Calculation of weighted moments and cumulants of probability distributions and samples

[Nematrian website page: WeightedMomentsAndCumulants, © Nematrian 2015]

Equally-weighted statistics

The aggregate characteristics of probability distributions and data samples are commonly analysed using a small number of statistics corresponding to their first few moments, namely:

(a) The mean of the distribution/sample, $\bar{x}$, see MnMean, where:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

(b) The 'sample' and the 'population' variance, v and v_p respectively and the corresponding 'sample' and 'population' standard deviation, s and s_p of the distribution/sample, see MnVariance, MnPopulationVariance, MnStdev and MnPopulationStdev, where:

$$s^2 = v = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

$$s_p^2 = v_p = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

In mathematical texts, s_p is often referred to as σ and v_p as σ^2.

(c) The skew (i.e. 'skewness') of the distribution/sample, γ_1, see MnSkew, where:

$$\gamma_1 = \frac{n}{(n-1)(n-2)}\sum\left(\frac{x_i - \bar{x}}{s}\right)^3$$

(d) The kurtosis (or more precisely the 'excess' kurtosis), γ_2, of the distribution/sample, see MnKurt, where:

$$\gamma_2 = \frac{n(n+1)}{(n-1)(n-2)(n-3)}\sum\left(\frac{x_i - \bar{x}}{s}\right)^4 - \frac{3(n-1)^2}{(n-2)(n-3)}$$

We see that the 'sample' variance differs from the 'population' variance by a factor of $n/(n-1)$ representing the loss of one degree of freedom when calculating the mean. This adjustment is needed to ensure that the sample variance is an unbiased estimate of the underlying population variance if the distribution is Normal, for finite sized samples. In the large sample limit, i.e. where $n \to \infty$, the two become equal.

The formula for the skew and kurtosis given above are properly 'sample' rather than 'population' measures. Both are dimensionless quantities, and thus invariant to changes in the 'scale' of the distribution (and its 'location') (i.e. if every element of the sample x_i was replaced by $y_i = ax_i + b$, where $a \neq 0$ (the a representing a change of scale, and the b representing a change in location)

then the skew and kurtosis would remain unaltered. We could if we wished also define 'population' equivalents, see MnPopulationSkew, and MnPopulationKurt, where:

$$\gamma_{1,p} = \frac{1}{n}\sum\left(\frac{x_i - \bar{x}}{s_p}\right)^3$$
$$\gamma_{2,p} = \frac{1}{n}\sum\left(\frac{x_i - \bar{x}}{s_p}\right)^4 - 3$$

Not equally-weighted statistics

In some circumstances different elements of a sample should be given different weights in the formulation of views regarding the overall probability distribution. For example, there may be greater errors known to be associated with some specific values used in constructing the sample, so less credibility should be attached to them when deciding on the overall shape of the distribution.

Given different weights, w_i to attach to each data point (which might, for example, be associated with the square of the standard error being ascribed to the relevant data point, say ε_i^2), derivation of corresponding weighted 'population' statistics is relatively simple, e.g. the weighted mean, $\tilde{x}$, the weighted population variance, $\tilde{v}_p$, the weighted population standard deviation $\tilde{s}_p$, the weighted population skew, $\tilde{\gamma}_{1,p}$, and the weighted population kurtosis, $\tilde{\gamma}_{2,p}$, see MnWeightedMean, MnWeightedPopulationVariance, MnWeightedPopulationStdev, MnWeightedPopulationSkew and MnWeightedPopulationKurt, may be defined as follows (dropping the explicit indexing of the summation element to simplify the formulae):

$$\tilde{x} = \frac{\sum w_i x_i}{\sum w_i}$$
$$\tilde{s}_p^2 = \tilde{v}_p = \frac{\sum w_i (x_i - \tilde{x})^2}{\sum w_i}$$
$$\tilde{\gamma}_{1,p} = \frac{\sum w_i \left(\frac{x_i - \tilde{x}}{\tilde{s}_p}\right)^3}{\sum w_i}$$
$$\tilde{\gamma}_{2,p} = \frac{\sum w_i \left(\frac{x_i - \tilde{x}}{\tilde{s}_p}\right)^4}{\sum w_i} - 3$$

More difficult is to identify the correct way to incorporate an appropriate small sample size adjustment to incorporate the right number of degrees of freedom. Different commentators (or at least software providers providing software downloadable from the Internet) appear to use different approaches, particularly when deriving a suitable measure of weighted 'sample' skew.

The Nematrian website adopts the approach that small sample size adjustment factors, at least for the simpler moments/cumulants, should involve scale invariant factors that are reciprocals of expressions taking the following form (for suitable combinations of $a, b, c, \ldots$ where $a + b + c + \cdots = n$, n being the 'order' of the relevant measure) which reproduce the equally-weighted adjustment factors for cases where $w_i = k$ for $1 \leq i \leq n$ and $w_i = 0$ for $i > n$:

$$\frac{\sum\left[\left(\sum w_i^a\right)\left(\sum w_i^b\right)\left(\sum w_i^c\right)\ldots\right]}{\left(\sum w_i\right)^n}$$

This implies that the most appropriate definitions of the weighted sample variance, $\tilde{v}$, the weighted sample standard deviation $\tilde{s}$ and the weighted sample skew, $\tilde{\gamma}_1$, see MnWeightedVariance, MnWeightedStdev and MnWeightedSkew, are:

$$\tilde{s}^2 = \tilde{v} = \frac{1}{F_v}\tilde{v}_p$$
$$\tilde{\gamma}_1 = \frac{1}{F_{\gamma_1}}\tilde{\gamma}_{1,p}$$

where

$$F_v = \frac{(\sum w_i)^2 - \sum w_i^2}{(\sum w_i)^2}$$
$$F_{\gamma_1} = \frac{(\sum w_i)^3 - 3(\sum w_i^2)(\sum w_i) + 2(\sum w_i^3)}{(\sum w_i)^3}\left(\frac{\tilde{s}_p}{\tilde{s}}\right)^3$$

Using this methodology it is less clear exactly what small sample size adjustment we should make when calculating a weighted (sample) kurtosis measure, but see Rimoldini (2013). In any case, some commentators such as Press *et al.* (2007) suggest that kurtosis (and skew) "should be used with caution, or better yet, not at all". Kemp (2009) also questions the appropriateness of using skew and kurtosis to identify how non-Normal is a distributional form, see also TVaRForCubicQuantileQuantileRelationships. The corresponding Cornish-Fisher approximation that might otherwise be used to extrapolate the shape of the distributional form seems in general to give inappropriate weight to the wrong parts of the distributional form when assessing the extent of non-Normality.

Some software systems also allow users to calculate sample 'moments' relative to a predefined value, rather than the sample mean, but this is not currently possible using existing pre-defined Nematrian web service functions. It is not obvious to us whether it would be particularly useful in practice. For example, Press *et al.* (2007) note that using the formula defined above for the (equally-weighted) sample skew has a standard error, if the sample is drawn from a Normal distribution, of approximately $\sqrt{6/n}$. However, if we replace the $\bar{x}$ in its definition by the true mean of the distribution then its standard error *rises* to approximately $\sqrt{15/n}$. The corresponding approximate standard errors for kurtosis are $\sqrt{24/n}$ and $\sqrt{96/n}$ respectively. Thus the computation of both skew and kurtosis becomes *less accurate* if we use the true mean in their formulae! For ease of reference, the $\sqrt{6/n}$ and $\sqrt{24/n}$ formulae are available directly using MnConfidenceLevelSkewApproxIfNormal and MnConfidenceLevelKurtApproxIfNormal.

Weighted correlation correficients, weighted covariances and weighted population covariances are defined in an equivalent manner, see MnWeightedCorrelations, MnWeightedCovariances and MnWeightedPopulationCovariances respectively.