

## Statistical and Mathematical Terms

### 1. Random variable

The outcome of a process that cannot be predicted with certainty.

### 2. Sample space

The set of all outcomes of a given random variable.

### 3. Population, Sample

The population is an ideal description of a random variable, i.e. its sample space and probability distribution. In contrast, a sample is a set of data or observations.

### 4. Continuous, discrete, quantitative, qualitative, ordinal, nominal, binary, dichotomous, polytomous, interval, ratio.

These terms all describe the type of value that is in the sample space. Continuous values are measurements made at infinite precision whereas discrete values are numerically separated (e.g. counts or measurements rounded to a given precision such as 1cm). A quantitative outcome measures the amount of something on a numerical scale (with or without units such as centimeters). Qualitative measurements are unordered labels such as gender and ethnicity. Ordinal outcomes can be ordered and may or may not be quantitative (e.g. a response of never/sometimes/always is ordinal but not quantitative). Nominal is a synonym for qualitative ("categorical" is yet another synonym). Binary and dichotomous are synonyms referring to an outcome that takes on only two values (e.g. yes/no, success/failure). Polytomous refers to a qualitative measure with more than two outcomes. An interval measure is a quantitative measure in which differences between two values have meaning but ratios may not (since the zero point on the scale is arbitrary). A ratio measure is a quantitative measure made on a scale where zero is fixed and meaningful. For example, a temperature measured in Fahrenheit or Celsius is an interval measure whereas a temperature measured in degrees Kelvin is both interval and ratio.

### 5. Scalar, vector, univariate, multivariate, dimension

A scalar is a single numerical value or variable. A vector is a finite, ordered list of values or variables. The dimension is the number of elements in a vector, or the smallest number of mathematically independent quantities needed to describe a continuous outcome. Univariate and multivariate are adjectives corresponding to scalar and vector, respectively.

#### 6. Event

An event is a subset of outcomes in the sample space. For example, if the measurement is a person's height, then 175cm is a point in the sample space and the set of heights greater than or equal to 175cm is an event.

#### 7. Probability

A probability is a number between 0 and 1 describing the likelihood that a given outcome or event will occur.

#### 8. Probability distribution

A probability distribution is the set of all probabilities corresponding to all outcomes in the sample space of a random variable. These probabilities must sum to 1.

#### 9. Simple random sample

A sample of measurements in which all observational units are equally likely to be included in the sample.

#### 10. Sample mean

The mathematical average of a set of numbers.

#### 11. Expected value

A characteristic of the population that describes the most typical value in the population. It only exists for quantitative measures. If the distribution has a density  $q(x)$  then the expected value is

$$\int xq(x)dx.$$

If the distribution has a mass function  $q(x)$  then the expected value is

$$\sum_{i=1}^m x_i q(x_i),$$

where  $x_1, \dots, x_m$  are the points in the sample space ( $m$  could be  $\infty$ ).

## 12. Sample variance

The sample variance is a measure of the dispersion or variability in a list of numbers  $X_1, \dots, X_n$ . It is notated  $\hat{\sigma}^2$ , and defined as:

$$\hat{\sigma}^2 = \frac{1}{n-1} \sum_i (X_i - \bar{X})^2.$$

## 13. Population variance

The population variance is an ideal measure of dispersion in the population. If a random variable  $X$  is described by a density function  $q(x)$  then the population variance is

$$\text{var}(X) = \int (x - EX)^2 q(x) dx,$$

where  $EX$  is defined as above. If  $X$  is described by a mass function  $q(x)$  then the population variance is

$$\text{var}(X) = \sum_i q(x_i) (x_i - EX)^2,$$

where  $x_1, x_2, \dots$  are the points in the sample space.

## 14. Statistic

A statistic is a summary quantity that aggregates many data values into a single summary value. The sample mean and sample variance are both statistics. Since the data used to define a statistic are random, a statistic is also random. As a random variable, a statistic has its own population, probability distribution, expected value, and so on.

## 15. Standard deviation

The standard deviation is the square root of the variance. The primary reason for taking the square root is to have the units of the standard deviation be the same as the units of the raw measurements. In contrast, for the variance the units are squared, so that if  $X$  is measured in cm,  $\text{var}(X)$  has units  $\text{cm}^2$ . This term can be qualified as “population” or “sample,” according to whether the square root of the population variance or the sample variance is taken.

#### 16. Standard error

The standard error is the standard deviation of a statistic.

#### 17. Accuracy

Accuracy refers to the distribution of measured values with respect to the true value of interest. If the values are not systematically in error (i.e. they fall approximately symmetrically around the true value), then they are accurate. If they are generally too high or too low, then they are not accurate.

#### 18. Precision

Precision refers to the variability, or reproducibility, of a measurement around its own mean. It does not consider how the values fall with respect to the true quantity of interest.

#### 19. Independent

Two random variables  $X$  and  $Y$  are independent if knowledge of the value of  $X$  does not allow one to predict anything about the value of  $Y$ , beyond what could be predicted when the value of  $X$  is unknown. This turns out to be a symmetric property of  $X$  and  $Y$ .

#### 20. Empirical

The term “empirical” describes any assertion or evidence based on experimentation, observation, or data analysis. In contrast, the term “theoretical” is used to describe an assertion or evidence based on deductions made from a set of principles or axioms, in the absence of direct observational data.

21. Deterministic

A deterministic process has no intrinsic randomness. If it is possible to replicate the process, then identical results will be obtained.

22. Stochastic

A stochastic process has some degree of intrinsic randomness and will not yield identical values if replicated.