Proceedings of the Iowa Academy of Science

Volume 41 | Annual Issue Article 73

1934

The Effect of Increasing the Size of Samples in Certain Distributions

Fred A. Brandner
Iowa State College

THE EFFECT OF INCREASING THE SIZE OF SAMPLES IN CERTAIN DISTRIBUTIONS

FRED A. BRANDNER

The problem of comparing two or more distributions has been under discussion for approximately 150 years. The classical method of comparing moments or functions of moments is without doubt the one which holds the favor of most of the eminent statisticians. It is not the purpose of this paper to enumerate the merits or demerits of any test of significance of differences of samples. Rather, a certain method of procedure is chosen, and in accordance with the hypotheses of this method the effect of increasing the size of the sample is shown. Comparative relations are easily shown with a few common statistics such as the standard deviations of standard deviations and means. For example if samples, Σ_1 with n observations, and Σ_2 with nk observations both have σ as the S. D., then the simple relation,

$$\sigma_{m_2} = \frac{\sigma_{m_1}}{\sqrt{k}}$$

exists, where m_1 and m_2 are the means of the samples. However, in very few if any cases, in statements of probabilities, has a simple functional relation been shown. For instance it is rather difficult to state the exact relation

$$p_2 = f(p_1),$$

where

$$p_1 = \sqrt{\frac{n}{2\pi}} \cdot \frac{1}{\sigma} \int_m^\infty e^{-\frac{n m^2}{2\sigma^2}} dm,$$

and

$$p_2 = \sqrt{\frac{kn}{2\pi}} \cdot \frac{1}{\sigma} \int_m^\infty e^{-\frac{knm^2}{2\sigma^2}} dm.$$

Consider with E. C. Rhodes[1] the frequency problem of two samples in alternate categories. These samples are drawn from two populations which need not even have the same characteristics.

1 Biometrika XVI, pp. 239-248. Are two given samples from the same population?

SAMPLE	CLASS A	CLASS NOT-A	TOTAL FREQ.	MEAN	S. D.
Σ_x	r	m-r	m	$\bar{x} = mp$	$\sigma_x^2 = mpq$
Σ_y	s	n-s	n	$\bar{y} = np$	$\sigma_y^2 = npq$

We will test the samples for a probability of no more likely existence.

With the hypothesis that m and n are large, and p and $q = 1 - p$ are not small, the binomial distribution of x and y can be replaced by the normal distribution. Then the simultaneous chance of the class A frequency being between x and $x + dx$ in Σ_x, and between y and $y + dy$ in Σ_y is given by

$$(1) \qquad dp_1 = \frac{1}{2\pi\sigma_x\sigma_y} e^{-\frac{1}{2}\left[\frac{(x-\bar{x})^2}{\sigma_x^2} + \frac{(y-\bar{y})^2}{\sigma_y^2}\right]} dxdy.$$

The likelihood of occurrence will be constant for equal values of the exponent. For the given samples a contour ellipse

$$(2) \qquad \frac{(x-x)^2}{\sigma_x^2} + \frac{(y-\bar{y})^2}{\sigma_y^2} = \frac{(r-mp)^2}{mpq} + \frac{(s-np)^2}{npq} = X_1^2$$

is determined, such that for values (x, y) outside, the occurence is less likely. Thus by finding the volume under the frequency surface, outside this contour, the probability of the occurence of no more likely doublets is given. In order to do this, change to polar coordinates and integrate. Thus

$$(3) \qquad p_1 = \frac{1}{2\pi}\int_{X_1}^{\infty}\int_0^{2\pi} e^{-\frac{X^2}{2}} XdX = e^{-\frac{X_1^2}{2}}.$$

Now consider two samples, Σ'_1 and Σ'_2, independent of the two given above. Suppose each of the frequencies in these samples are those of the table multiplied by a constant k. Then by following an exactly analogous process, the value of the probability of no more likely samples can be expressed as

$$(4) \qquad p_k = \frac{1}{2\pi}\int_{X_1^1}^{\infty}\int_0^{2\pi} e^{-\frac{X^2}{2}} XdX = e^{-\frac{X_1^{1^2}}{2}}$$

where

$$(5) \qquad X_1^{1^2} = \frac{(kr-kmp)^2}{kmpq} + \frac{(ks-knp)^2}{knpq} = kX_1^2.$$

Thus

$$p_k = e^{-\frac{kX_1^2}{2}} = (p_1)^k. \tag{6}$$

The following example clearly shows the effect of the increase. Suppose that an observer has two samples of 40, each having frequencies such that $p_1 = .1$. He has little reason to suppose that there is any significant difference in the two samples, since the probability of equally or less likely samples is only one in ten. A second observer independent of the first having the same ratios in samples of 80 each would procure a value $P_2 = (.1)^2 = .01$. With very little doubt he would infer that the samples were significantly different.

Now suppose B samples are given with A classes in each sample. The frequencies may be represented in the form:

CLASSES

Samples	C_1	C_2		C_A	Totals
Σ_1	X_{11}	X_{12}		X_{1A}	N_1
Σ_2	X_{21}	X_{22}		X_{2A}	N_2
Σ_B	X_{B1}	X_{B2}		X_{BA}	N_B

To any class C_i there is a probability p_i. Thus the mean and variance in the i-th class of j-th sample are $x_{ij} = N_j p_i$ and $\sigma^2_{ij} = N_j p_i q_i$, where

$$p_i + q_i = 1,$$

and

$$\sum_{i=1}^{A} p_i = 1.$$

As previously the chance of a simultaneous occurrence may be set up in the form

$$dp_1 = \frac{1}{(2\pi)^{\frac{B(A-1)}{2}} \prod_{j=1}^{B} \prod_{i=1}^{A-1} \sigma_{ij}} e^{-\frac{1}{2} \sum_{j=1}^{B} \sum_{h=1}^{A-1} \sum_{i=1}^{A-1} \frac{R_{hi}}{R} \cdot \frac{(x_{ij} - \bar{x}_{ij})(x_{hj} - \bar{x}_{hj})}{\sigma_{ij}\ \sigma_{hj}}} \prod_{j=1}^{B} \prod_{i=1}^{A-1} dx_{ij} \tag{7}$$

where $R_{hi} = R_{ih}$ is the minor of r_{hi} in the determinant

$$R = \begin{vmatrix} r_{11} & r_{12} & r_{1\overline{A-1}} \\ r_{21} & r_{22} & r_{2\overline{A-1}} \\ & & \\ r_{\overline{A-1}1} & r_{\overline{A-1}2} & r_{\overline{A-1}\ \overline{A-1}} \end{vmatrix} \tag{8}$$

The value dP_1 may be considered as an element of volume in $B(A-1)+1$ dimensions. If this is done, in order to evaluate the $B(A-1)$-fold integral derived from (7) it is only necessary to apply the method of n-dimensional geometry. By setting

$$(9) \qquad \sum_{j=1}^{B} \sum_{h=1}^{A-1} \sum_{i=1}^{A-1} \frac{R_{hi}}{R} \frac{(x_{ij}-\bar{x}_{ij})}{\sigma_{ij}} \frac{(x_{hj}-\bar{x}_{hj})}{\sigma_{hj}} = X^2$$

it is easily seen that the $B(A-1)$ dimensioned differential element if integrated out must reduce to the form $c_1 X^{B(A-1)-1} dX$. Thus

$$(10) \qquad p_1 = C \int_{X_1}^{\infty} e^{-\frac{X^2}{2}} X^{B(A-1)-1} dX$$

where by letting $X_1 = 0$ it is easy to show that $C = 1$.

If each cell of the A by B table is multiplied by a constant k, as in the previous argument, the value

$$(11) \qquad p_k = \int_{\sqrt{k}X_1}^{\infty} e^{-\frac{X^2}{2}} X^{B(A-1)-1} dX$$

is easily obtained.

Whenever the value $B(A-1)-1 = d$ is odd, that is for B even or for A odd, an expression of the form

$$(12) \qquad p_1 = p_{\frac{d-1}{2}}(X_1^2)\, e^{-\frac{X_1^2}{2}}$$

may be obtained by integrating by parts. Likewise

$$(13) \qquad p_k = p_{\frac{d-1}{2}}(kX_1^2)\, e^{-\frac{kX_1^2}{2}}$$

where $p_{\frac{d-1}{2}}(z)$ is a polynomial in z of degree $\frac{d-1}{2}$. Thus again for the more complex case a rather simple relation

$$(14) \qquad p_k = \frac{p_{\frac{d-1}{2}}(kX_1^2)}{\left[p_{\frac{d-1}{2}}(X^2)\right]^k} (p_1)^k$$

is established.

DEPARTMENT OF MATHEMATICS,
IOWA STATE COLLEGE,
AMES, IOWA.