STATISTICAL ANALYSIS OF HYDROLOGIC DATA

This handout summarizes the formulas for computing the statistics (sample estimators) commonly used in describing of rainfall or streamflow populations. These textbook formulas are useful for hand calculations, but you should note that roundoff problems (particularly the inability to retain a large enough number of significant digits) may lead to significant errors in the results. Properly designed computer algorithms avoid these problems. If possible, I recommend the use of a major statistics package (StatView, DataDesk, JMP, Systat or MINITAB on the Macintosh; MINITAB, SPSS, SAS , or Systat on other platforms) to carry out these calculations. Using spreadsheets for statistics is not generally advisable because of roundoff errors and use of inadequate algorithms.

KEY TO SYMBOLS USED:

n	total number of observations
$\bar{x}$	sample mean of x
s_x	sample standard deviation of x
P_k	k'th percentile (value of x for which k % of the data are smaller)
$\sum_{i=1}^{n} x_i$	sum of values of x from x_1 to x_n
$\sum_{i=1}^{n} x_i^2$	sum of squared x values from x_1 to x_n
$\sum_{i=1}^{n} x_i^3$	sum of cubed x values from x_1 to x_n
$\prod_{i=1}^{n} x_i$	product of x values from x_1 to x_n

I. MEASURES OF CENTRAL TENDENCY

These statistics are measures of where the data distribution is centered, i.e., they attempt in some way to describe quantitatively where the "middle" of the data lies. If the underlying data distribution is symmetrical, all these measures will give essentially the same result. Where the underlying distribution is asymmetric (skewed), each of these measures will yield a different value.

a. **sample arithmetic mean**

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

This is arithmetic average of the x values and is usually referred to simply as the *mean*.

b. **sample geometric mean**

$$\bar{x}_g = \left(\prod_{i=1}^{n} x_i \right)^{\frac{1}{n}}$$

This is the nth root of the product of n terms. Note that the logarithm of the geometric mean is equal to the arithmetic mean of the logarithms of the individual x values.

c. **sample harmonic mean**

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

This is the reciprocal of the mean value of the reciprocals of individual values.

d. **sample median**

$$x_{md} = \text{50th percentile of data}$$

This is the middle value of a data series: half of the values are larger than this number, and half are smaller. If the data series consists of an even number of values, the median is the average of the two middle values.

e. **sample mode**

$$x_{mo} = \text{most frequent value}$$

This is the value which occurs most frequently in a data series. If there is no most frequent value, the data series is *modeless*. If there are two or more most frequent values, the distribution is *bimodal* or *multimodal*.

The arithmetic mean is the most commonly used measure of central tendency on account of its computational simplicity and general sampling stability. The US Weather Bureau uses the mean as as the precipitation normal. However, in significantly skewed distributions the mean may be misleading and the median is a better indicator of the center of the distribution.

II. MEASURES OF VARIABILITY OR DISPERSION

a. **sample standard deviation**

$$s_x = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{\left[\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right]}{n-1}}$$

The sample standard deviation is the square root of the mean squared difference between each observation and the sample mean; it is defined by the left-most formula above. This formula is awkward for hand computations, but minimizes roundoff errors. The equation is usually reorganized into the form on the right for hand calculations.

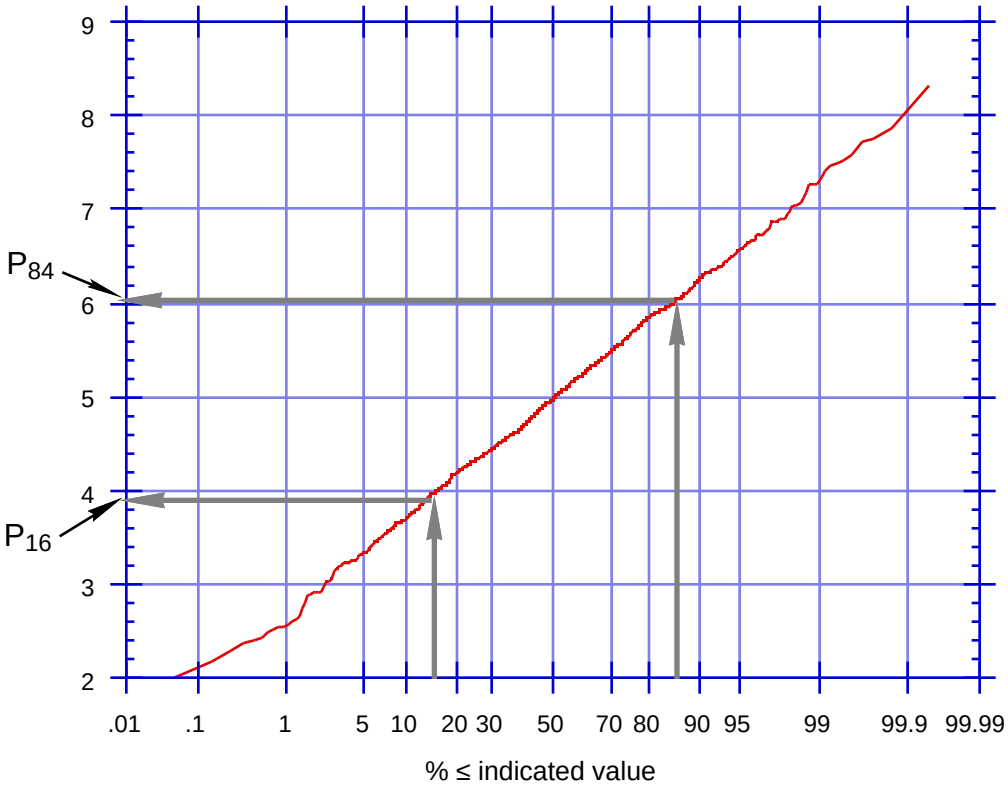
b. **sample variance**

$$s_x^2 = \text{square of the sample standard deviation}$$

c. **graphical standard deviation**

$$S_{gr} = \frac{P_{84} - P_{16}}{2}$$

The graphical standard deviation is a quick approximation of the sample standard deviation based on the cumulative curve. P_{84} and P_{16} are the 84th and 16th percentiles of the data as read from the cumulative frequency curve plotted on probability paper.



d. **sample range**

$$\text{sample range} = x_{max} - x_{min}$$

The sample range is the difference between the maximum and minimum values in a data set.

e. **sample coefficient of variation**

$$C_v = \frac{s_x}{\bar{x}}$$

The sample coefficient of variation is the ratio of the sample standard deviation to the sample mean. It is useful in comparing data sets where the variability of the data increases markedly with an increase in the mean.

III. MEASURES OF ASYMMETRY OR SKEWNESS

a. **sample skewness**

$$a = \frac{n}{(n-1)(n-2)} \sum_{i=1}^{n} (x_i - \bar{x})^3 = \frac{n^2}{(n-1)(n-2)} \left[\left(\frac{\sum_{i=1}^{n} x_i^3}{n} \right) - 3 \left(\frac{\sum_{i=1}^{n} x_i^2}{n} \right) \bar{x} + 2\bar{x}^3 \right]$$

The sample skewness measures the degree to which a distribution is asymmetric. Symmetric distributions have zero skewness. Distributions with a tail to the right yield positive (+) values of skew, while those with a left tail yield negative (–) values. Values of raw skewness are often very large and hard to interpret; the *sample coefficient of skewness*, below, is more commonly used.

b. **sample coefficient of skewness**

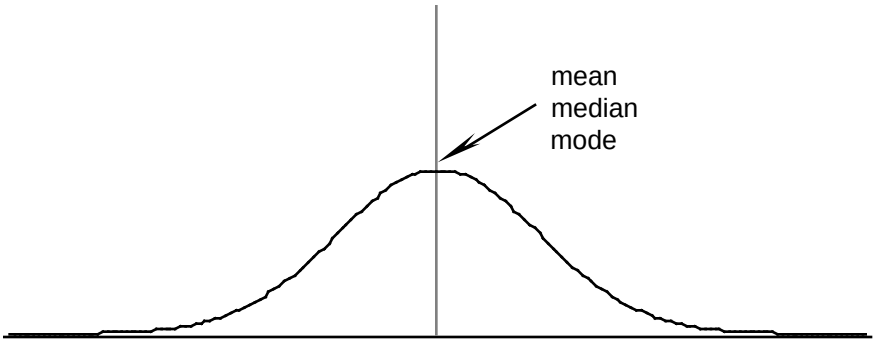
$$C_s = \frac{a}{s_x^3}$$

The sample coefficient of skewness is the ratio of the sample skewness to the cube of the sample standard deviation. For symmetrical distributions $C_s = 0$. If $C_s > 0$, the distribution is asymmetric with a tail extending to the right (*right skew*); if $C_s < 0$, the distribution is asymmetric with a tail extending to the left (*left skew*). The larger the absolute value of C_s, the more asymmetric the distribution.

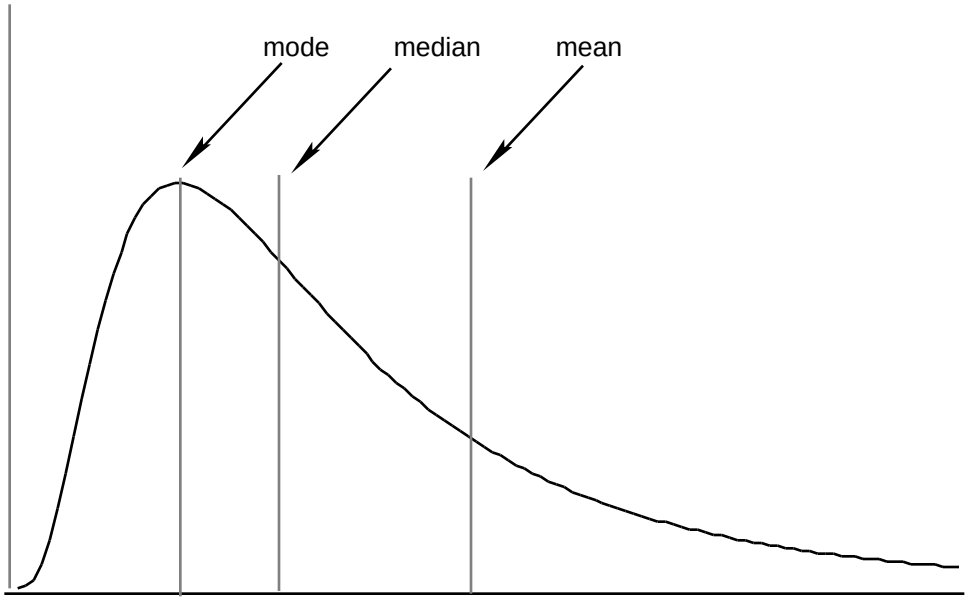
c. **Pearson's sample skewness**

$$S_k = \frac{\bar{x} - x_{mo}}{s_x}$$

This is another commonly used measure of skewness. It is much easier to calculate by hand than the ordinary sample skewness.



symmetrical distribution: skew = 0



asymmetric distribution: positive (right) skew

IV. CONFIDENCE LIMITS

a. **(1- α) x 100% confidence limits on the mean**

$$CI = \bar{x} \pm t_{n-1,\alpha/2} \frac{s_x}{\sqrt{n}}$$

$t_{n-1,\alpha/2}$ is the value of the t-statistic from tables, where n-1 is the degrees of freedom and (α/2) is typically 0.025 (i.e, α = 0.05). If n is "large enough" (typically n ≥ 25 -30) we can substitute $z_{\alpha/2}$ in place of the t - value.

Confidence limits on the mean allow us to assess the degree of uncertainty likely in our using the sample mean $\bar{x}$ to estimate the population mean μ. For example, if we construct a (1-α) = 95% confidence interval around the sample mean, we can have 95% confidence that the true population mean lies in that interval. That is, 95 times out of 100 an interval constructed from our sample in this fashion will contain the true population mean.