7.5 Great Expectations, 7.6 What the Average American Has

EXPECTED VALUE

If an experiment has n different numerical outcomes, $x_1, x_2, \ldots, x_n$ with probabilities $p_1, p_2, \ldots, p_n$, respectively, then the *expected value* of the experiment is

$$\sum_{i=1}^{n} p_i x_i = p_1 x_1 + p_2 x_2 + \cdots + p_n x_n.$$

Question. If a fair 6-sided die is rolled, what is the expected number to result?
Answer. The experiment has the following outcomes and probabilities:

x_i	p_i
1	1/6
2	1/6
3	1/6
4	1/6
5	1/6
6	1/6

So the expected value is

$$1 \times \frac{1}{6} + 2 \times \frac{1}{6} + 3 \times \frac{1}{6} + 4 \times \frac{1}{6} + 5 \times \frac{1}{6} + 6 \times \frac{1}{6} = 3.5.$$

Notice that the expected value is an average and not a possible outcome of any particular experiment.

Example. In the Virginia State Lottery game "Cash 5," you pick 5 numbers from a set of 34 and try to match the 5 numbers chosen at random by the state. As we saw in the previous lecture, the possible outcomes and probabilities are:

# matched	x_i	p_i
5	\$100,000	$1/278,256$
4	\$100	$145/278,256 \approx 1/1919$
3	\$5	$4060/278,256 \approx 1/69$
2	\$0	$36,540/278,256$
1	\$0	$118,755/278,256$
0	\$0	$118,755/278,256$

So the expected value of playing this game is

$$(\$100,000) \times \left(\frac{1}{278,256}\right) + (\$100) \times \left(\frac{145}{278,256}\right) + (\$5) \times \left(\frac{4060}{278,256}\right)$$

$$= \frac{\$100,000 + \$14,500 + \$20,300}{278,256} = \$0.48.$$

However, it costs \$1.00 to play, so you *expect* to loose 52 cents each time you play. These probabilities are from the Virginia State Lottery webpage (`www.valottery.com/cash5/howtoplay.asp`) and are typical of all state run lotteries (namely, that expected winnings are around -50 cents).

A game is a *fair game* if the expected winnings is 0. Cash 5 would be a fair game if it cost 48 cents to play.

Question. Suppose Cash 5 were to be made a fair game by increasing the "match 3 numbers" prize. What would it have to increase to?
Answer. \$40.33.

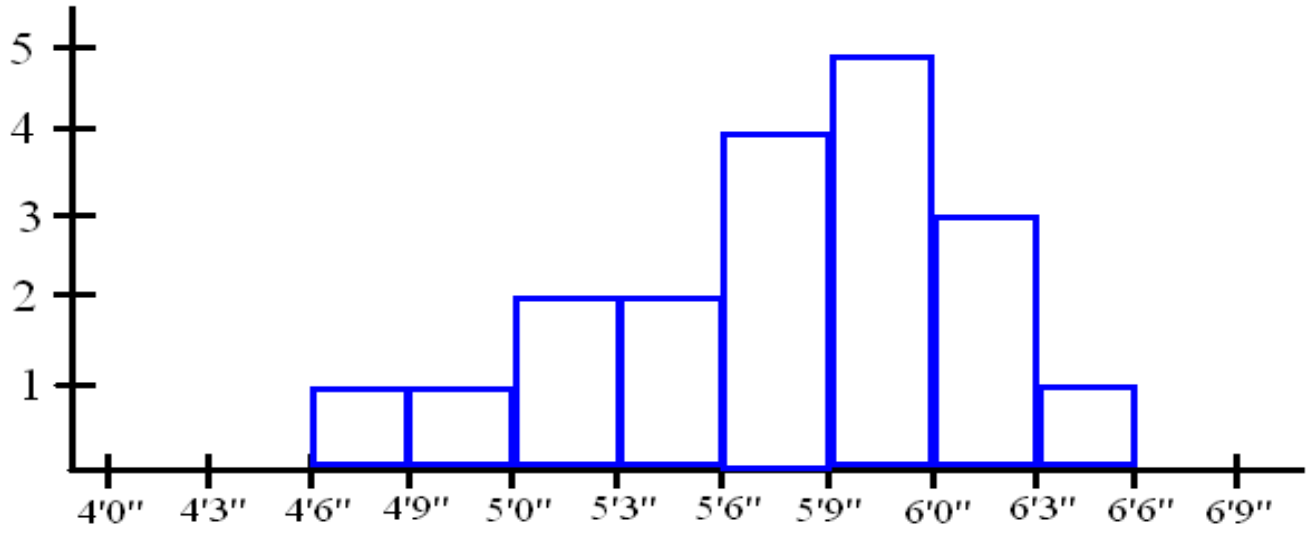
DISTRIBUTIONS

From the Law of Large Numbers, it follows that many quantitative things (such as weight, height, grades) have a distribution along a "bell curve." To find the distribution of a quantity, we might create a number of categories and produce a *histogram* of the data.

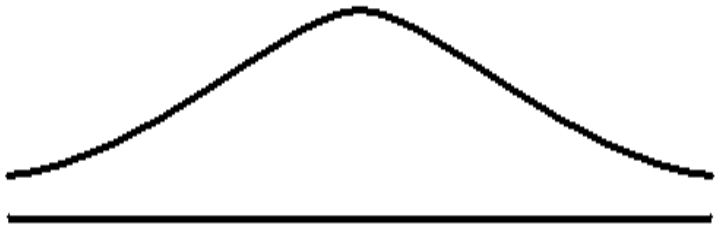
Example. Suppose we sample the height of 19 adult males and find:

height h	number
$4'0'' \leq h < 4'3''$	0
$4'3'' \leq h < 4'6''$	0
$4'6'' \leq h < 4'9''$	1
$4'9'' \leq h < 5'0''$	1
$5'0'' \leq h < 5'3''$	2
$5'3'' \leq h < 5'6''$	2
$5'6'' \leq h < 5'9''$	4
$5'9'' \leq h < 6'0''$	5
$6'0'' \leq h < 6'3''$	3
$6'3'' \leq h < 6'6''$	1
$6'6'' \leq h < 6'9''$	0

We make a histogram (or a "bar graph") by reflecting the number of individuals in each category as a function of the category.



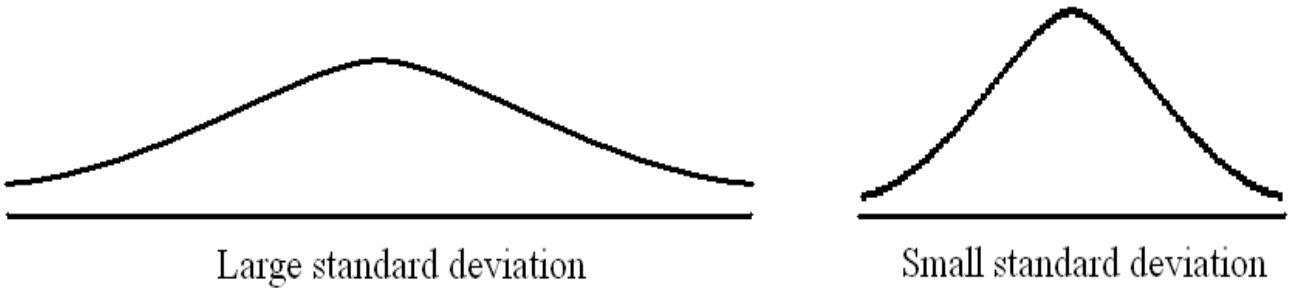
If we sample a huge number of individuals from the same population (such as adult American males), then we would find that the distribution approaches a "bell curve," or a *normal distribution*:



A normal distribution

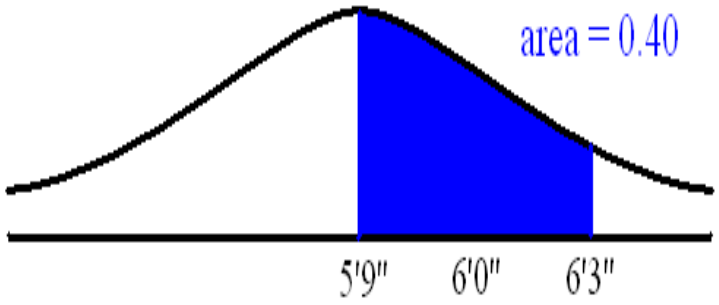
The Law of Large Numbers insures that a quantity which depends on many factors (height depends on several genetic factors, nutrition, etc.) will be approximated by a normal distribution. The function for a normal distribution is $f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\{(x-\mu)/\sigma\}^2/2}$ where μ is the *mean* (average) of the distribution and σ is the *standard deviation*. The standard deviation reflects how "spread out"

a distribution is:



In fact, the total area under a normal distribution is 1 and the normal distribution is an example of a *probability distribution*. In general, areas under the normal distribution represents probabilities.

Example. Here is a hypothetical normal distribution of height:



The area under the curve $y = f(x)$ for x between $5'9''$ and $6'0''$ is 0.40, so in this population, 40% of the individuals are in this height interval.

Question. In the above distribution, the mean is $5'9''$. What percentage of the population is less than $5'9''$? What percentage of the population is greater than $6'3''$? According to the distribution, what percentage of the population is exactly $5'9''$?

WHAT IS STATISTICS?

Statistics, in the broadest sense, deals with analyzing observed data quantitatively. The most common application we see day-to-day is the survey. Often we see (especially during election years) polls indicating opinions between two options (Bush vs. Kerry, pro-choice vs. pro-life, opinions on the economy improving/worsening). These polls (when done correctly) always come with a margin of error. The polls are performed as follows. First, a *random sample* of the population under study (such as likely voters) must be chosen. This, in practice, can be quite difficult and expensive. But, without a random sample, reliable conclusions cannot be drawn. Next, parameters

are calculated for the sample (such as a percentage that think the economy will improve in the next 6 months). Finally, a *level of confidence* is calculated for the sample results. This confidence depends primarily on sample size. If the sample is small, there is little confidence in the sample's reflection of the population as a whole. If the sample is large, then there is high confidence in the sample's parameters. This confidence reflects the probability that the sample is an accurate reflection of the population.

More generally, a *hypothesis test* deals with forming a null hypothesis, H_0. Data is then collected and the probability that we see the data which was collected under the assumption of the null hypothesis is calculated. THIS is where conditional probability (Bayesian probability) is used in statistics. The goal, surprisingly, is to reject H_0. If we find that, under the assumption H_0, the probability of us gathering the data we actually did gather is 50%, then we have learned nothing. If we find the probability is 99%, then we still do not know that our hypothesis is likely (after all, there could be other explanations for the trend we see in the data). But if we find that under the assumption of H_0, our data is unlikely to be observed (say 1% probability), then we can **reject the null hypothesis**.

Example. (From *An Introduction to Statistical Methods and Data Analysis* by L. Ott, 1988.) Suppose you speculate that the mean yield per acre of soybeans in a particular region is at least 520 bushels. You take 36 samples and find the average for these samples is 573 bushels/acre. How do you perform the statistical test?

Answer. You hypothesize that the population mean is less than 520 bushels/acre (remember that you want to reject this). Then you calculate the probability of getting a sample of size 36 with a mean of 573 bushels/acre (this requires knowledge of the standard deviation of the population — or at least of the sample). We find (using Ott's numbers) that the probability of these samples under the hypothesis H_0 is 0.3%. Hence we find the null hypothesis to be highly unlikely and reject it with great confidence (at a confidence level of $100\% - 0.3\% = 99.7\%$).

MEAN, MEDIAN, AND MODE

Suppose we have a list of numerical data $x_1, x_2, \ldots, x_n$. The *mean* (average) is the sum of the data divided by the number of data points:

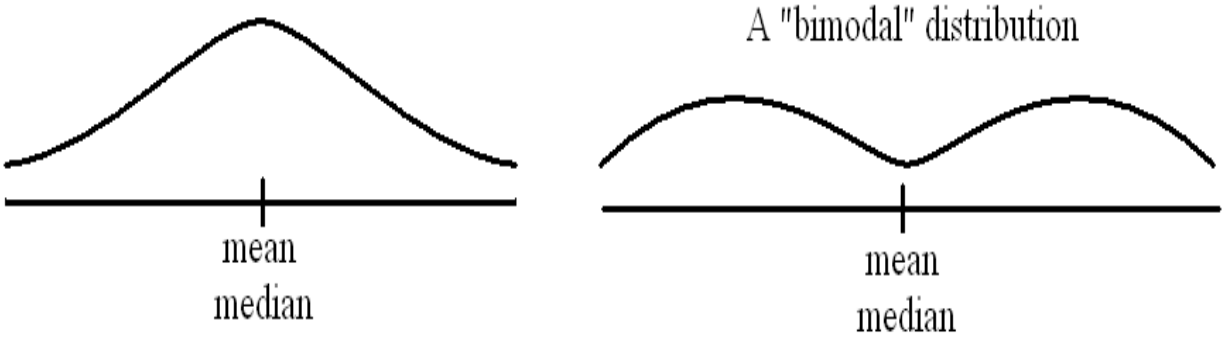
$$\text{mean} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n}.$$

The *mode* is the number that occurs most often in the list (and may not be unique). The *median* is the number in the middle when the data is listed from smallest to largest value (or if n is even, the average of the two numbers nearest the center).

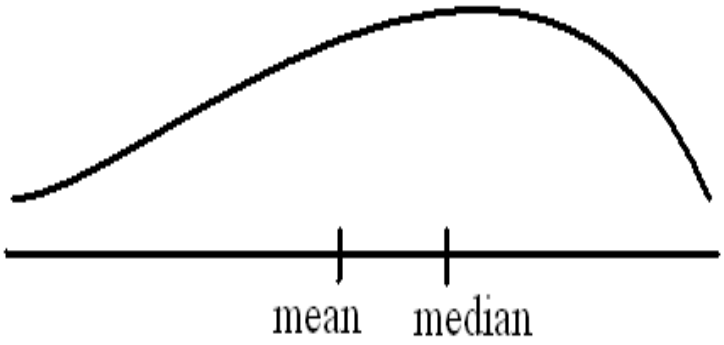
Example. Consider 11, 17, 18, 10, 22, 23, 15, 17, 14, 13, 10, 12, 18, 18, 11, 14. What are the mean, median, and mode?

Answer. Mean is 15.2, median is 14.5, mode is 18.

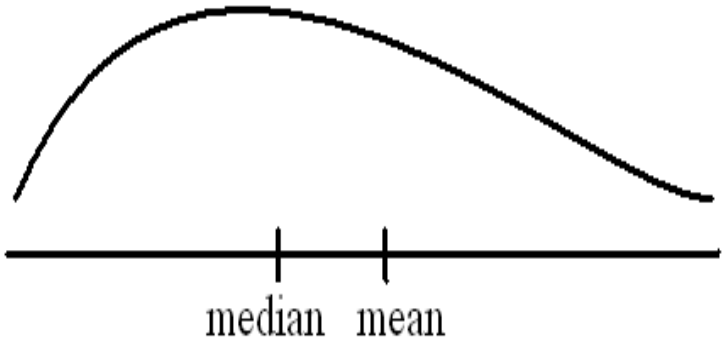
If we compare the mean to the median, then we can learn something about the data distribution. If the distribution is symmetric, then the mean and median are the same:



If the mean is less than the median, then the distribution is skewed to the left:



If the mean is greater than the median, then the distribution is skewed to the right:



Notice that we have a bit of data rather far from the mean when we considered skewed distributions. For example, if we plot average income in the U.S., we would see data skewed to the right since there are a few individuals making a large amount of money (for example, Bill Gates is an "outlier" since his income is so much greater than the rest of the population).