Appendix J

Review of Basic Concepts From Probability and Statistics

In this Appendix, basic probability and statistics concepts are reviewed that are considered for the safety analysis of Chapter 9 and the quality control charts of Chapter 21.

J.1 PROBABILITY CONCEPTS

The term *probability* is used to quantify the likely outcome of a random event. For example, if a fair coin is flipped, the probability of a head is 0.5, and the probability of a tail is 0.5. Let $P(A)$ denote the probability that a random event A occurs. Then $P(A)$ is a number in the interval $0 \leq P(A) \leq 1$, such that the larger $P(A)$ is, the more likely it is that A occurs. Let A' denote the *complement* of A, that is, the event that A does not occur. Then,

$$P(A') = 1 - P(A) \tag{J-1}$$

Now consider two events, A and B, with probabilities $P(A)$ and $P(B)$, respectively. The probability that one or both events occurs $(A \cup B)$ can be expressed as

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \tag{J-2}$$

If A and B are *mutually exclusive*, this means that if one event occurs, the other cannot; consequently, their intersection is the null set $A \cap B = \varnothing$. Then $P(A \cap B) = 0$ and Eq. (J-2) becomes

$$P(A \cup B) = P(A) + P(B)$$
$$\text{(for mutually exclusive events)} \tag{J-3}$$

Analogous expressions are available for the union of more than two events (Montgomery and Runger, 2007).

If A and B are *independent*, then the probability that both occur is

$$P(A \cap B) = P(A)\,P(B)$$
$$\text{(for independent events)} \tag{J-4}$$

Similarly, the probability that n independent events, $E_1, E_2, \ldots, E_n$, occur is

$$P(E_1 \cap E_2 \cap \cdots \cap E_n) = P(E_1)\,P(E_2) \cdots P(E_n)$$
$$\text{(for independent events)} \tag{J-5}$$

These probability concepts are illustrated in two examples.

EXAMPLE J.1

A semiconductor processing operation consists of five independent batch steps where the probability of each step having its desired outcome is 0.95. What is the probability that the desired end product is actually produced?

SOLUTION

In order to make the product, each individual step must be successful. Because the steps are independent, the probability of a success, $P(S)$, can be calculated from Eq. (J-5):

$$P(S) = (0.95)^5 = 0.77$$

EXAMPLE J.2

In order to increase the reliability of a process, a critical process variable is measured on-line using two sensors. Sensor A is available 95% of the time while Sensor B is available 90% of the time. Suppose that the two sensors operate independently, and that their periods of unavailability occur randomly. What is the probability that neither sensor is available at any arbitrarily selected time?

SOLUTION

Let A denote the event that Sensor A is not available and B denote the event that Sensor B is not available. The event that neither Sensor is available can be expressed as $(A \cup B)'$.

Then, from Eqs. (J-1) and (J-2),

$$P(A \cup B)' = 1 - P(A \cup B)$$
$$P(A \cup B)' = 1 - [P(A) + P(B) - P(A \cap B)]$$
$$P(A \cup B)' = 1 - [0.95 + 0.90 - (0.95)(0.90)] = 0.005$$

J.2 MEANS AND VARIANCES

Next, we consider two important statistical concepts, *means* and *variances*, and how they can be used to characterize both probability distributions and experimental data.

J.2.1 Means and Variances for Probability Distributions

In Section J.1, we considered the probability of one or more events occurring. The same probability concepts are also applicable for random variables such as temperatures or chemical compositions. For example, the product composition of a process could exhibit random fluctuations for several reasons, including feed disturbances and measurement errors. A temperature measurement could exhibit random variations due to turbulence near the sensor. Probability analysis can provide useful characterizations of such random phenomena.

Consider a continuous random variable, X, with an assumed probability distribution, $f(x)$, such as a Gaussian distribution. The probability that X has a numerical value in an interval [a, b] is given by (Montgomery and Runger, 2007),

$$P(a \le X \le b) \triangleq \int_a^b f(x)dx \qquad \text{(J-6)}$$

where x denotes a numerical value of random variable, X. By definition, the *expected value* of X, μ_X, is defined as

$$\mu_X \triangleq E(X) \triangleq \int_{-\infty}^{\infty} x f(x)dx \qquad \text{(J-7)}$$

The expected value is also called the *population mean* or *average*. It is an average over the expected range of values, weighted according to how likely each value is.

The *population variance* of X, σ_X^2, indicates the variability of X around its population mean. It is defined as:

$$\sigma_X^2 \triangleq E[(X - u_X)^2] \triangleq \int_{-\infty}^{\infty} (x - u_X)^2 f(x)dx \qquad \text{(J-8)}$$

The positive square root of the variance is the *population standard deviation*, σ_X.

These calculations are illustrated in Example J.3.

EXAMPLE J.3

A mass fraction of an impurity X varies randomly between 0.3 and 0.5 with a uniform probability distribution:

$$f(x) = \frac{1}{0.2}$$

Determine its population mean and population standard deviation.

SOLUTION

Substituting $f(x)$ into Eq. J-7 gives:

$$\mu_X = \int_{-\infty}^{\infty} x f(x)\,dx = \int_{0.3}^{0.5} x\left(\frac{1}{0.2}\right)dx$$

$$\mu_X = \left(\frac{1}{0.2}\right)\left(\frac{1}{2}x^2\right)\Bigg|_{0.3}^{0.5} = 0.4$$

Thus μ_X is the midpoint of the [0.3, 0.5] interval for X. To determine σ_x, substitute $f(x)$ into Eq. J-8:

$$\sigma_X^2 = \int_{-\infty}^{\infty} (x - \mu_X)^2 f(x)\,dx = \int_{0.3}^{0.5} (x - 0.4)^2\left(\frac{1}{0.2}\right)dx$$

$$\sigma_X^2 = \left(\frac{1}{0.2}\right)\left(\frac{1}{3}(x - 0.4)^3\right)\Bigg|_{0.3}^{0.5} = 0.00333$$

$$\sigma_X = 0.0577$$

J.2.2 Means and Variances for Experimental Data

A set of experimental data can be characterized by its *sample mean* and *sample variance* (or simply, its *mean* and *variance*). Consider a set of N measurements, $\{x_1, x_2, \ldots, x_N\}$. Its mean, $\bar{x}$, and variance s^2 are defined as (Montgomery and Runger, 2007)

$$\bar{x} \triangleq \frac{1}{N}\sum_{i=1}^{N} x_i \qquad \text{(J-9)}$$

$$s^2 \triangleq \frac{1}{N-1}\sum_{i=1}^{N} (x_i - x)^2 \qquad \text{(J-10)}$$

The standard deviation s is the positive square root of the variance.

The mean is the average of the dataset while the variance and standard deviation characterize the variability in the data.

J.3 STANDARD NORMAL DISTRIBUTION

The normal (or Gaussian) probability distribution plays a central role in both the theory and application of statistics. It was introduced in Section 21.2.1. For

probability calculations, it is convenient to use the *standard normal distribution*, $N(0, 1)$ which has a mean of zero and a variance of one. Suppose that a random variable X is normally distributed with a mean μ_X and variance σ_X^2. Then, the corresponding standard normal variable Z is

$$Z \triangleq \frac{X - \mu_X}{\sigma_X} \tag{J-11}$$

Statistics books contain tables of the *cumulative standard normal distribution*, $\Phi(z)$.

By definition, $\Phi(z)$ is the probability that Z is less than a specified numerical value, z (Montgomery and Runger, 2007; Ogunnaike, 2010):

$$\Phi(z) \triangleq P(Z \leq z) \tag{J-12}$$

Example 9.2 illustrates an application of $\Phi(z)$.

J.4 ERROR ANALYSIS

In engineering calculations, it can be important to determine how uncertainties in independent variables (or inputs) lead to even larger uncertainties in dependent variables (or outputs). This analysis is referred to as *error analysis*. Due to the uncertainties associated with input variables, they are considered to be random variables. The uncertainties can be attributed to imperfect measurements or uncertainties in unmeasured input variables. Error analysis is based on the statistical concepts of means and variances, considered in the previous section.

As an important example of error analysis, consider a linear combination of p variables,

$$Y = \sum_{i=1}^{p} c_i X_i \tag{J-13}$$

where X_i is an independent random variable with expected value μ_i and variance, σ_i^2. Then, Y has the following mean and variance (Montgomery and Runger, 2007):

$$\mu_Y = \sum_{i=1}^{p} c_i \mu_i \tag{J-14}$$

$$\sigma_Y^2 = \sum_{i=1}^{p} c_i^2 \sigma_i^2 \tag{J-15}$$

Equations J-14 and J-15 show how the variability of the individual X_i variables determines the variability of their linear combination, Y.

EXAMPLE J.3

Experimental tests are to be performed to determine whether a new catalyst A is superior to the current catalyst B, based on their yields for a chemical reaction. Denote the yields by X_A and X_B, and their standard deviations by 3% and 2%, respectively. What is the standard deviation for the difference in yields, $X_A - X_B$?

SOLUTION

Let $Y = X_A - X_B$, an expression in the form of (J-13) with $c_A = 1$ and $c_B = -1$. Thus Eq. (J-15) becomes

$$\sigma_Y^2 = \sigma_A^2 + \sigma_B^2$$

Thus,

$$\sigma_Y = \sqrt{\sigma_A^2 + \sigma_B^2} = \sqrt{(3\%)^2 + (2\%)^2} = 3.6\%$$

Thus, the standard deviation of the difference is larger than the individual standard deviations.

REFERENCES

Montgomery, D. C. and G. C. Runger, *Applied Statistics and Probability for Engineers, 4th ed*., John Wiley & Sons, Hoboken, NJ, 2007.

Ogunnaike, B. A., *Random Phenomena: Fundamentals of Probability and Statistics for Engineers*, CRC Press, Boca Raton, FL, 2010.