Student's z, t, and s: What if Gosset had R?

James A. Hanley[1] Marilyse Julien[2] Erica E. M. Moodie[1]

[1]Department of Epidemiology, Biostatistics, and Occupational Health,
McGill University, 1020 Pine Ave W., Montreal, Quebec, H3A 1A2, CANADA

[2]Department of Mathematics and Statistics,
McGill University, 805 Sherbrooke St. W., Montreal, Quebec, H3A 2K6, CANADA

Fisher's derivation of distribution of s,
referred to in our February 2008 American Statistician article

The following excerpt is from the article *Student and Small-Sample Theory*, by E. L. Lehmann, in *Statistical Science*, 1999, Vol. 14, No. 4, 418-426.

> In 1912 R. A. Fisher, then 22 years old and a Cambridge undergraduate, was put into contact with Gosset through Fisher's teacher, the astronomer F. J. M. Stratton. As a result, Gosset received from Fisher a proof of the z-distribution and asked Karl Pearson to look at it, admitting that he could not follow the argument (which was based on n-dimensional geometry) and suggesting, "It seemed to me that if its alright perhaps you might like to put the proof in a note [in Biometrika of which K. P. was the Editor]. It's so nice and mathematical that it might appeal to some people. In any case I should be glad of your opinion of it. . . ."
>
> Pearson was not impressed. "I do not follow Mr. Fisher's proof and it is not the kind of proof which appeals to me," he replied (Pearson 1990, page 47). As a result, the proof was only published in 1915 together with the corresponding proof for the distribution of the correlation coefficient that Student had conjectured in his second 1908 paper. In the correlation case, the n pairs of observations are considered as the coordinates of a point in $2n$-dimensional space, in which the two sample means, two sample variances, and the sample covariance have, as Fisher writes, "a beautiful interpretation," [Fisher, 1915] from which the desired density can be obtained.

The following is the first page of Fisher's 1915 article in *Biometrika*.

FREQUENCY DISTRIBUTION OF THE VALUES OF THE CORRELATION COEFFICIENT IN SAMPLES FROM AN INDEFINITELY LARGE POPULATION.

By R. A. FISHER.

1. My attention was drawn to the problem of the frequency distribution of the correlation coefficient by an article published by Mr H. E. Soper* in 1913. Seeing that the problem might be attacked by means of geometrical ideas, which I had previously found helpful in the consideration of samples, I have examined the two articles by "Student†," upon which Mr Soper's more elaborate work was based, with a view to checking and verifying the conclusions there attained.

"Student," if I do not mistake his intention, desiring primarily to obtain a just estimate of the accuracy to be ascribed to the mean of a small sample, found it necessary to allow for the fact that the mean square error of such a sample is not generally equal to the standard deviation of the normal population from which it is drawn. He was led, in fact, to study the frequency distribution of the mean square error. He calculated algebraically the first four moments of this frequency curve, both about the zero point, and about its mean, observed a simple law to connect the successive moments, and discovered a frequency curve, which fitted his moments, and gave the required law.

Thus if $x_1, x_2, \ldots x_n$ are the members of a sample,

$$n\bar{x} = x_1 + x_2 + \ldots + x_n,$$

and

$$n\mu^2 = (x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2,$$

the frequency with which the mean square error lies in the range $d\mu$ is proportional to

$$\mu^{n-2} e^{-\frac{n\mu^2}{2\sigma^2}} d\mu.$$

This result, although arrived at by empirical methods, was established almost beyond reasonable doubt in the first of "Student's" papers. It is, however, of interest to notice that the form establishes itself instantly, when the distribution of the sample is viewed geometrically.

* *Biometrika*, Vol. IX. p. 91. † *Ibid.* Vol. VI. pp. 1 and 302.

Biometrika, Vol. 10, No. 4. (May, 1915), p. 507.

The following is from the *third* page of Fisher's 1915 article in *Biometrika*. In section 2, he had defined the sample means, variances and covariance, and promised to show us that these quantities "have, in fact, an exceedingly beautiful interpretation in generalised space, which we may now examine."

3. Considering first the space of n dimensions in which the variations of x are represented, the mean and mean square error of n observations are determined by the relations of P, the point representing the n observations, to the line

$$x_1 = x_2 = x_3 = \ldots = x_n,$$

for the perpendicular PM drawn from P upon this line will lie in the region

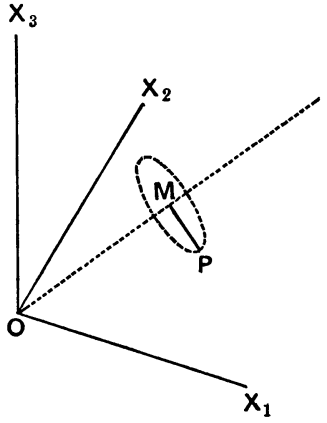
$$x_1 + x_2 + \ldots + x_n = n\bar{x},$$

and will meet it at the point M, where

$$x_1 = \bar{x}, \quad x_2 = \bar{x}, \quad \ldots x_n = \bar{x};$$

further, since, $$PM^2 = (x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2,$$

the length of PM is $\mu_1 \sqrt{n}$.



An element of volume in this n dimensional space may now without difficulty be specified in terms of $\bar{x}$ and μ_1; for, given $\bar{x}$ and μ_1, P must lie on a sphere in $n-1$ dimensions, lying at right angles to the line OM, and the element of volume is

$$C\mu_1^{n-2} d\mu_1 d\bar{x},$$

where C is some constant, which need not be determined.

65—2

from *Biometrika*, Vol. 10, No. 4. (May, 1915), bottom portion of page 509.

Working through Fisher's "*instant*" geometric insight...

Joint density, $g(s, \bar{x})$, of s & $\bar{x}$

Following the usual rules for the probability density of functions of random variables,

$$g(s, \bar{x}) ds d\bar{x} = \int f(x_1, \ldots, x_n)\ dx_1 \ldots dx_n = \int f(\underline{x})\ d\underline{x},$$

where $\underline{x}$ is one of the inverses of $(s, \bar{x})$, and the integration is over all $\underline{x'} = \{x'_1, \ldots, x'_n\}$ within a distance ds of $s = \{(1/n)\sum (x_i - \bar{x})^2\}^{1/2}$ and a distance $d\bar{x}$ of $\bar{x} = (1/n)\sum x_i$.
Now, if, w.l.o.g., $E[\underline{x}] = \underline{0}$, then

$$f(\underline{x}) = \prod f(x_i) \ \propto\ e^{-\sum x_i^2/2\sigma^2}.$$

As did Student, Fisher defined s^2 using a divisor of n. Thus, $\sum x_i^2 = ns^2 + n\bar{x}^2$, so that $f(\underline{x})$ factors into

$$f(\underline{x}) \ \propto\ e^{-ns^2/2\sigma^2} \times e^{-n\bar{x}^2/2\sigma^2}.$$

Therefore

$$g(s, \bar{x}) ds d\bar{x} \ \propto\ \int e^{-ns^2/2\sigma^2} \times e^{-n\bar{x}^2/2\sigma^2} dx_1 \ldots dx_n,$$

with the integral taken over the region described above.
The integrand is constant over this region. Thus, as can be inferred from the $n = 2$ and $n = 3$ cases shown in the Figures overleaf, $\int dx_1 \ldots dx_n$, is $\propto s^{n-2} ds d\bar{x}$. Therefore

$$g(s, \bar{x}) \ \propto\ s^{n-2} \times e^{-ns^2/2\sigma^2} \times e^{-n\bar{x}^2/2\sigma^2}\ ;$$

i.e., $g(s, \bar{x})$ factors into $g_s() \ \propto\ s^{n-2} \times e^{-(n/2)s^2/\sigma^2}$ and $g_{\bar{x}}() \ \propto\ e^{-\bar{x}^2/2(\sigma/\sqrt{n})^2}$.

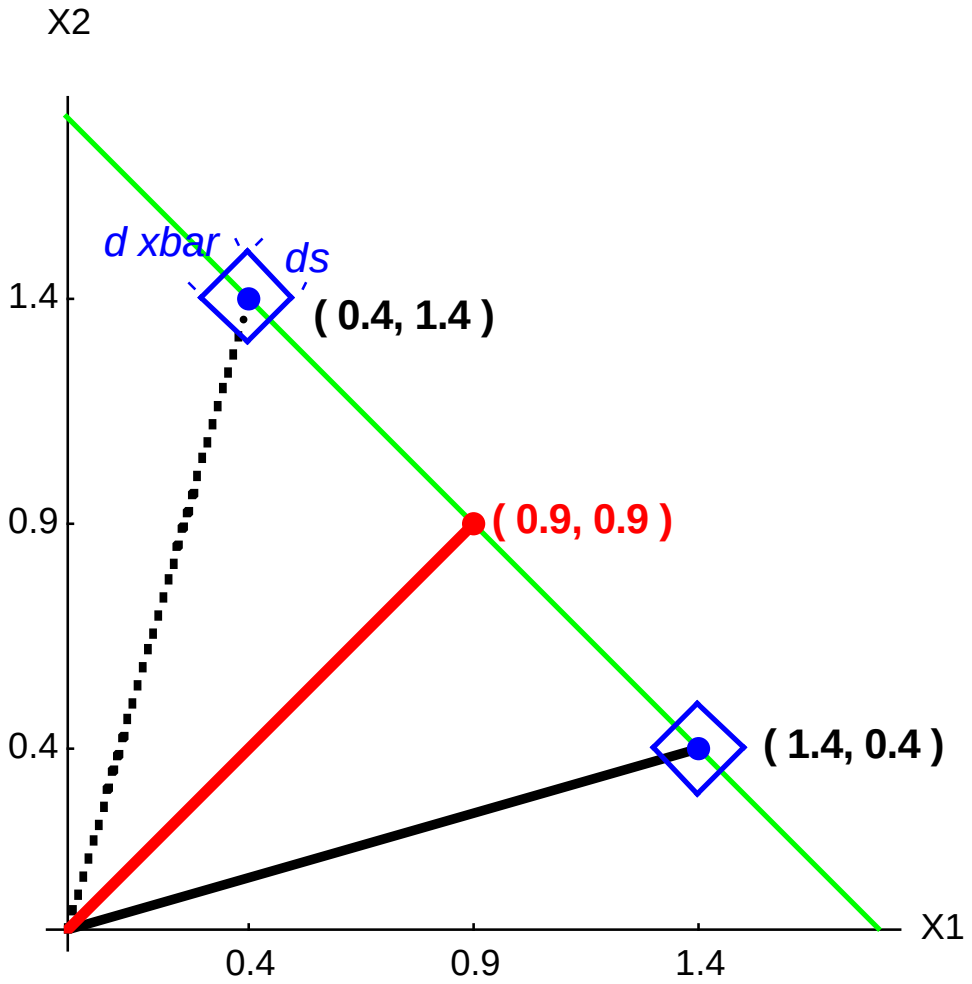


Figure 1: ***Expanded version of Fisher's diagram***: the $\boldsymbol{n = 2}$ case, with the concrete example $\underline{x} = \{1.4, 0.4\}$, so that $\bar{x} = 0.9,\ s = 0.5$. The other $\{X_1, X_2\}$'s within a distance $\{ds, d\bar{x}\}$ of these summary values lie within the two 2-D rectangles, each with area $ds \times d\bar{x}$. Thus the $dx_1 \times dx_2$ volume in R^2 corresponding to the area $ds \times d\bar{x}$ in $R^+ \times R$ is $2 \times s^{(n-2)} \times ds \times d\bar{x}$.

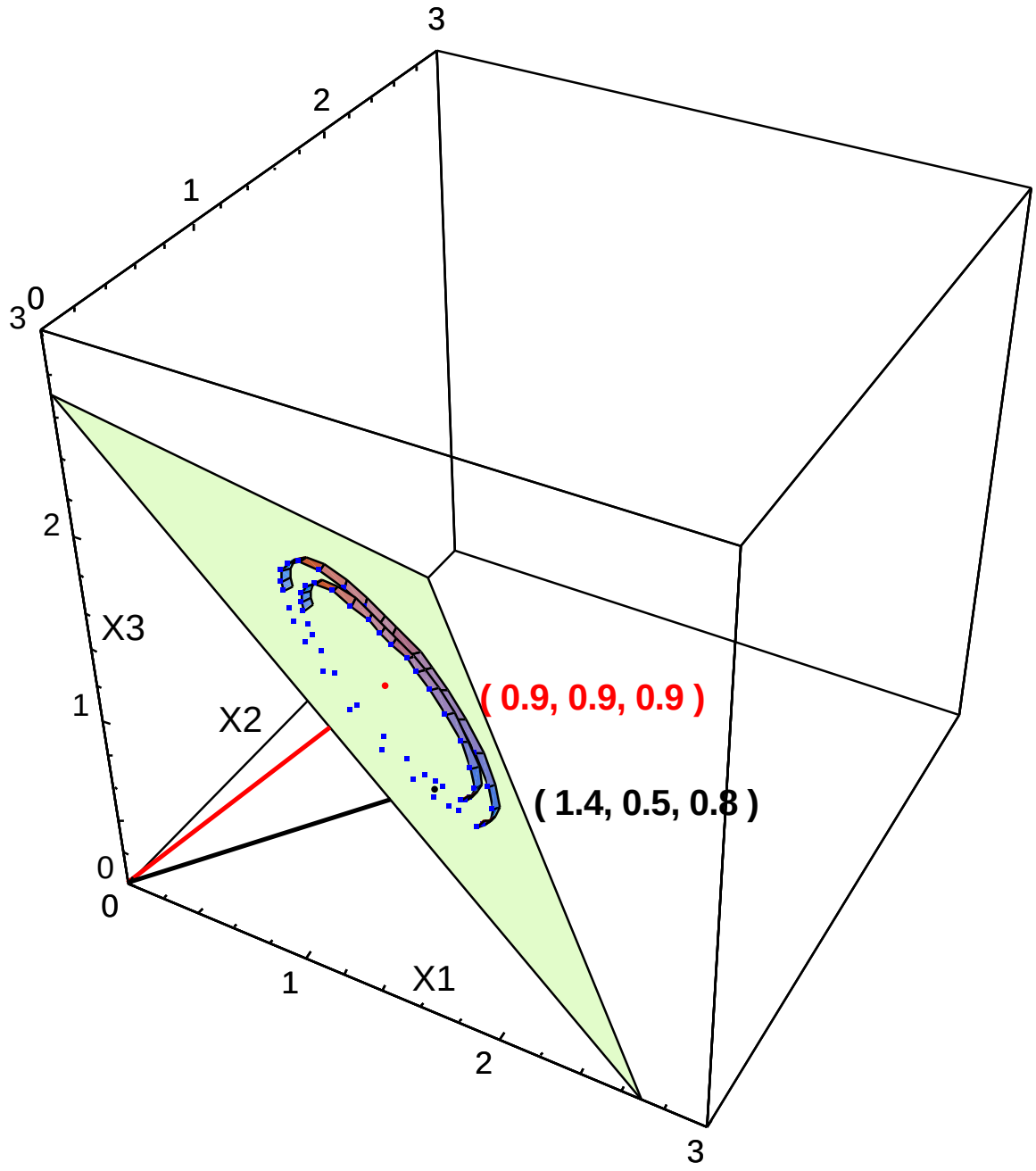


Figure 2: ***Expanded version of Fisher's diagram***: the $\boldsymbol{n = 3}$ case, with concrete example $\underline{x} = \{1.4, 0.5, 0.8\}$, so that $\bar{x} = 0.9$, $s = 0.37$. The other $\underline{X}$'s within a distance $ds, d\bar{x}$ of these summary values are contained in the 3-D region enclosed between the inner and outer surface of two cylinders, with inner and outer radii s and $s + ds$, and between the planes $(1/3)(X_1 + X_2 + X_3) = \bar{x}$ and $(1/3)(X_1 + X_2 + X_3) = \bar{x} + d\bar{x}$. Thus the $dx_1 \times dx_2 \times dx_3$ volume in R^3 corresponding to the area $ds \times d\bar{x}$ in $R^+ \times R$ is $s^{(n-2)} \times ds \times d\bar{x}$. *Not shown, but imaginable*: the $n = 4$ case, where the total probability mass associated with the $\underline{X}$'s that have close to the same summary values, lies 'near' the surface of a sphere with *surface area* $\propto s^2 = s^{n-2}$, and thus, within a volume $4\pi \times s^{n-2} \times ds \times d\bar{x}$. And so on for higher values of n.

Note

There is one confusing item in Fisher's 1915 article: in his section 1, he refers to the "frequency with which the *mean* square error [italics ours] lies in the range $d\mu$" (or, ds, as we would write it today. Given that the density he describes is proportional to μ^{n-2} (or s^{n-2}) he can only have meant the *root* mean square error square error. In his more expansive 1925 *Metron* paper, he again derives the joint distribution of s and $\bar{x}$, (as we do above) with the density of s proportional to $s^{(n-2)}$; from these he derives the joint distribution of s^2 and $\bar{x}$, with the density of s^2 proportional to $(s^2)^{(n-3)/2}$.

Incidentally

Although Gosset did not fully establish that s and $\bar{x}$ are statistically independent, Fisher's derivation of $g(s, \bar{x})$ shows that they are.