We use probability to help us look forward and to quantify the uncertainty of future events/outcomes. We use probability to make predictions. We use predictions to manage uncertainty, including the uncertainty created by random sampling error.

**Random Events**

Uncertainty derives from random events that derive from random processes (a process that produces variable outcomes), the results of which, cannot be known in advance. Random events can have either discrete or continuous units of measure and, thus the number of possible events can be discrete and enumerated, discrete and too large to be enumerated, or continuous and impossible to enumerate.

Measuring the weight of 100 randomly selected air travelers would be a random process as it will yield a variety of values that cannot be known in advance. Weight is a continuous numerical value, so it would be impossible to enumerate all the possible values for the weights of the 100 observations.

Determining the net worth of 500 randomly selected people attending the U.S. Open Tennis Tournament would be a random process as it will yield a variety of values that cannot be known in advance. Net worth is expressed in terms of money (currency) and so, is technically a discrete numerical value; however, as we learned in Module 3, the possible values in currency are too many and the increments between scale points (1 cent if U.S. dollars) is too small to make counting practical, so currency is usually treated as continuous numerical data.

Determining the number of sales made to qualified leads in a week for each of your sales people would be a random process as it will yield a variety of values that cannot be known in advance. The number of sales (not the sales amount) is a discrete numerical variable bounded on the low side at zero and on the high side at the number of qualified leads provided for the week.

**Probability of a Random Variable (*X*)**

The probability that some specific value (denoted as an italicized, lower case *x*) of a random variable (denoted as an italicized, capital *X*) is a number that represents the relative likelihood that *x* will be observed. Depending on whether *X* is a discrete or continuous variable the probability of observing *x,* denoted *P*(*X = x*), would be effectively zero (for continuous data) or a value between zero and 1 (for discrete data).

We will address probabilities for continuous random variables in the next module (8). We address probabilities for discrete random variables in this module (7).

A probability of 0 means that the event cannot occur and a probability of 1 means that the event must happen.

**Distributions**

The term *distribution* has been used in this course before. We used it in modules 3 and 4 as we covered the preliminary analysis and description of a data set, first visually and then numerically. We focused on the concept of a frequency distribution (discrete data) and of a normal distribution (continuous data).

In this module, we will build on the concept of a frequency distribution. Specifically, we will move from distributions created based on observed (empirical) frequencies to **probability distributions** created by the observations that are probable (predicted) based on a mathematical equation (classically assigned probabilities).

We will explore a variety of commonly used probability models. Probability models use mathematical equations to assign a probability to each possible outcome in the sample space. The sample space is defined by the random process. Mathematical models are very useful, but should be used with full awareness of their underlying assumptions.

Probability models should be reasonably realistic in their description of the stochastic (uncertain) characteristics of the random process. In other words, on average, over the course of many trials, the probability model should predict the frequency of occurrence for each possible outcome in the sample space reasonably well. However, it is important that the probability model be simple enough to analyze.

**Referring back to our conceptualization of critical thinking, which is at the core of this course, we will need to: 1) integrate all of these models and their underlying assumptions into our long-term memories (remembering), 2) fully understand these models (understanding) and, 3) use these models to inform our decision making (applying).**

**PDF (Probability Distribution Function)**

The PDF is the mathematical equation used by the model to assign probabilities for each of the possible outcomes in the sample space. It provides the probability that an observation is equal to a specific value.

**CDF**

The CDF is the mathematical equation used by the model to calculate the cumulative sum of probabilities. It provides the probability that an observation is less than or equal to a specific value.

**1 – CDF**

The complement of the CDF provides the probability of an observation being greater than some specific value.

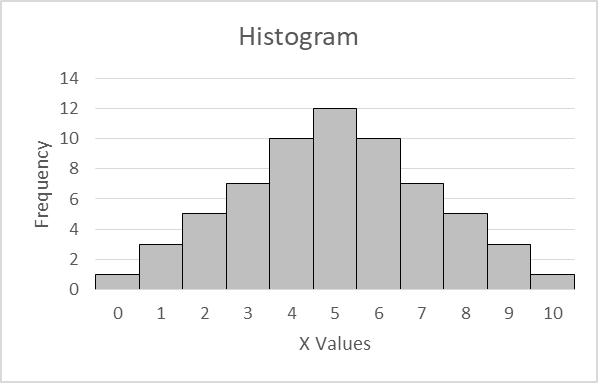
**Expected Value and Variance**

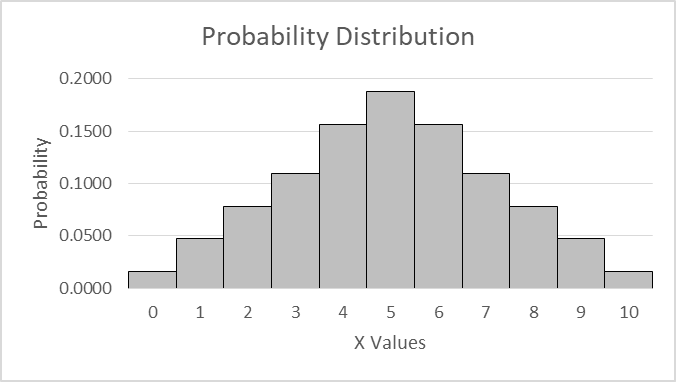
**Expected Value E(X)**

Keeping in mind that the possible outcomes within a sample space are discrete values, the expected value E(*X*) of a discrete random variable is the sum of all *x* values weighted by their probabilities. It is a weighted average. It is the population mean and is, thus, a measure of center. Because it is a calculated value, it does not have to be observable.

The table below is a frequency distribution that has been converted into a discrete probability distribution. The first column (*X*) provides the observed values within the data set. The second column (Freq.) provides the frequency of occurrence for each of those values. The third column (PDF) provides the empirically assigned probability of observing each of the *x* values. The fourth column (CDF) is the running total of the sum of the PDF values. The fifth column [*xP*(*x*)] provides the weighted PDF values. We calculate the expected value E(*X*) or the mean as the sum of the *xP*(*x*) column.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | PDF | CDF |  |
| *X* | Freq. | P(*X* = *x*) | P(*X* <= *x*) | *xP*(*x*) |
| 0 | 1 | 0.0156 | 0.0156 | 0.0000 |
| 1 | 3 | 0.0469 | 0.0625 | 0.0469 |
| 2 | 5 | 0.0781 | 0.1406 | 0.1563 |
| 3 | 7 | 0.1094 | 0.2500 | 0.3281 |
| 4 | 10 | 0.1563 | 0.4063 | 0.6250 |
| 5 | 12 | 0.1875 | 0.5938 | 0.9375 |
| 6 | 10 | 0.1563 | 0.7500 | 0.9375 |
| 7 | 7 | 0.1094 | 0.8594 | 0.7656 |
| 8 | 5 | 0.0781 | 0.9375 | 0.6250 |
| 9 | 3 | 0.0469 | 0.9844 | 0.4219 |
| 10 | 1 | 0.0156 | 1.0000 | 0.1563 |
|  | 64 | 1.0000 | E(X) | 5.0000 |
|  |  |  | Mean |  |
|  |  |  |  |  |



****

In this distribution, the probability of observing a 7 is *P*(*X* = 7) .1094. Although Excel displays a leading zero as in 0.1094, we do not because a probability cannot be greater than 1. Displaying a leading zero indicates that the value could be greater than 1. We use the symbol to denote **approximately equal to**. We do that to indicate that .1094 is not necessarily the exact value. In this case, the exact value is .109375. Rounding is a very important consideration when calculating and when comparing calculation methods.

The probability of observing a 7 or less is *P*(*X*7) .8594 and the probability of observing a value greater than 7 is *P*(*X* > 7) 1 - .8594 .1406.

The probability of seeing an observation in the interval [3, 6], between 3 and 6 inclusive, is the sum of the individual probabilities for each of the values 3, 4, 5, and 6 = *P*(3 (.1094 + .1563 + .1875 + .1563) .6095. We can also use the CDF to calculate the probability of a value lying within an interval. The *P*(3 the CDF of 6 minus the CDF of 2 would be an easier way to make our calculation. Make certain you understand why we subtract the CDF for 2 from the CDF for 6 rather than subtracting the CDF for 3 from the CDF for 6. *P*(*X* 6) – *P*(*X* 2) .7500 - .1406 .6094. The difference between these two approximate probabilities is due only to the effect of rounding.

As you collect more and more observations and update your frequency distribution and your probability distribution, your predictions for future observations will become increasingly more accurate (the law of large numbers).

**Variance and Standard Deviation**

Expanding our probability distribution further, we can calculate the variance and standard deviation for the distribution.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| *X* | Freq. | PDF | CDF |  | *xP*(*x*) |  | |  | | --- | |  | |  |  |
| 0 | 1 | .0156 | .0156 |  | 0.0000 |  | -5.0000 | 25.0000 | 0.3906 |
| 1 | 3 | .0469 | .0625 |  | 0.0469 |  | -4.0000 | 16.0000 | 0.7500 |
| 2 | 5 | .0781 | .1406 |  | 0.1563 |  | -3.0000 | 9.0000 | 0.7031 |
| 3 | 7 | .1094 | .2500 |  | 0.3281 |  | -2.0000 | 4.0000 | 0.4375 |
| 4 | 10 | .1563 | .4063 |  | 0.6250 |  | -1.0000 | 1.0000 | 0.1563 |
| 5 | 12 | .1875 | .5938 |  | 0.9375 |  | 0.0000 | 0.0000 | 0.0000 |
| 6 | 10 | .1563 | .7500 |  | 0.9375 |  | 1.0000 | 1.0000 | 0.1563 |
| 7 | 7 | .1094 | .8594 |  | 0.7656 |  | 2.0000 | 4.0000 | 0.4375 |
| 8 | 5 | .0781 | .9375 |  | 0.6250 |  | 3.0000 | 9.0000 | 0.7031 |
| 9 | 3 | .0469 | .9844 |  | 0.4219 |  | 4.0000 | 16.0000 | 0.7500 |
| 10 | 1 | .0156 | 1.0000 |  | 0.1563 |  | 5.0000 | 25.0000 | 0.3906 |
|  | 64 | 1.0000 | E(X) = | | 5.0000 |  |  |  | 4.8750 |
|  |  |  |  | |  |  | |  | | --- | |  | |  | 2.2079 |

The variance of the distribution denoted Var(*X*) or is the weighted average of the squared deviations from the mean (both below and above), where the weights are the probabilities of observing each value of *X*. The standard deviation denoted is the square root of the variance.

**Simple Application**

A real estate developer is considering a condominium project. The project would yield different levels of return depending on the demand for condominium units. The three levels of demand used for analysis are low, medium, and high. The four sizes of development being analyzed are small, medium, large, and extra large. The probability of low demand is 20%, medium demand is 50%, and high demand is 30%. The levels of return for the different sized projects are provided below. What is the expected value for each project size?

|  |  |  |  |
| --- | --- | --- | --- |
|  | Demand Probabilities & Payoffs | | |
| Project Size | Low (.2) | Medium (.5) | High (.3) |
| Small | $7,000,000 | $8,000,000 | $9,000,000 |
| Medium | $5,000,000 | $12,000,000 | $14,000,000 |
| Large | $1,000,000 | $10,000,000 | $19,000,000 |
| Extra Large | -$4,000,000 | $5,000,000 | $25,000,000 |

E(Small project) = (7,000,000 x .2) + (8,000,000 x .5) + (9,000,000 x .3) = $8,100,000

Etc.

Etc.

Etc.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Demand Probabilities & Payoffs | | |  |
|  | Low | Medium | High |  |
| Project Size | 0.2 | 0.5 | 0.3 | EV |
| Small | $7,000,000 | $8,000,000 | $9,000,000 | $8,100,000 |
| Medium | $5,000,000 | $12,000,000 | $14,000,000 | $11,200,000 |
| Large | $1,000,000 | $10,000,000 | $19,000,000 | $10,900,000 |
| Extra Large | ($4,000,000) | $5,000,000 | $25,000,000 | $9,200,000 |

We can see from our calculations of expected value, that on average, the medium sized project is optimum – it yields the highest weighted average return.

**The module companion contains detailed calculators for all the probability models discussed below. PLEASE review the calculators carefully and “play” with them to expand your understanding of the distributions.**

**Uniform Discrete Model**

The **uniform distribution** describes a random variable with a known number of consecutive integers from *a* to *b.* Each of the integer values is equally likely, i.e., their probabilities are identical or uniform. The entire model relies on just these two values (parameters).

**We use this model** when we know that the probability of each possible outcome is identical or when we have no reason to believe that any outcome is more likely to occur than another.

Excel uses this model to generate random numbers.

**Binomial Distribution**

A Bernoulli experiment is a random process that has only two possible outcomes, it is a binary model. The outcome of interest is labeled “success” (*X* = 1) and the other outcome is labeled “failure” (*X* = 0). Success has nothing to do with the desirability of the outcome, it only indicates which of the two outcomes is being studied. The probability of success is denoted (the Greek letter pi) and the probability of failure is 1 - . The symbol is used for several different parameters in statistics, none of which is the mathematical constant 3.14159 used in geometry. The trials are independent and, thus, the probability of success remains constant for each trial. The mean of the Bernoulli experiment has mean and variance .

Many business activities can be viewed as binary … the activity occurred or it did not: a target was hit or it was not, a sale was made or it was not, a deadline was met or it was not, a goal was achieved or it was not, etc.

A **binomial distribution** is created when **Bernoulli experiments** are repeated *n* times. Each Bernoulli trial is independent and the probability of success remains constant for each trial. In a binomial distribution, *X* is the number of success in *n* trials, the mean is , the variance is and, thus, the standard deviation is . The domain (sample space, possible outcomes) is *x* = 0, 1, 2, …, *n.*

**Recognizing Binomial Applications**

* The number of trials (*n*) is fixed and known.
* There are only two possible outcomes for each trial
* The trials are independent
* The probability of success remains constant for all trials
* The random variable (*X*) is the number of successes in *n* trials.

You need to understand the information above to recognize when to use the binomial distribution.

**Application**

You are the sales manager for your company. Your sales model is to use only qualified leads, no cold selling. Your sales team has a collective closing rate of 39% and an average sales amount of $250 per sale. From what you can determine from the data you have collected for years, the probability of making a sale is constant for each attempt. You have 117 qualified leads for the upcoming week. What is the expected value of sales volume for next week?

Sales can be conceptualized as a binary event, the sale is made or it is not. We understand that the sales attempts are independent (due to the constant ) and that the probability of success is 39% or .39. If we use the binomial distribution, the expected value is or 117 x .39 = 45.63 or 46 closed sales next week. As the mean sales amount is $250, the expected sales volume for next week = 45.63 x 250 = $11,407.50

What is the interval of expected sales volume using 1 standard deviation?

The standard deviation for a binomial distribution with *n* = 117 and = .39 is = . The expected value of $11,407.50 $1,320 creates an interval of [$10,087.50, $12,727.50].

What is the probability that sales volume will fall within that interval?

The CDF for X = (45.63 + 5.28 = 50.91 or 51) .8668 and the CDF for X = (45.63 – 5.28 = 40.35 or 39) .1221. Subtracting .1221 from .8668 gives us the probability of seeing an observation within the interval approximately equal to .7447 and so we are approximately 74.47% confident that our sales volume for next week will be between $10,087.50 and $12,727.50.

**Poisson Distribution**

The **Poisson distribution** describes the number of occurrences (*X*) within a randomly selected period of time (seconds, minutes, hours, days, weeks, etc.) or space (square inch, square foot, acre, linear mile). The occurrences must be random and independent and the period of time must be a continuum or the area of space must be contiguous (no gaps or spaces within the time period or within the space). The random variable *X* depends not just on the length of time or size of the space you use to count the occurrences, but also when or where you begin counting. Thus, you must select when or where you begin counting randomly.

The Poisson distribution relies on only one parameter, the mean occurrence rate per randomly selected unit of time or space. The variance of the distribution is also and the standard deviation is . There is no theoretical boundary for *X*, the number of occurrences within the randomly selected time or space, however, Poisson probabilities taper off rapidly as *X* increases, so the effective range of *X* is from 0 to 20, but 10 is probably a more useful limit. Thus, the time periods or areas of space must be sufficiently small to keep *X* within its effective range. If = 90 occurrences per hour, we can reformulate to = 1.5 occurrences per minute. Then, *X* must be in terms of per minute.

**Recognizing Poisson Applications**

* An event of interest occurs randomly over time or within space
* The occurrences are independent of each other
* The average occurrence rate remains constant
* The random variable (*X*) is the number of occurrences observed within a randomly selected unit of time or space

You need to understand the information above to recognize when to use the Poisson distribution.

**Application**

You are the floor manager in a large retail store. Among other things, you are responsible to assure that cashier checkout lines are reasonable. You have 10 cashier stations, but it would be wasteful to keep them all staffed for the full hours during which the store operates. Thus, you must schedule cashiers to optimize the desire to maintain short checkout lines and the desire to keep labor costs low.

You have collected data for 90 days on the precise day and time that each customer enters the queue to pay for their purchases. The mean number of arrivals on Mondays between 10 AM and 11 AM is 75. What is the probability that more than 7 customers arrive within a randomly selected 5-minute period during that period on any given Monday?

We immediately suspect that this process is reasonably well described by the Poisson distribution because it involves the arrival of customers (occurrences) within a randomly selected period of time. We know that the mean arrival rate is 75 customers per hour during the period within which we will randomly select a 5-minute period. So, we reformulate as . We can reasonably assume that the customer arrivals are reasonably random and that they are reasonably independent (they are not shopping together in large groups, etc.). Using our Excel calculator, we calculate the probability of seeing more than 7 customers arrive at the queue for checkout in any randomly selected 5-minute period on a Monday between 10 AM and 11 AM [1 – CDF(7)] is .2911.

**Hypergeometric Distribution**

The **hypergeometric distribution** is very similar to the binomial distribution because it describes a random process that generates binary outcomes, success and failure. The difference from the binomial is that sampling is *without* replacement (like for permutations and combinations in Module 6), thus the probability of success will not be constant.

We need to know the population size *N* from which we will draw a sample of *n* items, and we need to know the number of successes (*s*) within the population. The random variable (*X*) is the number of successes within the sample.

**Recognizing Hypergeometric Applications**

* The process has a known population size (*N*) with a known number of successes (*s*) – this will be the most obvious characteristic
* There is a known sample size that was drawn without replacement
* The probability of success is not constant – you will have to surmise this from the sampling without replacement

**Application**

You are the operations manager of a manufacturing company making widgets, buggy whips, baby bumpers, and other mythical items. A delivery truck has just unloaded 500 units of Makebelievium Z. The delivery driver wants your signature accepting the delivery. Your supplier guarantees that each shipment will have no more than 5% defective units and authorizes the use of acceptance sampling through the use of hypergeometric probabilities calculated with sample sizes of a minimum of 10% of the delivery size. You may refuse the delivery if the cumulative probability of the observed number of defects within the sample is greater than 95%. Should you accept the delivery?

In compliance with the supplier’s guarantee, you randomly sampled 50 of the 500 units and found 6 defective units. The CDF for 6 items is .9923 or 99.23%, which is more than 95% so you should refuse the delivery based on the supplier’s guarantee. (You might decide to accept the delivery if the lead time on reordering them is too long for you to meet your production requirements AND under the condition that the supplier will pick up the defective units and credit your account for them.)

**Geometric Distribution**

The **geometric distribution** is related to the binomial distribution except that the random variable (*X*) is the number of trials until the first success. The trials are independent and, thus the probability of success is constant for all trials. There must be at least one trial, but there is no sample size limit.

**Recognizing Geometric Applications**

* The random variable (*X*) of interest is the number of trials before the first success
* The trials are independent
* The probability of success remains constant for each trial
* The sample size is a minimum of 1 and has no maximum

**Application**

You are the sales manager for XYZZZ Telemarketing Company. Your current sales script for product A has a success rate of 0.3%. In an effort to improve your results, you modified the script for product A. You decided to test the script and had your entire sales team use the new script. The first closed sale required 1,017 telephone calls. Should you reconsider using the new script?

The use of the geometric distribution is clearly warranted here as we are using the number of calls until the first success to evaluate the effectiveness of the new sales script. We have the for the original sales script. We know the intent of creating a new script was to increase sales, so we would be looking for a higher success rate (fewer than typical number of calls to the first success based on the success rate for the original script).

Using our Excel calculator, we calculated the CDF for *X* = 1,016 calls when = .003 to be .9528 meaning that we would expect to see success before 1,017 calls 95.28% of the time and we would expect to see success after 1,016 calls .0472 or 4.72% of the time. It seems unlikely that the new script is more effective than the original one.

**Linear Transformation**

Linear transformation of a random variable (*X*) is achieved by either adding a constant, multiplying by a constant, or both.

Where *X* is the random variable, *a* and *b* are constants and .

Rule 1:

Rule 2:

Adding a constant (*b*) shifts the mean, but does not affect the standard deviation.

Multiplying by a constant (*a*) shifts the mean and affects the standard deviation.

**Summing Random Variables**

Rule 3:

Rule 4:

Rule 5: )

Where:

The

The

The

**End of Module**