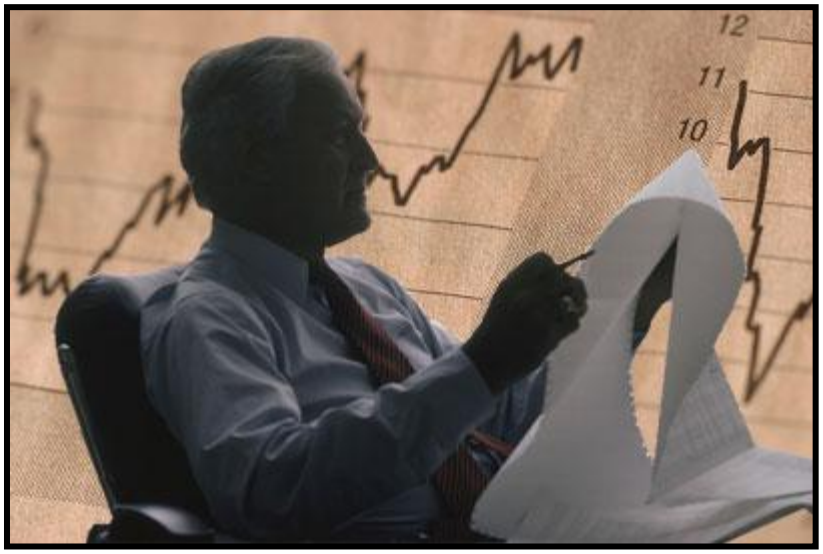
Data Analysis



BCF106

Fundamentals of Cost Analysis

June 2009

Defense Acquisition University

Chapter 5

Data Analysis

Data Analysis

5.0 Introduction

How can I summarize the data I've collected, and what conclusions can I draw from it?

Our purpose in collecting data is to develop an understanding of what took place in the past so that we might better predict or forecast what will take place in the future. The previous chapter on inflation suggested that after we collect the data, we should adjust the data to a common economic year so that as we compare one value to another we have a more consistent comparison. We should also adjust or "normalize" the data so that it is consistent in content and so that the impact of quantity has been addressed as well. Having made these adjustments we are better able to make statements about, and draw conclusions from, the data.

These "statements about the data" are really nothing more than the questions you would have in planning to purchase something for yourself. What's the typical price? How much do the prices vary? What are the odds that you will be paying more than or less than a particular price? This information in itself may meet your needs, or you may find yourself needing to do more analysis.

Let's look at a cost estimating example. You're estimating the cost of computer support for your installation. You check with a number of similar installations and find that everyone is paying about the same price. In this case using the average price would probably be adequate.

But, what if on the other hand you saw a significant variation in the price of computer support from one installation to the next? You might need to re-examine the data to see if it was truly similar and to ensure that it had been properly normalized. It might lead you to consider the use of another estimating technique like regression, where we try to relate the variation in the prices with those things that drive computer support such as the number of users, the number of computers, the number of software applications on the servers, etc. Or perhaps you conclude that computer support varies so much from one location to another that using a single-point analogy (picking the installation most like yours) would be more useful.

Our discussion of data analysis will not only help us address the questions we have noted above, but will also provide us with a foundation for our discussions in later chapters on regression, learning curves, and risk analysis among others.

Our objectives, from a cost estimating perspective, will be to develop descriptive and inferential statistics from one variable data; or more specifically to:

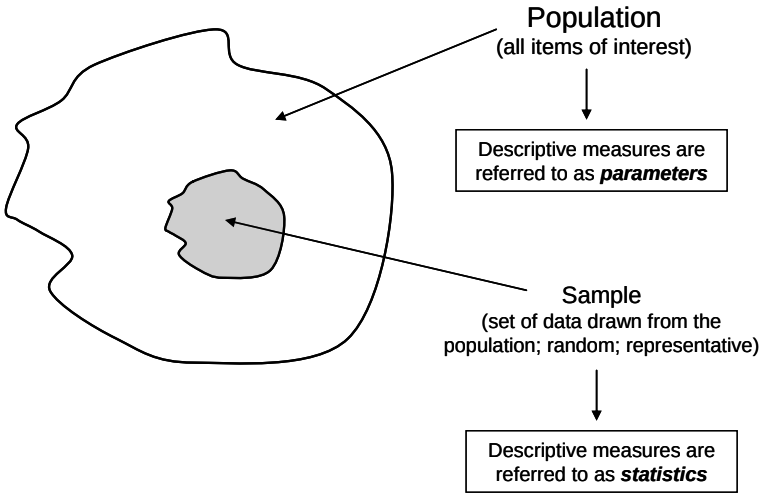
1. Define and calculate the measures of central tendency (i.e. the mean, median, and mode).
2. Define and calculate the measures of dispersion (i.e. the range, variance, standard deviation, and coefficient of variation).
3. Determine an area of probability under a normal distribution.
4. Calculate confidence intervals for both small and large sample sizes.
5. Perform one-tailed and two-tailed hypothesis tests.

5.1 Terminology

The general use of the word *statistics* involves the observation, recording, processing, and analyzing of data. The word *statistic* is used in this course as a number calculated from sample data. Statistics is sometimes broadly classified into two distinct areas known as descriptive statistics and inferential statistics. Descriptive statistics describe or summarize the data (e.g. on average it takes 65 hours to install the CFX modification kit). Inferential statistics are usually associated with using descriptive statistics in an attempt to make predictions or inferences about a given item (e.g. we are 90% confident that it will take between 60 and 70 hours to install the next CFX modification kit).

A *variable* is some characteristic of a product, service or activity; and is usually designated or named with a letter to make it more convenient to refer to in a formula. We could use X to represent the CFX modification install hours. If the first mod required 62 hours and the second required 67 hours we could write this as "$X_1 = 62$" and "$X_2 = 67$". More generically we could refer to each of these values as X_i or the i-th observation of X.

Populations and *samples* are basic terms in statistics. Populations can be *finite* (e.g. there were 82 CFX mod kits installed) or populations can be *infinite* (e.g. while we can refer to the hours required for each of the 82 mod kits that were installed, these hours only represent what did happen, not all of the things that could have happened). [We will leave more in-depth discussions of the concepts of a universe, a population, and a sample to other courses.]



If the average install hours for the population of 82 kits were 67 hours, the 67 hours would be referred to as a population *parameter*. If we took a sample of 10 kits from the 82 kits installed and the average was 65 hours, then we would refer to the 65 hours as a sample *statistic*. Unfortunately, it is nearly always too expensive or in some cases impossible to examine the entire population and compute the descriptive parameters. Therefore, samples are taken.

A valid sample has the following characteristics:

- First, the sample should be a random sample. This means that every member of the population should have an equal chance of being selected for the sample. This reduces the possibility of getting a *biased* sample.
- Secondly, the sample should be representative of what the population contains. A non-representative sample will obviously yield a distorted picture of the population (e.g. the 10 kits were installed by trainees as part of maintenance training).

5.2 Measures of Central Tendency

The base commander is considering the construction of a new base auditorium and has asked you what the "typical" cost is for an auditorium. You contact a number of military installations which have constructed auditoriums in the last five years and come up with the following costs (shown in Table 5.1) which you have normalized to constant year (CY) dollars in millions.

Base Auditorium Construction Cost (CY$M)				
4.66	3.44	2.77	3.85	4.15
2.75	2.71	4.25	3.60	3.26
3.68	3.26	2.31	2.15	4.75
4.21	4.98	5.70	5.92	3.65
4.58	3.11	3.37	4.55	3.26

Table 5.1

Now, for purposes of discussion, let's assume that these 25 observations or data points represent the relevant *population* of base auditoriums. Three *measures of central tendency* that might be used to describe the "typical" cost are the mean, the median, and the mode.

a. The **mean** or average, is the best known and most commonly used measure of central tendency.

The formula for the *population* mean is

$$\mu = \frac{\sum_{i=1}^{N} X_i}{N} = \frac{X_1 + X_2 + X_3 + ... + X_N}{N}$$

where X_i represents the various members of the population,

N is the number in the population,

$\sum$ (uppercase sigma) signifies summation (add all the X_i's), and

μ (mu, pronounced "mū") symbolizes the population mean.

Throughout the remainder of this lesson, we will use an abbreviated form of the summation formula, omitting variable subscripts and indexing on $\sum$ signs. In other words:

$$\mu = \frac{\sum X}{N} \text{ is understood to mean } \mu = \frac{\sum_{i=1}^{N} X_i}{N} = \frac{X_1 + X_2 + X_3 + ... + X_N}{N}$$

$$\mu = \frac{4.66+2.75+3.68+4.21+ ... +3.26}{25} = \frac{94.92}{25} = 3.7968 \text{ (3.80 rounded)}$$

So, the average or mean cost of an auditorium is $3.80M.

b. The **median** is the middle value when you arrange the data in either ascending or descending order. If the population size (N) is an odd number, the median is simply the middle value. If N is an even number, the median is defined as the average of the two middle values. Since it only considers the middle values, the median is not affected by extreme values (e.g. in the example on the right, whether the highest value is 5.92 or whether the value was 59.20, it will not impact the median).

5.92
5.70
4.98
4.75
4.66
4.58
4.55
4.25
4.21
4.15
3.85
3.68
3.65
3.60
3.44
3.37
3.26
3.26
3.26
3.11
2.77
2.75
2.71
2.31
2.15

The ordered population data for the example appears to the right. Since there are 25 observations included in the population, the median is determined by the middle value, which in this case is the 13^{th} observation of \$3.65M. Half of the auditoriums cost more than \$3.65M and half of the auditoriums cost less than \$3.65M.

c. The **mode** is the value that occurs most frequently in a data set. There can be more than one mode for a given set of data or no mode at all.

Referring to the ordered data on the right, we would determine the mode to be \$3.26M since this value appeared three times, more than any other value.

So, how would you answer the question as to the "typical" cost for an auditorium? The mean is \$3.80M, the median is \$3.65M, and the mode is \$3.26M.

We could say that the most common cost is \$3.26M (the mode), but that would seem somewhat misleading since only three of the twenty-five auditoriums cost that amount and since the mode seems to occur more in the lower half of the data rather than in the middle of the data.

Given that the mean and median are fairly close together, it doesn't appear that we have any "extreme" values affecting the average (mean) cost. This, along with the general use of the "average" by people, would probably lead us to use the mean cost of \$3.80M as a representative cost for an auditorium. Notice, however, that none of the auditoriums actually cost \$3.80M.

Using Sample rather than Population Data

5.70
4.66
4.58
3.60
3.44
3.37
3.26
3.11
2.77
2.31

The 10 data points shown represent a randomly drawn sample from our population of 25 auditoriums. How would we determine the mean, median, and mode?

For the sample, the mean is defined as "X-bar": $\overline{X} = \frac{\sum X}{n} = \frac{36.80}{10} = 3.68$

Notice in this case that 7 of the 10 auditoriums actually cost less than the mean.

The ordered data on the right has an even number of data points so we will determine the median by averaging the middle two data points: $\frac{3.44 + 3.37}{2} = \frac{6.81}{2} = 3.405$ or 3.41

There is no mode for the sample since each number occurs only once.

Our estimate would either be the \$3.68M (mean) or \$3.41M (median).

5.3 Measures of Dispersion

Let's return now to our base commander. Using the *population* data, we report that the average cost or price of an auditorium is $3.80M. The base commander responds by asking if most installations pay right around $3.80M or if there has been a lot of variability in the costs. What are some of the ways that we could describe the amount of variability in the costs?

Measures of dispersion give us an indication as to whether the data is tightly grouped or more widely spread around the center of the data. These measures are used with measures of central tendency to better describe the data. The measures we will be considering are the range, variance, standard deviation, and the coefficient of variation. Additionally, we will look at frequency distributions for a graphical depiction of the data.

a. **Range**. The best known and easiest to calculate measure of dispersion is the **range**. The range is defined as the highest value minus the lowest value.

(1) For population data the range is 5.92 – 2.15 = 3.77
(2) Or, alternatively, we could express the range as [2.15, 5.92]

Putting this in words we could say that there is a range in the costs of $3.77M, or that the auditorium costs range from $2.15M to $5.92M.

b. **Variance**. The range is a useful measure, but it simply indicates the distance from the lowest to highest value; it does not give us an indication as to how the data is grouped around the population mean. You can see that while the range is identical in Figures 5.1 and 5.2, the variability in the two is very different.

Low Variability

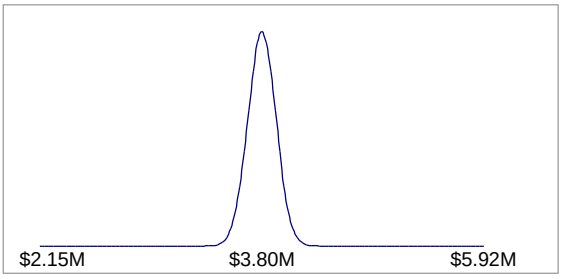


Figure 5.1

High Variability

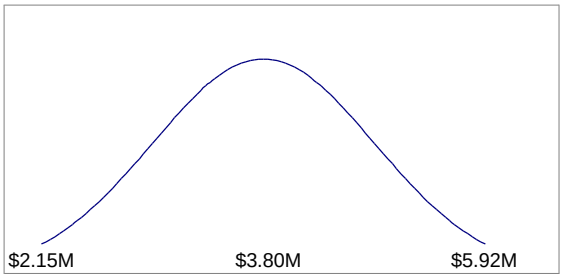


Figure 5.2

We need a measure that indicates the average distance that a data point falls from the middle of the data. In other words, on average do the auditoriums cost right around the average or mean cost (Figure 5.1), or is there a lot of variability in the cost of an auditorium (Figure 5.2)?

The **variance** is a measure of how far the data points fall away from the mean. It directly measures the distance that each X value is from the mean, "μ" in the case of the population.

The formula is: $\sigma^2 = \frac{\sum(X - \mu)^2}{N}$ (σ^2 is lowercase "sigma squared")

Variance Calculations

X_i	μ	$X_i - \mu$	$(X_i - \mu)^2$
4.66	3.7968	0.8632	0.75
2.75	3.7968	-1.0468	1.10
3.68	3.7968	-0.1168	0.01
4.21	3.7968	0.4132	0.17
4.58	3.7968	0.7832	0.61
3.44	3.7968	-0.3568	0.13
2.71	3.7968	-1.0868	1.18
3.26	3.7968	-0.5368	0.29
4.98	3.7968	1.1832	1.40
3.11	3.7968	-0.6868	0.47
2.77	3.7968	-1.0268	1.05
4.25	3.7968	0.4532	0.21
2.31	3.7968	-1.4868	2.21
5.70	3.7968	1.9032	3.62
3.37	3.7968	-0.4268	0.18
3.85	3.7968	0.0532	0.00
3.60	3.7968	-0.1968	0.04
2.15	3.7968	-1.6468	2.71
5.92	3.7968	2.1232	4.51
4.55	3.7968	0.7532	0.57
4.15	3.7968	0.3532	0.12
3.26	3.7968	-0.5368	0.29
4.75	3.7968	0.9532	0.91
3.65	3.7968	-0.1468	0.02
3.26	3.7968	-0.5368	0.29
		0.0000	22.84

Table 5.2

If we wanted to know the average distance that the X values lie from "μ", one approach would be to sum the 25 distances $(X_i - \mu)$ and divide by 25. However, the reason the mean of 3.80 was carried to four decimal places (3.7968) was to illustrate the problem with this approach. The $(X_i - \mu)$ values sum to zero. One solution is to square the values $(X_i - \mu)$ which results in a column of all positive numbers.

The resulting calculations are:

$$\sigma^2 = \frac{\sum(X_i - \mu)^2}{N}$$

$$\sigma^2 = \frac{22.84}{25}$$

$$\sigma^2 = .9136 \text{ or } .91$$

So how do we interpret the variance of .91? Well, the X values are \$M, therefore the mean (μ) is in terms of \$M, and the difference between the two $(X - \mu)$ is in \$M. We then squared the values and took the average by dividing by 25. We could say then that the variance is the average *squared* distance that the X values lie from the middle, or that the average variation in the costs is $\$.91M^2$. Not very intuitive is it?

c. **Standard Deviation**. Since we are interested in the *average* variation in the auditorium costs and not the *average squared* variation, we want to take the square root of the variance. We refer to the square root of the variance as "σ" (sigma), the standard deviation.

$$\sigma = \sqrt{\frac{\sum(X_i - \mu)^2}{N}} = \sqrt{\frac{22.84}{25}} = \sqrt{.9136} = .9558 \text{ or } .96$$

The result of this calculation is in \$M, so we can say that the average variation in the auditorium costs is \$.96M. We could tell the base commander that the average cost of an auditorium is \$3.80M and that the costs typically vary from that by plus or minus \$.96M. What does that imply? Consider this, in the column of $(X - \mu)$ values above:

- if we had budgeted \$3.80M for the \$5.92M stadium, we would have been off by \$2.12M
- if we had budgeted \$3.80M for the \$3.85M stadium, we would have been off by \$ .05M

The standard deviation represents "on average" how much we would expect "to be off by". The \$.96M represents the average estimating error if we used the mean of \$3.80M as our estimate.

d. **Coefficient of Variation (CV)**. The standard deviation gives us a measure of dispersion or variability that is in the same units as our data (dollars, hours, etc.). It would also be useful to have a *relative* measure of dispersion to give us a sense of the size of the standard deviation. The CV is a ratio of the standard deviation (average error) to the mean (average value). For the auditorium data set it would be calculated:

$$CV = \frac{\sigma}{\mu} = \frac{.96}{3.80} = .2526 \text{ or } 25.26\%$$

We could say that if we used the mean or average cost of $3.80 as our budget or estimate, that we would typically or on average expect to be off by plus or minus 25% of the mean. A good question to ask at this point is, "Would you be willing to use $3.80M as your estimate, knowing that you are likely to be off by ± 25%?" Perhaps the $3.80M would be reasonable to use if you were doing a long range affordability assessment, while on the other hand, if you were programming funds for the actual construction of the auditorium you would feel the need for more confidence in your estimate. Keep in mind that estimating is somewhat subjective in nature, requiring judgment and an awareness of the purpose of the estimate.

Another benefit of the CV is that since it is a relative measure of dispersion it can be used to compare variability between data sets. Consider the following:

a) The average auditorium cost is $3.80M and the standard deviation is $.96M.
b) Let's say that the average parking lot cost for auditoriums is $125K with a standard deviation of $50K.

Is there greater variability is the cost of an auditorium, or an auditorium parking lot?

$$CV = \frac{\sigma}{\mu} = \frac{.96}{3.80} = .2526 \text{ or } 25.26\% \qquad CV = \frac{\sigma}{\mu} = \frac{50}{125} = .40 \text{ or } 40\%$$

While the auditorium costs typically vary by ± $.96M (or ± $960K) and the parking lot costs only vary by ± $50K, there is greater *relative* variation (as a percentage of the mean) in the parking lot costs (40%) than the auditorium costs (25%).

Using Sample rather than Population Data

How would we calculate the measures of dispersion for our *sample* that was drawn from the population of auditorium costs?

5.70
4.66
4.58
3.60
3.44
3.37
3.26
3.11
2.77
2.31

a. **Range**. The difference between the highest and lowest value can be represented:

(1) For the sample data as: 5.70 – 2.31 = 3.39
(2) Or, alternatively, we could express the range as [2.31, 5.70]

Notice that our sample range (3.39) is smaller than the population range (3.77) since our sample did not happen to include the endpoints in the population.

b. **Variance**. The population variance (the average squared variability) was calculated:

$$\sigma^2 = \frac{\sum (X_i - \mu)^2}{N}$$

The sample variance will be calculated:

$$s^2 = \frac{\sum (X_i - \bar{X})^2}{n-1}$$

$$s^2 = \frac{9.22}{10-1} = 1.02$$

Why did we divide by "n-1" as opposed to dividing by "n" as we did for the population variance?

Variance Calculations Using the Sample Mean

X_i	$\bar{X}$	$(X_i - \bar{X})$	$(X_i - \bar{X})^2$
5.70	3.68	2.02	4.08
4.66	3.68	0.98	0.96
4.58	3.68	0.90	0.81
3.60	3.68	-0.08	0.01
3.44	3.68	-0.24	0.06
3.37	3.68	-0.31	0.10
3.26	3.68	-0.42	0.18
3.11	3.68	-0.57	0.32
2.77	3.68	-0.91	0.83
2.31	3.68	-1.37	1.88
			9.22

Table 5.3

First, we need to keep in mind that the sample statistics are estimators of the population parameters, and we want them to be "unbiased" estimators.

In Table 5.3 you can see that the total squared distance that the X_i values lie from $\bar{X}$ is 9.22.

However, if we had used the population mean of 3.80 in these calculations, as shown in Table 5.4, the total squared distance would have been 9.36, a higher value (which will always be the case).

The sample mean ($\bar{X}$) minimizes the squared distances and results in a biased calculation of the population variance. To correct for that bias we divide the squared distances by "n-1" rather than dividing by "n".

Variance Calculations Using the Population Mean

X_i	μ	$(X_i - \mu)$	$(X_i - \mu)^2$
5.70	3.80	1.90	3.61
4.66	3.80	0.86	0.74
4.58	3.80	0.78	0.61
3.60	3.80	-0.20	0.04
3.44	3.80	-0.36	0.13
3.37	3.80	-0.43	0.18
3.26	3.80	-0.54	0.29
3.11	3.80	-0.69	0.48
2.77	3.80	-1.03	1.06
2.31	3.80	-1.49	2.22
			9.36

Table 5.4

The "n-1" is referred to as the *degrees of freedom*. A simple rule is that we will "lose" one degree of freedom for each population parameter estimated with a sample statistic. In the variance calculation we are using the sample mean (a sample statistic) as an estimate of the population mean (a population parameter).

c. **Standard Deviation**. The sample standard deviation is determined by taking the square root of the sample variance:

$$s = \sqrt{\frac{\sum (X - \bar{X})^2}{n - 1}} = \sqrt{1.02} = \$1.01M$$

If we used the sample mean of $3.68M as our estimate, we would typically expect to be off by give or take (±) $1.01M.

d. **Coefficient of Variation** (CV). The sample CV is calculated as:

$$CV = \frac{s}{\bar{X}} = \frac{1.01}{3.68} = .2745 \text{ or } 27.45\%$$

If we used the mean of $3.68M as our estimate, we would typically expect to be off by give or take (±) 27% of the sample mean.

Take this opportunity to complete the on-line practical exercises and knowledge reviews for Descriptive Statistics before proceeding.

Following the knowledge reviews on the Mean, Median, and Mode, you will find a video that reviews the Variance, Standard Deviation, and CV calculations using the same examples as in this part of the text. If you would like to walk through an explanation of these concepts before attempting the knowledge reviews on these concepts, then take the opportunity to view the video.

5.4 Frequency Distributions

Frequency distributions are a graphical way to depict the central tendency and dispersion of data. Rather than providing a direct numerical measurement of the data, frequency distributions provided a visualization of the data. A histogram is constructed by dividing the data range into a number of equal intervals, commonly called bins or classes. The data is then distributed into the bins, ensuring that each item is in only one bin or class.

Let's use our population of 25 auditoriums as an example. We first need to decide how many bins or intervals we want. Some texts provide suggestions like "at least six, but no more than 15 bins". Other references provide formulas, sometimes elaborate, for calculating the number of bins or classes. Sometimes the nature of the data will suggest a logical bin width (e.g. data occurring over time might be grouped by week, month, or quarter). And many suggest that it is a matter of judgment and trial and error to determine the number of bins. We are going to use one of the more simple rules of thumb:

Number of bins or classes = $\sqrt{N}$ or $\sqrt{n}$, so in our example: # bins or classes = $\sqrt{25} = 5$

Now, the costs ranged from \$5.92M to \$2.15M with a range of \$3.77M which we will now divide into 5 bins of equal width. The 3.77 ÷ 5 = .754, our bin width. In our example we will start the first bin at the lowest value (2.15) plus the bin width (.754) to give us a value of 2.90. Each successive bin will be the value of the previous bin plus .754. This gives us:

Bins	Frequency
2.15 - 2.90	5
2.91 - 3.66	8
3.67 - 4.41	5
4.42 - 5.17	5
5.18 - 5.92	2

Frequency: the number of data points within a given bin

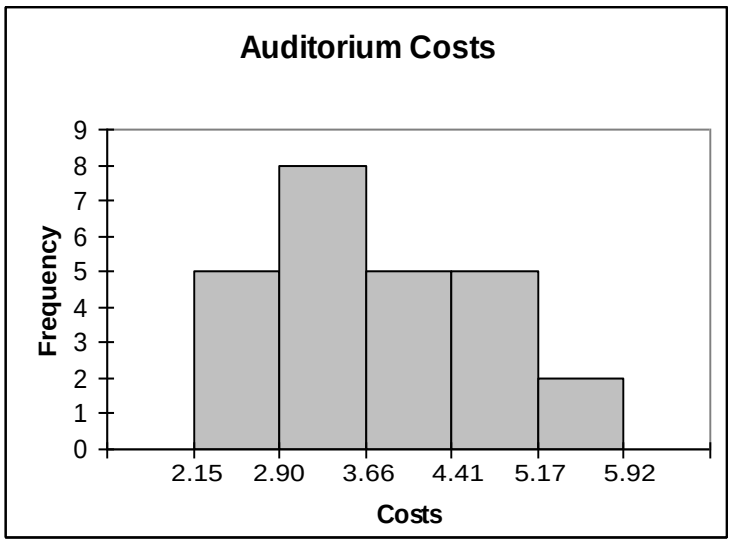


Figure 5.3

We would interpret the histogram such that 5 auditoriums cost less than \$2.90M, 8 auditoriums cost between \$2.90M and \$3.66M, etc. It appears that the center of the data is somewhere around \$3.66M and that the costs are fairly dispersed, not tightly grouped around any particular value (as suggested by the CV of ± 25%). We could take this one step further and say that 8 out of 25, or 32%, of the auditoriums cost between \$2.90M and \$3.66M. We might then infer that there is a 32% probability or likelihood that an auditorium will cost between \$2.90M and \$3.66M.

5.5 Probability Distributions

Just as frequency distributions are pictures of data behavior, probability distributions are pictures of probability behavior. Probability distributions are generally classified as either discrete or continuous.

a. The **discrete** probability distribution applies to events for which probabilities can take on only certain discrete values. To illustrate this type of distribution, the rolling of two dice will be considered. The probabilities associated with the different possible occurrences are listed below.

Outcome	2	3	4	5	6	7	8	9	10	11	12
Probability	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36

Each of these possible outcomes has one discrete probability value associated with it. These probabilities are plotted against their respective outcomes to give the discrete probability distribution. This is shown in Figure 5.4.

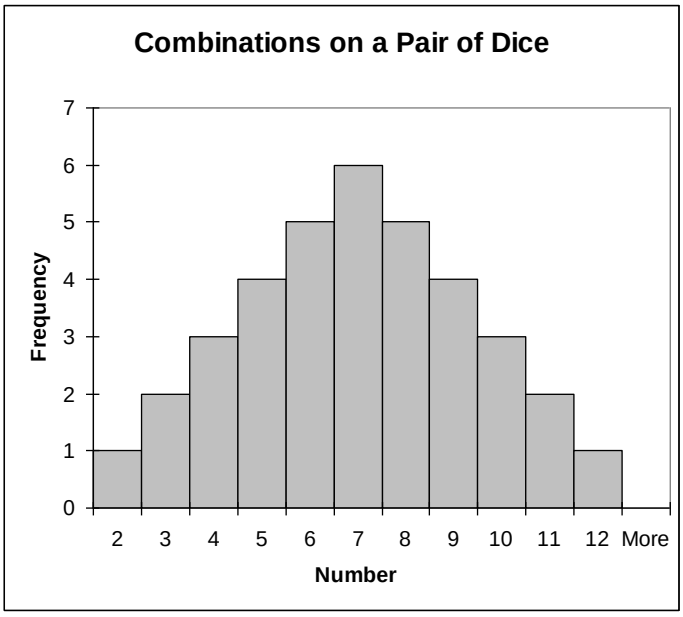


Figure 5.4

b. The **continuous** probability distribution describes probability behavior that doesn't take on specific values for specific events. It is drawn so that the area contained under this curve equals 1.00 or 100%, i.e. every possible outcome is contained under the curve. The probability of any specific value under the curve occurring is zero; however, we can make use of the continuous distribution by finding the probability of an event falling within a certain interval as illustrated in Figure 5.5. This probability is equal to the area under the curve between the two end points of the interval as in this diagram.

Continuous distributions can take on an infinite number of shapes. Some of the more common shapes belong to the Normal, Chi-square, F, Student-T, and Uniform distributions. However, for the purposes of this lesson, only the Normal and Student-T distributions will be used.

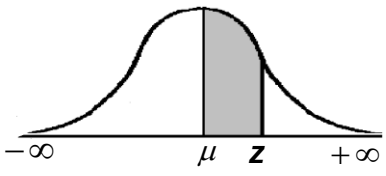


Figure 5.5

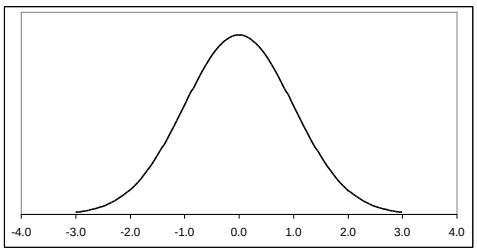
5.6 The Normal Distribution

Before we delve further into distributions, let's take a step back and look at the broader picture of cost estimating. In a 1978 report[1] to Congress, the Comptroller General of the United States stated, "Cost estimating is more art than science. Cost estimates are not statements of fact; rather, they are judgments of the cost to perform work under specified conditions. For programs that span years from the drawing board to completion, economic uncertainties and technological risks are inherent. The single-point or specific-dollar estimate assumes a certainty as to cost that does not exist."

In short, there is not a cost per se, but rather there exist distributions of cost. Analysts over the years have determined many different types of distributions that apply to cost estimating, one of the most common and most useful being the normal distribution. In fact, we will discover later in the course in our discussion of Risk Analysis that the total cost distribution tends toward a normal distribution regardless of the type of distributions associated with the lower cost elements. We will be using the normal distribution to assess the likelihood of a cost overrun and the funds required to achieve a certain likelihood of success. For this reason, and for a foundation of our discussion on statistics and regression, we will spend some time discussing the nature and application of the normal distribution.

The normal distribution, commonly referred to as the "bell-shaped curve", is best described by listing its properties.

(a) It is symmetric about its mean. This says that if the normal distribution is divided in half at the mean, the two halves are mirror images of each other.



(b) The normal distribution is continuous.

(c) The range of the normal distribution from - ∞ to + ∞ . This says that the two tails of the distribution approach the horizontal axis without ever reaching it. This is also known as approaching the axis asymptotically.

(d) The normal distribution is defined completely by the mean "μ" and standard deviation "σ" parameters. Therefore, anything you need to know about a normal distribution can be found using "μ" and "σ".

(e) A given percentage of the outcomes falls between "μ" and a certain number of "σ's". This allows the use of the standard normal distribution tables to determine probabilities of events occurring within certain limits.

[1] A Range of Cost Measuring Risk and Uncertainty in Major Programs, Comptroller General, PSAD-78-12.

In Figure 5.6 you can see that the area under a normal curve that falls within 1 standard deviation "σ" of the mean "μ" is approximately 68.26%. At 2 σ the area is about 95.5% and at 3 σ the area is around 99.75%.

Normal Distribution

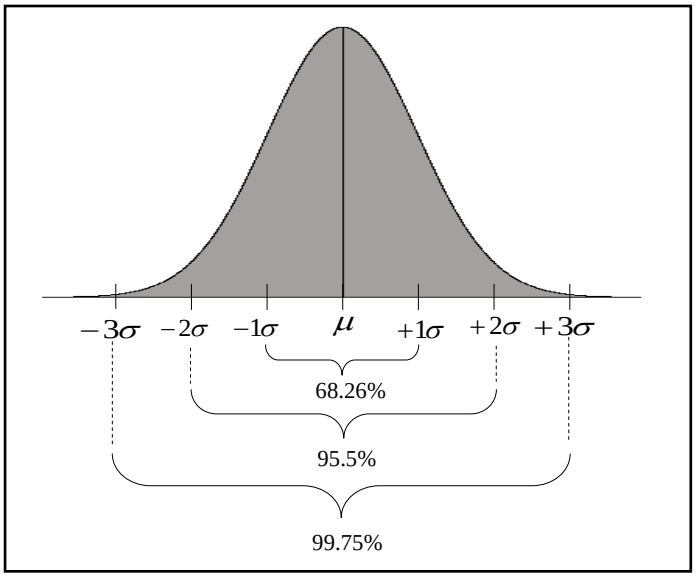


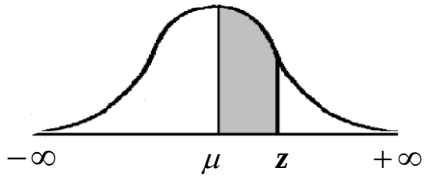
Figure 5.6

(f) Finally, the normal distribution is conveniently tabled for μ = 0 and σ = 1. When these two conditions hold, the distribution is known as the standard normal distribution. Any normal distribution can be converted to this form if you know μ and σ for the distribution. Table 5.5, the standard normal distribution (also known as the Z table) is on the following page.

What if we wanted to find the area under the curve between μ (which would be 0 standard deviations) and 1.00 standard deviation? In Table 5.5 we would look in the Z column for the row with 1.00 and then go to the column with .00 to find .3413 or 34.13%. So, there is a 34.13% probability of a value following between 0 and 1.00 standard deviation. The area under the curve between μ and a standard deviation or Z value of 1.01 is .3438 or 34.38%. The area under the curve between μ and a Z value of 1.09 is .3621 or 36.21%. Since the *total* area under the curve is 1.000 or 100.00%, the area to either the left or right of μ would be .5000 or 50%.

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.00	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
0.10	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
0.20	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.30	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
0.40	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
0.50	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
0.60	.2257	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2517	.2549
0.70	.2580	.2611	.2642	.2673	.2704	.2734	.2764	.2794	.2823	.2852
0.80	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3133
0.90	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.00	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621

The Standard Normal Distribution



z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.00	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
0.10	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
0.20	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.30	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
0.40	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
0.50	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
0.60	.2257	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2517	.2549
0.70	.2580	.2611	.2642	.2673	.2704	.2734	.2764	.2794	.2823	.2852
0.80	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3133
0.90	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.00	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.10	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.20	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.30	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.40	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.50	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441
1.60	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4545
1.70	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.80	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.90	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4767
2.00	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817
2.10	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.20	.4861	.4864	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890
2.30	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.40	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936
2.50	.4938	.4940	.4941	.4943	.4945	.4946	.4948	.4949	.4951	.4952
2.60	.4953	.4955	.4956	.4957	.4959	.4960	.4961	.4962	.4963	.4964
2.70	.4965	.4966	.4967	.4968	.4969	.4970	.4971	.4972	.4973	.4974
2.80	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4979	.4980	.4981
2.90	.4981	.4982	.4982	.4983	.4984	.4984	.4985	.4985	.4986	.4986
3.00	.4987	.4987	.4987	.4988	.4988	.4989	.4989	.4989	.4990	.4990

Table 5.5

How do we apply this? Suppose that the costs for the auditoriums are normally distributed. Our population mean (μ) was $3.80M and the standard deviation (σ) was $.96M. Given these assumptions, what would be the likelihood of an auditorium costing more than $4.86M?

1. We need to determine the distance between the mean (μ) of $3.80M and the X value of $4.86M in terms of standard deviations, referred to in the following equation as "Z".

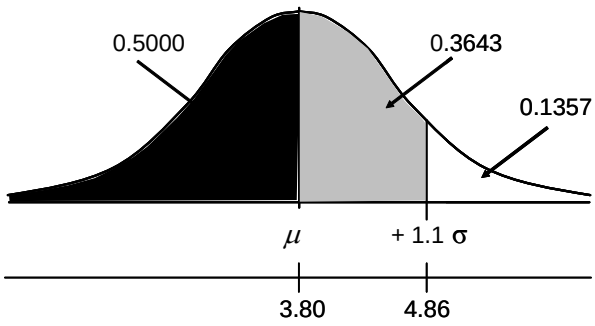
$$Z = \frac{X - \mu}{\sigma} = \frac{4.86 - 3.80}{.96} = \frac{1.06}{.96} = 1.1042 \text{ or } 1.10 \text{ standard deviations}$$

2. How much area (probability) is between μ and 1.10 σ's? Referring to Table 5.5, if we locate the 1.10 row in the Z column and then go to the right to the .00 column, we find .3643, which is the probability between 0 and 1.10 standard deviations. We would say that 36.43% of the area is between 0 and 1.10 standard deviations.

3. Since we are interested in the likelihood of an auditorium costing more than $4.86M, we need to ask how much of the area under the curve is actually to the right of + 1.10 σ's. Since the total area to the right of μ is .5000, we need to subtract the area between μ and +1.10 σ's (which is .3643).

   $.5000 - .3643 = .1357$

   Therefore, there is a 13.57% chance an auditorium will cost more than $4.86M.

   We could have also looked at the area to the left of +1.10 σ's, which is:

   .5000 + .3643 = .8643

   and concluded there is an 86.43% chance that an auditorium will cost less than $4.86M.

What is the likelihood that an auditorium will cost between $2.50M and $4.86M?

1. The distance between the mean (μ) of $3.80M and the X value of $2.50M is:

$$Z = \frac{X - \mu}{\sigma} = \frac{2.50 - 3.80}{.96} = \frac{-1.30}{.96} = (-)1.3542 \text{ or } (-)1.35 \text{ standard deviations}$$

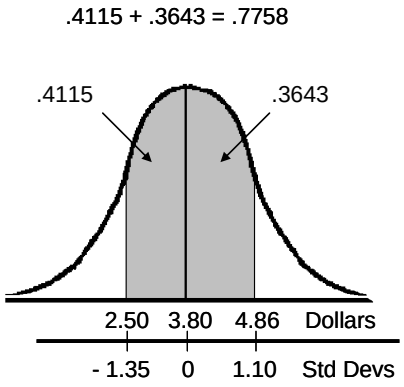
2. Using Table 5.5, we see that .4115 or 41.15% of the area is between μ and 1.35 σ's (between $2.50M and $3.80M).

3. Since we know the area between $3.80M and $4.86M is .3643, and the area between $2.50 and $3.80M is .4115, then the area between $2.5M and $4.86M is:

   .4115 + .3643 = .7758 (see diagram on next page)

There is a 77.58% likelihood that an auditorium will cost between $2.50M and $4.86M.

What is the likelihood that an auditorium will cost between $2.50M and $4.86M?



There is a 77.58% likelihood that an auditorium will cost between $2.50M and $4.86M.

Before proceeding, take this opportunity to view a video on determining the probability under a normal distribution,

and

to complete the on-line practical exercises and knowledge reviews on frequency distributions and applications of the normal distribution.

5.7 The Student t-Distribution

From our earlier discussion of the properties of the normal distribution, we would say that if we had a population of 500 observations or data points, we would expect 68.26% of the observations to lie within ± 1.00 standard deviation (σ) of the mean (μ). But what if we drew a sample of 20 observations out of that population; would we still expect 68.26% of the observations to lie within ± 1.00 standard deviation (s) of the sample mean ($\overline{X}$) given that each successive sample would result in a different sample mean and standard deviation? And what if we only drew a sample of 10 items; wouldn't we be even more uncertain than with the sample of 20 items? If we were to treat a small sample with the same level of confidence as a population would we not risk drawing the wrong conclusion about the population simply due to the chance of sampling error?

Recognizing this dilemma, W.S. Gosset, publishing under the name of "Student", developed a distribution with the characteristics of a normal distribution, but that took into consideration the sample size and number of population parameters being estimated by sample statistics (degrees of freedom). This became known as the Student t-distribution or simply the t-distribution.

The t distribution has nearly the same properties as the normal distribution.

(a) It is symmetric about its mean, ($\overline{X}$).

(b) The t distribution is continuous.

(c) The t distribution ranges from - ∞ to + ∞

(d) The t distribution is defined totally by the mean, $\overline{X}$; the sample standard deviation, s; and the degrees of freedom.

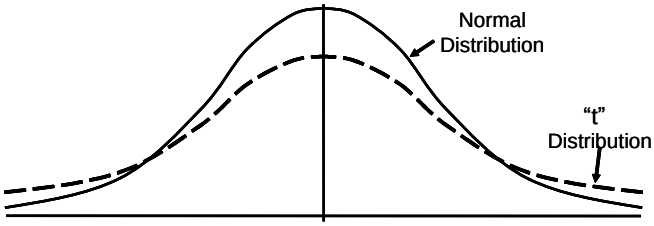


Figure 5.7

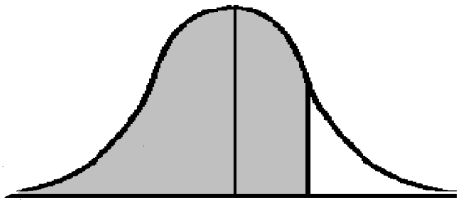
(e) Given the degrees of freedom, a percentage of the outcomes fall between $\overline{X}$ and a certain number of standard deviations.

As depicted in Figure 5.7, in relation to the normal distribution, the t distribution is flatter and less peaked. This reflects the increased uncertainty due to the use of sample statistics instead of population parameters. As the degrees of freedom (df) increase, the t-distribution approaches the normal distribution. The normal distribution is generally used when dealing with the population or a large sample (n > 30). The t-distribution is recommended for small samples (n ≤ 30).

An example of a one-tailed t-table in shown in Table 5.6. The left-hand column represents degrees of freedom (df). In situations where we estimate the population mean with the sample mean we will have "n-1" degrees of freedom.

The values across the top of the columns represent the *level of confidence* (e.g. 60%, 70%, 80%) and are depicted as the shaded section on the graphic. The un-shaded "tail" is referred to as the *level of significance* (or "α" pronounced "alpha"). The level of significance is equal to 1.00 minus the level of confidence, and vice-versa. Let's look at an application of the t-distribution.

Percentiles of the Student t-Distribution



t_p

df	$t_{.60}$	$t_{.70}$	$t_{.80}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$
1	.325	.727	1.376	3.078	6.314	12.706	31.821	63.656
2	.289	.617	1.061	1.886	2.920	4.303	6.965	9.925
3	.277	.584	.978	1.638	2.353	3.182	4.541	5.841
4	.271	.569	.941	1.533	2.132	2.776	3.747	4.604
5	.267	.559	.920	1.476	2.015	2.571	3.365	4.032
6	.265	.553	.906	1.440	1.943	2.447	3.143	3.707
7	.263	.549	.896	1.415	1.895	2.365	2.998	3.499
8	.262	.546	.889	1.397	1.860	2.306	2.896	3.355
9	.261	.543	.883	1.383	1.833	2.262	2.821	3.250
10	.260	.542	.879	1.372	1.812	2.228	2.764	3.169
11	.260	.540	.876	1.363	1.796	2.201	2.718	3.106
12	.259	.539	.873	1.356	1.782	2.179	2.681	3.055
13	.259	.538	.870	1.350	1.771	2.160	2.650	3.012
14	.258	.537	.868	1.345	1.761	2.145	2.624	2.977
15	.258	.536	.866	1.341	1.753	2.131	2.602	2.947
16	.258	.535	.865	1.337	1.746	2.120	2.583	2.921
17	.257	.534	.863	1.333	1.740	2.110	2.567	2.898
18	.257	.534	.862	1.330	1.734	2.101	2.552	2.878
19	.257	.533	.861	1.328	1.729	2.093	2.539	2.861
20	.257	.533	.860	1.325	1.725	2.086	2.528	2.845
21	.257	.532	.859	1.323	1.721	2.080	2.518	2.831
22	.256	.532	.858	1.321	1.717	2.074	2.508	2.819
23	.256	.532	.858	1.319	1.714	2.069	2.500	2.807
24	.256	.531	.857	1.318	1.711	2.064	2.492	2.797
25	.256	.531	.856	1.316	1.708	2.060	2.485	2.787
26	.256	.531	.856	1.315	1.706	2.056	2.479	2.779
27	.256	.531	.855	1.314	1.703	2.052	2.473	2.771
28	.256	.530	.855	1.313	1.701	2.048	2.467	2.763
29	.256	.530	.854	1.311	1.699	2.045	2.462	2.756
30	.256	.530	.854	1.310	1.697	2.042	2.457	2.750
40	.255	.529	.851	1.303	1.684	2.021	2.423	2.704
60	.254	.527	.848	1.296	1.671	2.000	2.390	2.660
120	.254	.526	.845	1.289	1.658	1.980	2.358	2.617
∞	.253	.524	.842	1.282	1.645	1.960	2.326	2.576

Table 5.6

5.8 Confidence Intervals

Whether we are dealing with small or large samples, generally the purpose in drawing a sample is to make a statement about the population from which it came. The purpose of our sample of 10 auditoriums was to make a statement about the average cost of an auditorium and the typical variation in the cost. Our best guess of the average cost of an auditorium would be the sample mean of $3.68M. We really wouldn't expect the population mean to be exactly $3.68M, but we would hope that it is somewhere within that ballpark. We can easily see the reason for our skepticism by looking at 5 random samples of 10 items from our "population" of 25 auditoriums.

Random Samples from Population

Observations	A	B	C	D	E
1	5.70	4.15	4.98	3.26	2.71
2	4.66	2.77	3.85	3.65	2.15
3	4.58	3.68	4.75	3.85	4.66
4	3.60	3.37	3.26	5.70	5.70
5	3.44	5.92	4.25	4.66	4.25
6	3.37	2.15	4.15	3.68	2.77
7	3.26	2.71	3.44	4.75	3.65
8	3.11	3.11	2.77	3.37	4.21
9	2.77	4.98	4.66	4.21	2.75
10	2.31	3.26	4.58	4.15	3.85
Sample Mean	3.68	3.61	4.07	4.13	3.67

Table 5.7

The idea behind a confidence interval is that we acknowledge the variability in sampling, and instead of making a statement that the population mean is a specific value, we make a statement that we are 80% or 90% confident that the population mean is within a specific range.

Small samples. When $n \leq 30$ we use the t-distribution, and the confidence interval is determined:

$$\bar{X} \pm (t_p)\left(\frac{s}{\sqrt{n}}\right)$$

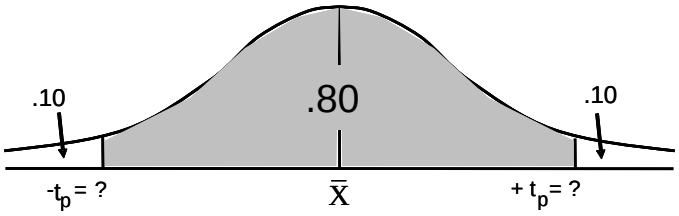
How would we calculate an 80% confidence interval for the average cost of an auditorium?

Given:

the sample mean ($\bar{X}$) = $3.68M,
the standard deviation (s) = $1.01M, and
the sample size (n) = 10;

the only piece of information we lack in order to calculate the confidence interval is (t_p). This value is the number of standard deviations under a t-distribution associated with a given level of confidence for a given number of degrees of freedom. Since we have estimated the population mean with a sample statistic we will have "n-1" degrees of freedom.

Looking at this graphically…

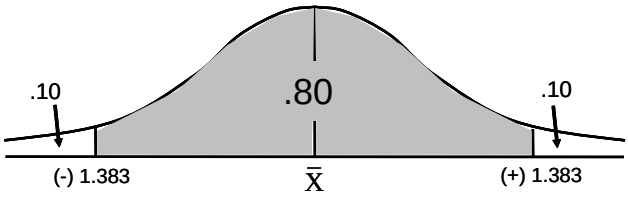


An 80% level of confidence means a 20% level of significance or α. Since this is a confidence *interval*, there would be .80 in the middle of the curve and the α of .20 would be split between the two tails, so one-half of α or α/2 would be in each tail.

We want to use our t-table (Table 5.6) to determine how many standard deviations (t_p) will be required to bound the .80 level of confidence. Unfortunately, our table has been calibrated based on *one-tail*, so we need to treat our two-tailed confidence interval as if there was only one-tail. Since we have .10 in the right tail, then .90 of the area lies to the left, so we would use the .90 column in the table. A helpful reminder sometimes used in interval notation is $t_{p\left(1-\alpha/2,\ n-1\right)}$.

Our sample size (n) is 10, the degrees of freedom (df) = n – 1 = 9; so we use row 9 on the table.

The calculations would be:



$\bar{X} = 3.68$

$s = 1.01$

$n = 10$

$df = (n-1)$

$$\bar{X} \pm t_{p\left(1-\alpha/2,\ n-1\right)} \left(\frac{s}{\sqrt{n}}\right)$$

$$3.68 \pm t_{p\left(1-.20/2,\ 10-1\right)} \left(\frac{1.01}{\sqrt{10}}\right)$$

$$3.68 \pm t_{p(1-.10,\ 9)} \left(\frac{1.01}{3.16}\right)$$

$$3.68 \pm t_{p(.90,\ 9)} \left(.32\right)$$

$$3.68 \pm \left(1.383\right)\left(.32\right)$$

$$3.68 \pm .44$$

Now, after taking 3.68 minus .44, and 3.68 plus .44, we can now make the statement:

We are 80% confident that the average cost of an auditorium is between \$3.24M and \$4.12M, or

$$P\left(\$3.24M \le \mu \le \$4.12M\right) = .80$$

[the probability that μ is between 3.24 and 4.12 is 80%]

How would the problem change for a 90% confidence interval? There would now be .05 in each tail, and we would use the .95 column $\left(1-.10/2\right)$ for a t_p = 1.833.

Large samples. As the degrees of freedom increase, the t-distribution approaches the normal distribution. Generally, when n > 30, the normal distribution is used to support the calculations for a confidence interval.

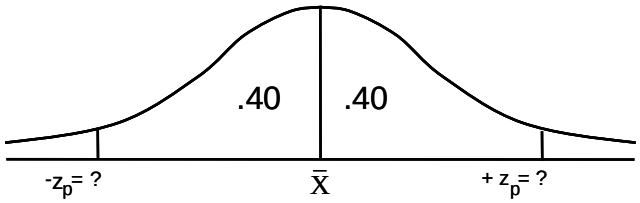
What if we were to compute an 80% confidence interval, as in the previous example, with the only difference being that the sample size (n) was now 36 rather than 10?

Given: $\bar{X} = 3.68$ s = 1.01 n = 36

The confidence interval would be calculated: $\bar{X} \pm (z_p)\left(\frac{s}{\sqrt{n}}\right)$

The only difference in the formula is the use of "z_p" instead of "t_p". How do we determine "z_p"?

The Z table (Table 5.5) reflects the area under *one side* of the distribution between 0 and a specific number of standard deviations. So we need to treat our confidence interval as if we are only looking at one side of the distribution.



The 80% confidence interval would have 40% (.40) of the area on either side of $\bar{X}$. We want to find the number of standard deviations associated with this .40 of the area. On Table 5.5, the area under the curve is represented by the values in the body of the table. Looking for a number as close to .40 as possible, we find a value of .3997 in row 1.20 and column .08. This would be read as 1.28 standard deviations and is the "z_p" value. The area within ± 1.28 standard deviations is .7994 (.3997 x 2) or approximately 80%.

Returning to our calculations:

$$\bar{X} \pm (z_p)\left(\frac{s}{\sqrt{n}}\right)$$

$$3.68 \pm (1.28)\left(\frac{1.01}{\sqrt{36}}\right)$$

$$3.68 \pm (1.28)\left(\frac{1.01}{6}\right)$$

$$3.68 \pm (1.28)\ (.17)$$

$$3.68 \pm .22$$

After taking 3.68 minus .22, and 3.68 plus .22, we can now make the statement:

We are 80% confident that the average cost of an auditorium is between \$3.46M and \$3.90M, or

$P(\$3.46M \le \mu \le \$3.90M) = .80$ [the probability that μ is between 3.46 and 3.90 is 80%]

How would the problem change for a 90% confidence interval? There would now be .45 on either side of $\bar{X}$, and we could use either .4495 (for a z_p of 1.64) or .4505 (for a z_p of 1.65).

Before proceeding, take the opportunity to view a video on constructing confidence intervals,

and

to complete the on-line practical exercises and knowledge reviews for confidence intervals.

5.9 Hypothesis Testing

Have you ever assumed something to be false, only to find out that it was actually true (statisticians call this a Type 1 error); or you assumed something to be true, but found out later that it was actually false (a Type 2 error)? Hypothesis tests allow us to make statements of probability or likelihood to reduce our chances of making these types of errors. We will look at some examples of their use here, and then revisit them later in our regression discussion.

What if we were working base budget issues and the communication shop said that a significant portion of their budget was associated with equipment repair, and that they had budgeted 8.0 hours for the typical repair call. In order to test that assumption we collected data on equipment repairs for the last quarter. We found that there had been 25 repairs made with an average repair time of 7.0 hours and a standard deviation of .75 hours. Our supervisor tells us that we had better not challenge the communications shop budget unless we can be 90% confident in our position. How do we test the assumption that it typically takes 8.0 hours for a repair?

We will start by assuming that it does typically take 8.0 hours for a repair. This will be called our *null* hypothesis (H_o). We think there is a possibility that it actually takes less than 8.0 hours, so we will call this our *alternate* hypothesis (H_a). These statements are written:

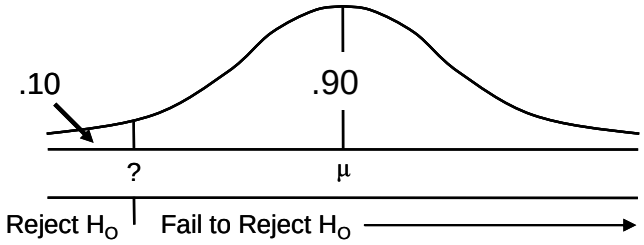
H_o: μ_0 repair = 8.0 hours (i.e. the population average repair takes 8.0 hours)

H_a: μ_0 repair < 8.0 hours (i.e. the population average repair takes less than 8.0 hours)

Much like our criminal justice system, we will assume that the null hypothesis (H_o) is true (not guilty) unless we can provide evidence beyond a reasonable doubt to the contrary (guilty). In this case the "reasonable doubt" is our 90% level of confidence.

Visually the test will look like this:

Keep in mind that our H_o is that $\mu = 8.0$. If our sample mean ($\bar{X}$) is significantly less than 8.0 (such that it falls into the .10 region) then we would conclude that there is less than a 10% chance that the average repair is 8.0 hours or more.



Based on our t-table (Table 5.6) how many standard deviations would we have to go out in order to have .90 of the area on one side and .10 on the other side. Using "n-1" or "25-1" degrees of freedom, we would go to row 24, and across to the column with .90 in the heading and locate 1.318 standard deviations. Since the rejection region is to the left of μ, this will be a (-) 1.318.

The next step will be to determine how far (in standard deviations) the sample mean of 7.0 is from the hypothesized mean of 8.0. We will designate this as "t_c" or "t_{calc}" and calculate it as:

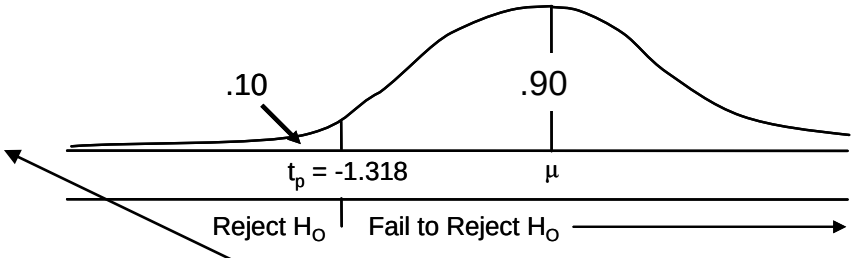
$$t_C = \frac{\bar{X} - \mu_0}{\frac{s}{\sqrt{n}}}$$

Pulling this all together the problem would be worked like this:

H_o: μ_0 repair = 8.0 hours

H_a: μ_0 repair < 8.0 hours

$\bar{X} = 7.0$
$s = .75$
$n = 25$
$df = 24$



$$t_C = \frac{\bar{X} - \mu_0}{\frac{s}{\sqrt{n}}} = \frac{7.0 - 8.0}{\frac{.75}{\sqrt{25}}} = \frac{-1.00}{.15} = \boxed{(-)6.67}$$

The sample mean of 7.0 hours falls -6.67 standard deviations from 8.0 hours, which is well beyond the -1.318 standard deviations. Thus, based on our sample, we would reject the H_o and conclude at the 90% level of confidence that the average repair takes less than 8.0 hours.

> **Note**: The hypothesis test could have been based on a given level of significance. A 90% level of confidence is equivalent to a .10 level of significance ($\alpha = .10$).

Since one of the regression statistics we will be looking at is evaluated based on a "two-sided" hypothesis test, let's take a look at an example here.

You are working at a depot and have been asked to review the fee for service rate for auxiliary power unit (APU) overhauls. Your supervisor said to use the existing rate of $1280 unless you are 80% confident that the rate should be changed. Since there are two possibilities (i.e. the rate should be higher or lower) we will need a "two-sided" test. The hypothesis statements would be:

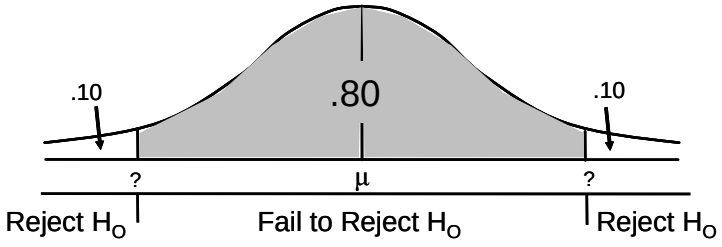
H_o: μ_0 rate = $1280 (i.e. the population average cost is equal to $1280)

H_a: μ_0 rate ≠ $1280 (i.e. the average cost is not equal to $1280, its actually higher or lower)

The average actual cost for the last 18 overhauls is $1235 with a standard deviation of $175.

What would be the "t_p" value for a two-sided test with a confidence of .80?

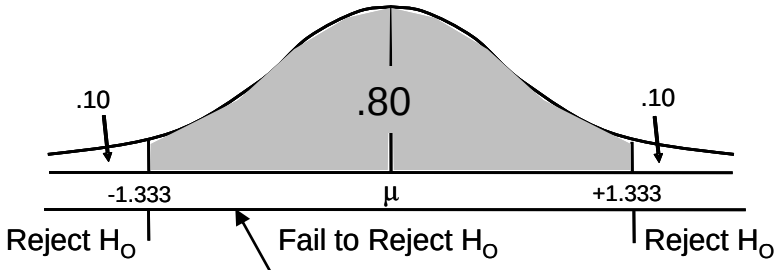
Using Table 5.6, and remembering the discussion on confidence intervals, we will focus on the right "tail" being .10, and treat this as the "t_p" associated with a level of confidence of .90. We will be on row 17 (i.e. 18-1) and column .90 for a $t_p = 1.333$.



Putting this all together:

H_o: μ_0 rate = \$1280
H_a: μ_0 rate $\neq$ \$1280

$\bar{X} = 1235$
s = 175
n = 18
df = 17



$$t_C = \frac{\bar{X} - \mu_0}{\frac{s}{\sqrt{n}}} = \frac{1235 - 1280}{\frac{175}{\sqrt{18}}} = \frac{-45}{41.27} = \boxed{(-)1.09}$$

Since the sample mean of \$1235 fell only 1.09 standard deviations from the current rate of \$1280, we cannot reject the current rate at the 80% level of confidence, so we will continue to use the \$1280 rate.

An alternative approach we could use in this case is to construct an 80% confidence interval.

$$\bar{X} \pm t_{p\left(1-\frac{\alpha}{2}\right)} \left(\frac{s}{\sqrt{n}}\right)$$

Based on these calculations, we would be 80% confident that the average overhaul cost was between \$1180 and \$1290 (i.e. the \$1235 minus and plus the \$55). Since the current price of \$1280 falls within this range, we cannot reject the possibility that the average cost is equal to \$1280.

$$1235 \pm t_{p\left(1-\frac{.20}{2}\right)} \left(\frac{175}{\sqrt{18}}\right)$$

$$1235 \pm t_{p(1-.10)} \left(\frac{175}{4.24}\right)$$

$$1235 \pm t_{p(.90)} \left(41.27\right)$$

$$1235 \pm \left(1.333\right)\left(41.27\right)$$

$$1235 \pm 55$$

5.10 Conclusion

Descriptive and inferential statistics are powerful tools for summarizing data and associating a likelihood or probability to events taking place. It's no wonder that many statistics books coin the phrase "statistics for decision making" in their titles. We have developed some useful techniques in and of themselves, and also laid an important foundation for our discussion on regression.

On-line you will find videos on one and two tailed hypothesis tests,

and

practical exercises and knowledge reviews for hypothesis testing.