

Environmental Hydrology Chapter 1 Equations:

Descriptive Statistics, Data Transformations, Multiple Regression Analysis

Sample Mean

The *sample mean* (\bar{y}) is the most commonly used and is represented by the equation:

$$\bar{y} = \frac{\sum y_i}{n} \quad (1.1)$$

where \bar{y} is the sample mean, $\sum y_i$ is the sum of all measurements (i denotes any one measurement in a series), and n is the number of measurements made.

Sample Variance

The *sample variance* is described by the equation:

$$s^2 = \frac{\sum (y_i - \bar{y})^2}{n - 1} \quad (1.2)$$

where n is the size of sample and y (*bar*) = sample average

Sample Standard Deviation

Sample standard deviation is the square root of the sample variance and is described by the equation:

$$s = \sqrt{\frac{\sum (y_i - \bar{y})^2}{n - 1}} \quad (1.3)$$

where $n - 1$ = *degrees of freedom* for a sample or number of independent observations, n , in a dataset.

Standard Error

Standard error is the estimate of the variation of a statistic. The estimate of the *standard error of the mean (SEM)* is described by the following equation:

$$s_y = \frac{s}{\sqrt{n}} \quad (1.4)$$

where s is the sample standard deviation and n is the number of samples.

Another measure of statistical variation is the *coefficient of variation (CV)*, also known as the relative standard deviation. The coefficient of variation is the standard deviation expressed as a percentage mean.

$$CV = \frac{s}{\bar{y}} \cdot 100 \quad (1.5)$$

Confidence Intervals

A way to include the true value of the mean by relating it to an expression of the sample mean is by expressing the estimated value, (\bar{y}) , as a range (eg. $\bar{y} \pm 10$). Therefore the *confidence interval* of the true population mean estimated by the value of the sample mean is represented by the equation:

$$CI = \bar{y} \pm t_{n-1} \left(\frac{s}{\sqrt{n}} \right) \quad (1.6)$$

where t_{n-1} is the chosen t for $n-1$ degrees of freedom. When calculating the *CI*, the t -values are chosen for the desired level of significance.

t-distribution

The *t-distribution* is also known as the *student's t-distribution* and is a way of creating a distribution for samples collected from a population when actual population values are unknown. Standardizing a normal random variable requires that the mean, η , and the standard deviation, σ , are known. When these values are known, a normal distribution can be used, the data are scaled in terms of standard deviation as defined above, and are represented by the following equation:

$$z = \frac{y - \eta}{\sigma} \quad (1.7)$$

where z is the standardized normal random variable, y is a data value from the sample, η is the mean of the population, and σ is the standard deviation of the population.

Statistical Hypotheses

The *null hypothesis* is the hypothesis to be tested, and is generally a statement that a population parameter has a specified value or parameters from two or more populations are similar as in:

$$H_o : \mu_1 = \mu_2 \quad (1.8)$$

For this example, the alternative hypothesis would be:

$$H_A : \mu_1 \neq \mu_2 \quad (1.9)$$

Another way to state this alternative hypothesis is as a *two-sided hypothesis* in which case the alternative hypothesis is true if:

$$\mu_1 > \mu_2 \quad \text{or} \quad \mu_1 < \mu_2 \quad (1.10)$$

If, for example, the null hypothesis was:

$$H_o : \mu_1 \leq \mu_2 \quad (1.11)$$

Then the alternative hypothesis is represented by a *one-sided hypothesis* and will be defined as:

$$H_A : \mu_1 > \mu_2 \quad (1.12)$$

Linear Regression Equations

Applying a *linear regression* to a data set is a very common way to determine the presence or absence of a correlation between the data points. Linear regression applies to *bivariate data*, where two variables are related systematically such that one is a fairly constant multiple of the other. Linear regression, also known as the least-squares best-fit line, can be calculated for bivariate data that have a linear appearance when plotted in scatter plots. The best-fit line through these data is defined by the *straight-line equation*:

$$y = a + bx \quad (1.13)$$

This equation represents the relationship between an independent variable (x) and dependent variable (y). Given a value for the independent variable, use of this equation allows for the prediction of the dependent variable. The statistical parameters of this equation are the intercept of the line, a , and the slope of the line, b .

To calculate the slope of the line, the following equation is used:

$$b = \frac{n(\sum xy) - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2} \quad (1.14)$$

where n is the number of measurements.

Once the slope of the line is calculated, the intercept can be calculated using the equation:

$$a = \bar{y} - b\bar{x} \quad (1.15)$$

where \bar{y} is the sample mean of the y values sampled, and \bar{x} is the sample mean of the x values sampled.

The correlation coefficient, r , can be calculated using the following equation:

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{n(\sum x^2) - (\sum x)^2} \cdot \sqrt{n(\sum y^2) - (\sum y)^2}} \quad (1.16)$$

The coefficient of determination, r^2 (commonly reported as R^2), is the fraction or percentage of the total variation in the data that is accounted for by the regression equation. For example a regression equation with an R^2 of 0.8 accounts for 0.8 or 80% of the variability in the data. The 20% that is not accounted for and could be due to measurement error and/or factors that are not included in the equation.

Data Transformations

One common method to make power and exponential relationships linear by performing a logarithmic transformation. For example, consider the equation:

$$y = ax^b \quad (1.17)$$

If we take logarithms (to the base 10 in this case) of both sides of the equation we obtain:

$$\log_{10} y = \log_{10} a + b \log_{10} x \quad (1.18)$$

we now have a linear equation.

Multiple regression analysis

Multiple regression analysis is similar to simple linear regression models except multiple regressions contain more terms and can be used to represent more complex relationships. An example is the quadratic model also known as a second-order model:

$$y = ax^2 + bx + c \quad (1.19)$$

Often in watershed, hydrology, and ecological studies the data are related by power functions such as:

$$y = ax_1^b x_2^c x_3^d \dots \quad (1.20)$$

A logarithmic transformation of Equation 1.18 results in the linear equation:

$$\log_{10} y = \log_{10} a + b \log_{10} x_1 + c \log_{10} x_2 + d \log_{10} x_3 + \dots \quad (1.21)$$