Statistics Preliminaries

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. In some cases, the data consists of pairs of values (X and Y), or **tuples** of values.

- **Average/Mean**, $\mu_X = (\sum X)/N$. Wherever the context is clear, μ_X will be replaced by μ. It is often also denoted by $\overline{X}$.
- **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 the value from the mean value.
- **Variance** (σ^2) is the mean of the square of the deviation. Also,

$$E(X^2) = \sigma^2 + \mu^2$$

- **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 the data varies around its mean.
- **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.
- **Random Variable** is a numerical quantity that exhibits some degree of randomness in the set of possible values that it assumes in an experiment. Random variables can be **discrete** or **continuous**. They are often associated with **events** that take values from an **event space**.
- **Independence** is a concept defined on random variables or events. Two random variables are independent if the value of one does not affect the value of the other. Two events are independent if the outcome of one does not affect the outcome of the other.
- **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.
- **Chebychev Inequality**:

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

- **Markov Inequality**: If X is a nonnegative random variable with mean μ, then for any positive constant c,

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

- 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.

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

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

$$\textbf{Correlation}(X, Y) = \frac{\textbf{Covariance}(X, Y)}{S_X \cdot S_Y}$$

.

- $Pr(\text{event})$ is a real number between 0 and 1.
- Event; complementary event; event A and B; event A or B; independent events;
- **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.
- **Entropy** If Y is a discrete random variable, then its entropy is given by

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

  The **relative entropy** of two probability distributions measures the amount of dissimilarity between them.

- **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.