UNIT III STUDY GUIDE

Binomial and Normal Distributions

Probability Distributions

To look at probability distributions, one should define a random variable as an unknown that could be any real number, including decimals or fractions. Many problems in life have real numbers of any value of a whole number and fraction or decimal as the value of the random variable amount. *Discrete random variables* will have a certain limited range of values, and *continuous random variables* may have an infinite range of possible values. These continuous random variables could be any value at all (Render, Stair, Hanna, & Hale, 2015).

One true tendency is that events that occur in a group of trials tend to cluster around a middle point of values as the most occurring, or highest probabilities they will occur. They then taper off to one or both sides as there are lower probabilities that the events will be very low from the middle (or zero) and very high from the middle.

This middle point is called the *mean* or *expected value E(X):*

$$E(X) = \sum_{i=1}^{n} X_i P(X_i)$$

Where X_i is the random variable value, and the summation sign Σ with n and i=1 means you are adding all n possible values (Render et al., 2015).

The sum of these events can be shown as graphs. If the random variable has a discrete probability distribution (e.g., cans of paint that can be sold in a day), then the graph of events may look like this:

Figure 2.4 –Class Distribution
(Render et al., 2015)
The bar heights show the probability for any X (or, P(X)) along the y-axis, given the discrete number for X along the x-axis and no fractions for discrete variables (no half-cans of paint).
The *variance* (σ_2) is the spread of the distribution of events in a probability distribution (Render et al., 2015). The variance is interesting because a small variance may indicate that the event value will most likely be near the mean most of the time, and a large variance may show that the mean is not all that reliable a guide of what the event values will be, as the spread is all over a charted curve of trials.
n
Variance = $\sigma_2 = \Sigma\ [X_i - E(X)]_2\ P(X_i)$
i=1
Because of mathematical relationships, often in probability calculations the *standard deviation* is also important, which is the square root of the variance:
Standard deviation = $\sigma = \sqrt{\sigma 2}$
Then, there are *probability distributions of continuous random variables.* These probabilities of whole numbers and fractions or decimals are important to researchers too, as probabilities of temperatures, rainfall, achievement scores, engine horsepower, and many other things do not generally measure to be a whole or discrete number. As is the case for what discrete probability distributions resemble if their levels or bars are connected, most continuous probability distributions naturally appear in curves and can be described in mathematics as a probability function f(X):
(Render, et al., 2015)

On the graph, any value for probability X, or P(X), can be estimated by following up the y-axis for any value X on the x-axis to the curve. Importantly, the probability of X being between two points on the x-axis is the area under the curve described by drawing lines up from the x-axis at those two points until they meet the curve. Because this is a graph of a continuous probability distribution, there are infinite points along the curve and the endpoints may or may not be counted—they don't change the probability solutions. These mathematical facts give us a couple of possible processes to calculate probability of continuous random variables:

The Binomial Distribution

The *Bernoulli Process* is used to determine a *Binomial Distribution*, or the probability of a certain number of outcomes in a given amount of trials *n* if the outcomes are binomial or can be only one of two outcomes. There are more conditions for using the Bernoulli Process: the trials are independent, have the same probability, and are positive in number (Render et al., 2015).

If you recall from mathematics that the symbol ! indicates a factorial, or factors, of the number indicated multiplied by each other, as in 4! = (4)(3)(2)(1) = 24, then for p = probability of a trial success, $q = 1-p$ = probability of a trial failure, and r = number of successes, then the binomial formula for the probability of r successes in n trials is:

n!

----

r!(n – r)! prqn-r

You can calculate probability with this formula, and tables have been developed for binomial distributions, so these can be consulted if you know *r* and *n*.

The Normal Distribution is often used because it is predictably symmetrical in its graphed probability that peaks in the curve's middle (at the mean, μ) and tapers off to $p = 0$ at the extreme high and low value ends. Also, when the mean is changed, the probability curve moves on the x-axis to the new mean value, and when the standard deviation σ changes, the probability range spreads (for larger σ) or shrinks (for smaller σ). The formula for the probability function is a large one (see on page 40 of the textbook).

There is an easier way to work with normal distributions, with the help of normal distribution tables which have been created to provide a simple way to look up the desired factor or probability. These tables require that you convert the normal distribution to a *standard normal distribution* by setting the mean (μ) to 0 and standard deviation (σ) to 1. You establish a new variable, the standard random variable Z, which will be one of the terms in normal distribution tables; to find it the mathematical relationship is:

X – μ

Z = σ

with X being the value of the random variable or range limit desired, and Z being the number of standard deviations σ from the mean μ. If the number of standard deviations σ is a positive number, the area under the curve with X as its limit will be to the right of the mean μ.

In the Haynes Construction Company example on page 43 of the textbook, see that this standard normal distribution formula and the normal distribution tables helped find the area inside the curve with a limit of X = 125 days as 0.89435 in the table, meaning Hayes has about an 89% chance of not violating the contract by going over 125 days.

Persistent mathematical research and working of the normal distribution tables led to rules researchers found that tend to be true. These are useful when your research does not need several-places-of-decimals precision but more of an overall estimate. For a normal distribution (Render et al., 2015):

1. Approximately 68% of the values will be within ± 1 standard deviation of the mean.
2. Approximately 95% of the values will be within ± 2 standard deviations of the mean.
3. Almost all (about 99.7%) of the values will be within ± 3 standard deviations of the mean.

(Render, et al., 2015)

When managers brief ideas and offer a 95% or 99% confidence intervals that an event will occur as predicted, they are referring to a calculation like those illustrated here; that the probability is the large amount under the respective curve, leaving only a small probability on the ends beyond one, or two, or three standard deviations from the mean.

The F Distribution is a continuous probability distribution that estimates the ratio of two variances, so it is used when you have them and the ratio is important. Note the symmetrical curve that skews to the right:

(Render, et al., 2015)

The y-axis is probability, so in an F-distribution the probability is high that the ratio of the degrees of freedom of the numerator and denominator (F) is about 1, and low that F is either near 0 (at the curve's beginning on the left) or very high (on the right).

Appendix D of the textbook has standard F-distribution tables. Given variance degrees of freedom as a numerator (df_1) and denominator (df_2), use the table for the significance (α) you want, $\alpha = .05$ or 5% probability or $\alpha = .01$ or 1% probability furnished in Appendix D, and find the F-ratio for that numerator and denominator. If your calculated value of F is equal to or larger than the F-ration in the table, then your F value is significant to the probability of the table (95% for $\alpha = .05$ or 99% for $\alpha = .01$). The F-distribution will be of some use to us in Unit VII.

What are the benefits of probability distribution? You can work them if, and only if, you have values for the terms in their respective formulas. These may have to be estimated, or an investment made to run experiments to determine them to a certain degree of confidence. Remember, a number that is guesswork

and perhaps not near what is actually happening will lead to a probability determination that is not too realistic either.