A note on the calculation of entropy from histograms

Wallis, Kenneth

University of warwick

October 2006

A note on the calculation of entropy from histograms

Kenneth F. Wallis

Department of Economics
University of Warwick
Coventry CV4 7AL, UK
[K.F.Wallis@warwick.ac.uk]

October 2006

Abstract An expression for the entropy of a random variable whose probability density function is reported as a histogram is given. It allows the construction of time series of entropy from responses to density forecast surveys such as the US Survey of Professional Forecasters or the Bank of England Survey of External Forecasters, where the questionnaire provides histogram bins whose width changes from time to time.

The derivation presented in this note was prompted by discussion with Robert Rich and Joseph Tracy about the entropy-based measures of uncertainty and disagreement used in their study of the US Survey of Professional Forecasters (Rich and Tracy, 2006). They cite no source for the expression for entropy which they implement, simply giving a general reference for entropy. I could find no discussion of histograms in the literature, hence this note. A useful recent reference is the "Entropy" entry by Harris in the new edition of the *Encyclopedia of Statistical Sciences*. Harris begins by considering the entropy of a discrete random variable, but to treat the discrete probability distribution as analogous to the histogram of a continuous random variable needs some care.

For a continuous random variable *X*, such as inflation or growth in the SPF data, with probability density function $f(x)$, the definition of its entropy is

$$H(X) = -E\left[\log f(X)\right] = -\int f(x)\log f(x)dx\,.$$

To relate this definition to the situation in which the density is represented as a histogram, we divide the range of the variable into *n* intervals (l_k, u_k), *k*=1,…,*n*, so that

$$H(X) = -\sum_{k=1}^{n}\int_{l_k}^{u_k} f(x)\log f(x)dx\,.$$

We then relate the *k*th term in this summation to the *k*th bin of a histogram, with width

$$w_k = u_k - l_k\,.$$

When the range of the variable is unbounded, so that $l_1 = -\infty$ and $u_n = \infty$, it is customary for computational purposes to assume l_1 and u_n finite, for example by assuming that the first and last intervals have width similar to that of the interior intervals.

The bin probabilities p_k, *k*=1,…,*n*, defined as

$$p_k = \int_{l_k}^{u_k} f(x)dx\,,$$

can be approximated as $w_k f(x_k)$, the area of a rectangle of height $f(x_k)$, where x_k is a representative value within the interval (l_k, u_k). Similarly the *k*th integral in the above summation can be approximated as $w_k f(x_k)\log f(x_k)$. Rewriting this expression in terms of the bin probabilities then gives the entropy as

$$H(X) = -\sum_{k=1}^{n} p_k \log\left(p_k / w_k\right). \tag{1}$$

This corresponds to the expression given by Harris (2006) for a discrete distribution, and given by Rich and Tracy (2006) for a histogram, only if $w_k = 1$. In the typical case in which w_k is constant, but not necessarily equal to 1, we have

$$H(X) = -\sum_{k=1}^{n} p_k \log p_k + \log w. \tag{2}$$

If w_k is not constant, equation (1) calls for bin-by-bin adjustments before comparisons of entropy between histograms with different bin configurations can be made. Correction (2) is required in constructing time series of entropy from responses to density forecast surveys such as the US Survey of Professional Forecasters or the Bank of England Survey of External Forecasters, where the questionnaire provides bins whose width changes from time to time.

Note that entropy is only defined up to a scale factor, since the base of the logarithms has not been specified. For continuous random variables it is convenient to work with natural logarithms, and for a normal distribution we have

$$\begin{aligned} H(X) &= -\int f(x)\left(\ln\frac{1}{\sigma\sqrt{2\pi}} - \frac{(x-\mu)^2}{2\sigma^2}\right)dx \\ &= \ln\left(\sigma\sqrt{2\pi}\right) + 0.5 = \ln\sigma + 1.42\,. \end{aligned}$$

This suggests that, when comparing entropy-based and moment-based uncertainty measures, natural logarithms should be used throughout. The pattern of their respective variation over time will be clear from comparison of the time series of log standard deviation and the time series of entropy, perhaps without the $\log w$ adjustment in the case of constant bin widths. However the levels of the two measures are not comparable unless all the above adjustments are made, the last adjustment being subject to the acceptability of the normality assumption.

References

Harris, B. (2006). Entropy. In N. Balakrishnan, C.B. Read and B. Vidakovic (eds), *Encyclopedia of Statistical Sciences* (2nd edn, vol.3), pp.1992-1996. New York: Wiley.

Rich, R. and Tracy, J. (2006). The relationship between expected inflation, disagreement, and uncertainty: evidence from matched point and density forecasts. Staff Report No.253, Federal Reserve Bank of New York. (Revised version published in *Review of Economics and Statistics*, 92 (2010), 200-207.)