Probability

A **probability** is a mathematical measurement of the likelihood that an event will occur. This gives a quantitative description for the chance that the event will happen - the higher the probability, the more likely that you will see the event.

Mathematical measures have several common properties:

1. A measure is a non-negative number;
2. A measure of zero indicates emptiness of some kind;
3. The total measure of several non-overlapping objects is the sum of the measures of each object.

Additionally, probability is a normed measure, that is, it has an upper boundary. For probability,

4. The highest probability possible for an event is 1 (100%).

An **event** is something that may occur when you perform a certain action, like obtaining an even number when you roll a die. The probability of any event is a number between 0 and 1. If the probability is 1 (likelihood is 100%), we say the event is **certain**. If the probability is 0 (no chance of happening) the event is **impossible**. If you can break down a complex event into simpler, non-overlapping occurrences (rolling an even number with a die consists of rolling a 2, rolling a 4, rolling a 6) then you can compute the probability of the entire event by adding the probabilities of each of the simpler events.

To calculate the probability of an event we start with a **sample space**. A sample space is a list of all the possible things that may happen when you perform an action of some kind. In order for the sample space to help you find probabilities effectively, you should make your sample space the list of **elementary outcomes**, the simplest occurrences that can happen when you perform the action. For example, “getting an even number when you roll a die” is not an elementary outcome because it consists of simpler outcomes - rolling a 2, rolling a 4, rolling a 6. Your sample space for one roll of a single die is {1, 2, 3, 4, 5, 6}.

Your next step is to assign probabilities to each of the elementary outcomes. This must reflect their relative likelihoods. In many (but not all) cases, each of the elementary events are equally likely, as they would be for a fair die. Since the total probability for the entire space is 1 (since something in the space must happen), the probability for each elementary outcome in the equally likely case will be 1 out of the total number of elementary outcomes (1/6 for each of the possible outcomes for the die). If the elementary outcomes are not equally likely, you must adjust these probabilities to show the correct relative likelihoods but also to make sure that the total probability is still 1.

The sample space with the assigned probabilities is a **probability space**. If the probability space is small enough, it is a good idea to write it out because you will then be able to find the probability of any event by just counting up the probabilities of all the elementary outcomes that make up the event. In this way we find the probability of rolling an even number with a single fair die is \( \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{3}{6} = \frac{1}{2} \). This works because the elementary outcomes making up the event will be non-overlapping so all you have to do is add up their probabilities.

When you are writing out a sample space, make sure you understand what an elementary outcome is. Be sure you understand exactly what you would see when you perform the action. If you roll a die twice, you would have two numbers. If three coins are tossed, you would have three heads or tails. If you draw a ball from balls labeled A, B, C, D and also roll a die you would have a letter and a number.
If the sample space is too large to write out, there are some other techniques you can use. You can count the number of elementary outcomes and the number of outcomes that satisfy a particular event by using combinatorial methods.

Briefly, a permutation is an arrangement of objects in which each object is distinct. The distinction could be from an ordering (first, second, etc.) or simply because the two objects are different (in a blackjack card hand one card is dealt face up the other face down - the up and down cards are different). In this case you calculate the total number of possibilities by the product of the number for each item (in the blackjack hand there would be 52 possibilities for the down card, then 51 for the up card, $52 \cdot 51 = 2652$ possibilities) because for each of the first possibilities there would be the number for the second - think of a tree diagram.

A combination is an arrangement in which the objects are not distinct (as when three balls are drawn from an urn together). In this case you can find the number of possibilities by dividing the previous answer by the number of ways those items could be arranged among themselves - which would be the factorial of that number. If three balls were drawn from seven in an urn that would be $\frac{7 \cdot 6 \cdot 5}{3 \cdot 2 \cdot 1} = \frac{7 \cdot 6 \cdot 5}{3! \cdot 4!}$. This computation is represented by the symbol $\binom{7}{3}$, the number of combinations found by selecting 3 items from 7 items without distinction.

There are also some useful probability rules of computation.

As with the Monte Hall problem, when an event is complicated because there are many ways it could happen, the complementary event (the event that it does not happen) may be simpler. You can count all the ways the event can happen, or you can count all the ways it cannot happen (the rest). The probability that an event will happen is then $1$ - the probability that it does not happen.

We have seen that if you can list all the (non-overlapping) ways an event can occur, the total number of ways will be the sum of each of these. So, if an event can happen one way or a second way or a third way, etc. and these do not overlap, you can simply add the probability of each of these together to get the total probability.

If for an event to occur, one thing followed by another followed by a third, etc. have to happen, then, like a permutation, the probability of the entire event will be the product of the probabilities. Note that frequently once one part occurs, the sample space changes for the second and further steps. For example, the probability of getting two kings in a blackjack hand is $\frac{4 \cdot 3}{52 \cdot 51}$ because after the down card is dealt it remains in the end, so the up card comes from the remaining 51 cards and only three of these could be kings if the down card is a king.

The probability $\frac{3}{51}$ is not just the probability that the second card is a king. It is the probability that the up card is a king given that the down card was a king and was not returned to the deck. We call this changed probability a conditional probability. It is affected by the condition that the down card is a king and kept out of the deck before the up card is dealt.
If you are only interested in one aspect of the result when you perform an action, and you can describe that aspect with a number, then a random variable may be useful. A random variable assigns a number to each outcome and also keeps the probability for that number that is assigned to each elementary outcome leading to that number.

The random variable $S$, which gives the sum on two dice that are rolled, can have the values 2, 3, 4, 5 ... 11, 12. The probability that $S = 6$ is the probability of (1, 5), (2, 4), (3, 3), (4, 2), (5, 1) or $\frac{5}{36}$.

The probability distribution for a random variable is the list of the probabilities for each of the values of that random variable.

For example, the probability of the random variable $S$, described above, is

<table>
<thead>
<tr>
<th>S</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
</tr>
</thead>
</table>

There are many types of distributions that we encounter. For example, the uniform distribution in which each value of the random variable has the same probability and the binomial and Poisson distributions which were mentioned in class. These are examples of discrete distributions - the possible values of the random variable are separate. You can get 2 heads or 3 heads when you flip a coin 5 times but you can't get 2.3 heads.

Other random variables are continuous, for which all values in a range of numbers may be used, such as the variable that assigns height to each person. Heights can be any number in some reasonable range. An important continuous distribution is the Gaussian or normal distribution. The graph of this distribution has a symmetric, bell-shaped appearance, showing maximum likelihood around the center of the distribution and rapidly tailing off at the sides. Many natural measurements have this distribution such as heights, weights, IQ scores, the lives of light bulbs, etc. This accounts for the importance of the distribution.

The center of the normal distribution is the mean (or average), which is also the most likely value of this distribution, and the spread is measured by the standard deviation. The symbol used for the mean is the Greek letter $\mu$, and the symbol for standard deviation is the Greek letter $\sigma$.

A normal distribution has 68% of the probability within 1 standard deviation of the mean, 95% within 2 standard deviations of the mean, and 99.7% within 3 standard deviations of the mean. There is very little likelihood that values that are more than 4 standard deviations from the mean will occur.

When working with the normal distribution, you usually use a standardized form that has the mean adjusted to 0 and standard deviation adjusted to 1. This done by converting your data (X-values) to standardized Z-scores. You do this by subtracting the mean from your value and then dividing by the standard deviation. This would always be your first step, using the formula $Z = \frac{X - \mu}{\sigma}$.

Some important values for the normal distribution are given in the table below:
<table>
<thead>
<tr>
<th>Standard deviations</th>
<th>Amount of probability past this value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.28</td>
<td>.10</td>
</tr>
<tr>
<td>1.65</td>
<td>.05</td>
</tr>
<tr>
<td>1.96</td>
<td>.025</td>
</tr>
<tr>
<td>2.33</td>
<td>.01</td>
</tr>
<tr>
<td>2.57</td>
<td>.005</td>
</tr>
</tbody>
</table>

Because of the symmetry of the distribution, the values for the left side are the same, just negative.

We have 99% likelihood of being below 2.33; 90% of being above –1.28; 95% of being between –1.96 and 1.96.

**Statistics**

A statistic is a measurement of your data. The mean, median, standard deviation, IQR, fences, maximum value are all statistics. The subject of statistics has as its purpose to help people make decisions in situations in which they do not have complete knowledge. This decision-making is based on likelihoods (that is, probabilities) rather than certainties (which are unknown).

**Exploratory Analysis**

Exploratory data analysis is used to organize a set of data into a form from which you can discover useful information about your data.

Your first step will be to create a stem-and-leaf diagram of your data. Use the most natural stems for the numbers in your data. Spread your data out at least once or twice until it is too spread out to be useful. While you are doing that you should also put the numbers in order. If the data is too spread out, then you should compress it. See what you can tell about your data from looking at the diagram. Do you see one homogeneous population or are there different groups in the diagram? What might the different groups represent?

Next, calculate some statistics that help measure important characteristics of the data. Usually the most important statistics are the ones that find where the “middle” is and those that the spread of the data.

The traditional measure of the center is the average, or mean. However, the mean is easily influenced by just a few values which are much smaller or larger than the others. This can bias your conclusions about the data. Preferable is a statistic that is robust, or resistant to the influence of these extreme values. A robust statistic that accurately finds the center is the median, the value that is exactly in the middle. (Note that for 15 numbers, the 8th is the median and for 12 numbers, the average of the 6th and the 7th gives the median).
The **standard deviation**, which is the average distance of each data value from the mean, is commonly used to measure the spread of the data. However, the standard deviation is also easily influenced by outlying values. A resistant measure of spread is given by the **interquartile range**, or IQR, the distance between the quarter values (each of which is halfway between an end of the data and the media).

To identify values that are truly different in value from the rest of the data, or **outliers**, we want an objective measurement. Outliers are defined as values that lie outside boundaries, or **fences**, that depend on how spread out the set of data is. The fences are located at a distance of 1.5 times the IQR from the quarters (below the lower quarter and above the higher quarter). Values beyond the fences are so far from the rest of the data that they appear to be different in some way from the rest of the data.

Combining what you learn from the appearance of the stem-and-leaf diagram, the values of the resistant statistics, and what makes the outliers different from the rest of the data should give you enough insights to get you started on an analysis of your data.

**Inference**

Statistical inference proceeds from information given by a sample from the population under study to conclusions based on the probabilities of what the sample represents about the population.

It is common for there to be questions about the average measurement of a population. When the entire population is not available, you would a (hopefully random, representative) sample drawn from the population. You can calculate the average obtained from the sample (symbol $\bar{x}$). If the sample is truly random and representative of the population, then $\bar{x}$ should be approximately equal to the true population mean, $\mu$.

It is important to realize that $\mu$, although unknown, is a constant value. However, $\bar{x}$ will vary according to the sample that is used, with some answers being more likely than others. $\bar{x}$ is a random variable and has a probability distribution.

The distribution of $\bar{x}$ is normal, with mean equal to $\mu$ (that is, on average the sample mean will equal the true population mean), and standard deviation equal to $\frac{\sigma}{\sqrt{n}}$, where $n$ is the size of the sample. This shows that larger samples will show less variability in their averages than small samples.

For example, if a brand of tire claims to have an average tread life with an average, $\mu = 40,000$ miles, and a standard deviation, $\sigma = 5000$ miles, then if you had a tire wear out after 35,000 miles $z = \frac{35000 - 40000}{5000} = -1$ and more than 10% of these tires could be expected to wear out this soon. So, although your tire lasted less than average, it’s not a surprising result.

However, if all four of your tires needed replacement by 35,000 miles, the standard deviation for a sample of four tires is $\frac{5000}{\sqrt{4}} = 2500$ and $z = \frac{35000 - 40000}{2500} = -2$. Only about 2.5% of samples of four tires wear out this fast, so you should suspect that there is something wrong about the claim that these tires average 40,000 miles.
Another type of problem solving using statistical inference is finding estimates for a statistic. These estimates are given as intervals of values, called **confidence intervals**.

If a sample of 100 of the tires from the previous example (same standard deviation of 5000 miles) is tested for tread wear and the average obtained is 41,500 miles \(( \bar{x} = 41,500 \quad n = 100 )\), then the standard deviation for this sample is \( \frac{5000}{\sqrt{100}} = 500 \) miles. For 95% confidence, you would only exclude the lowest and highest 2.5%, that is, the values below \(-1.96\) or above 1.96 standard deviations. For convenience, let’s round off 1.96 to 2. Then a 95% confidence interval for the true mean tread ware is \( 41,500 \pm 2(500) \) or between 40,500 and 42,500 miles. This would certainly be convincing evidence that these tires do have an average tread wear of at least 40,000 miles.

Understand that 95% confidence means that 95% of the time to find the interval this way, you will have the true average value within your interval. Of course, you will not know whether or not this is true this time. The only way to know for sure would be to test every tire. The result you found using statistics is the best you can do if you only test 100 tires.

The numbers included in the table for the normal distribution also allow you to find 90%, 98%, or 99% confidence intervals. Just realize that if you want more confidence you will get wider intervals (99% requires 2.57 standard deviations instead of 1.96 for 95% confidence). Although you certainly would not want less than 90% confidence, you will have to decide for your problem whether high confidence or high precision is more important.