Discrete and Continuous Variables

The probability models we have looked at so far have involved discrete events. You can think of discrete as meaning the variable can only have certain numerical values, with no intermediate values in between.

Example Rolling a die leads to a discrete distribution, since the outcomes are $S = \{1, 2, 3, 4, 5, 6\}$, ie, we can't get an outcome between 1 and 2.

A variable is continuous if between any two values for the variable there exists another possible value for the variable.

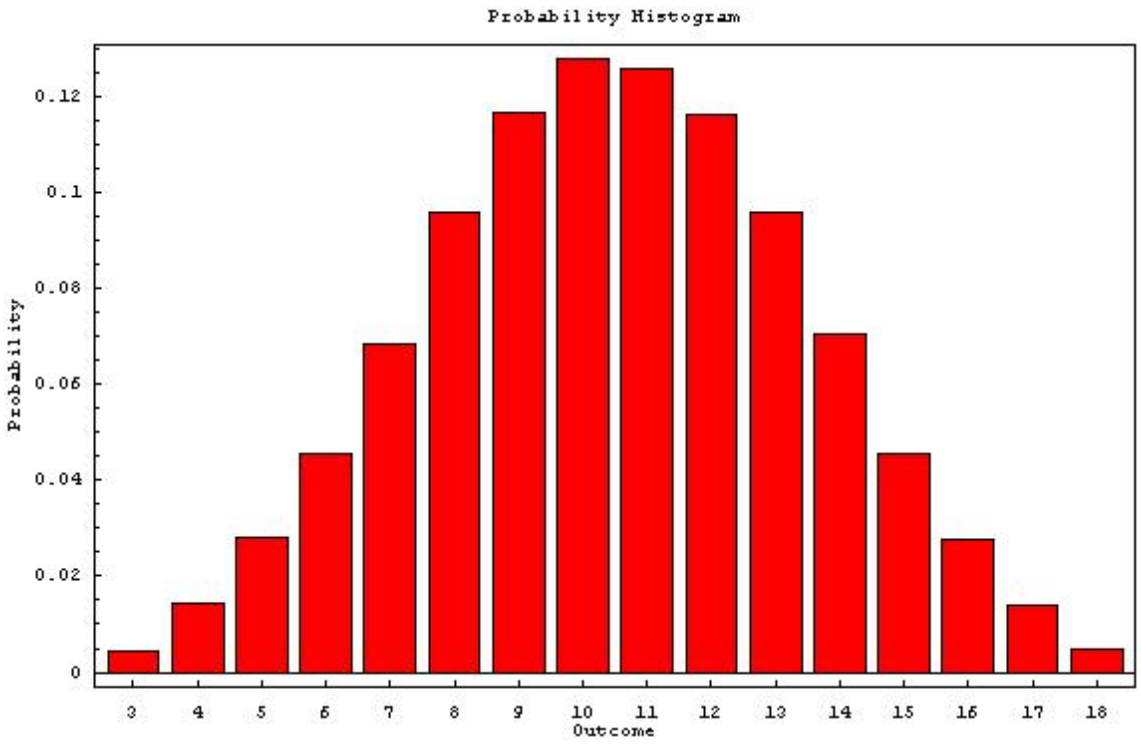
Example The height of students attending colleges in the USA is continuous, since for any two heights there could be a height between them (there is actually an infinite number of heights between any two heights!).

Notice that there doesn't *have* to be a value of the variable between the two values, but there could be.

Discrete distributions are described by histograms, and continuous distributions are described by areas under curves.

Example of Rolling Three Fair Dice

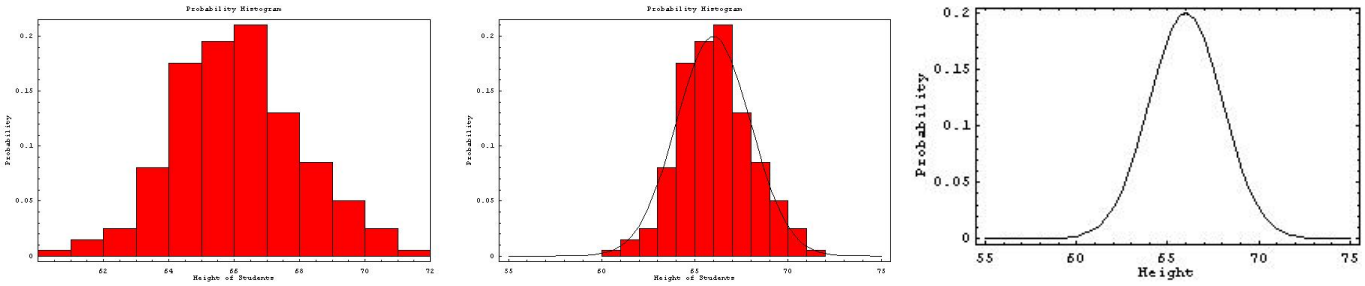
Here is the histogram that was created by rolling three fair dice 100,000 times and recording the occurrences (computer experiment).



- The height of each bar is the probability of rolling the number labelled at the base of the bar.
- If the width of the base is 1, then the area of each rectangle is the probability of rolling the number at the base.
- Since the sample space is discrete, there is no way to get any other outcomes, like 8.5 for example.
- The sum of all the areas in all the rectangles is 1.

We can extend these ideas to a continuous distribution, where intermediate values are always possible.

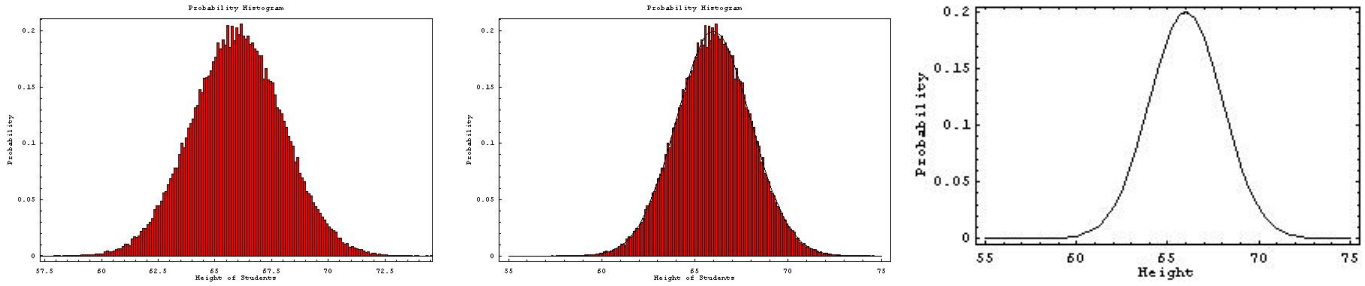
Example of Continuous Distribution Consider the distribution mentioned earlier, the height of students at colleges in the USA. Say we sample the population by measuring the height (in inches) of 100 students at UMM, and we get the following histogram (measurements done by a computer simulation).



The heights of the bars represent the percentage of students that we measured in the range given at the bottom of the bar.

In the graph in the middle I have drawn a smooth curve, guided by the heights of the bars. Since the height can be any number, this curve is a more appropriate way to describe the distribution than the histogram itself.

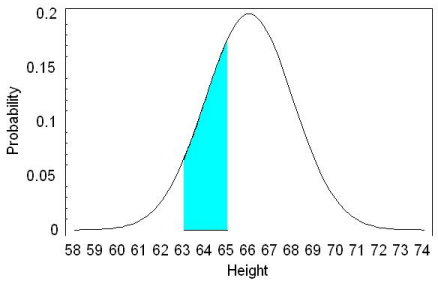
If we could measure more students, we would expect that the curve would better approximate the height of the bars. Also, if we measure more students we could make the width of the bars smaller. We should also measure the heights of students at colleges other than UMM. We might expect something like the following (I have measured 100,000 students by computer simulation):



Here are the important properties of the distribution using the curve:

- The total area under the curve is 1.
- The probability in any interval of outcomes is the area under the curve above that interval.

For example, the probability that a student will have height between 63 and 65 inches is given by the shaded area, which is 0.24:

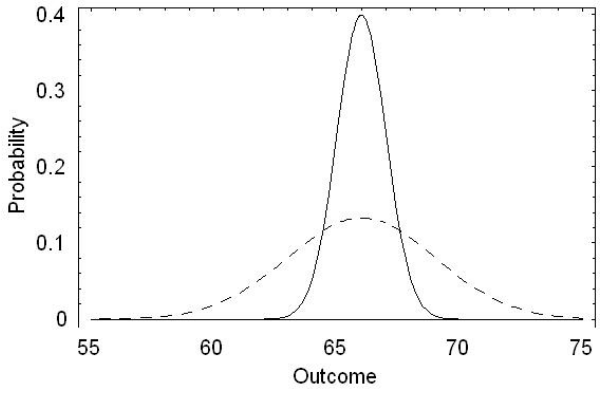


Normal Distribution

The distribution we saw above is a normal distribution (sometimes called the bell curve), which has the properties:

- The tails of the curve fall off rapidly.
- The distribution is symmetric.
- The mean lies at the center of symmetry.
- The mean is the same as the median (true for any symmetric distribution).
- Since it is symmetric, it is well described by the mean and standard deviation.
- In fact, a normal distribution is completely described by the mean and standard deviation.
- The spread is entirely measured by the standard deviation.
- Where the curve changes from curving down to curving up is one standard deviation away from the mean.
- It has no outliers.

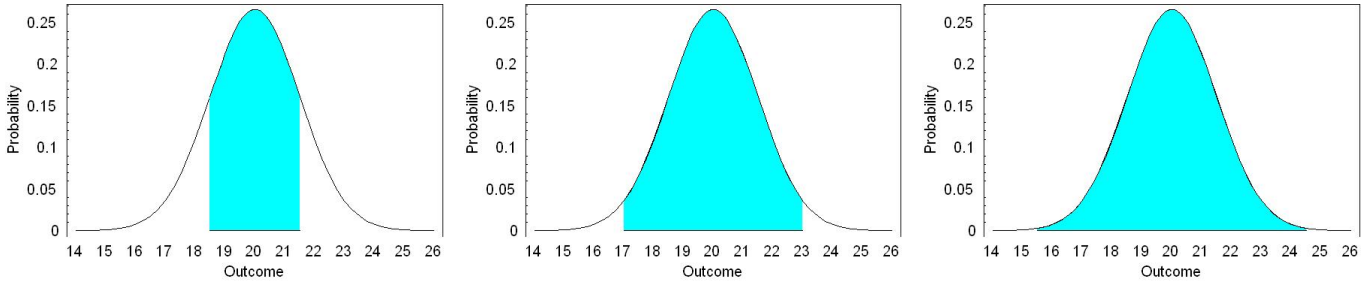
Here is a sketch of a couple of normal distributions. Notice that is we make the height smaller, the width must become greater, since the area under the distribution is always 1!



The 68-95-99.7 Rule

All normal distributions behave in certain regular ways. For example,

- the probability that a measurement falls within 1 standard deviation of the mean is 0.68.
- the probability that a measurement falls within 2 standard deviations of the mean is 0.95.
- the probability that a measurement falls within 3 standard deviations of the mean is 0.997.

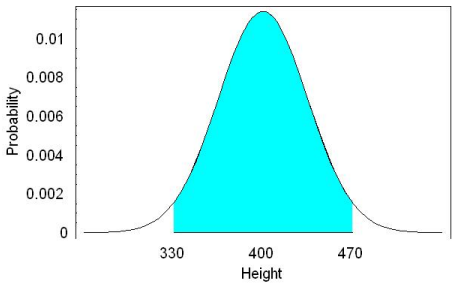


The above graphs show the 68-95-99.7 rule for a normal distribution with mean 20 and standard deviation of 1.5.

Example The distribution of the scores on a standardized exam is approximately normal with mean 400 and standard deviation 35. Between what two values do 95% of the scores lie? What percentage of students score above 435?

Since 95% of the values will lie within 2 standard deviations of the mean, 95% of the scores will lie between $400-2(35) = 330$ and $400 + 2(35) = 470$.

Pictorially, this looks like the following:

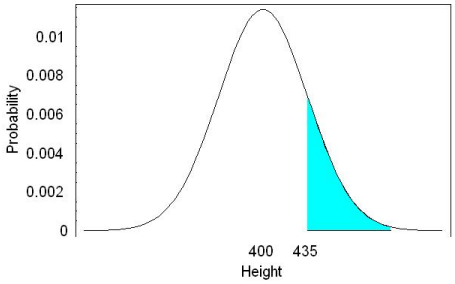


Since 435 is one standard deviation away from the mean, we would expect 68% to score between 365 and 435.

We would expect 100%-68% = 32% of the students to score below 365 or above 435

Since the distribution is symmetric, we would expect half of 32%, or 16%, of the students to score above 435 on the exam.

Pictorially, this looks like the following:



For Your Information: Areas under curves are closely related to the idea of an integral which is studied in calculus. To verify the 68-95-99.7 rule you would need to study calculus. You can also use the ideas of calculus to answer more general problems involving continuous probability, where the area under the curve you seek is not related to 1, 2, or 3 standard deviations away from the mean.

The Central Limit Theorem

The Central Limit Theorem states in part that a distribution on n random trials has a distribution that is approximately normal when n is large (this is what we see if we roll n dice and take n very large). This fact combined with the fact that a normal distribution is completely described by its mean and standard deviation is what leads to the use of the mean and standard deviation to describe so many distributions we see. People assume the distribution would be a normal distribution if they took enough measurements. But we must be careful–in the real world, not all distributions underlying a probability event are normal distributions! So the mean and standard deviation might not be the best way to describe the distribution you find in a particular situation. You really need to look at each on a case-by-case basis.