statistic

Canonical name	Statistic
Date of creation	2013-03-22 14:46:18
Last modified on	2013-03-22 14:46:18
Owner	CWoo (3771)
Last modified by	CWoo (3771)
Numerical id	11
Author	CWoo (3771)
Entry type	Definition
Classification	msc 62A01
Defines	sample mean
Defines	sample variance

A *statistic*, or *sample statistic*, S is simply a function, usually real-valued, of a set of (sample) data or observations $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$: $S = S(\boldsymbol{X})$. More formally, let Ω be the sample space of the data $\boldsymbol{X}$, then S is a function from Ω to some set T, usually a subset of $\mathbb{R}^k$. The data $\boldsymbol{X}$ is usually considered as a vector of iid random variables X_i.

Examples

1. 100 light bulbs out of 1,000,000 are tested for their functionality. Then the number n, of defective light bulbs in the 100 samples is a statistic. To see this, define, for each i from 1 to 100,

$$x_i = \begin{cases} 1 & \text{if the event } X_i = \{\text{the } i\text{th light bulb is defective}\} \\ 0 & \text{otherwise.} \end{cases}$$

   Then $n = \sum_{i=1}^{100} x_i$, a function of the data. Similarly, the number of operating light bulbs is also a statistic if we switch the 1 and 0 in the above definitions for the x_i's. If we make all $x_i = 1$, then n is just the count of the observations, one of the simplest forms of sample statistics. If we make all $x_i = 0$, then $n = 0$ is a statistic that is not at all useful.

2. Let $w_1, w_2, \ldots, w_{20}$ be the weights of 20 students from a particular college. Then the average weight defined by

$$\overline{w} = \frac{1}{20}\sum_{i=1}^{20} w_i$$

   is a statistic. It is commonly called the *sample mean*. It is often used to estimate $\mathrm{E}[X]$, the expectation of a particular random variable, which, in this case, is the weight of a student in the college. Of course, other averages, such as medians, mode, trimmed mean, are also examples of (sample) statistics.

3. Using the same example as in (2), we can define

$$s^2 = \frac{1}{20-1}\sum_{i=1}^{20}(w_i - \overline{w})^2.$$

   This is also a statistic, for, after some substitution and rewriting,

$$s^2 = \frac{1}{20-1}\Big[\sum_{i=1}^{20} {w_i}^2 - \frac{1}{20}(\sum_{i=1}^{20} w_i)^2\Big],$$

which is a function explicitly in terms of the w_i's. This statistic is known as the *sample variance*, which is a common estimate of $\mathrm{Var}[X]$, the variance of the random variable X. Again, in this example, the X is the weight of a student in the college.

4. Again, borrowing from the same example above, we can simply order the weights of the 20 students in an ascending order, so we get a vector of 20 real numbers $(w_{(1)}, w_{(2)}, \ldots, w_{(20)})$. This is also a statistic, called an order statistic. It is not real-valued and its range is a subset of $\mathbb{R}^{20}$.

5. Given a set of numeric observations $X_1, X_2, \ldots, X_n$, without knowing the distribution of these observations, one can define what is known as the empirical distribution function $\hat{F}$, which is a real-valued function, based on the observations. This is an example of a statistic whose range is a function space.

Remarks.

- Any function of a statistic is again a statistic.
- Since the underlying data is assumed to be random, a statistic is necessarily a random variable.
- Although mostly real-valued, a statistic can be vector-valued, or even function-valued as we have seen in earlier examples.