EM375 STATISTICS AND MEASUREMENT UNCERTAINTY
INTRODUCTION TO THE STATISTICS MODULE

Statistics will probably be the single topic taught in your USNA engineering program that you will use most often after graduation. The methods presented are directly applicable to many engineering subjects. But they are equally useful in a wide range of technical and non-technical areas.

In this module on statistics we do not cover all the topics that are in complete statistics text books, or are used in the real world of applied statistics. If you are looking at these notes and you are not studying the EM375 course at the United States Naval Academy, they a) may or may not have the content you need, and b) may or may not focus on your applications. *Caveat Emptor*. The topics we cover include:

Histograms, terminology, central tendency and dispersion

Concepts of probability and probability functions

Binomial distribution (for yes/no situations)

Poisson distribution (for queuing)

Normal distribution

Design of sampling experiments

Elimination of bad data

Tests between two samples

Correlations of data

Linear and nonlinear regression

For this course we concentrate on problems associated with measurement uncertainty.

The transducer we use to measure something will give a slightly different output when used for repeated measurements of the same object. For example, if we use GPS to locate a fixed position on the earth, the observed latitude and longitude will vary slightly with time.

Not all ‘things’ have the same measurements. For example, if you buy a box of 2-inch screws they are not all exactly 2 inches long. Indeed, you will find out later in the course that *none* of the screws is *exactly* 2 inches long!

Later in the EM375 course we will see that measurement errors are typically classified as either bias errors or as precision errors. Bias errors are consistent and repeatable. They cannot be analyzed using statistics. Precision errors, though, are generally random, and are amenable to a statistical analysis.

Let us consider an experiment where we measure the speed of an aircraft. While the pilot may try to keep his speed constant, external influences and measurement errors mean the recorded speed will change. The following table shows a set of 60 observations of airspeed.

Airspeed (knots)	Number of observations
434	1
440	1
444	4
446	5
448	6
450	10
452	10
454	8
456	7
458	2
460	5
466	1

We can learn a lot of information directly from the table. For example, during the period we measured the airspeed it never dropped below 434 knots, and never exceeded 466 knots. It also looks like an airspeed of 450 or 452 knots is “most likely” to be measured.

We can often better visualize the data by plotting it in a graph. The first part of the data reduction is to arrange the data into groups. We call each group a ***bin***. The following table shows the same data grouped into 9 bins:

Bin (knots)	Number of observations
$432 \le v < 436$	1
$436 \le v < 440$	0
$440 \le v < 444$	1
$444 \le v < 448$	9
$448 \le v < 452$	16
$452 \le v < 456$	18
$456 \le v < 460$	9
$460 \le v < 464$	5
$464 \le v < 468$	1

When organizing data into bins, note the following:

Each bin has the same 'width' (in this case, range of airspeed). Even though it is not essential to have all the bins the same size, it simplifies the later analysis.

It is customary to have somewhere between 5 and 15 bins covering the range of the data. Special analysis may require fewer bins. There are some situations that require a very large number of bins. Unless instructed otherwise, you should stay with the 5-to-15 rule for this course.

The bins must cover the entire range of the data. Every single datum must go into a bin.

There must be no overlap of the bin ranges. Each datum must go in one, and only one, bin.

A ***continuous random variable*** can take any value in the data range. For example, someone's height could be 69 inches. Or 69.1 inches. Or 69.123456789 inches. Bins for continuous random variables must make sure each datum gets assigned to just one bin.

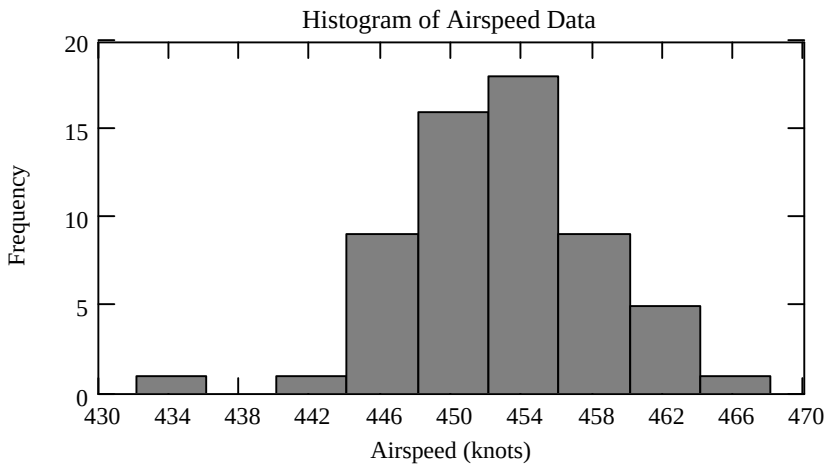
A ***discrete random variable*** can only take set values. For example, when throwing a die (that's one dice!) the only possible outcomes are 1, 2, 3, 4, 5 or 6. You can't throw a 3.25! Similarly, if you are using a digital thermometer that only shows integer temperatures, you can never measure 71.5^{O}F. The *actual* temperature may be a continuous random variable, and it might be 71.5^{O}F. But after you have measured it, the temperature becomes a discrete random variable and it will be either 71°F or 72°F.

Bins for discrete random variables may have their boundaries set half-way between the possible outcomes. For the discrete thermometer temperature problem, either of the following bin sets would produce the same result:

$69 \le T < 70$		$68.5 < T < 69.5$
$70 \le T < 71$	or	$69.5 < T < 70.5$
$71 \le T < 72$		$70.5 < T < 71.5$
etc		*etc*

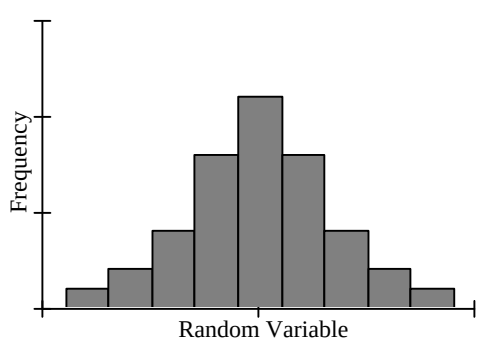
Empty bins are acceptable and quite normal.

Once the data have been organized into bins, one of the most common ways of showing the data is in a ***histogram***. The histogram for the airspeed data looks like this:

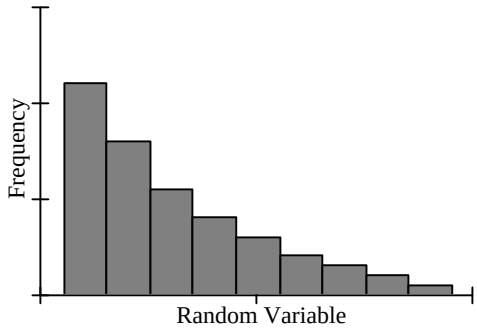


This particular histogram has a single ***mode*** or "peak". The distribution looks reasonably symmetric about the highest peak, but the distribution is not exactly symmetric.

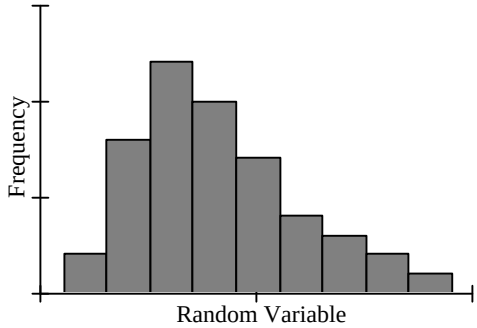
In engineering applications we see several different distributions. The following figures show some of the more common distribution shapes found in engineering problems.



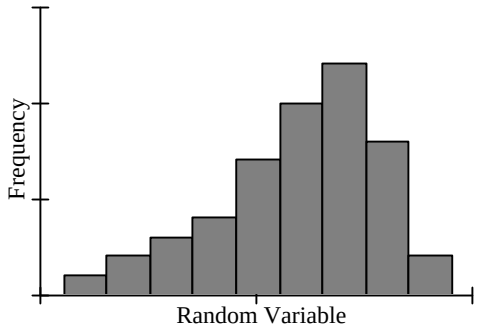
The Normal, Gaussian, or Bell-shaped distribution is one of the most often seen distributions of random variables. At the extremes, the frequency of occurrence for very large or very small values of the random variable is small, but not zero.



This distribution is J-shaped. Distributions similar to this are often used for failure analysis. For example, medical research[1] found that the relationship between the blood pressure level achieved through antihypertensive treatment and the incidence of coronary heart disease showed a J-shaped distribution in relation to achieved treated systolic and diastolic blood pressure levels.

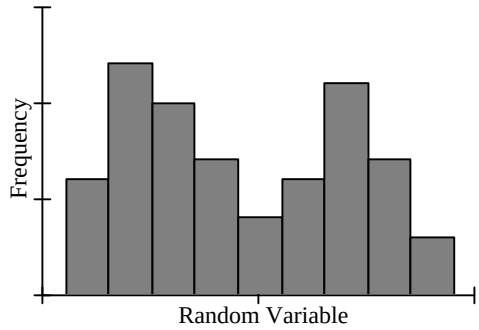


This distribution is right-skewed. The term “right” refers to where the long tail of the distribution is in comparison to the mode. Distributions of this type can be used for some failure analyses. Also, the distribution is often representative of wage and salary distributions (more low-paid employees).

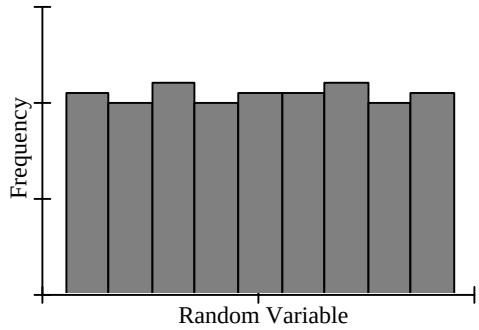


This distribution is left-skewed since the long tail is to the left of the mode. Student grades in class often follow this distribution. We just give too many high grades!

[1]“The J-shaped relationship between coronary heart disease and achieved blood pressure level in treated hypertension: further analyses of 12 years of follow-up of treated hypertensives in the Primary Prevention Trial in Gothenburg, Sweden,” Samuelsson OG, Wilhelmsen LW, Pennert KM, Wedel H, Berglund GL., J Hypertens. 1990 Jun;8(6):547-55



This is a bimodal distribution, most commonly seen when we sample a population that is really a mixture of two different populations. For example, if we measure the percentage of carbon in steel, we may get a bimodal distribution if the steel samples were manufactured in two different foundries. You might also see this distribution in class grades if the class combines a mixture of 'interested' students as well as those forced to take the course.



A perfect uniform distribution has all the bars at exactly the same height. This distribution is widely used in gaming, and is used to introduce variability (randomness) into the way a computer plays a game.

SOME STATISTICAL DEFINITIONS

Statistics has a language of its own. Some "common language" words that are also used in statistics can have very precisely defined meanings. The following are some of the definitions you are expected to know.

Population: The *population* is the entire set of possible objects, measurements, etc. In engineering measurements we rarely know the entire population.

Parameter: A *parameter* is a numerical attribute of the population. For example, the mean (or average value) of a population is a parameter.

Sample: A *sample* is a representative subset of the population. A sample can be a single measurement, or a number of measurements.

Statistic: A *statistic* is a numerical attribute of the sample. If we wish to determine (for example) the average of a population, we do this by taking a sample set of measurements. We calculate the average of the sample, and this is a statistic. The sample statistic is an estimate of the population parameter.

Sample space: All possible outcomes from an experiment.

Random variable: When repeated measurements or experiments give different answers, the measurements are considered to be a *random variable*. Random variables can be *discrete* or *continuous.*

Distribution function: A *distribution function* is a mathematical (equation) or graphical representation of a random variable.

Event: The outcome of a random experiment is called an *event.*

CENTRAL TENDENCY

The term ***central tendency*** means the tendency of random variables to cluster around a "central" value. For example, if we measure the temperature at various locations in a room, there will be a variety of readings. These readings, though, will tend to group around a value. Obviously there isn't a single value that exactly gives the temperature of the room, but we commonly refer to this central value as "the room's temperature".

There are several different measures of central tendency. For this course we use the most common parameter, the ***mean.*** The mean is the same as the everyday concept of average. The sample mean is defined as:

<u>SAMPLE MEAN:</u>

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \cdots\cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

<u>EXAMPLE:</u>
The diameters of five bolts were measured as follows:

0.253, 0.247, 0.246, 0.251, 0.248 inches

The average diameter is:

$$\bar{x} = \frac{0.253 + 0.247 + 0.246 + 0.251 + 0.248}{5} = 0.2490 \text{ inches}$$

Notes:

In the equation, x_1, x_2 etc refer to the 1st, 2nd, etc values in the sample set.

Thus, for the above sample, the symbol x_1 and the value 0.253 mean the same thing.

POPULATION MEAN:
Some populations have a finite number of elements (for example, the score from throwing two dice). If the number of elements in the population is N, the population mean can be calculated as:

$$m = \frac{x_1 + x_2 + x_3 + \cdots\cdots + x_n}{N} = \frac{\sum_{i=1}^{N} x_i}{N}$$

Note the use of different symbols for the population and sample means. We use $\bar{x}$ (x-bar) for the sample mean, and m (Greek symbol mu) for the population mean.

MEDIAN
The median is the "middle" value of a set if the data are sorted in increasing order. If there is an even number of data values, the median is the average of the two central values.

For the bolt data, we first put the measurements in increasing order:

0.246, 0.247, 0.248, 0.251, 0.253

There are 5 values, so the median is the middle, or 3rd, value: 0.248 inches

MODE
The mode of a data set is the one "most likely" to happen. Thus, if you plot a histogram, the mode is represented by the tallest bar.

MEASURE OF DISPERSION

As we described earlier, repeated measurements will give different results. The measurements can be widely spaced, or they may be closely spaced. For example, if we take three readings of room temperature, we could get either of the following data sets:

Set 1:	69, 73, 71 ^{O}F
Set 2:	62, 72, 79 ^{O}F

Both sets have the same average (71 ^{O}F), but Set 1 has measurements closer together than Set 2. In statistics we have several measures of the *dispersion* (or spread or variability) of the data.

The ***deviation*** of an individual measurement is defined as:

$$d_i = x_i - \bar{x}$$

For data set 1 above, the deviations are:

Data value (^{O}F)	Deviation $d_i = x_i - \bar{x}$
69	-2 ^{O}F
73	+2 ^{O}F
71	0 ^{O}F

If we add up (sum) all the deviations, we get zero – *always!* Here's why.

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

$$\text{now} \quad d_i = x_i - \bar{x}$$

$$\text{so} \quad \sum_{i=1}^{n} d_i = \sum_{i=1}^{n} \left(x_i - \bar{x} \right)$$

$$= \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \bar{x}$$

$$= \frac{n}{n} \sum_{i=1}^{n} x_i - n\bar{x}$$

$$= n\bar{x} - n\bar{x} = zero!$$

So if we calculated the mean of the deviations, we would get zero. To eliminate this problem, we calculate the mean absolute value. This ***mean deviation*** is defined as:

$$\bar{d} = \frac{\sum_{i=1}^{n} |d_i|}{n}$$

For data set 1, the mean deviation is:

$$\bar{d} = \frac{\sum_{i=1}^{n} |d_i|}{n} = \frac{|-2| + |+2| + |0|}{3} = \frac{4}{3} \approx 1.333 \ ^{O}F$$

And for data set 2 the mean deviation is:

$$\bar{d} = \frac{\sum_{i=1}^{n} |d_i|}{n} = \frac{|62 - 71| + |72 - 71| + |79 - 71|}{3} = \frac{18}{3} = 6 \ ^{O}F$$

These two deviations clearly show that data set 2 has a higher deviation (or spread, or variability) than data set 1.

However, in the field of statistics rather than using the deviation defined above, it is more common to use a standardized deviation. For a population, the ***population standard deviation*** is defined as:

$$s = \sqrt{\sum_{i=1}^{N} \frac{(x_i - m)^2}{N}}$$

More often, we do not know the entire population, and only have access to a sample. The sample standard deviation calculation depends on the number of ***degrees of freedom*** in the data set. The number of degrees of freedom is the number of independent measurements minus the minimum number of measurements theoretically necessary to estimate a statistical parameter. For the standard deviation calculation where there are *n* measurements, the number of degrees of freedom[2] is (*n*-1). The ***sample standard deviation*** is now defined as:

$$S = \sqrt{\sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{(n-1)}}$$

We often use another measure of dispersion, the ***variance***. Variance does not have a separate symbol. Variance and standard deviation are directly related.

Population variance = s^2

Sample variance = S^2

Here is a summary of the statistics for data sets 1 and 2:

Statistic	Data Set 1	Data Set 2
Mean, $\bar{x}$	71 ^{O}F	71 ^{O}F
Mean deviation, $\bar{d}$	1.333 ^{O}F	6 ^{O}F
Standard Deviation, S	2 ^{O}F	8.54 ^{O}F
Variance, S^2	4 (^{O}F)^2	73 (^{O}F)^2

Question: Why does the table show *statistics* and not *parameters*?

[2] Why (n-1)? The reason is based on the fact that we use the sample standard deviation to obtain an estimate of the population standard deviation. Put in simpler terms, the standard deviation is derived from the mean and from the deviations of each measurement from the mean. If we lacked any one of these measurements (the mean or a single deviation value) we could calculate it from the other information. So, with *n* measurements only (*n*-1) of them are free to vary since we could calculate the missing one. (*n*-1) is therefore the number of degrees of freedom of the data set.

Summary:

POPULATION: Population size is N

Average value = $m = \dfrac{\sum_{i=1}^{N} x_i}{N}$ MathCAD function: *mean()*

Variance = $s^2 = \sum_{i=1}^{N} \dfrac{(x_i - m)^2}{N}$ MathCAD function: *var()*

Standard Deviation = $s = \sqrt{\sum_{i=1}^{N} \dfrac{(x_i - m)^2}{N}}$ MathCAD function *stdev()*

SAMPLE: Sample size is n

Average value = $\bar{x} = \dfrac{\sum_{i=1}^{n} x_i}{n}$ MathCAD function: *mean()*

Variance = $S^2 = \sum_{i=1}^{n} \dfrac{(x_i - \bar{x})^2}{(n-1)}$ MathCAD function: *Var()*

Standard Deviation = $S = \sqrt{\sum_{i=1}^{n} \dfrac{(x_i - \bar{x})^2}{(n-1)}}$ MathCAD function *Stdev()*

Remember that, by default, MathCAD uses subscript 0 to (n-1). Also remember that Stdev and stdev are different MathCAD functions!

PROBLEMS

Intro-1:
A novice was asked to cut pieces of wood into 4-foot lengths. The following are the actual lengths in inches:

48.3, 49.1, 47.9, 48.2, 48.3, 49.5, 48.9, 48.2, 48.8, 49.2

a) Arrange the data into bins with a width of 0.2-inches. The first bin should have a lower bound of 47.0-inches.

b) Plot a histogram

Intro-2:
Calculate the following for the data in problem Intro-1:

Mean, median, mode, mean deviation, variance, and standard deviation