Correspondence

Smoking in pregnancy and low weight babies: a statistical consideration

Sir, In an interesting article on smoking and pregnancy published in March (Goldstein, 1977), Professor Harvey Goldstein makes a point concerning the distribution of birthweights that I should like to criticise.

Professor Goldstein tries to account for the higher mortality of low birthweight babies born to mothers who do not smoke by a hypothesis concerning the mean birthweights below an arbitrary level, based upon the nature of the normal distribution. He claims that if the birthweights of babies born to mothers, either smokers or non-smokers, are distributed as normal distributions of equal variance, then, under certain conditions, one would expect, purely on mathematical grounds, that the average weight of low weight babies would be greater if their mothers smoked than if they did not, despite the fact that overall the mean birthweight of babies born to mothers who do not smoke is highest. Naturally if low weight babies born to mothers who smoke are on average heavier than those born to mothers who do not smoke, other things being equal, we would expect the mortality of such a group to be lower.

Professor Goldstein's argument was, however, not mathematically justified and is based on a false premise, as is demonstrated below, since, if we have two normal distributions of equal variance but different means, the average value of the variable in question in any given fixed range will always be greater in the distribution with the higher mean.

Consider two normal distributions of equal variance σ^2 and with means μ and $\mu + \theta$, $\theta > 0$. The situation is illustrated in the Figure. The means of the two distributions over any given range will be linear combinations of the same values but in different proportions. That is to say, the same values will appear in any such calculation, only the weights attached to them representing the relative frequencies of these values will differ.

Let us take any two such values within the given range. Let these two values be x and $x + \Delta$, $\Delta > 0$. In the Figure the probability densities of the distribution with the higher mean are labelled a_1 and a_2 where a_2 is the density corresponding to the higher

Figure Two normal distributions of equal variance.

value $(x + \Delta)$. b_1 and b_2 are the corresponding values for the distribution with the lower mean.

The values of a_1 , a_2 , b_1 , b_2 can be determined from the equation of the normal distribution and it can be

shown that $\frac{a_2}{a_1} = \frac{b_2}{b_1} e^{\frac{\Delta \theta}{\sigma^2}}$

since $\frac{\Delta \theta}{\sigma^2} > 0$ then $e^{\frac{\Delta \theta}{\sigma^2}} > 1$

therefore $\frac{a_2}{a_1} > \frac{b_2}{b_1}$

This means that the ratio with which the value $x + \Delta$ is combined with the value x is greatest for the distribution with the highest mean. The logic can be extended to all the values in the calculation of the means and thus we can say that the higher the overall mean, the greater the mean in a given fixed range will be.

In case the problem of estimating the expected mean in a given range of the normal distribution is one of general interest to epidemiologists, a formula for calculating such values is given below.

For any normal distribution of mean μ and variance σ^2 the mean, M , in a given range Z_1 to Z_2

is given by the formula:

$$M = \mu - \sigma \frac{\phi\left(\frac{z_2 - \mu}{\sigma}\right) - \phi\left(\frac{z_1 - \mu}{\sigma}\right)}{\Phi\left(\frac{z_2 - \mu}{\sigma}\right) - \Phi\left(\frac{z_1 - \mu}{\sigma}\right)}$$

where $\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$ the ordinate of the normal frequency curve

and $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{1}{2}t^2} \cdot dt$ the normal distribution function

Thus the mean in a given range can be calculated with the aid of statistical tables (for example, the *Cambridge Elementary Statistical Tables* by D. V. Lindley and J. C. P. Miller in which the notation used is the same as that above).

If we are interested only in the mean of all values less than a given value, our range is $-\infty$ to z_2 and the formula simplifies to:

$$M = \mu - \sigma \frac{\phi\left(\frac{z_2 - \mu}{\sigma}\right)}{\Phi\left(\frac{z_2 - \mu}{\sigma}\right)}$$

Using this formula we can estimate, for example, that if babies born to mothers who do not smoke have a mean weight of 3350 g with a standard deviation of 500 g, then, if the distribution is normal, the mean weight of babies under 2500 g will be 2296 g. If the weights of babies of mothers who do smoke are normally distributed with the same standard deviation and with an overall mean weight

of 3150 g the average birth weight of all babies under 2500 g will be 2265 g. Thus in practice the difference in mean birthweights of low weight babies is likely to be small (in our example approximately 30 g), and Professor Goldstein is correct in stating that this will be the case, though not in proposing that a reversal of mean weights may in certain conditions occur if birthweights are distributed normally.

S. J. SENN

*Tunbridge Wells Health District,
Sherwood Park,
Pembury Road,
Tunbridge Wells,
Kent*

Editor's note:

This letter was shown to Professor Goldstein who agrees that Mr Senn is correct and wishes to thank him for pointing this out. The substance of Professor Goldstein's comments concerning the 'paradoxical' findings of Yerushalmy (1972) is however not seriously affected.

References

- Goldstein, H. (1977). Smoking in pregnancy: some notes on the statistical controversy. *British Journal of Preventive and Social Medicine*, **31**, 13-17.
- Yerushalmy, J. (1972). Infants with low birthweight born before their mothers started to smoke cigarettes. *American Journal of Obstetrics and Gynecology*, **112**, 227-284.