Notes on Statistics for Bioinformatics: *by* Giri Narasimhan

Note: This is an evolving document. The current draft was last modified on January 7, 2013.

1 Introduction

We use **statistics** to analyze data that involves randomness in its generation. In *bioinformatics*, statistical methods are used for *estimation* and *hypthesis testing*.

Statistical methods for hypothesis testing can be *frequentist* or *Bayesian*. In the frequentist approach, the question asked is: "What is the probability of the observed data, assuming that the hypothesis is true?" In the Bayesian approachm the question asked is: "Given the data, what is the probability that the hypothesis is true?" The biggest advantage of Bayesian methods is that *prior* knowledge can be used effectively.

Basic Local Alignment Search Tool (or BLAST) is a procedure that searches for high-scoring alignments between two sequences. Typically, it takes a query sequence and finds all high-scoring alginments with sequences in a database. Its strength and utility lies in the fact that it (a) it is very fast and (b) it estimates the statistical significance of getting the score for an alignment. BLAST uses a Bayesian approach for its estimation. It is a search *heuristic*, which implies that it is not guaranteed to be correct, i.e., find all significant hits. However, the search as well as the P-value computations are very fast.

Here are some basic concepts in *Statistics* for us to get started. Assume that you are given a set consisting of N data values X, generated from an underlying "distribution". In some cases, the data consists of pairs of values (X and Y), or **tuples** of values.

2 Basic Definitions and Concepts

- **Random Variable** is a variable that exhibits some degree of randomness in the set of possible values that it assumes in an experiment. A random variable does not have a fixed value, but may take values from a set of different outcomes. Random variables can be **discrete** or **continuous**. They are often associated with **events** that take values from an **event space**. For example, if X is a discrete random variable representing the day of the week today, then it has 7 possible outcomes. If Y is a random variable representing the height of a person, then it it is a continuous random variable with a infinite number of different outcomes.
- An **event** is a set of outcomes of a random variable. Pr(event) is a real number between 0 and 1. For example, Pr(today is a Tuesday) $= 1/7$.
- Since events are outcomes, two events are **complementary** if one of them must happen. For example, the events "today is a Wednesday" and "today is not a Wednesday" are complementary events. Similarly, "person A has height at least 6 ft" and "person A has height less than 6 ft" are complementary events.
- **Independence** is a concept defined on random variables or events. Two or more random variables are independent if the value of one does not affect the values of the others. Two or more events are independent if the outcome of one does not affect the outcomes of the others.
- For a set of data values, $x_1, \ldots, x_n$, the **Average** value (also called the **Arithmetic Mean**, or simply the **Mean**) is defined as

$$\mu = \big(\sum_i x_i\big)/N. \tag{1}$$

For a random variable X, the mean is equal to its **Expected Value**, denoted by μ_X or $E[X]$. For a discrete random variable, this is defined as

$$\mu_X = E[X] = \sum_x xPr(x), \tag{2}$$

where the summation is over *all* possible outcomes x of the random variable X. For a continuous random variable, the summation is replaced by an integral. Wherever the context is clear, μ_X will be replaced by μ. It is sometimes also denoted by $\overline{X}$. (Do you know what is the *geometric mean*?)

- **Sample Mean** vs **Population Mean**: Typically, it is not practical to compute the average precisely. For e.g., the average height of humans on this planet, or the average length of the genomes of all the individual bacteria in your lungs. Instead, it is often possible to get a series of observations by some "fair" sampling from the population, and the arithmetic mean of the measurements is then computed using Eq. 1. This quantity is termed as the "sample mean" and is different from the real "population mean", which is equal to the arithmetic mean for the whole population, if it is possible to measure it. The *law of large numbers* (see below) states that as the size of the sample gets closer to the population size, the sample mean tends to get closer to the population mean.
- **Median** is the middle value, i.e., the data value from X with equal number of data values larger and smaller than it.
- **Mode** is the most frequent value.
- **Deviation/Residual** is the difference of a value from the mean value or its expectation.
- **Variance** (σ^2) is the mean of the square of the deviation. Also,

$$E(X^2) = \sigma^2 + \mu^2 \tag{3}$$

- **Standard Deviation**, σ, is the square root of the variance. It is often also denoted by S_X.
- **Variance** and **Standard Deviation** measure how much a random variable varies around its mean, i.e., its spread.
- **Range** is the distance between the smallest and largest value. **Interquartile range** is the difference between the first quartile and the third quartile, i.e., the range of the middle half of the data.
- **Probability Distribution** of a random variable is the set of possible values that the random variable can take along with their associated probabilities. They can be discrete or continuous.
- The value of the **cumulative distribution function** (cdf) at x is the probability that the variable takes a value less than or equal to x. The value of the **probability density function** (pdf) is the derivative of the cdf at x. The area under the pdf curve in the range $[a, b]$ is the probability that the variable takes values in that range.
- A set of random variables is independent and identically distributed (**iid**) if each random variable has the same probability distribution as the others and all are mutually independent.
- **P-value** of an event represents the probability that the event occurred by pure chance. In many areas of research, the p-value of .05 is customarily treated as a "border-line acceptable" error level. Also, p-value represents a decreasing index of the reliability of a result.

3 Distributions

- **Important discrete probability distributions:**

  Binomial number of successes in n independent Bernoulli trials, i.e., 0/1 outcome events

  Uniform equiprobable events

Geometric number of successes before failure in independent Bernoulli trials; variants of this are important to understand BLAST.

Negative Binomial number of trials to have m successes

Generalized Geometric number of trials to have k failures

Poisson Limiting form of binomial distribution where n is large and probability of success (p) is small.

- **Important continuous probability distributions:**

  Uniform equiprobable events

  Normal/Gaussian bell-shaped probability distribution with the peak at the average value, and exponentially tapering off in both directions. The **standard** normal distribution has $\mu = 0$ and $\sigma = 1$. Normal distribution is the limit of (discrete) binomial distribution, as n gets large. Therefore it also generalizes the Poisson distribution.

  Exponential It generalizes the (discrete) geometric distribution.

  Gamma It generalizes the exponential distribution, and is given by the sum of k Poisson distribution terms.

  Beta It generalizes the uniform distribution.

- If a random variable X has a normal distribution, then the random variable $Z = \frac{X-\mu}{\sigma}$ (z-score) has a standard normal distribution.
- **Outliers** An outlier is a data point that emanates from a different model than do the rest of the data.
- **Central Limit Theorem** Assume that $X_1, \ldots, X_n$ are *iid*, each with finite mean μ and finite variance σ^2. As $n \to \infty$, the random variable $\frac{(\overline{X}-\mu)\sqrt{n}}{\sigma}$ converges in distribution to a random variable having the standardized normal distribution. The theorem holds regardless of the common distribution function of the X_is. Therefore, for large enough sample size n, the sample average $\overline{X}$ and the sample sum are approximately normally distributed.
- The square of a standard normal random variable has a gamma distribution. The sum of squares of a standard normal random variable has a chi-squared distribution. The sum and the average of n iid random variables each having the exponential distribution also has an exponential distribution.
- **Law of Large Numbers** As the sample size n becomes large, the sample average $\overline{X}$ concentrates more and closely around its mean μ.

4 Tail Inequalities

- **Markov's Inequality** estimates the tail probability without any knowledge of the distribution of the random variable. If X is a nonnegative random variable with mean μ, then for any constant $c > 0$,

$$P\{X \geq c\mu\} \leq \frac{1}{c} \tag{4}$$

  For example, if $c = 2$, then Markov's inequality says that at least half of the values are smaller than twice the mean, regardless of how "skewed" the distribution might be.

- **Chebyshev's Inequality** is stronger than Markov's tail inequality and estimates the probability that a random variable is away (above or below) from the mean by more than t standard deviations.

$$Pr\{|X - \mu_x| \geq t \cdot \sigma_x\} \leq \frac{1}{t^2} \tag{5}$$

  This estimate is independent of the distribution of the random variable and is bounded by $\frac{1}{t^2}$. In other words, for any distribution, at most $\frac{1}{t^2}$ of the distribution values are more than t standard deviations away from the mean. For example, if $t = \sqrt{2}$, then at least half of the values lie in the range $(\mu_x - \sqrt{2}\sigma_x, \mu_x + \sqrt{2}\sigma_x)$.

5 Line Fitting

- Given two-variable data, **Linear Regression** is the method of fitting a line to the data. A regression line passes through (μ_X, μ_Y). **Regression** is the task of fitting a curve through a set of data points, while satisfying some goodness-of-fit criteria.
- **Mean Error** for a regression line is the average vertical distance from data points to that line. **Least Mean Error Linear Regression** finds the regression line that minimizes mean error. **Least Squares Linear Regression** finds the regression line that minimizes mean square error, or the **root mean square error** (RMSE).
- **Covariance** is a measure of how closely two variables vary.

$$\textbf{Covariance}(X, Y) = \frac{\sum(X - \mu_x)(Y - \mu_y)}{N} \tag{6}$$

- **Correlation** is a measure of how well a straight line fits data.

$$\textbf{Correlation}(X, Y) = \frac{\textbf{Covariance}(X, Y)}{S_X \cdot sS_Y}. \tag{7}$$

6 Entropy

- **Entropy** If Y is a discrete random variable generated from some probability distribution, P, then its entropy is given by

$$H(P) = -\sum_y Pr\{Y = y)\} \log Pr\{Y = y\} \tag{8}$$

- Several procedures in bioinformatics use the **relative entropy** of two probability distributions, P_0 and P_1, assuming that they are distributions over the same range of values. Relative entropy measures the amount of dissimilarity between the distributions and is defined as follows. Let Y_0 and Y_1 be two random variables over the tw distributions.

$$H(P_0||P_1) = -\sum_y Pr\{Y_0 = y\} \log \frac{Pr\{Y_0 = y\}}{Pr\{Y_1 = y\}} \tag{9}$$

  This quantity is not symmetric and $H(P_0||P_1)$ and $H(P_1||P_0)$ need not be equal.