STATISTICAL ANALYSIS DESIGN INCLUDING BIOSTATISTICS

Kaustubh D. Bhalerao
Department of Agricultural and Biological Engineering, The University of Illinois at Urbana-Champaign, 1304 W Pennsylvania Ave, Urbana, IL 61801, USA.

Keywords: Probability distribution, statistical estimation, biostatistics, probabilistic data analysis

Contents

Summary

Statistics as a field of study is a branch of applied mathematics dealing with the collection, organization, analysis and interpretation of quantitative data. Systems designed to interface with and support life must be designed to account for stochasticity inherent in biological processes. A statistical framework can be used to, a) quantify the degree of variability is stochastic processes, and b) make quantitative predictions on the future behavior of these processes. The field of statistical analysis is vast. This article provides some mathematical background and guidelines for developing analytical procedures to explore, visualize and model biological data. Typical problems associated with interpretation and appropriateness of statistical analyses with relevance to biological data are discussed.

1. The Need for Statistical Data Analysis

Much of our understanding of the physical world comes from observing the behavior of material systems. The observations are called *data*. Assuming that the system obeys a certain set of immutable physical laws, our observations of that system will follow certain distinguishable *trends*. A *statistic* is the measure of such a trend. Statistics are

commonly used in day-to-day conversation, without the explicit knowledge of the field. One of the most commonly used statistics is the *average* or *mean*. Averages are a measure of the centre of any data. Numbers representing, for example, the average requirement of oxygen per person per day, or the average life of an air-conditioner unit provide a preliminary insight into the behavior of the respective systems. Statistics as a field of study helps us identify which of these numbers provide the most insight into the behavior of the system, how they may be calculated by designing experiments and collecting data, and how one may make meaningful *predictions* about the behavior of the system in the future. A statistical analysis generally involves two related modes of analysis: descriptive statistics and inferential statistics.

Statistics provides the framework for quantifying uncertainty in the predicted behavior of a system. This is its defining aspect. Any real-life, non-trivial system will exhibit some level of uncertainty through random fluctuations in its operation over time. Life support systems are no exception. Statistics plays a vital role in understanding the science behind, as well as engineering technological advances of life support systems. As the name suggests, life support systems interact with life itself. The systems need to posses a level of robustness to contend with the large variability characteristic of living processes.

Characterizing the trends and variability in a living process helps us make projections for the demands that are likely to occur upon the life support systems. The design of the systems must reliably and sufficiently handle these predicted demands. Models for the projected demand are inherently *probabilistic* in nature, i.e. they are constructed to account for every possible situation likely to occur with any regularity. Every estimate is based on a probabilistic model comprising of two parts – the *expectation* i.e. the prediction based on the underlying central trend, and its *confidence*, an estimate of the error in prediction due to the inherent, unaccountable fluctuations in the system.

2. Principles of Statistical Analysis

2.1. Probability – The Foundation of Statistics

Scientific knowledge and engineering design are both based upon finite assumptions and therefore incorporate some level of uncertainty. Relationships that are generally accepted as cause-and-effect are accepted at a defined level of likelihood. There is always a small but positive likelihood that the observed effect may not be due to the cause. Understanding the basic concepts of probability is fundamental to computing and interpreting statistics. The concept of probability is fairly intuitive. It refers to the likelihood with which the outcome of an experiment will be one of its certain outcome possibilities. The first instances of a formalized approach of using probability to solving certain problems involving dice and gambling are generally attributed to Blaise Pascal (1623 – 1662) and Pierre Fermat (1601 – 1665), although numerical probabilities for dice had been calculated previously by Galileo Galilei (1564 – 1642).

Although the formalization has come a long way in over 350 years, and has found application in the sciences, social sciences, medicine and engineering, a clear interpretation of the word *probability* is still somewhat contentious. Probability can be

interpreted as a relative frequency. For example, in a coin toss, the statement "The probability of getting a heads in one-half" can be interpreted to mean that if one were to toss a coin a very large number of times, half the times it would result in heads. Another interpretation relies on the concept of "equally likely outcomes". Since a coin toss can have two outcomes, heads and tails, and assuming these outcomes are equally likely, we can say that the probability of each is one-half. This is somewhat convoluted since "equally likely" implies "equally probability", thereby making the above statement a circular argument. Fortunately, the mathematical development of probability theory and statistics is consistent and provides sufficient tools for interpretation of results relevant in science and engineering.

2.2. Basic Axioms of Probability Theory Based on Set Theory.

An *experiment* in probability theory refers to a process whose outcome is not known in advance with certainty but the set of all possible outcomes is known. Thus a coin toss is an experiment with two possible outcomes, but a coin toss result cannot be predicted with certainty. Similarly biomass yield in an agricultural production system can be considered an experiment with a non-negative value, but the exact value cannot be predicted. The set of all possible outcomes is known as the *sample space*; each specific outcome is called a *point* or an *element* in the sample space. In addition, the probability of observing the occurrence of a specific point or element within the sample space is also assumed defined. The basic objectives of probability theory are:

1. To enable the computation of the probabilities of combinations of events, and
2. To enable the revision of the probabilities of events when additional information is available.

The computation of combination of events is enabled by basic relationships in set theory pertaining to union, intersection and complements of sets. In addition, the following axioms are necessary:

1. For any event A within a sample space S, $\Pr(A) \geq 0$. i.e. if an event has zero probability, it is not legitimately a part of the sample space.
2. $\Pr(S) = 1$. For every experiment, some event from S must occur, i.e. the sample space S is exhaustive of all the possible outcomes.
3. $\Pr\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \Pr(A_i)$ For a set of *disjoint* events $A_1, A_2, \ldots, A_i$, the probability of the union of all events is the sum of probabilities of individual events.

Thus a mathematical definition of probability or more specifically a *probability distribution* is a set of numbers $\Pr(A_i)$ associated with events $(A_i), i = 1, 2, \ldots, n$ in a sample space S that satisfy Axioms 1, 2 and 3.

If some variable X can take values over a sample space S, and there exists a real valued function f that assigns a real value $f(X = A_i)$ to each possible outcome $A_i \in S$, then X is known as a *random variable*. The space S can be discrete-valued space (e.g. number of

tomatoes per plant) or continuous (e.g. the weight of an individual tomato). The function *f* is called the probability mass function for a discrete sample space or a probability density function (abbreviated as p.d.f.) for a continuous sample space. The important corollary of the second axiom mentioned above, i.e. $\Pr(S)=1$, is that the sum of the p.d.f. over all possible values taken by the random variable is unity. For discrete sample spaces, $\sum_{-\infty}^{\infty} f(x)=1$, and for continuous functions, $\int_{\infty} f(x)dx=1$. The function *f* can also be mixed distribution having discrete points and p.d.f. values over continuous intervals. Another important definition is the *cumulative density function F(x)*. (also known as the *distribution function*). The p.d.f. *f*(*x*) and the c.d.f. *F*(*x*) are mathematically related as follows:

$$F(x)=\int_{-\infty}^{x} f(x)dx \qquad (1)$$

2.3. Types of Probability Distributions

There are several different types of probability density functions used to model process often encountered in engineering and naturally occurring situations. Such distributions are well-defined mathematical expressions with well-defined properties such as mean and variance. .Some basic distributions commonly used in engineering applications are described below:

2.3.1. The Uniform Distribution

Consider a process such as rolling a fair die. It will result in one of six possible outcomes, each equally likely. The outcome will follow a uniform distribution. If there exists a random variable *X* that is equally likely to take on values from 1,2,…*k*, then the p.d.f. of *X* is given by:

$$f(x)=\begin{cases}\frac{1}{k} & \text{for } x=1,2,\ldots,k\\ 0 & \text{otherwise}\end{cases} \qquad (2)$$

This is a discrete distribution called the uniform distribution on integers. Since any distribution must add up to 1, (by Axiom 2), this implies that the sample space must be finite.
A continuous distribution over a finite interval (*a, b*) can be represented as:

$$f(x)=\begin{cases}\frac{1}{(b-a)} & \text{for } x\in(a,b)\\ 0 & \text{otherwise}\end{cases} \qquad (3)$$

The uniform distribution is the simplest of distributions that can be used to generate random numbers for simulation and design of experiments.

2.3.2. The Binomial Distribution

An experiment where only two outcomes are possible is known as a Bernoulli trial.

Examples of such a process include a coin toss, a quality control process where an item is judged defective or not. When such Bernoulli trials occur in batches, i.e. the total number of heads in 10 coin tosses, or the number of defective parts per container of 100, the process follows a binomial distribution.

Consider a machine that produces parts that have a probability $0<p<1$ of being defective. If n items are independently produced by the machine and if X is the number of defectives it has produced (where X is between 0 and n), then the probability of X taking a specific value x is given by:

$$\Pr(X = x) = \binom{n}{x} p^x q^{n-x} \tag{4}$$

The p.d.f. becomes

$$f(x) = \begin{cases} \binom{n}{x} p^x q^{n-x} & \text{for } x = 0,1,2,\ldots n \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

2.3.3. The Poisson Distribution

In the binomial distribution, if n is large and p is small, then the number of defectives x in a process will have a distribution that approaches a Poisson's distribution where $\lambda = np$. The Poisson's distribution is a natural model for discrete processes that have small probabilities of an undesirable outcome. Such process, e.g., include the relatively rare occurrence of a defective unit in a manufacturing process. The mathematical expression of a Poisson's p.d.f. is:

$$f(x) = \begin{cases} \frac{e^{-\lambda}\lambda^x}{x!} & \text{for } x = 0,1,2,\ldots \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

The Poisson distribution is used in several engineering and industrial applications, including quality control and queuing theory in communications.

2.3.4. The Exponential Distribution

An exponentially distributed, continuous random variable has a distribution defined as:

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{for } x \geq 0 \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

As can be seen, the distribution is defined on non-negative values for the random variable. This distribution is very useful, as it requires only one parameter; λ in order to completely define it. The exponential distribution is the special case of a family of distributions that are widely used in reliability studies. It is related to the Poisson

distribution in that the interval between two failures in a Poisson process is exponentially distributed.

2.3.5. The Normal Distribution

The normal distribution is perhaps the most widely-known and used distribution of all. It is a continuous distribution defined over the entire number line (-∞, ∞). The p.d.f. is given by

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\mu)^2}{2\sigma^2}} \quad \text{for } -\infty < x < \infty \qquad (8)$$

The normal distribution is completely defined by two parameters, μ and σ, the mean and standard deviation of the distribution. The function represents the familiar bell-shaped curve shown in Figure 1. The mean indicates the central tendency while the standard deviation indicates the spread of the distribution.

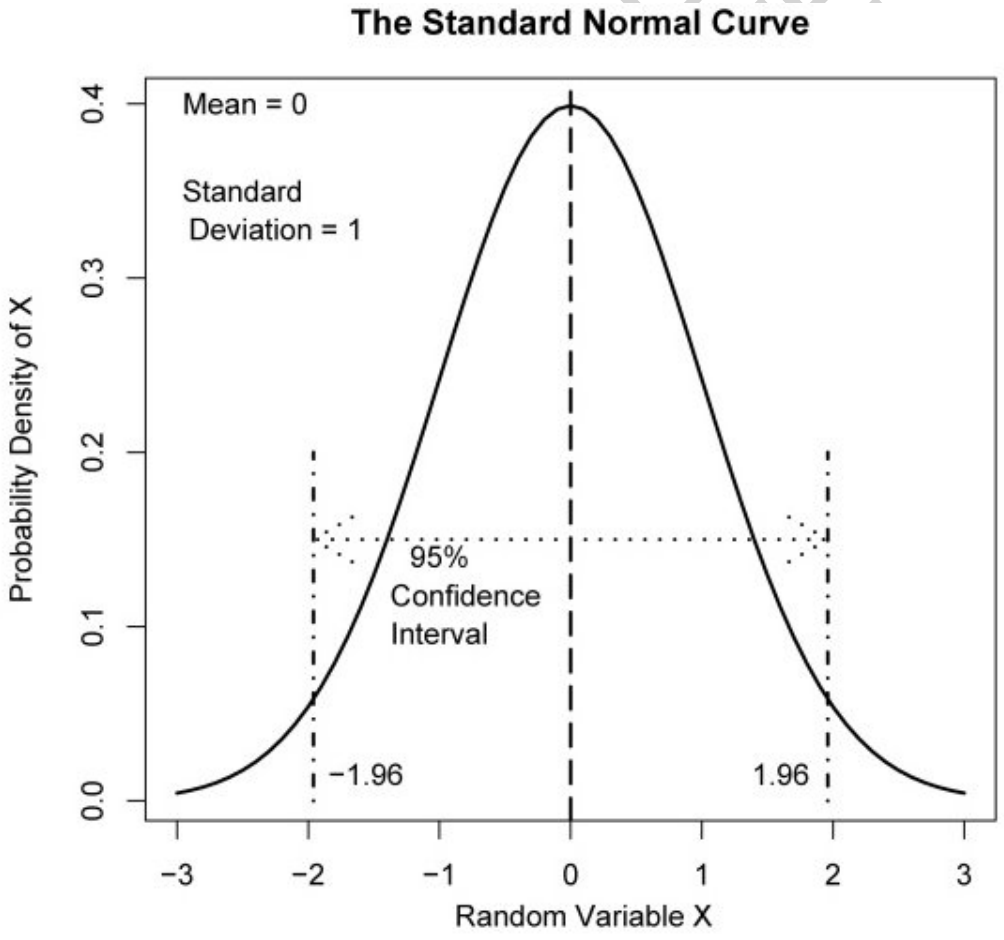

Figure 1: The normal distributions. In this plot, the mean $\mu = 0$ and the standard deviation $\sigma = 1$. When μ and σ are 0 and 1 respectively, the distribution is called the *standard normal distribution*.

The normal distribution is widely used owing to its mathematical convenience. Various functions of normally distributed random variables can be explicitly shown to be normally distributed. Many natural processes and measurements themselves have distributions that can be approximated to be normal. Finally, the *central limit theorem*, which states that if a large sample is observed from any normal or non-normal distribution, the important properties (i.e. the statistics) of that sample will be approximately normally distributed. These reasons make the normal distribution the most useful and versatile for computing statistics, making predictions and performing simulations.

One of the most convenient tools used in the analysis of normally distributed variable is the process of mapping any normal distribution $N(\mu, \sigma)$ with mean μ and standard deviation σ to the *standard normal distribution* $N(0, 1)$ with mean 0 and standard deviation of value 1. The mapping function is known as the *normal transformation* or the Z-*score* and is defined as $z = (x - \mu)/\sigma$. The convenience of this transformation stems from the fact that the area contained under all probability distributions has a value of one, i.e., $\left(\int_{\infty} f(x)\,dx = 1\right)$. The consistency in mapping from the variable x under $N(\mu, \sigma)$ to z under $N(0, 1)$ is mathematically expressed as $F_x(x) = F_z(z)$. This implies that the distribution functions at x and its corresponding *z-score* have equal magnitudes. This property allows for the convenience of publishing statistical tables of only the standard normal distribution tables. The distribution function values for any other normal distribution can be computed simply form the normal transformation.

2.4. Outcome and Expectation

Consider a discrete random variable X having a probability mass function $p(x)$. The *expected value* $(E[X])$ is the average of all the possible outcomes, weighted according to their respective occurrence probabilities. Thus for all x such that $p(x) > 0$, the expected value becomes:

$$E[X] = \sum_{x:p(x)>0} xp(x) \tag{9}$$

If X is a continuous variable with a probability density function $f(x)$, the expected value is:

$$E[X] = \int_{-\infty}^{\infty} xf(x)\,dx \tag{10}$$

The expected value can be interpreted as the centre of gravity of the probability density function. Expectations of functions of x can also be computed.

If there exists a function $g(x)$, its expected value can be computed as:

$$E\left[g(X)\right] = \int_{-\infty}^{\infty} g(x) f(x) dx \qquad (11)$$

If the function $g(x)$ is of the form $|X|^k$ and if $|X|^k < \infty$, then $E\left[|X|^k\right]$ is known as the k^{th} moment of the distribution. The basic statistic, the mean, is the first moment of a distribution, with $k = 1$. The variance (the second central moment) is the difference between the second moment and the square of the mean μ. i.e. $\mathrm{Var}[X] = E\left[(X-\mu)^2\right]$.

-
-
-

Bibliography

DeGroot M.H. (2002). *Probability and Statistics*, 3rd edition, Addison-Wesley (Boston) 816 pages. [This provides a comprehensive description of the mathematical nature of statistics, including properties of distributions and algebraic treatment of probability and statistics.]

Kunter M.H., C.J. Nachtshein and J. Neter. (2004) *Applied Linear Regression Models*, 4th edition, McGraw-Hill (Irwin) 701 pages. [This book describes the application of linear regression to model quantitative data. It describes the formal approach to building and interpreting uni- and multivariate regression models.]

Brockwell P.J. and R.A. Davis (1996). *Introduction to Time Series and Forecasting*, Springer (New York), 420 pages. [Biological systems are inherently dynamic. This text provides a foundation to model the evolution of processes in time, including how to distinguish between periodic and aperiodic aspects in their variation.]

Biographical Sketch

Kaustubh Deepak Bhalerao was born in Mumbai (formerly Bombay), Maharashtra, India on February 9th 1978. He studied at the Government College of Engineering, Pune and received the B.E. degree in Civil Engineering from University of Pune in 1999. He received his M.S. and Ph.D. degrees from The Ohio State University in Columbus, Ohio, USA in the Department of Food, Agricultural and Biological Engineering in 2001 and 2004 respectively in the area of probabilistic mechanics and statistical modeling. Between July 2004 and September 2005 he spent a year as a post doctoral researcher on a research grant funded by the National Aeronautics and Space Administration developing a statistical framework to model the reliability of advanced life support systems for space exploration.
Since September 2005 he has been on the faculty in the Department of Agricultural and Biological Engineering at the University of Illinois at Urbana-Champaign, Urbana, Illinois, USA. His current research interests include biological nanotechnology, biotechnology, bioinformatics and probabilistic methods. He is a member of the American Society for Agricultural and Biological Engineers and is currently serving on a technical committee for biological nanotechnology. He has authored several peer reviewed and conference publications including a cover article in *Applied Physics Letters* on the need for a new informatics framework for organizing nanoscale biological information required for device design.