/g does not exceed, in absolute value, the amount K. We use the identity

$$\frac{1}{g} - \frac{1}{a_2} = \frac{(g - a_2)^2}{g a_2^2} - \frac{g - a_2}{a_2^2} .$$

Multiplying both sides by $f\,dV$ and integrating, we obtain

$$\int \frac{f}{g}\,dV - \frac{a_1}{a_2} = \frac{1}{a_2{}^2}\int \frac{f}{g}(g - a_2)^2\,dV - \frac{1}{a_2{}^2}\left[\int fg\,dV - a_1a_2\right].$$

The first integral on the right-hand side cannot exceed, in absolute value, $Ks_2{}^2$, while the expression in brackets according to (55) has an absolute value not exceeding s_1s_2. Thus

$$\left| E\left[\frac{f}{g}\right] - \frac{E[f]}{E[g]} \right| \leqslant \frac{1}{E[g]^2}(Ks_2{}^2 + s_1s_2). \tag{56}$$

In particular, if $f = 1$ and, therefore, $E[f] = 1$, $s_1{}^2 = 0$, one has

$$\left| E\left[\frac{1}{g}\right] - \frac{1}{E[g]} \right| \leqslant \frac{Ks_2{}^2}{a_2{}^2} = \left| \frac{1}{g} \right|_{\max} \frac{\mathrm{Var}[g]}{(E[g])^2}. \tag{57}$$

The equality sign holds only if g = constant and both sides vanish.

5.4. *Particular types of distributions.* Other inequalities, often sharper than those valid for all $V(x)$, can be derived if the distribution $V(x)$ is in some way restricted. A case of some interest in the theory of errors and which was investigated by Gauss is that of a *convex distribution*, where $V(x)$ has a derivative $v(x)$ *monotonically decreasing* (or at least not increasing) with increasing $x \geqslant 0$.[2] [This means that higher values of x have smaller probability densities; the graph of $V(x)$ is convex upward.] We assume knowledge of M_ν and write

$$M_\nu = \int x^\nu v(x)\,dx = \frac{x^{\nu+1}}{\nu + 1}v(x)\Big|_0^\infty - \frac{1}{\nu + 1}\int x^{\nu+1}\,dv(x). \tag{58}$$

The first term to the right vanishes since $v(\infty) = 0$ and even $x^{\nu+1}v(x)$ must go to zero with $x \to \infty$ if M_ν is supposed to exist. Now, we introduce a new distribution $V_1(x)$ by

$$dV_1(x) = -\frac{dv(x)}{v(0)}, \qquad x \geqslant 0, \qquad V_1(0) = 0. \tag{59}$$

It is seen that (59) defines a distribution since $dv(x)$ is known to be always negative and

$$\int dV_1(x) = -\frac{1}{v(0)}\int dv(x) = -\frac{v(\infty) - v(0)}{v(0)} = 1.$$

[2] Again $V(x)$ is $\neq 0$ for $x > 0$ only. It is, however, admissible that $V(0) > 0$, i.e., that at the point $x = 0$ and only there, there is a point-probability $v(0) > 0$.

From (58) we have, with $dv(x) = -v(0)\, dV_1(x)$,

$$M_\nu = \frac{v(0)}{\nu + 1} \int x^{\nu+1}\, dV_1(x)\,. \tag{60}$$

This shows that, except for a positive constant factor $v(0)$, the expression $(\nu + 1)\, M_\nu$ is the $(\nu + 1)$th moment of the distribution $V_1(x)$. It follows that the curve representing $y = \log\,[(x + 1)\, M_x]$ as a function of x is convex downward, and since $y = 0$ for $x = 0$, we can apply (50):

$$\frac{\log[(\beta + 1)M_\beta]}{\beta} \geqslant \frac{\log[(\alpha + 1)M_\alpha]}{\alpha} \qquad \text{for} \quad \beta \geqslant \alpha > 0. \tag{61}$$

For example, if $\alpha = 2$, $\beta = 4$, this gives $(5M_4)^2 \geqslant (3M_2)^4$ or

$$5M_4 \geqslant 9M_2^2 \qquad \text{for convex distributions} \tag{62a}$$

as against

$$M_4 \geqslant M_2^2 \qquad \text{for all distributions.} \tag{62b}$$

Problem 17. Find the limits of the expectation of $\sqrt[3]{x}$ for a distribution of positive x-values with the mean $a = 4$ and the variance $s^2 = 2$.

Problem 18. Give a lower limit for the skewness of a distribution and show that a negative skewness is only possible if $s^2 < a^2$.

Problem 19. The variable x is normally distributed with mean value zero and variance s^2. Give upper and lower limits for the expectation of $1/g(x)$ where $g(x) = |\, x\,| + c$ (c = positive constant).

Problem 20. The quantity $x\,(>0)$ is distributed with the density e^{-x}. Find upper and lower limits for the expectation of fg and f/g where $f = 2 + x$, $g = 1 + x$. In the first case compare your result with the exact value of $E[fg]$.

C. Various Distributions Related to Normal Distributions (Sections 6 and 7)

6. The Chi-Square Distribution. Some Applications

In Chapter III we introduced the normal distribution, whose important role in probability theory appeared particularly in Chapter VI. In

Chapter V we proved that the sum (and the average) of normally distributed variates is normally distributed. In the following two sections we shall introduce some distributions which appear as the distributions of other functions of normally distributed variates. (See Sections 6.1 and 6.3.) Statistical applications of these distributions will follow in Chapter IX.

6.1. Introduction of the chi-square distribution. Let z be a random variable which is normally distributed with mean value 0 and variance 1. *The probability density of* $t = z^2$ *is then*

$$\frac{1}{\sqrt{2\pi}} t^{-\frac{1}{2}} e^{-\frac{t}{2}}$$

for $t > 0$, and zero elsewhere. In fact (see also Chapter III, p. 146) if $P_1(x) = \Pr\{z^2 < x\}$, then

$$P_1(x) = \frac{1}{\sqrt{2\pi}} \int_{-\sqrt{x}}^{\sqrt{x}} e^{-z^2/2}\, dz = \frac{\sqrt{2}}{\sqrt{\pi}} \int_0^{\sqrt{x}} e^{-z^2/2}\, dz$$

and with $z^2 = t$

$$P_1(x) = \frac{1}{\sqrt{2\pi}} \int_0^x t^{-\frac{1}{2}} e^{-\frac{1}{2}t}\, dt = \frac{1}{\Gamma(\frac{1}{2})\sqrt{2}} \int_0^x t^{\frac{1}{2}-1} e^{-\frac{1}{2}t}\, dt\,. \tag{63}$$

The distribution $P_1(x)$ is known as the *chi-square distribution with* 1 *degree of freedom.*

The *chi-square distribution with "m degrees of freedom"* is defined as follows:

$$P_m(x) = \frac{1}{\Gamma\left(\frac{m}{2}\right) 2^{\frac{m}{2}}} \int_0^x t^{\frac{m}{2}-1} e^{-\frac{t}{2}}\, dt, \qquad x \geqslant 0. \tag{64}$$

We shall also call it *Helmert's distribution,* since Helmert[1] was probably the first one to study it. In (64) the *gamma function* $\Gamma(t)$ appears; it is defined as

$$\int_0^\infty e^{-x} x^t\, dx = \Gamma(t+1).$$

For a positive integer $t = \nu$ this integral has the value $\nu!$. This covers the

[1] F. R. Helmert, *Z. Math. u. Phys.* 21 (1876), pp. 192–218.

case of even $m \geqslant 2$ in (64). For odd m, we find through integration by parts ($m = 2n + 1$, and n integer)

$$\Gamma\left(\frac{m}{2}\right) = \Gamma\left(\frac{2n+1}{2}\right) = \int_0^\infty e^{-x}x^{n-\frac{1}{2}}\,dx = (n - \tfrac{1}{2})\int_0^\infty e^{-x}x^{n-\frac{3}{2}}\,dx = \cdots$$

$$= (n - \tfrac{1}{2})(n - \tfrac{3}{2}) \cdots \tfrac{1}{2}\int_0^\infty e^{-x}x^{-\frac{1}{2}}\,dx\,,$$

and by the substitution $x = z^2$, the last integral transforms into the Gaussian integral whose value is $\sqrt{\pi}$. Hence,

$$\Gamma\left(\frac{m}{2}\right) = \frac{2n-1}{2}\cdot\frac{2n-3}{2}\cdots\frac{3}{2}\frac{1}{2}\sqrt{\pi}, \qquad \text{for odd} \quad m = 2n + 1\,.$$

In particular, $\Gamma(\frac{1}{2}) = \sqrt{\pi}$.

Thus we obtain from (64) the density

$$p_m(x) = \frac{1}{\Gamma\left(\frac{m}{2}\right)2^{\frac{m}{2}}}x^{\frac{m}{2}-1}e^{-\frac{x}{2}} = \frac{1}{2}\frac{e^{-\frac{x}{2}}x^{\frac{m}{2}-1}}{2\cdot 4\cdots(m-2)}, \qquad m \text{ even}$$

$$= \frac{1}{\sqrt{2\pi}}\frac{e^{-\frac{x}{2}}x^{\frac{m}{2}-1}}{1\cdot 3\cdots(m-2)}, \qquad m \text{ odd.} \tag{65}$$

In Fig. 30 the chi-square distribution is plotted for $m = 4,\ 10,\ 20$.

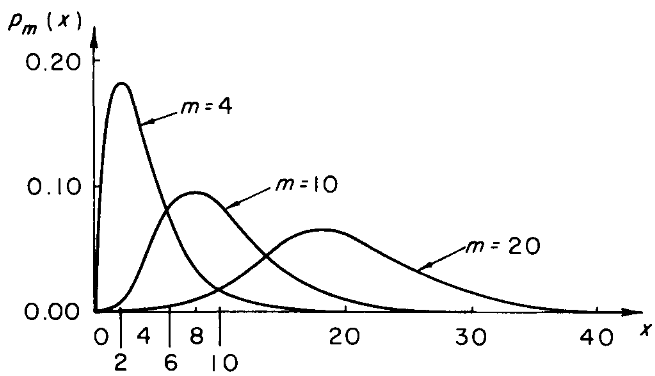

FIG. 30. Helmert's distribution for $m = 4, 10, 20$ degrees of freedom.

Denote by $\chi_p{}^2$ a value of χ^2 such that the probability that χ^2 exceeds $\chi_p{}^2$ is $p/100$; tables of the probability $\Pr\{\chi^2 > \chi_p{}^2\} = p/100$ have been given by Fischer [8] for $m = 1$ to $m = 30$ and 14 values of p while

Hald [12] gives tables for 23 values of p and for values of m from 1 to 100. We find, for example, for $p = 95$, $m = 10$, $\Pr\{\chi^2 < 18.3\} = 95\%$. Bilateral limits for χ^2 are obtained accordingly from $\Pr\{\chi^2_{p_1} < \chi^2 < \chi^2_{p_2}\} = p_2/100 - p_1/100$. For $p_2/100 = 97.5\%$, $p_1/100 = 2.5\%$ we have, for example, $\Pr\{3.25 < \chi^2 < 20.5\} = 95\%$ for $m = 10$. We give an abbreviated table, Table VI, for $p = 5\%, 2.5\%, 1\%, 0.5\%$.

It is very easy to compute the moments of any order of the chi-square distribution (65). Multiplying $p_m(x) = [1/x\Gamma(m/2)](x/2)^{m/2}\, e^{-x/2}$ by x^k means increasing the exponent of $(x/2)$ in $p_m(x)$ by k and multiplying by the factor 2^k. Thus, whether m is odd or even, we obtain $M_k{}^0$, the zero moments of order k:

$$M_k{}^0 = \int_0^\infty x^k p_m(x)\, dx = \frac{\Gamma\left(\frac{m}{2} + k\right)}{\Gamma\left(\frac{m}{2}\right)} 2^k = (m - 2 + 2k)(m - 4 + 2k) \cdots m\,. \tag{66}$$

This gives

$$a = M_1{}^0 = m, \qquad M_2{}^0 = (m + 2)m, \qquad s^2 = m(m + 2) - m^2 = 2m.$$

A few more of the moments are, m denoting the number of degrees of freedom, M_k the kth moment about the mean, $M_3 = 8m$, $M_4 = 12(4m + m^2)$.

The cumulative d.f. (64) is zero for $x \leqslant 0$ and equals $P_m(x)$ for $x > 0$. The indefinite integral in Eq. (64) for $P_m(x)$, is the *incomplete gamma function* for which tables are available.[2] It also may be computed by integration by parts to give

$$P_m(x) = 1 - e^{-\frac{x}{2}}\left[1 + \frac{x}{2} + \frac{x^2}{2\cdot 4} + \frac{x^3}{2\cdot 4\cdot 6} + \cdots + \frac{x^{\frac{m-2}{2}}}{2\cdot 4\cdots(m-2)}\right] \quad \text{for even } m, \tag{67}$$

$$P_m(x) = 2\Phi(\sqrt{x}) - 1 - \sqrt{\frac{2x}{\pi}}e^{-\frac{x}{2}}\left[1 + \frac{x}{3} + \frac{x^2}{3\cdot 5} + \cdots + \frac{x^{\frac{m-3}{2}}}{1\cdot 3\cdots(m-2)}\right] \quad \text{for odd } m.$$

6.2. *Stability of the chi-square distribution.* In Chapter V we saw that the sum of normally distributed variables is normally distributed. To prove here a similar theorem we compute *the characteristic function of Helmert's*

[2] K. Pearson, *Tables of the Incomplete Gamma Function.* London, 1922. See also R. A. Fisher and F. Yates, *Statistical Tables*, 2nd ed. London, 1943.

distribution. (See Chapter V, Problem 17.) We write $m = 2\lambda$ and C for the constant before the integral sign in (64), and $P(x)$ for $P_m(x)$. Thus,

$$P(x) = C \int_0^x t^{\lambda-1} e^{-t/2}\, dt\,.$$

The characteristic function $f(u)$ equals $f(u) = C \int_0^\infty t^{\lambda-1}\, e^{-\frac{1}{2}t+itu}\, dt$. With $(\frac{1}{2} - iu)\, t = \tau$ we obtain $f(u) = 2^\lambda(1 - 2iu)^{-\lambda} \cdot C \int_0^\infty \tau^{\lambda-1}\, e^{-\tau}\, d\tau$; the integration in the last integral is along the straight line $\tau = (\frac{1}{2} - iu)\, t$, u fixed, from $t = 0$ to $t = \infty$ and this is the same as integration along the real τ axis from $\tau = 0$ to $\tau = \infty$; hence the integral gives $\Gamma(\lambda)$ and as $C = [\Gamma(\lambda)\, 2^\lambda]^{-1}$ we obtain the characteristic function

$$f(u) = (1 - 2iu)^{-\lambda} = (1 - 2iu)^{-m/2}\,. \tag{68}$$

Consider now two independent random variables x_1 and x_2, each having χ^2-distribution with m_1 and m_2 degrees of freedom, respectively. The characteristic functions are $f(u) = (1 - 2iu)^{-\lambda_1}$ and $g(u) = (1 - 2iu)^{-\lambda_2}$ and their product is $(1 - 2iu)^{-(\lambda_1+\lambda_2)}$. Reasoning in the usual way we see that $x_1 + x_2$ have Helmert's distribution with $m_1 + m_2$ degrees of freedom. Hence, *if n independent random variables x_i, $i = 1, 2, \ldots, n$ each have χ^2-distributions with m_i, $i = 1, 2, \ldots, n$, degrees of freedom, respectively, their sum $x_1 + \cdots + x_n$ has a χ^2-distribution with $m_1 + \cdots + m_n$ degrees of freedom.*

6.3. *Distribution of the sum of squares of normally distributed variates.* We now consider n independently and normally distributed variables x_i, each with mean value zero and variance one. Then by the result of Section 6.1, each x_i^2 has a χ^2-distribution with one degree of freedom. Using the last result in Section 6.2 we obtain with an obvious generalization:

If $x_1, x_2, \ldots, x_n$ are independently and normally distributed, each with mean value α and variance σ^2, then the sum

$$\chi^2 = \frac{(x_1 - \alpha)^2 + (x_2 - \alpha)^2 + \cdots + (x_n - \alpha)^2}{\sigma^2} \tag{69}$$

has a χ^2-distribution with n degrees of freedom[3]:

$$\text{Prob}\left\{\frac{\sum_{\nu=1}^n (x_\nu - \alpha)^2}{\sigma^2} \leqslant x\right\} = \frac{1}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)} \int_0^x t^{\frac{n}{2}-1} e^{-\frac{t}{2}}\, dt\,. \tag{70}$$

[3] The χ^2-distribution may be *defined* as the distribution of (69); this is justified since (69) is an exact distribution valid for any n.

We note also that the distribution of $\chi^2/n = (1/n\sigma^2) \sum_{\nu=1}^{n} (x_\nu - \alpha)^2$ equals

$$\operatorname{Prob}\left\{\frac{\chi^2}{n} \leqslant x\right\} = (n)^{\frac{n}{2}} \frac{1}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)} \int_0^x t^{\frac{n}{2}-1} e^{-\frac{t}{2}}\, dt\,. \tag{70'}$$

In fact, $\operatorname{Prob}\{\chi^2/n \leqslant x\} = \operatorname{Prob}\{\chi^2 \leqslant nx\}$. Using the new variable $\tau = t/n$ we obtain (70′).

We use the result expressed in (70) together with the limit theorem of Chapter VI to find an *asymptotic formula for Helmert's distribution* as the number of degrees of freedom tends to infinity. Consider $x_1^2 + x_2^2 + \cdots + x_m^2$ where all x_i are normally distributed with mean 0 and variance 1. Then x_i^2 has mean 1 and variance 2 and the $\sum_{\mu=1}^{m} x_\mu^2$ has mean value m and variance $2m$. From the limit theorem for the sum of independent variates it follows that $\chi^2 = x_1^2 + \cdots + x_m^2$ is *asymptotically normal*, $N(m, 2m)$. (See also Problem 23.) Hence χ^2/m is asymptotically normal, $N(1, 2/m)$.

The χ^2-distribution converges slowly toward the normal distribution. Attempts have been made, therefore, to find transformations of χ^2 which are approximately normally distributed for moderate values of m. A very simple transformation is due to Fisher who has proved that for $m > 30$, $\sqrt{2\chi^2}$ is asymptotically normal with mean value $\sqrt{2m-1}$ and unit variance. (See Cramér, [4], p. 251.)

6.4. *Distribution of s^2.* Let us consider a sample of independent results each from a normal population with mean value α and variance σ^2. We wish to find the distribution of s^2, or rather, for reasons of homogeneity, of

$$\chi^2 = \frac{ns^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_{\nu=1}^{n} (x_\nu - \bar{x})^2\,, \tag{71}$$

where

$$\bar{x} = \frac{1}{n}\sum_{\nu=1}^{n} x_\nu\,, \qquad s^2 = \frac{1}{n}\sum_{\nu=1}^{n} (x_\nu - \bar{x})^2\,. \tag{72}$$

This is a more complex problem than that of Section 6.3 since the $\bar{x}$ in s^2 is no longer a constant. If we introduce $y_i = (x_i - \alpha)/\sigma$, each y_i is normally distributed with mean value 0 and variance 1 and

$$\chi^2 = \sum_{\nu=1}^{n} (y_\nu - \bar{y})^2 = \sum_{\nu=1}^{n} y_\nu^2 - \frac{1}{n}\left(\sum_{\nu=1}^{n} y_\nu\right)^2. \tag{73}$$

For $n = 3$ it is immediately seen that the quadric $\chi^2 = u$ is a cylinder in the y_1, y_2, y_3-space whose equation can, therefore, be reduced by an orthogonal transformation to the sum of *two* squares equal to a constant. We shall see that the analogous result holds for any n. By a suitable orthogonal transformation of the y_i into z_i, the right-hand side reduces to the sum of $n - 1$ squares in the new coordinates.[4] Let us show this.

For the first line of the orthogonal transformation we take

$$z_1 = \frac{1}{\sqrt{n}} y_1 + \frac{1}{\sqrt{n}} y_2 + \cdots + \frac{1}{\sqrt{n}} y_n , \tag{74}$$

where the sum of the squares of the coefficients equals one. *One can then find* $(n - 1)$ *further rows with coefficients* $a_{i1}, \ldots, a_{in}$, $i = 2, 3, \ldots, n$, *such that the complete matrix is orthogonal.*[5] The very simple proof of this proposition follows at the end of this subsection.

Since the transformation is orthogonal we have $\sum_{\nu=1}^n y_\nu^2 = \sum_{\nu=1}^n z_\nu^2$ and from (73) and (74),

$$\chi^2 = \sum_{\nu=1}^n y_\nu^2 - \frac{1}{n}\Big(\sum_{\nu=1}^n y_\nu\Big)^2 = \sum_{\nu=1}^n z_\nu^2 - z_1^2 = z_2^2 + z_3^2 + \cdots + z_n^2 , \tag{75}$$

which demonstrates that $\chi^2 =$ constant is a cylinder in n-space whose axis is in the z_1-direction.

Each $z_i = \sum a_{i\nu} y_\nu$ is a linear form of n independently and normally distributed variates y_ν. *Such a linear form is itself normally distributed.*[6] A proof will be given in Section 9.1. Let us now show that the z_i are mutually independent. The y_i are mutually independent normal variates, each with zero mean and unit variance. It follows that the $z_i = \sum_{\nu=1}^n a_{i\nu} y_\nu$ also have zero mean and that

$$E[z_i z_j] = \sum_{\nu=1}^n a_{i\nu} a_{j\nu} = \begin{cases} 1 & \text{for } i = j \\ 0 & \text{for } i \neq j , \end{cases}$$

i.e., that the new variables z_i are "uncorrelated." From the analytic form of a multi-variate normal distribution it follows (see Section 8 or 9) that n uncorrelated and normally distributed variates are also independent. Thus, the z_i are $N(0, 1)$ and mutually independent. Hence, it follows from the result of Section 6.3 that the sum $z_2^2 + z_3^2 + \cdots + z_n^2$ has a

[4] The n variates $y_\nu - \bar{y}$ contained in (73) for χ^2 are linearly dependent: $\sum_{\nu=1}^n (y_\nu - \bar{y}) = 0$.

[5] The proposition holds for *any* first line $z_1 = \sum_1^n a_{1\nu} y_\nu$, where $\sum_{\nu=1}^n a_{1\nu}^2 = 1$.

[6] That the *sum* of n independent, normally distributed variates is normally distributed was seen in Chapter V.

χ^2-distribution with $m = n - 1$ degrees of freedom. We then have the following result.

If the $x_1, \ldots, x_n$ *are normally and independently distributed,* $N(\alpha, \sigma^2)$, *then* $\chi^2 = (1/\sigma^2)\sum_{\nu=1}^{n}(x_\nu - \bar{x})^2 = ns^2/\sigma^2$, *where* $\bar{x} = (1/n)\sum_{\nu=1}^{n} x_\nu$ *has a* χ^2 *distribution with* $(n-1)$ *degrees of freedom.*[7]

COROLLARY. From Eq. (74) we had $\bar{y} = z_1/\sqrt{n}$. Hence, $\bar{y}$ depends on z_1 only, and the sum $\chi^2 = z_2{}^2 + \cdots + z_n{}^2$ is independent of $\bar{y}$. Hence, the result.

If $x_1, \ldots, x_n$ *is a sample of a normal distribution* $N(\alpha, \sigma^2)$, *then the* $\chi^2 = \dfrac{ns^2}{\sigma^2}$ of Eq. (71) *and the mean* $(\bar{x} - \alpha)/\dfrac{\sigma}{\sqrt{n}}$ *are independently distributed.* The first distribution is the χ^2-distribution (64) with the number of d.o.fr. equal to $n - 1$, the second is normal $N(0, 1)$. (Note that $\sigma/\sqrt{n}$ is the standard deviation of $\bar{x}$.)

We have still to prove the proposition of p. 408. We denote by $\mathbf{a}_1$ the unit vector in n-space with coordinates $1/\sqrt{n}, 1/\sqrt{n}, \ldots, 1/\sqrt{n}$. We have to find a vector $\mathbf{b}_2$ non-zero and normal to $\mathbf{a}_1$. If $n > 2$ this is possible in infinitely many ways and $\mathbf{b}_2/|\mathbf{b}_2| = \mathbf{a}_2$ is a unit vector. Next we find $\mathbf{a}_3$ such that $\mathbf{a}_1 \cdot \mathbf{a}_3 = 0$, $\mathbf{a}_2 \cdot \mathbf{a}_3 = 0$, and $\mathbf{a}_3 =$ unit vector. This is possible in at least one way if $n \geqslant 3$. Finally we have to find $\mathbf{a}_n$, as a unit vector in n-space, normal on $n - 1$ given mutually normal vectors; this is possible in one way and our lemma is proved.

6.5. *Direct computation of* $P(x) = Pr\{\chi^2 \leqslant x\}$. It is of interest to obtain our result of Sect. 6.4 also by direct evaluation of the defining integral. Since the y_i are all independent, $N(0, 1)$, their joint distribution is

$$\phi(y_1, \ldots, y_n) = (2\pi)^{-n/2} e^{-\frac{1}{2}(y_1{}^2 + \ldots + y_n{}^2)} \tag{76}$$

and

$$P(x) = \Pr\{\chi^2 \leqslant x\} = \int_{\chi^2 \leqslant x} \cdots \int \phi(t_1, \ldots, t_n)\, dt_1 \cdots dt_n\,. \tag{77}$$

The coefficient determinant Δ of the orthogonal transformation from the y_i to the z_i, viz., $\Delta = 1$, is at the same time the Jacobian for the transformation of the last integral which, therefore, becomes

$$P(x) = (2\pi)^{-n/2} \int_{\chi^2 \leqslant x} \cdots \int e^{-\frac{1}{2}(z_1^2 + \ldots + z_n{}^2)}\, dz_1 \cdots dz_n\,. \tag{77'}$$

[7] This result was first proved by Helmert, *Astron. Nachr.* **88** (1876), pp. 113–132.

As the domain of integration does not depend on z_1 we obtain for $P(x)$

$$P(x) = (2\pi)^{-\frac{n}{2}}\int_{-\infty}^{\infty} e^{-z_1^2/2}\, dz_1 \int_{\chi^2 \leqslant x}\cdots\int e^{-\frac{1}{2}(z_2^2+\ldots+z_n^2)}\, dz_2 \ldots dz_n$$

$$= (2\pi)^{-\frac{n-1}{2}}\int_{\chi^2 \leqslant x}\cdots\int e^{-\frac{1}{2}(z_2^2+\ldots+z_n^2)}\, dz_2 \ldots dz_n . \tag{77''}$$

To evaluate the last integral we use polar coordinates with $r^2 = z_2^2 + z_3^2 + \cdots + z_n^2$. The integrand then becomes $e^{-r^2/2}$. The general p-dimensional volume element $dx_1\, dx_2 \cdots dx_p$ is in polar coordinates of the form $r^{p-1}\, dr\, d\omega$ where $d\omega$ is a function of the $p-1$ polar angles $\phi_1, \phi_2, \ldots, \phi_{p-1}$ multiplied by $d\phi_1\, d\phi_2 \cdots d\phi_{p-1}$. (For $p = 2 : dx_1\, dx_2 = r\, dr\, d\phi$. For $p = 3 : dx_1\, dx_2\, dx_3 = r^2\, dr \sin\phi_1\, d\phi_1\, d\phi_2$.) The integral (77'') thus takes the form

$$P(x) = (2\pi)^{-\frac{n-1}{2}} \int_{\chi^2 \leqslant x}\cdots\int e^{-r^2/2}\, r^{n-2}\, dr\, d\omega . \tag{77'''}$$

Since in (77''') the integration region depends on r^2 only, the integration over $d\omega$ can be carried out and supplies merely a constant, so that

$$P(x) = C\int_{r^2 \leqslant x} e^{-r^2/2}\, r^{n-2}\, dr = C\int_0^{\sqrt{x}} e^{-r^2/2} r^{n-2}\, dr . \tag{78}$$

The constant may be evaluated either by carrying out the integration in (77''') over $d\omega$ (this leads to $\int d\omega = \dfrac{2}{\Gamma\left(\frac{n-1}{2}\right)}\pi^{\frac{n-1}{2}}$ or making use of $P(\infty) = 1$. Introducing finally $r^2 = t$, we obtain

$$P(x) = \frac{1}{\Gamma\left(\frac{n-1}{2}\right)2^{\frac{n-1}{2}}}\int_0^x t^{\frac{n-3}{2}}\, e^{-\frac{t}{2}}\, dt , \tag{64'}$$

i.e., the result found in the preceding subsection.

Note that this is not an asymptotic result; it is valid for any n.

6.6. *The χ^2-distribution as a limit distribution.* In Sections 6.3 and 6.4 we have found that Helmert's distribution is the exact distribution (i.e., valid for any n) of important statistics. Let us recall that in Chapter

VII this distribution appeared as a "limit distribution." There *Helmert's distribution appeared as an asymptotic expression for the probability of inference, in the case of so-called rare events, taking the place of the Gaussian distribution* (44) of Chapter VII. This result is parallel to the well-known result in the "direct" problem where, in the case of rare events, the Poisson distribution holds asymptotically (Chapter IV, Section 5) in place of the Gaussian distribution (Chapter VI, Section 2.2). In Chapter IX, Section 4, Helmert's distribution will appear as the asymptotic distribution of an important test function.

Problem 21. What is Helmert's distribution for 2 degrees of freedom?

Problem 22. Compute the mode, the skewness, and the kurtosis of Helmert's distribution with m degrees of freedom. Study the inequalities of Section 5 for Helmert's distribution.

Problem 23. Prove that Helmert's distribution approaches a normal distribution as m goes to infinity. Introduce instead of x the new variable $z = (x - m)/2\sqrt{m}$ and proceed in the same way as in the derivation of the Stirling formula of Chapter VI.

Problem 24. Prove the stability of Helmert's distribution without the use of the characteristic function by applying the formula for the distribution of a sum (Chapter V).

Problem 25. The probability of

$$Z^2 = \frac{1}{\sigma^2}\sum_{\nu=1}^{n}(x_\nu - \alpha)^2$$

was given by Eq. (70). Prove that the distribution of Z is

$$\text{Prob}\{Z \leqslant x\} = \frac{2}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)}\int_0^x t^{\frac{n}{2}-1}e^{-\frac{t^2}{2}}\,dt$$

and that of $Z/\sqrt{n}$ is given by

$$\text{Prob}\left\{\frac{Z}{\sqrt{n}} \leqslant x\right\} = \frac{\sqrt{2n}}{\Gamma\left(\frac{n}{2}\right)}\left(\frac{n}{2}\right)^{\frac{n-1}{2}}\int_0^x t^{n-1}e^{-\frac{nt^2}{2}}\,dt\,.$$

7. Student's Distribution and F Distribution

7.1. *Student's distribution.* Let us consider the quotient of two variables u and z, where u is normally distributed with mean 0 and variance 1 and z has Helmert's distribution with m degrees of freedom. Specifically, we shall obtain the distribution of $t = \frac{u}{\sqrt{z/m}}$.

The joint density of u and z is

$$p(u, z)\, du\, dz = \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} \frac{1}{2^{\frac{m}{2}} \Gamma\left(\frac{m}{2}\right)} z^{\frac{m}{2}-1} e^{-\frac{z}{2}} du\, dz\,,$$

since z and u are independent variables by hypothesis. Introducing t as a new variable instead of u we have $u = \sqrt{z/m}\, t$, $du = \sqrt{z/m}\, dt$ and we obtain for the density element in z and t

$$\bar{p}(t, z) = \frac{1}{\sqrt{2\pi}} \exp\left[-\tfrac{1}{2}\left(t^2 \frac{z}{m} + z\right)\right] \frac{1}{2^{\frac{m}{2}} \Gamma\left(\frac{m}{2}\right)} z^{\frac{m}{2}-1} \sqrt{\frac{z}{m}}\,.$$

To find the distribution of t we have merely to compute the marginal density $p_m(t)$ by integration of $\bar{p}(t, z)$ over z from 0 to ∞:

$$p_m(t) = \frac{1}{\sqrt{2}\sqrt{2\pi}\sqrt{m}\,\Gamma\left(\frac{m}{2}\right)} \int_0^\infty \left(\frac{z}{2}\right)^{\frac{m-1}{2}} \exp\left[-\tfrac{1}{2} z\left(1 + \frac{t^2}{m}\right)\right] dz \tag{79}$$

and with $(z/2)(1 + (t^2/m))$ a new variable we obtain by simple computation

$$p_m(t) = \frac{\Gamma\left(\frac{m+1}{2}\right)}{\sqrt{m\pi}\,\Gamma\left(\frac{m}{2}\right)} \left(1 + \frac{t^2}{m}\right)^{-\frac{m+1}{2}} \tag{80}$$

The distribution defined by (80) is known as *Student's distribution* or the *t-distribution.* The parameter m in (80) is called the number of degrees of freedom. It was first used by W. S. Gosett, better known under his pen name, Student, in an important problem which we shall consider in Chapter IX, Section 3.

For $p_m(t)$ of (80) the mean value is zero since it is an even function, and the same is true for all moments of odd order. The curve is bell shaped like the normal density curve and for $m \to \infty$ it converges toward the normal curve $\phi(t)$. In fact, $\Gamma\left(\frac{m+1}{2}\right)\Big/\Gamma\left(\frac{m}{2}\right)$ is of the same order of infinity as $\sqrt{\frac{m}{2}}$, and $\lim_{m\to\infty}\left(1 + \frac{t^2}{m}\right)^{-\frac{m+1}{2}} = e^{-\frac{t^2}{2}}$. Figure 31 shows

Student's distribution for $m = 5$. For the variance of $p_m(t)$ we find

$$\int_{-\infty}^{+\infty} x^2 p_m(x)\,dx = \frac{m}{m-2}\,. \tag{81}$$

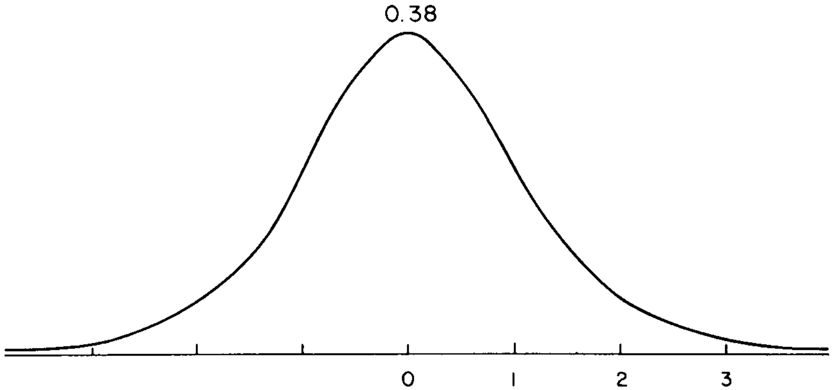

FIG. 31. Student distribution for $m = 5$.

It is seen that for $m < 3$ no finite variance exists. For $m = 1$ we have

$$p_1(t) = \frac{1}{\pi(1+t^2)}\,, \qquad P_1(t) = \frac{1}{\pi}\arctan t + \frac{1}{2}\,, \tag{82}$$

the Cauchy distribution. For $m = 2$, the definition (80) gives

$$p_2(t) = (2+t^2)^{-3/2}, \qquad P_2(t) = \frac{t}{2\sqrt{2+t^2}} + \frac{1}{2}\,. \tag{83}$$

By integration of (80) the general expression for the c.d.f. can be found as follows (with $t' = t/\sqrt{m}$):

$$\begin{aligned} P_m(t) &= \frac{1}{2} + \frac{t'}{2\sqrt{1+t'^2}}\left[1 + \frac{1}{2}(1+t'^2)^{-1} + \frac{1\cdot 3}{2\cdot 4}(1+t'^2)^{-2} + \cdots \right. \\ &\qquad \left. + \frac{1\cdot 3\cdots(m-3)}{2\cdot 4\cdots(m-2)}(1+t'^2)^{-\frac{m-2}{2}}\right] \\ &\qquad\qquad \text{for even } m, \\ P_m(t) &= \frac{1}{2} + \frac{1}{\pi}\arctan t' + \frac{t'}{\pi(1+t'^2)} \\ &\qquad \times\left[1 + \frac{2}{3}(1+t'^2)^{-1} + \cdots + \frac{2\cdot 4\cdots(m-3)}{3\cdot 5\cdots(m-2)}(1+t'^2)^{-\frac{m-3}{2}}\right] \\ &\qquad\qquad \text{for odd } m. \end{aligned} \tag{84}$$

A table[1] for $P_m(t)$ is given here as Table VII. The table shows for several values of m the quantity x for which $\Pr\{|t| > x\} = 2\int_x^\infty p_m(t)\,dt = 2 - 2P_m(x)$ assumes the value p indicated in the respective line; for example, with $m = 9$, $x = 0.398$, the probability that $|t| \geqslant x$, i.e., $1 - P_9(x) + P_9(-x) = 2 - 2P_9(x)$, equals 70%. In particular, the 50% limits (or "probable limits") for t in the case $m = 9$ are seen to be ± 0.703. The symmetrical form of the distribution shows that (for a given m)

$$\Pr\{t \leqslant x\} = P \qquad \text{implies} \qquad \Pr\{t < -x\} = 1 - P,$$

and

$$\Pr\{|t| \leqslant x\} = P - (1 - P) = 2P - 1;$$

for example, for $m = 8$: $\Pr\{t \leqslant 1.86\} = 95\%$, $\Pr\{t \leqslant -1.86\} = 5\%$, $\Pr\{|t| \leqslant 1.86\} = 90\%$, $\Pr\{|t| > 1.86\} = 10\%$. We shall return to the Student distribution in Chapter IX, Section 3.1.

Problem 26. Prove, using the theorems of Chapter VI, Section 1, that the Student distribution approaches the Gaussian distribution $N(0, 1)$ as m increases indefinitely.

Problem 27. Let $p_1(\xi)\,p_2(\eta)$ be the joint density of ξ and η, where $-\infty < \xi < +\infty$, $0 < \eta < \infty$. Put $t = \xi/\eta$ and prove that $P(x) = \operatorname{Prob}\{t < x\} = \int_0^\infty P_1(xz)\,p_2(z)\,dz$, where $P_1(\xi)$ is the distribution of ξ, and that

$$p(x) = \frac{dP}{dx} = \int_0^\infty z p_1(xz) p_2(z)\,dz\,.$$

(*Hint*: See Chapter III, p. 147.)

Problem 28. Use the result of the preceding problem and a result of Problem 25 to recover Student's distribution.

7.2. *The F distribution.* We were led to the Student distribution as the distribution of the quotient of two independent variates—the first normally distributed and the second having a χ^2-distribution. We wish now to determine *the distribution of the ratio of two quantities z_1 and z_2, both distributed independently in χ^2-distributions with m_1 and m_2 degrees of freedom, respectively.* The density of z_1 is then,

$$g_1(z_1) = d_1 z_1^{\frac{m_1}{2}-1} e^{-\frac{z_1}{2}} \qquad \text{with} \qquad d_1 = 2^{-\frac{m_1}{2}} \Gamma\left(\frac{m_1}{2}\right)^{-1}$$

and similarly for z_2, and the joint distribution of (z_1, z_2) is the product $g_1(z_1)\,g_2(z_2)$.

[1] See Fisher [8].

We introduce

$$F = \frac{z_1/m_1}{z_2/m_2} \tag{85}$$

and make a change of variables introducing the new variables

$$F = \frac{z_1/m_1}{z_2/m_2}, \qquad u = z_2$$

then

$$z_1 = u\frac{m_1}{m_2}F, \qquad z_2 = u.$$

The Jacobian of the transformation is $u\,(m_1/m_2)$ and the new density of F and u is

$$c \cdot F^{\frac{m_1}{2}-1} u^{\frac{m_1+m_2}{2}-1} \exp\left[-\frac{1}{2}\left(u + \frac{m_1}{m_2}Fu\right)\right], \qquad c = d_1 d_2 \left(\frac{m_1}{m_2}\right)^{\frac{m_1}{2}}.$$

Integrating with respect to u from 0 to ∞ we obtain the density of the distribution of F, namely, for $0 \leqslant F < \infty$,

$$h(F) = K \cdot F^{\frac{m_1}{2}-1}\left(1 + \frac{m_1}{m_2}F\right)^{-\frac{m_1+m_2}{2}}, \tag{86}$$

$$K = \frac{\Gamma\left(\frac{m_1+m_2}{2}\right)}{\Gamma\left(\frac{m_1}{2}\right)\Gamma\left(\frac{m_2}{2}\right)}\left(\frac{m_1}{m_2}\right)^{\frac{m_1}{2}}.$$

This is the density of *Snedecor's F-distribution*,[2] for m_1 and m_2 degrees of freedom. The mean value of F is $m_2/(m_2 - 2)$, $m_2 > 2$, (see problem 30) which for large values of m_2 is approximately equal to one. The cumulative d.f. is

$$H(w) = K\int_0^w F^{\frac{m_1}{2}-1}\left(1 + \frac{m_1}{m_2}F\right)^{-\frac{m_1+m_2}{2}} dF. \tag{86'}$$

If we introduce $y = \dfrac{m_1}{m_2}F$ as a new variable we obtain

$$K(v) = C\int_0^v y^{\frac{m_1}{2}-1}(y+1)^{-\frac{m_1+m_2}{2}} dy, \qquad C = \frac{\Gamma\left(\frac{m_1+m_2}{2}\right)}{\Gamma\left(\frac{m_1}{2}\right)\Gamma\left(\frac{m_2}{2}\right)}. \tag{87}$$

[2] F. Snedecor, *Statistical Methods*, 5th ed. Iowa, 1962.

The integral is an incomplete Beta function; it can be easily computed for integral m_1 and m_2. It has been tabulated in K. Pearson, *Tables of the Incomplete Beta Function*, London 1934.

Formula (86′) has been tabulated by Snedecor for certain values of w and certain combinations of m_1 and m_2.

The so-called F-test based on the F-distribution will be considered in Chapter IX, Section 3.4.[3]

Problem 29. Show that by the change of variable

$$x = \frac{\frac{m_1}{m_2}F}{\left(1 + \frac{m_1}{m_2}F\right)} = \frac{m_1 F}{m_2 + m_1 F}$$

the density $h(F)$ of Eq. (86) transforms into

$$\frac{1}{B\left(\frac{m_1}{2}, \frac{m_2}{2}\right)} x^{\frac{m_1}{2}} (1 - x)^{\frac{m_2}{2} - 1}, \quad \text{where} \quad B\left(\frac{m_1}{2}, \frac{m_2}{2}\right) = \frac{\Gamma\left(\frac{m_1 + m_2}{2}\right)}{\Gamma\left(\frac{m_1}{2}\right)\Gamma\left(\frac{m_2}{2}\right)},$$
$$0 \leqslant x \leqslant 1.$$

Problem 30. Prove that the mean of the F distribution is $m_2/(m_2 - 2)$. (*Hint*: Use the result of Problem 29.)

D. Multivariate Normal Distribution (Sections 8 and 9)

8. Normal Distribution in Three Dimensions

8.1. *Definition.* In this and the following section we shall study some properties of the *normal* or *Gaussian distribution in n variables.* The results of this section hold for any number of variables. We shall write them, however, for three variables only, for reasons of simplicity and concreteness. In Section 9 we shall establish some properties of the n-dimensional normal distribution in terms of matrix notation.

We consider a density of the form

$$p(x, y, z) = Ce^{-F(x,y,z)}, \tag{88}$$

[3] Since we hardly apply this test, used mainly in the "analysis of variance," no table is given in this book. Extensive tabulations are in the work of Snedecor quoted above.

where C is a constant and $F(x, y, z)$ a polynomial of second degree in x, y, z. Such a polynomial is the sum of second order terms, of linear terms, and of a constant (absolute term). The latter can be omitted and absorbed in the C. We therefore set

$$F(x, y, z) = Q(x, y, z) + L(x, y, z) \tag{89}$$

with

$$Q = \tfrac{1}{2}(a_{11}x^2 + a_{22}y^2 + a_{33}z^2 + 2a_{12}xy + 2a_{23}yz + 2a_{31}zx) \tag{89'}$$

$$L = b_1x + b_2y + b_3z\,. \tag{89''}$$

The first condition $p(x, y, z)$ has to fulfill is that, with $dx\,dy\,dz = dS$:

$$\int p\,dS = 1, \tag{90}$$

hence $\int e^{-F}\,dS$ must have a finite value. This imposes a restriction upon Q only, rather than on $F = Q + L$, since for large values of x, y, z the second order terms prevail over the linear ones. Now, if Q has the value Q_1 at some point P_1 at a distance 1 from the origin, its value at the point P lying on the line OP_1 at a distance r is r^2Q_1 since each coordinate is multiplied by r: the Q-value does not change its sign on a line passing through O. Were the minimum of all Q_1-values negative, there would be a portion of the space in which e^{-Q}, and therefore e^{-F}, tends toward infinity as r increases. In these circumstances, the above integral could not be finite. Were Q to vanish at some point P_1, it would be zero at all points of the straight line OP_1. Taking this line as the x'-axis of a new coordinate system, the expression for Q in x', y', z' would be independent of x' (vanishing wherever $y' = z' = 0$). This too would make the three-dimensional integral infinite since $\int \text{const.}\,dx'$ is infinite. We therefore make the assumption that Q is positive on *all* lines passing through O, that is everywhere except at the origin itself. An expression of the form (89′), which takes only positive values except at the origin, is called a positive definite quadric or a positive definite quadratic form.

DEFINITION. *A distribution in n dimensions is called a normal distribution in n dimensions if its density function has the form* (88) *with the second order terms of F being a positive definite quadric in n variables.*

A set of necessary and sufficient conditions for Q to be positive definite has already been given in Section 4.3, in (33). They read (if a_{ik} and a_{ki} are used indiscriminately)

$$a_{11} > 0, \qquad \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} > 0, \qquad \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} > 0\,, \tag{91}$$

and so forth in the case of more than 3 variables. If (91) holds, the matrix of the a_{ik} *is of rank three.* If Q is non-negative but vanishes at points other than the origin it is called *semi-definite.* In this case the last of the determinants (91) must be zero and the matrix of the a_{ik} is of rank two or of rank one. As noted before, the corresponding distribution cannot exist as a three-dimensional distribution. If the degenerate matrix is of rank two, it reduces, in appropriate coordinates, to a positive definite quadratic form in two variables. It reduces to one in one variable if the rank of the matrix is one. The corresponding degenerate three-dimensional normal distribution may be regarded as an ordinary normal distribution in two variables or in one variable, according to whether the rank of the matrix of the a_{ik} is two or one.

We return to the non-degenerate case of positive definite Q, where all three conditions (91) hold. In this case, the value of Q at all points P_1 at the distance $OP_1 = 1$ is a positive quantity q. Therefore, for sufficiently large r the e^{-Q} and e^{-F} approach zero at least as e^{-qr^2}. It follows that not only the integral of $p\,dS$ but also, more generally, all integrals of the form $\int Pp\,dS$ where P is any polynomial in x, y, z, exist. All products of the form Pp vanish strongly at infinity.

8.2. *Mean value and variance.* The *mean value* of a three-dimensional distribution such as (88) is a point with the coordinates a, b, c (or a vector with the components a, b, c) where

$$a = \int xp\,dS, \qquad b = \int yp\,dS, \qquad c = \int zp\,dS\,. \tag{92}$$

The *variance* is given by a quadratic symmetrical scheme, a *symmetrical matrix*

$$\begin{pmatrix} s_{11} & s_{12} & s_{13} \\ s_{21} & s_{22} & s_{23} \\ s_{31} & s_{32} & s_{33} \end{pmatrix}, \tag{93}$$

where

$$\begin{gathered} s_{11} = \int (x-a)^2\,p\,dS, \qquad s_{22} = \int (y-b)^2\,p\,dS, \ldots, \\ s_{12} = s_{21} = \int (x-a)(y-b)\,p\,dS, \ldots\,. \end{gathered} \tag{94}$$

The fact that we write these formulas for $n = 3$ variables is obviously immaterial.

The remark at the end of Section 8.1 leads to a simple method of computing mean values and variances. From

$$\int \frac{\partial p}{\partial x}\,dx = p\Big]_{-\infty}^{\infty} = 0, \ldots, \int \frac{\partial}{\partial x}(xp)\,dx = xp\Big]_{-\infty}^{\infty} = 0, \quad \int \frac{\partial}{\partial y}(xp)\,dy = 0, \qquad \text{etc.,}$$

it follows that the triple integrals

$$\int \frac{\partial p}{\partial x}\,dS = \int \frac{\partial p}{\partial y}\,dS = \int \frac{\partial p}{\partial z}\,dS = 0 \tag{95}$$

and

$$\int \frac{\partial}{\partial x}(xp)\,dS = \int \frac{\partial}{\partial y}(xp)\,dS = \int \frac{\partial}{\partial z}(xp)\,dS = 0\,. \tag{96}$$

By differentiation, we have from (88), (89), (89′), and (89″)

$$\frac{\partial p}{\partial x} = -Ce^{-F}\frac{\partial F}{\partial x} = -p\,\frac{\partial F}{\partial x} = -p\,(a_{11}x + a_{12}y + a_{13}z + b_1)\,. \tag{97}$$

Introducing this into the first Eq. (95) and applying (92) and (89), we find

$$\begin{aligned} 0 = \int \frac{\partial p}{\partial x}\,dS &= -a_{11}\int xp\,dS - a_{12}\int yp\,dS - a_{13}\int zp\,dS - b_1\int p\,dS \\ &= -a_{11}a - a_{12}b - a_{13}c - b_1\,. \end{aligned}$$

In exactly the same way, differentiation with respect to y (or z) leads to two analogous relations. The set of three equations for the unknowns a, b, c reads

$$\begin{aligned} a_{11}a + a_{12}b + a_{13}c &= -b_1 \\ a_{21}a + a_{22}b + a_{23}c &= -b_2 \\ a_{31}a + a_{32}b + a_{33}c &= -b_3 \end{aligned} \tag{98}$$

and by the last condition (91) the determinant of these equations

$$D = \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} \tag{99}$$

is not zero. If the cofactor of the element a_{ik} in (99) is called D_{ik}, e.g., with $D_{ik} = D_{ki}$:

$$D_{11} = \begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix}, \quad D_{12} = -\begin{vmatrix} a_{11} & a_{13} \\ a_{31} & a_{33} \end{vmatrix}, \quad D_{13} = \begin{vmatrix} a_{21} & a_{22} \\ a_{31} & a_{32} \end{vmatrix}, \qquad \text{etc.,} \tag{100}$$

the solution of (98) is given by

$$a = -\frac{1}{D}\sum_{i=1}^{3} b_i D_{i1}\,, \qquad b = -\frac{1}{D}\sum_{i=1}^{3} b_i D_{i2}\,, \qquad c = -\frac{1}{D}\sum_{i=1}^{3} b_i D_{i3}\,. \tag{101}$$

It is seen from (101) that $a = b = c = 0$ if and only if $b_1 = b_2 = b_3 = 0$, i.e., if all linear terms of F vanish. (The three cofactors, D_{i1} , $i = 1, 2, 3$ or D_{i2} or D_{i3} , cannot vanish simultaneously without D itself vanishing.)

If the number of variables is greater or smaller than 3, the modification of Eqs. (101) is obvious. For example, in the case of one variable x, we have

$$p(x) = Ce^{-\frac{1}{2}(a_{11}x^2+b_1x)}; \qquad D = a_{11}\,, \quad D_{11} = 1, \quad a = -\frac{b_1}{a_{11}}\,, \tag{102}$$

where $a_{11} > 0$.

We introduce new coordinates

$$x' = x - a, \qquad y' = y - b, \qquad z' = z - c, \tag{103}$$

in other words, we shift the origin to the mean. If we introduce these x', y', z' into F, the coefficients of the second-order terms x'^2, $x'y'$, ... are the same as in the original a, namely, a_{11} , $2a_{12}$, ..., and the linear terms are all zero since in the new coordinate system the mean coincides with the origin; there is now an absolute term, $F(a, b, c)$. Thus, setting

$$C' = Ce^{-F(a,b,c)} \tag{104}$$

the expression for p in the new coordinates is

$$p = C'e^{-Q(x',y',z')} = C'e^{-Q(x-a,y-b,z-c)}\,. \tag{105}$$

The first component of the variance is given by

$$s_{11} = \int (x-a)^2\, p\, dS = C' \int x'^2 e^{-Q(x',y',z')}\, dS' = C' \int x^2 e^{-Q(x,y,z)}\, dS\,. \tag{106}$$

since $dS = dS'$. Since C' is determined by the condition $\int p\, dS = 1$ and therefore according to (105), $1/C'$ equals $\int e^{-Q}\, dS$, Eq. (106) shows that the variance does not depend on the linear terms in F. Hence, we may compute the s_{11} , s_{12} , ... as if the distribution were given by the expression

$$P(x, y, z) = C'e^{-Q(x,y,z)} \tag{107}$$

We now proceed as in the computation of the mean, using Eqs. (96) instead of the Eqs. (95). Note that

$$\begin{aligned}
\frac{\partial}{\partial x}(xp) &= p + x\frac{\partial p}{\partial x} = p[1 - a_{11}x^2 - a_{12}xy - a_{13}xz] \\
\frac{\partial}{\partial y}(xp) &= \phantom{p + {}} x\frac{\partial p}{\partial y} = p[- a_{21}x^2 - a_{22}xy - a_{23}xz] \\
\frac{\partial}{\partial z}(xp) &= \phantom{p + {}} x\frac{\partial p}{\partial z} = p[- a_{31}x^2 - a_{32}xy - a_{33}xz] \,.
\end{aligned} \tag{108}$$

Introducing this into (96), and using the expression (107) so that

$$s_{11} = \int x^2 p \, dS, \qquad s_{12} = \int xyp \, dS, \qquad s_{13} = \int xzp \, dS, \tag{109}$$

we find the following three equations for the three unknowns s_{11}, s_{12}, s_{13}

$$\begin{aligned}
a_{11}s_{11} + a_{12}s_{12} + a_{13}s_{13} &= 1 \\
a_{21}s_{11} + a_{22}s_{12} + a_{23}s_{13} &= 0 \\
a_{31}s_{11} + a_{32}s_{12} + a_{33}s_{13} &= 0 \,.
\end{aligned} \tag{110}$$

The determinant of these equations is the same as that of Eqs. (98), but on the right-hand side of (110) we have 1, 0, 0 instead of $-b_1$, $-b_2$, $-b_3$. Thus, the solution of (110) is

$$s_{11} = \frac{D_{11}}{D}, \qquad s_{12} = \frac{D_{12}}{D}, \qquad s_{13} = \frac{D_{13}}{D}. \tag{111}$$

If we choose y (or z) instead of x on the left-hand side of (108) we obtain in the same way

$$\begin{aligned}
s_{21} &= \frac{D_{21}}{D}, \qquad s_{22} = \frac{D_{22}}{D}, \qquad s_{23} = \frac{D_{23}}{D}, \\
s_{31} &= \frac{D_{31}}{D}, \qquad s_{32} = \frac{D_{32}}{D}, \qquad s_{33} = \frac{D_{33}}{D}.
\end{aligned} \tag{111'}$$

This solves our problem. In algebra one calls a matrix composed of the ratios D_{ki}/D, where D_{ik} is the cofactor of a_{ik} in D, the "inverse" matrix or the matrix "reciprocal" to the original matrix. Note that here Q is symmetric and $D_{ik} = D_{ki}$. Our result can, therefore, be expressed in the simple statement: *The matrix of the variance components of a normal distribution is the inverse (reciprocal) to the matrix of the quadratic form in the exponent of the distribution.*

In the case of one dimension, see Eq. (102), we have $D = a_{11}$, $D_{11} = 1$, $s_{11} = 1/a_{11}$, which agrees with our earlier formula $s^2 = 1/2h^2$. In two dimensions, (see Chapter III, Section 4) with

$$D = a_{11}a_{22} - a_{12}^2, \qquad D_{11} = a_{22}, \qquad D_{22} = a_{11}, \qquad D_{12} = -a_{12} \tag{112}$$

the explicit formulas can be written as

$$\begin{aligned} p &= C \exp\left[-\tfrac{1}{2}(a_{11}x^2 + 2a_{12}xy + a_{22}y^2) - \tfrac{1}{2}(b_1x + b_2y)\right], \\ a &= \frac{b_2a_{12} - b_1a_{22}}{D}, \qquad b = \frac{b_1a_{12} - b_2a_{11}}{D}, \\ s_{11} &= \frac{a_{22}}{D}, \qquad s_{12} = -\frac{a_{12}}{D}, \qquad s_{22} = \frac{a_{11}}{D}. \end{aligned} \tag{113}$$

Note the minus sign in s_{12}. Ordinarily, in the case $n = 2$, the density function of a normal distribution for given a, b, s_{11}, s_{12}, s_{22} is given as follows:

$$\begin{aligned} p(x, y) &= \frac{1}{2\pi \sqrt{s_{11}}\sqrt{s_{22}}\sqrt{1 - r^2}} \\ &\times \exp\left(-\frac{1}{2(1 - r^2)}\left[\frac{(x - a)^2}{s_{11}} - 2r\frac{x - a}{\sqrt{s_{11}}}\frac{y - b}{\sqrt{s_{22}}} + \frac{(y - b)^2}{s_{22}}\right]\right) \end{aligned} \tag{113'}$$

where

$$r = \frac{s_{12}}{\sqrt{s_{11}}\sqrt{s_{22}}}.$$

8.3. *The constant.* There remains the question of the determination of the C in (88) or the C' in (107), the two being connected by (104). The answer is readily found if in the expression for Q the bilinear terms are absent, i.e., if $a_{12} = a_{23} = a_{31} = 0$. Then the triple integral of $p\,dx\,dy\,dz$ reduces to the product of three simple integrals. Writing $Q = \frac{1}{2}(\lambda_1x^2 + \lambda_2y^2 + \lambda_3z^2)$, we obtain

$$\begin{aligned} 1 &= \int p\,dS = C' \int e^{-(\lambda_1/2)x^2}\,dx \int e^{-(\lambda_2/2)y^2}\,dy \int e^{-(\lambda_3/2)z^2}\,dz \\ &= C'\sqrt{\frac{2\pi}{\lambda_1}}\sqrt{\frac{2\pi}{\lambda_2}}\sqrt{\frac{2\pi}{\lambda_3}}, \quad C' = \sqrt{\frac{\lambda_1\lambda_2\lambda_3}{(2\pi)^3}}, \end{aligned} \tag{114}$$

and a corresponding relation holds for any number of dimensions.

In the case of a general Q including all 6 terms, it is shown in algebra and analytic geometry that at least one system of reference can always be found so that by rotating the original coordinate axes, in the new coordinates $\bar{x}$, $\bar{y}$, $\bar{z}$, called *principal axes*, the quadric Q becomes

$$2Q = \lambda_1 \bar{x}^2 + \lambda_2 \bar{y}^2 + \lambda_3 \bar{z}^2 ,$$

where λ_1, λ_2, λ_3 are the roots of the equation

$$\begin{vmatrix} a_{11} - \lambda & a_{12} & a_{13} \\ a_{21} & a_{22} - \lambda & a_{23} \\ a_{31} & a_{32} & a_{33} - \lambda \end{vmatrix} = 0 . \tag{115}$$

In expanding this determinant, it is seen that the highest term is $-\lambda^3$ and the absolute term (i.e., the value of the left-hand expression for $\lambda = 0$) is the determinant D defined in (99). Thus (114) can be written as

$$-\lambda^3 + A\lambda^2 + B\lambda + D = 0 . \tag{116}$$

We are not interested here in A and B. We know that if this equation has three *real* roots λ_1, λ_2, λ_3 (and their existence is included in the assumption that three principal axes exist) their product must equal the absolute term D. This is all we need. In formula (114) for C', the product $\lambda_1\lambda_2\lambda_3$ has to be replaced by D in order to cover the general case. The result is

$$C' = \sqrt{\frac{D}{(2\pi)^3}} , \qquad C = e^{F(a,b,c)} \sqrt{\frac{D}{(2\pi)^3}} . \tag{117}$$

(In the case of n dimensions, the exponent of 2π is n instead of 3.)

The density function of a normal distribution (in the positive definite case) can now be written in either one of the forms

$$p(x, y, z) = \sqrt{\frac{D}{(2\pi)^3}} \, e^{-Q(x-a, y-b, z-c)} = e^{F(a,b,c)} \sqrt{\frac{D}{(2\pi)^3}} \, e^{-F(x,y,z)} , \tag{118}$$

with Q, F defined in (89) and (89′) and D given in (99).

Problem 31. Show that in the case $n = 2$, the density function of a normal distribution for given a, b, s_{11}, s_{12}, s_{22} can be written as in (113), (113′). Take in particular $a = b = 0$, $s_{11} = s_{22} = 1$, $s_{12} = \frac{1}{2}$. What are the lines $p =$ constant in this case? Show that the assumption $s_{11} = s_{22} = s_{12} = 1$ would not lead to a normal distribution in two variables.

Problem 32. The probability of x, y, z is normally distributed with $Q = \frac{1}{2}(2x^2 + y^2 + 2z^2 + xy + yz)$. Using a Tchebycheff inequality, find the lower bound for the probability that the distance of the point x, y, z from its mean position is not greater than 10. (See Chapter III, Section 8.4.)

Problem 33. Prove that if the variance components s_{11}, s_{22}, ..., s_{12}, ... are given, the coefficients a_{ik} of the corresponding Q are $a_{ik} = S_{ik}/S$ where S is the determinant

$$S = \begin{vmatrix} s_{11} & s_{12} & s_{13} \\ s_{21} & s_{22} & s_{23} \\ s_{31} & s_{32} & s_{33} \end{vmatrix}$$

and S_{ik} the cofactor of s_{ik} in this determinant; show that $S = 1/D$ (theorem on reciprocal matrices).

Problem 34. Using the formula $V = \frac{4}{3}abc\pi$ for the volume of an ellipsoid with axes a, b, c find the volume of the quadric $2Q = K^2$.

9. Properties of the Multivariate Normal Distribution

9.1. *Linear combination of normal variates.* In this section it will be useful to introduce and use some matrix notation. Denote now by A the matrix of a_{ik}, $i = 1, 2, ..., n$, $k = 1, 2, ..., n$, by $|A|$ its determinant, by $X = \begin{pmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ x_n \end{pmatrix}$ the column vector of the x_i and by $X' = (x_1, x_2, ..., x_n)$ their row vector. The density function of a n-variate normal distribution can then be written as follows:

$$p(x_1, ..., x_n) = \frac{\sqrt{|A|}}{(2\pi)^{n/2}} e^{-\frac{1}{2}(X-M)'A(X-M)}, \tag{119}$$

where

$$M = E[X] = \begin{pmatrix} m_1 \\ \vdots \\ m_n \end{pmatrix}, \quad m_i = E[x_i] = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} x_i p(x_1, ..., x_n) \prod_{j=1}^{n} dx_j$$

and

$$(X - M)'A(X - M) = \sum_{i,j,=1}^{n} a_{ij}(x_i - m_i)(x_j - m_j).$$

The components of the mean of X are (just as in 2 or 3 dimensions) the means of the components of X; and the variance matrix, denoted by S, has the components $s_{ij} = E[(x_i - m_i)(x_j - m_j)]$. We will denote the variance matrix by $S = E[(X - M)(X - M)']$. A diagonal element of this matrix, $s_{ii} = E[(x_i - m_i)^2]$, is the variance of x_i; a non-diagonal element, $s_{ij} = E[(x_i - m_i)(x_j - m_j)]$, is the covariance of x_i and x_j.

It can be shown by a direct generalization of the result for $n = 3$ that *if the density of* X *is* (119), *the mean value of the distribution is* M (as we had anticipated by our notation) *and the variance matrix* $S = A^{-1}$. Hence,

$$p(X) = (2\pi)^{-n/2} \mid S \mid^{-1/2} e^{-\frac{1}{2}(X-M)'S^{-1}(X-M)} . \tag{120}$$

Conversely, *given an n-dimensional vector* M *and a positive definite symmetric* $n \times n$ *matrix* S, *there is an n-variate normal density whose mean value and variance matrix are* M *and* S, *respectively*. The density (120) will be briefly denoted by $N(X \mid M, S)$.

We are going to show: *If we subject the* $x_1, x_2, \ldots, x_n$ *to a linear non-singular transformation* C

$$y_i = \sum_{j=1}^{n} c_{ij} x_j , \qquad i = 1, 2, \ldots, n \tag{121}$$

then the $y_1, \ldots, y_n$ *are again normally distributed.* More precisely, *if* X (with n components) *is distributed as* $N(X \mid M, S)$, *then* $Y = CX$ *is distributed as* $N(Y \mid CM, CSC')$ where C' denotes the transposed matrix of C.

In fact, the density of Y is obtained from that of X, by replacing X by

$$X = C^{-1}Y$$

and multiplying by the Jacobian of the transformation (121) which is $\mid C^{-1} \mid$. This is

$$\mid C^{-1} \mid = \frac{1}{\mid C \mid} = \sqrt{\frac{1}{\mid C \mid^2}} = \sqrt{\frac{\mid S \mid}{\mid C \mid \mid S \mid \mid C' \mid}} = \frac{\mid S \mid^{1/2}}{\mid CSC' \mid^{1/2}} .$$

The quadratic form $Q = (X - M)'S^{-1}(X - M)$ in the exponent of $N(X \mid M, S)$, will be carried into the following expressions by the transformation (121):

$$\begin{aligned}
Q &= (C^{-1}Y - M)'S^{-1}(C^{-1}Y - M) \\
&= (C^{-1}Y - C^{-1}CM)'S^{-1}(C^{-1}Y - C^{-1}CM) \\
&= [C^{-1}(Y - CM)]'S^{-1}[C^{-1}(Y - CM)] \\
&= (Y - CM)'(C^{-1})'S^{-1}C^{-1}(Y - CM) = (Y - CM)'(CSC')^{-1}(Y - CM),
\end{aligned}$$

where $(C^{-1})' = (C')^{-1}$, has been used.

Hence, the density of Y is

$$\begin{aligned}
&N(C^{-1}Y \mid M, S) \mid C^{-1} \mid \\
&\quad = (2\pi)^{-n/2} \mid CSC' \mid^{-1/2} \exp\left[-\tfrac{1}{2}(Y - CM)'(CSC')^{-1}(Y - CM)\right] \\
&\quad = N(Y \mid CM, CSC').
\end{aligned}$$

Thus, our statement is proved.

9.2. *Remarks on independence.* Let us divide X into two vectors $X^{(1)}$ and $X^{(2)}$ with r and $n - r$ components, respectively:

$$X = \begin{pmatrix} X^{(1)} \\ X^{(2)} \end{pmatrix}, \qquad X^{(1)} = \begin{pmatrix} x_1 \\ \vdots \\ x_r \end{pmatrix}, \qquad X^{(2)} = \begin{pmatrix} x_{r+1} \\ \vdots \\ x_n \end{pmatrix} \tag{122}$$

and similarly for M and S

$$M = E[X] = \begin{pmatrix} M^{(1)} \\ M^{(2)} \end{pmatrix}, \qquad S = E[(X - M)(X - M)'] = \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix}, \tag{122'}$$

where

$$M^{(1)} = \begin{pmatrix} m_1 \\ \vdots \\ m_r \end{pmatrix}, \qquad M^{(2)} = \begin{pmatrix} m_{r+1} \\ \vdots \\ m_n \end{pmatrix}, \qquad \text{with} \qquad m_i = E[x_i],$$

$$S_{11} = (s_{ij}), \quad 1 \leqslant \begin{smallmatrix} i \\ j \end{smallmatrix} \leqslant r, \qquad S_{12} = (s_{ij}), \quad 1 \leqslant i \leqslant r, \quad r + 1 \leqslant j \leqslant n.$$

$$S_{22} = (s_{ij}), \quad r + 1 \leqslant \begin{smallmatrix} i \\ j \end{smallmatrix} \leqslant n.$$

If $S_{12} = 0$, $S_{21} = 0$, hence

$$S = \begin{pmatrix} S_{11} & 0 \\ 0 & S_{22} \end{pmatrix}, \tag{123}$$

then the determinant of S is $\mid S \mid = \mid S_{11} \mid \cdot \mid S_{22} \mid$ and the inverse of S is

$$S^{-1} = \begin{pmatrix} S_{11}^{-1} & 0 \\ 0 & S_{22}^{-1} \end{pmatrix}.$$

Hence the quadratic form in the exponent of $N(X \mid M, S)$ is

$$\begin{aligned} Q &= (X - M)'S^{-1}(X - M) \\ &= [(X^{(1)} - M^{(1)}), (X^{(2)} - M^{(2)})] \begin{pmatrix} S_{11}^{-1} & 0 \\ 0 & S_{22}^{-1} \end{pmatrix} \begin{bmatrix} X^{(1)} - M^{(1)} \\ X^{(2)} - M^{(2)} \end{bmatrix} \\ &= (X^{(1)} - M^{(1)})'S_{11}^{-1}(X^{(1)} - M^{(1)}) + (X^{(2)} - M^{(2)})'S_{22}^{-1}(X^{(2)} - M^{(2)}) \\ &= Q_1 + Q_2 , \end{aligned}$$

where

$$Q_1 = (X^{(1)} - M^{(1)})'S_{11}^{-1}(X^{(1)} - M^{(1)}), \qquad Q_2 = (X^{(2)} - M^{(2)})'S_{22}^{-1}(X^{(2)} - M^{(2)}) .$$

The density of X can, therefore, be written

$$\begin{aligned} N(X \mid M, S) &= (2\pi)^{-\frac{n}{2}} \mid S \mid^{-\frac{1}{2}} e^{-\frac{Q}{2}} \\ &= (2\pi)^{-\frac{r}{2}} \mid S_{11} \mid^{-\frac{1}{2}} e^{-\frac{Q_1}{2}} (2\pi)^{-\frac{n-r}{2}} \mid S_{22} \mid^{-\frac{1}{2}} e^{-\frac{Q_2}{2}} \qquad (124) \\ &= N(X^{(1)} \mid M^{(1)}, S_{11}) \cdot N(X^{(2)} \mid M^{(2)}, S_{22}) . \end{aligned}$$

We have proved: *If* $S_{12} = 0$, $S_{21} = 0$, *hence*, $S = \begin{pmatrix} S_{11} & 0 \\ 0 & S_{22} \end{pmatrix}$ *then* $X^{(1)}$ *and* $X^{(2)}$ *are independently distributed* as in (124).

Conversely, *if* $X^{(1)}$ *and* $X^{(2)}$ *are independent*, as in (124), *the variance matrix* S *for* $X = \binom{X^{(1)}}{X^{(2)}}$ *is given by* (123).[1]

9.3. Marginal distributions and conditional distributions. It can be shown that *if* X *is distributed as* $N(X \mid M, S)$, *and* B *is* an r by n matrix ($r \leqslant n$) *of rank* r, *then the* r*-dimensional vector* $Z = BX$ *is distributed as* $N(Z \mid BM, BSB')$ (for $r = n$ this result has been shown in Section 9.1). From this result, which we are not going to prove, it follows immediately that *any marginal distribution of a multivariate normal distribution is again normal*:

Let $X^{(1)} = \begin{pmatrix} x_1 \\ \cdot \\ \cdot \\ x_r \end{pmatrix}$ as in (122). If we let $B = (I, 0)$ where I is the r times r identity matrix and 0 is the r by $(n - r)$ zero matrix, then

$$Z = BX = X^{(1)} .$$

[1] In two dimensions, if $X = \binom{x_1}{x_2}$ this theorem would read as follows: x_1 and x_2 are independent if and only if they are uncorrelated. If the joint distribution is not assumed normal, it is possible for two variates to be uncorrelated and yet not independent. (See Chapter XI for further details.)

Hence, $X^{(1)}$ is normally distributed with *the mean value* $BM = M^{(1)}$ *and the variance matrix* $BSB' = S_{11}$. Because the numbering of the components of X is arbitrary, we can state that the marginal distribution of any set of components of X is multivariately normal with mean and variance matrix obtained by taking the proper components of M and S, respectively.

With the same notation as in (122) and (122′), the following result will be proved.

If the distribution of X *is normal, the conditional distribution of* $X^{(1)}$, *given* $X^{(2)}$, *is normal (in r-dimensions) with mean* $M^{(1)}+S_{12}S_{22}^{-1}[X^{(2)} - M^{(2)}]$ *and variance matrix* $S_{11} - S_{12}S_{22}^{-1}S_{21}$.

Let us define $Y^{(1)} = X^{(1)} - S_{12}S_{22}^{-1}X^{(2)}$ and $Y^{(2)} = X^{(2)}$. Then $\begin{pmatrix}Y^{(1)}\\Y^{(2)}\end{pmatrix} = Y = \begin{pmatrix}1 & -S_{12}S_{22}^{-1}\\0 & I\end{pmatrix} \cdot X$. The vector Y is a non-singular transform of X, and, therefore, has a normal distribution (as proved in 9.1) with mean

$$E\begin{bmatrix}Y^{(1)}\\Y^{(2)}\end{bmatrix} = \begin{bmatrix}1 & -S_{12}S_{22}\\0 & 1\end{bmatrix} E[X]$$

$$= \begin{bmatrix}1 & -S_{12}S_{22}\\0 & 1\end{bmatrix}\begin{bmatrix}M^{(1)}\\M^{(2)}\end{bmatrix} = \begin{bmatrix}M^{(1)} - S_{12}S_{22}^{-1}M^{(2)}\\M^{(2)}\end{bmatrix}.$$

Therefore

$$E[Y^{(1)}] = [M^{(1)} - S_{12}S_{22}^{-1}M^{(2)}] \equiv N_1, \qquad E[Y^{(2)}] = M^{(2)} \equiv N_2,$$

where $E[Y] = N$ has been introduced. The variance matrix

$$E[(Y - N)(Y - N)']$$
$$= \begin{pmatrix}E[(Y^{(1)} - N^{(1)})(Y^{(1)} - N^{(1)})'] & E[(Y^{(1)} - N^{(1)})(Y^{(2)} - N^{(2)})']\\E[(Y^{(2)} - N^{(2)})(Y^{(1)} - N^{(1)})'] & E[(Y^{(2)} - N^{(2)})(Y^{(2)} - N^{(2)})']\end{pmatrix}$$

will be proved to be simply:

$$\begin{pmatrix}S_{11} - S_{12}S_{22}^{-1}S_{21} & 0\\0 & S_{22}\end{pmatrix}.$$

In fact,

$$E[(Y^{(1)} - N^{(1)})(Y^{(1)} - N^{(1)})']$$
$$= E[(X^{(1)} - M^{(1)}) - S_{12}S_{22}^{-1}(X^{(2)} - M^{(2)})][(X^{(1)} - M^{(1)})' - S_{12}S_{22}^{-1}(X^{(2)} - M^{(2)})']$$
$$= S_{11} - S_{12}S_{22}^{-1}S_{21} - S_{12}S_{22}^{-1}S_{21} + S_{12}S_{22}^{-1}S_{22}S_{22}^{-1}S_{21}$$
$$= S_{11} - S_{12}S_{22}^{-1}S_{21}.$$

The computation of the three other terms is similar. Therefore, $Y^{(1)}$ and $Y^{(2)}$ are independent, and $X^{(2)} = Y^{(2)}$ has the marginal distribution $N(X^{(2)} \mid M^{(2)}, S_{22})$. The joint density of $Y^{(1)}$ and $Y^{(2)} = X^{(2)}$ is

$$N(Y^{(1)} \mid M^{(1)} - S_{12}S_{22}^{-1}M^{(2)}, S_{11} - S_{12}S_{22}^{-1}S_{21}) \cdot N(Y^{(2)} \mid M^{(2)}, S_{22}). \qquad (125)$$

Now, to find the conditional distribution of $X^{(1)}$ given $X^{(2)}$, the joint density of $X^{(1)}$ and $X^{(2)}$ is obtained from (125) by substituting $X^{(1)} - S_{12}S_{22}^{-1}X^{(2)}$ for $Y^{(1)}$ and $X^{(2)}$ for $Y^{(2)}$ and multiplying by the Jacobian of the transformation (which is 1). The joint density of $X^{(1)}$ and $X^{(2)}$ is, therefore,

$$\begin{aligned} &p(X^{(1)}, X^{(2)}) \\ &= (2\pi)^{-\frac{r}{2}} \mid S_{11\cdot 2}\mid^{-\frac{1}{2}} \cdot \exp\{-\tfrac{1}{2}[(X^{(1)} - M^{(1)}) - S_{12}S_{22}^{-1}(X^{(2)} - M^{(2)})]' \\ &\quad \cdot S_{11\cdot 2}^{-1}[(X^{(1)} - M^{(1)}) - S_{12}S_{22}^{-1}(X^{(2)} - M^{(2)})]\} \, . \\ &\quad \cdot (2\pi)^{-\frac{n-r}{2}} \mid S_{22}\mid^{-\frac{1}{2}} \exp\{-\tfrac{1}{2}(X^{(2)} - M^{(2)})' S_{22}^{-1}(X^{(2)} - M^{(2)})\}, \end{aligned} \qquad (126)$$

where $S_{11\cdot 2} = S_{11} - S_{12}S_{22}^{-1}S_{21}$. Note that $p(X^{(1)}, X^{(2)})$ must be identical with $N(X \mid M, S)$, the density of X.

The conditional density of $X^{(1)}$ given $X^{(2)}$ is the quotient of (126) and the marginal density of $X^{(2)}$, which is the second factor in (126). Hence, the conditional density is

$$\begin{aligned} p(X^{(1)} \mid X^{(2)}) &= (2\pi)^{-r/2} \mid S_{11\cdot 2}\mid^{-1/2} \\ &\times \exp\{-\tfrac{1}{2}[(X^{(1)} - M^{(1)}) - S_{12}S_{22}^{-1}(X^{(2)} - M^{(2)})]' \\ &\quad \cdot S_{11.2}^{-1}[(X^{(1)} - M^{(1)}) - S_{12}S_{22}^{-1}(X^{(2)} - M^{(2)})]\} \, . \end{aligned}$$

The density $p(X^{(1)} \mid X^{(2)})$ is, therefore, an r-variate normal density with mean $M^{(1)} + S_{12}S_{22}^{-1}(X^{(2)} - M^{(2)})$ and the covariance matrix $S_{11\cdot 2} = S_{11} - S_{12}S_{22}^{-1}S_{21}$.

As an example of this result, let us consider the bivariate normal distribution and find the conditional distribution of x given y. In this case

$$M^{(1)} = a, \quad M^{(2)} = b, \quad S_{11} = s_{11}, \quad S_{12} = r\sqrt{s_{11}}\sqrt{s_{22}},$$
$$S_{22} = s_{22}, \quad S_{11\cdot 2} = s_{11}(1 - r^2).$$

Hence, the conditional density of x, given y, is a univariate normal distribution with mean $a + r\sqrt{s_{11}/s_{22}}(y - b)$ and variance $s_{11}(1 - r^2)$.

Problem 35. Show by direct computation that, given he bivariate normal distribution [(113′), p. 422], the marginal density of x (or y) is the above univariate normal density.

Problem 36. Show by direct computation that the conditional density of x given y [i.e., $p(x \mid y) = p(x, y)/p_2(y)$ where $p(x, y)$ is a bivariate normal distribution and $p_2(y)$ its marginal distribution with respect to y] is again a normal density in x, with mean and variance as in the general statement p. 428.