MGM600/Phase 3
Article 1

Distributions and Probability

Distributions and probability provide the basic foundation on which most inferences are made.

Probability can be thought of as the laws of chance. It is the likelihood that something will happen, *based on what you already know*. The simplest example, often used to illustrate these concepts, is the probability that a tossed coin will land heads or tails. There are only two possible **outcomes**, and each outcome has an equal chance of occurring; therefore the probability that a coin will land on heads is 0.5, and the probability that it will land on tails is also 0.5. If we use dice as the example, the probability that a die lands on any particular number is 0.167, or 1/6. The probability that an event will occur is calculated as the number of ways that an event can occur divided by the total number of outcomes that are possible for discrete variables (variables that can be counted). Note that the probabilities for all possible outcomes of an event always sum to 1.00.

Some terminology:

P (event A) = the probability that event A will occur

Using our examples, then:

P (fair tossed coin landing on heads) = ½ = 0.5
Numerator = 1 = number of ways to get heads
Denominator = 2 = total possible outcomes = 1 way to get heads + 1 way to get tails

P (2 coins landing on heads) = ¼ = 0.25
Numerator = 1 = number of ways to get heads = HH = coin 1 is heads and coin 2 is heads
Denominator = 4 = total possible outcomes = HH HT TH TT

P (one die landing on 4) = 1/6 = .167
Numerator = 1 = number of sides of a die that have 4
Denominator = 6 = total possible outcomes = 1 side for each number 1-6

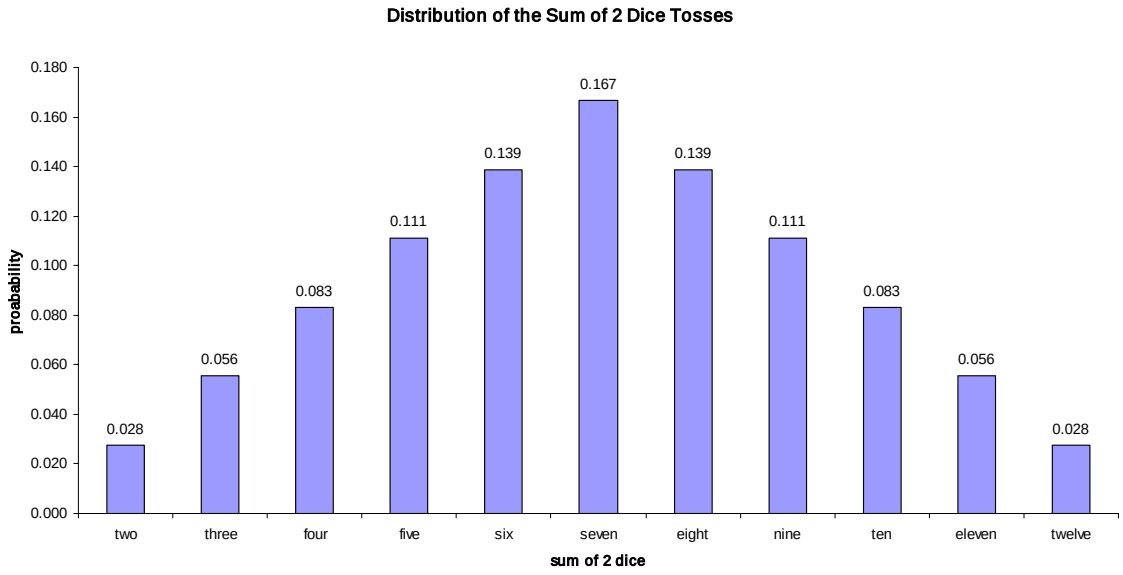
P (two dice show a total of 4) = 3/36 = 0.083 = 8.3%
Numerator = 3 = ways to get 4 (1 3, 2 2, 3 1)
Denominator = 36 = total possible outcomes (The following chart shows the 36 possible outcomes of tossing two dice. The right side represents the 36 outcomes, the left side shows the sum of the two dice)

1 1	1 2	1 3	1 4	1 5	1 6	→	2	3	4	5	6	7
2 1	2 2	2 3	2 4	2 5	2 6	→	3	4	5	6	7	8
3 1	3 2	3 3	3 4	3 5	3 6	→	4	5	6	7	8	9

4 1	4 2	4 3	4 4	4 5	4 6	→	5	6	7	8	9	10
5 1	5 2	5 3	5 4	5 5	5 6	→	6	7	8	9	10	11
6 1	6 2	6 3	6 4	6 5	6 6	→	7	8	9	10	11	12

2, 12 = 1 possible way each
3, 11 = 2 ways each
4, 10 = 3 ways each
5, 9 = 4 ways each
6, 8 = 5 ways each
7 = 6 possible ways

If we make a bar chart or histogram of the frequency of the possible outcomes for the tossing of two dice, it looks like this:



The distribution of possible outcomes is charted above. Notice that it resembles a bell, so this shape curve is often called a "bell curve." A **distribution** is defined by its mean and standard deviation, and can be thought of as a summary of the probability of all possible outcomes for an event. Many different kinds of data look like this when graphed, assuming a random, representative sample has been taken. Examples include the IQ of all people in the U.S., the height or weight of a large population, the number of weekly calls to a computer help desk, the production output of a large manufacturing business, etc. A distribution is **normal** (its mean, median and mode are all the same) when it is "taller in the middle" with symmetrically decreasing tails at either end. Data that fits this distribution is often rescaled to have a mean of zero and a standard deviation of 1.0 because this shape is so common. This **standard normal distribution** is frequently used in statistical tests such as Hypothesis Testing and ANOVA. This distribution only applies to ratio and interval, not categorical, data. The distribution becomes a line, instead of individual bars, when data is continuous. The standard normal distribution (plus some others mentioned below) is referenced to determine the probability that an outcome would occur by pure chance.

There are other common distributions used in business decision making. The **binomial distribution** is used with discrete, as opposed to continuous, variables. It looks more and more like the normal distribution the more possible outcomes there are. The **t-distribution** also resembles the normal distribution and has a mean of zero. It is used for small samples sizes (< 30) and is almost the same as the standard normal distribution for sample sizes beyond that. The **chi-square distribution** is used to ascertain if two variables are independent. The **F-distribution** is used in regression analysis to establish the significance of the coefficients.