Review of Probability and Statistics

We muddle through life making choices based on ***incomplete information.***

Why P&S for simulation?

- Understand how to model a probabilistic system
- Validate the simulation model
- Choose the input probability distributions
- Generate random samples from these distributions
- Perform statistical analyses of simulation output data
- Design the simulation experiments
- Evaluate & compare alternatives

Random Variables

- **Experiment**: a process whose outcome is not known with certainty
- **Sample space *S***: a set of all possible outcomes of an experiment
- **Sample points**: the possible outcomes (values) in the sample space
- **Random variables**: a function/rule that assigns a real number to each point in the sample space *S*.
  - **Notation**: uppercase letters (X, Y) for RVs, lowercase for the values.

Random Variables (cont'd)

- **Distribution function (cumulative dis. function):** F(x) of the RV is defined for each x as
  $F(x) = P(X \leq x)$ for $\infty < x < -\infty$
  where $P(X \leq x)$ is the probability associated with the event $\{X \leq x\}$
- **Properties** of a distribution function or cdf:
  - $0 \leq F(x) \leq 1$
  - F(x) is nondecreasing, i.e., if $x1 < x2$, then $F(x1) \leq F(x2)$
  - $\lim_{x \to \infty} F(x) = 1$ and $\lim_{x \to -\infty} F(x) = 0$

Random Variables (cont'd)

- **Discrete** RV X: countable number of values.
  - Properties of a discrete RV:
  - **Probability mass function**: $p(x_i)$ of X
- **Continuous** RV X: uncountably infinite number of different values (an interval or a collection of intervals).
  - Properties of a continuous RV:
  - **Probability density function** : f(x) of X
- pmf vs. pdf

Random Variables (cont'd)

- Expected value or mean of RV X: central tendency (CT)
  - $\mu = \Sigma x_i p(x_i)$ if X is discrete
  - $\mu = ? x f(x) dx$ if X is continuous
- Median of a RV X: alternative measure of CT
- Variance of a RV X: variation or spread
  - $V(X) = s^2 = E(X^2) - [E(X)]^2$
- Standard deviation of RV X: spread
  - *P(X) between* $\mu \pm 1.96$s is 0.95 if X has normal distribution
- Mode:

Random Variables (cont'd)

- RVs may be related.
  - Example: In an inventory simulation, the lead time and demand may be related. An increase in demand results in an increase in lead time.
- Covariance & Correlation of 2 RVs: how well the relationship between X_1 and X_2.
  - cov $(X_1, X_2) = E[(X_1 - \mu 1)(X_2 - \mu 2)] = E(X_1 X_2) - \mu_1 \mu_2$
  - *corr* $(X_1, X_2) =$ cov $(X_1, X_2) / s_1 s_2$
  - If cov $(X_1, X_2) = 0$, X_1 and X_2 are uncorrelated.
  - The closer *corr* (X_1, X_2) is to –1 or 1, the stronger the linear relationship is between X_1 and X_2.
  - Example 9.19.

Popular Distributions & Typical Applications

- Bernoulli
  - Bernoulli trial:
    - The result of each trial may be either a success or a failure.
    - The probability p of success is the same in every trial.
    - The trials are independent: the outcome of one trial has no influence on later outcomes.
  - Examples: Toss a coin or a die.

Popular Distributions & Typical Applications

- Binomial
  - Application: Gives probability of exactly successes in n independent trials, when probability of success p on single trial is a constant. Used frequently in quality control, reliability, survey sampling, and other industrial problems.
  - Example: What is the probability of 7 or more "heads" in 10 tosses of a fair coin?
  - Comments: Can sometimes be approximated by normal or by Poisson distribution.

More on Popular Distributions

- Geometric
  - Application: Gives probability of requiring exactly x binomial trials before the first success is achieved. Used in quality control, reliability, and other industrial situations.
  - Example: Determination of probability of requiring exactly five tests firings before first success is achieved.

More on Popular Distributions

- Poisson
  - Application: Gives probability of exactly x independent occurrences during a given period of time if events take place independently and at a constant rate. May also represent number of occurrences over constant areas or volumes. Used frequently in quality control, reliability, queuing theory, and so on.
  - Example: Used to represent distribution of number of defects in a piece of material, **customer arrivals**, insurance claims, incoming telephone calls, alpha particles emitted, and so on.
  - Comments: Frequently used as approximation to binomial distribution.

More on Popular Distributions

- Normal
  - Application: A basic distribution of statistics. Many applications arise from **central limit theorem** (average of values of n observations approaches normal distribution, irrespective of form of original distribution under quite general conditions). Consequently, appropriate model for many, but not all, physical phenomena.
  - Example: Distribution of physical measurements on living organisms, intelligence test scores, product dimensions, average temperatures, and so on.
  - Comments: Many methods of statistical analysis presume normal distribution.
  - In a normal distribution, about 68% of the scores are within one standard deviation of the mean and about 95% of the scores are within two standard deviations of the mean.

More on Popular Distributions

- Gamma
  - Application: A basic distribution of statistics for variables bounded at one side - for example x greater than or equal to zero. Gives distribution of time required for exactly k independent events to occur, assuming events take place at a constant rate. Used frequently in queuing theory, reliability, and other industrial applications.
  - Example: Distribution of time between re calibrations of instrument that needs re calibration after k uses; time between inventory restocking, time to failure for a system with standby components.
  - Comments: Erlangian, exponential, and chi- square distributions are special cases. The Dirichlet is a multidimensional extension of the Beta distribution.

More on Popular Distributions

- Exponential
  - Application: Gives distribution of time between independent events occurring at a constant rate. Equivalently, probability distribution of life, presuming constant conditional failure (or hazard) rate. Consequently, applicable in many, but not all reliability situations.
  - Example: Distribution of **time between arrival** of particles at a counter. Also life distribution of complex nonredundant systems, and usage life of some components - in particular, when these are exposed to initial burn-in, and preventive maintenance eliminates parts before wear-out.
  - Comments: Special case of both Weibull and gamma distributions.

More on Popular Distributions

- Uniform
  - Application: Gives probability that observation will occur within a particular interval when probability of occurrence within that interval is directly proportional to interval length.
  - Example: Used to generate random valued.
  - Comments: Special case of beta distribution.

More on Popular Distributions

- Log-normal
  - Application: Permits representation of random variable whose logarithm follows normal distribution. Model for a process arising from many small multiplicative errors. Appropriate when the value of an observed variable is a random proportion of the previously observed value.

    In the case where the data are lognormally distributed, the geometric mean acts as a better data descriptor than the mean. The more closely the data follow a lognormal distribution, the closer the geometric mean is to the median, since the log re-expression produces a symmetrical distribution.

  - Example: Distribution of sizes from a breakage process; distribution of income size, inheritances and bank deposits; distribution of various biological phenomena; life distribution of some transistor types. The ratio of two log-normally distributed variables is log-normal.

More on Popular Distributions

- Chi-square
  - The probability density curve of a chi-square distribution is asymmetric curve stretching over the positive side of the line and having a long right tail. The form of the curve depends on the value of the degrees of freedom.
  - Applications: The most widely applications of Chi-square distribution are:
    - Chi-square Test for Association is a (non-parametric, therefore can be used for nominal data) test of statistical significance widely used bivariate tabular association analysis. Typically, the hypothesis is whether or not two different populations are different enough in some characteristic or aspect of their behavior based on two random samples. This test procedure is also known as the Pearson chi-square test.
    - Chi-square Goodness-of-fit Test is used to test if an observed distribution conforms to any particular distribution. Calculation of this goodness of fit test is by comparison of observed data with data expected based on the particular distribution.

More on Popular Distributions

- Weibull
  - Application: General time-to-failure distribution due to wide diversity of hazard-rate curves, and extreme-value distribution for minimum of N values from distribution bounded at left.

    The Weibull distribution is often used to model "time until failure." In this manner, it is applied in actuarial science and in engineering work.

    It is also an appropriate distribution for describing data corresponding to resonance behavior, such as the variation with energy of the cross section of a nuclear reaction or the variation with velocity of the absorption of radiation in the Mossbauer effect.
  - Example: Life distribution for some capacitors, ball bearings, relays, and so on.

More on Popular Distributions

- t distributions

  The t distributions were discovered in 1908 by William Gosset who was a chemist and a statistician employed by the Guinness brewing company. He considered himself a student still learning statistics, so that is how he signed his papers as pseudonym "Student". Or perhaps he used a pseudonym due to "trade secrets" restrictions by Guinness.

  Note that there are different t distributions, it is a class of distributions. When we speak of a specific t distribution, we have to specify the degrees of freedom. The t density curves are symmetric and bell-shaped like the normal distribution and have their peak at 0. However, the spread is more than that of the standard normal distribution. The larger the degrees of freedom, the closer the t-density is to the normal density.

Estimation of Means and Variances

- Suppose X_1, X_2, ..,. X_n are independent and identically distributed RVs (sharing the same probability distributions) with mean μ and variance s^2.
- To estimate μ, we use the sample mean:
- An unbiased estimator of s^2 is:
- To access the precision of the sample mean as an estimator of μ is to construct a confidence interval of μ, which involves estimating the variance of the sample mean.

Confidence Intervals for the Mean

- X_i is the same as before; Z is RV:
- The classical central limit theorem:
  - If n is "sufficient large", the random variable Z will be approximately distributed as a standard normal random variable.
  - Difficult to use the above, since the variance is generally unknown.
    - Use the sample variance
  - For n sufficiently large, we can construct an approximate confidence interval for *c*.
  - If one constructs a very large # of independent $(1-\alpha)*100\%$ CIs, each based on n observations, where n is sufficiently large, the proportion of these CIs that contain (cover) μ should be $1 - \alpha$.

More on Confidence Intervals

- What does "n is sufficiently large" mean?
- Alternate CI expression
  - Begins with the assumption that the observations are normally distributed and then uses the t distribution:
  - The new CI will also be approximate in coverage but it will less peaked and has longer tails than the normal distribution.

Hypothesis Testing

- Assuming the observations are normally distributed (or are approximately so) then test the null hypothesis H_0 that $\mu = \mu_0$, for some fixed, hypothesized μ_0.
- If $|X(n) \text{ bar} - \mu_0|$ is large, (recall that X(n) bar is the point estimator for μ), H_0 is not likely to be true.
- The form of two-tailed hypothesis test for $\mu = \mu_0$:

  If $|t_n| > t_{n-1, 1-\alpha/2}$ reject H_0

  If $|t_n| <= t_{n-1, 1-\alpha/2}$ accept H_0

Empirical Distributions & Summary

- Used when a RV has no known distribution.
- Discrete: histogram and cdf
- Continuous: generate cdf by defining a continuous, pairwise-linear distribution function.

A major task in simulation is the collection and analysis of input data. One of the first steps in this task is hypothesizing a distributional form for the input data.

How to do this?

- Compare the shape of the pdf or pmf to a histogram of the data.
- Understand that certain physical processes give rise to specific distributions.