Statistics in Traffic Engineering Application

Because traffic engineering involves the collection and analysis of large amounts of data for performing all types of traffic studies, it follows that statistics is also an important element in traffic engineering.

Statistics helps us determine how much data will be required, as well as what meaningful inferences can confidently be made based on that data.

Because of this, traffic engineers often observe and measure the characteristics of a finite sample of vehicles in a population that is effectively infinite.

Statistical analysis is used to address the following questions:

- How many samples are required (i. e., how many individual measurements must be made)?
- What confidence should I have in this estimate (i. e., how sure can I be that this sample measurement has the same characteristics as the population)'!
- What statistical distribution best describes the observed data mathematically?
- Has a traffic engineering design resulted in a change in characteristics of the population? (For example, has a new speed limit resulted in reduced speeds?).

Probability Functions and Statistics

Discrete versus Continuous Functions

They can assume only specific whole values and not any value in between. Continuous functions, made up of continuous variables, in contrast, can assume any value between two given values.

For example, Let N = the number of children in a family. N can equal 1, 2, 3, and so on, but not 1.5, 1.6, and 2.3. Therefore it is a discrete variable. Let H = the height of an individual. H can equal 5 ft, 5.5 ft, 5.6 ft, and so on, and, therefore, is a continuous variable.

Examples of discrete probability functions are:

- The Bernoulli,
- Binomial, and
- Poisson distributions.

Some examples of continuous distributions are:

- The normal,
- Exponential,
- And chi-square distributions.

Common Statistical Estimators

In dealing with a distribution, two key characteristics are of interest. These are discussed in the following subsections.

- **Measures of Central Tendency**
- **Measure of Dispersion**

Measures of Central Tendency

Measures of central tendency are measures that describe the center of data in one of several different ways. The arithmetic mean is the average of all observed data. The true underlying mean of the population, μ, is an exact number that we do not know, but can estimate as:

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

where: $\bar{x}$ = arithmetic average or mean of observed values
x_i = ith individual value of statistic
N = sample size

Consider the following example: Estimate the mean from the following sample speeds in mi/h: (53, 41, 63, 52, 41, 39, 55, and 34). Using Equation above:

$$\bar{x} = \frac{1}{8}(53 + 41 + 63 + 52 + 41 + 39 + 55 + 34)$$
$$= 47.25$$

Because the original data had only two significant digits, the more correct answer is 47 mi/h.

For grouped data, the average value of all observations in a given group is considered to be the midpoint value of the group. The overall average of the entire sample may then be found as:

$$\bar{x} = \frac{\sum_j f_j m_j}{N}$$

where: f_j = number of observations in group j

m_j = middle value of variable in group j

N = total sample size or number of observations

$$\bar{x} = \frac{61(5) + 64(18) + 67(42) + 70(27) + 73(8)}{100}$$

$$= 67.45 = 67$$

The median is the middle value of all data when arranged in an array (ascending or descending order). The median divides a distribution in half: Half of all observed values are higher than the median, and half are lower. For no grouped" data, it is the middle value; for example, for the set of numbered (3, 4, 5, 5, 6, 7, 7, 7, 8), the median is 6. It is the fifth value (in ascending or descending order) in an array of 9 numbers.

For grouped data, the easiest way to get the median is to read the 50% percentile point off a cumulative frequency distribution.

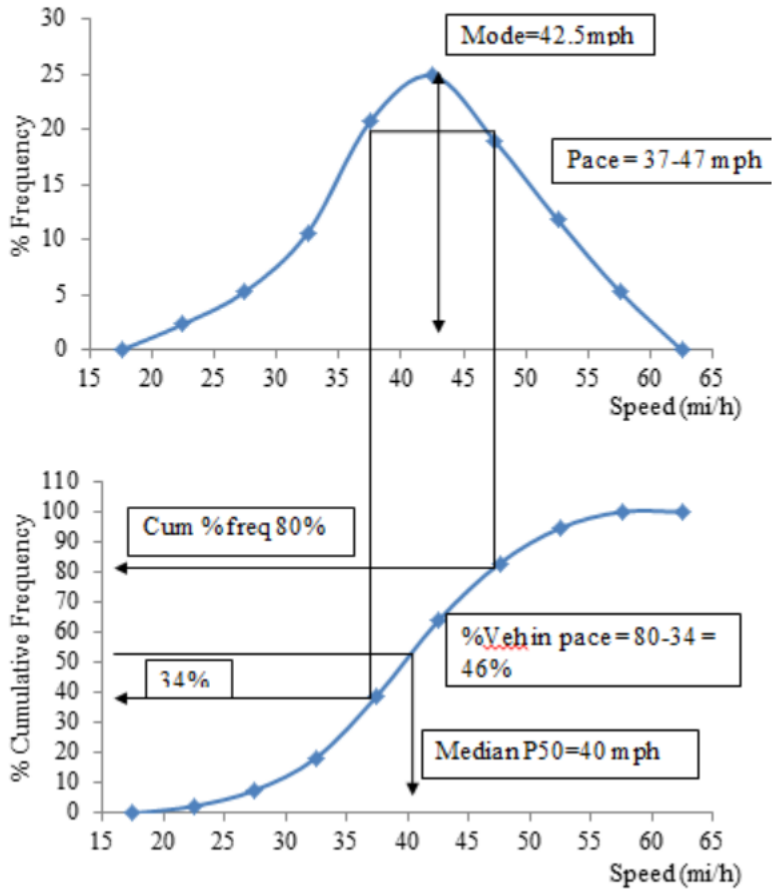
The mode is the value that occurs most frequently-that is, the most common single value.

For example, in non-grouped data, for the set of numbers (3, 4, 5, 5, 6, 7, 7, 7, 8), the mode is 7.

For the set of numbers (3, 3, 4, 5,5,5,6, 7, 8, 8, 8, 9), both 5 and 8 are modes, and the data are said to be bimodal.

For grouped data, the mode is estimated as the peak of the frequency distribution curve.

For a perfectly symmetrical distribution, the mean, median, and mode are the same.



Measures of Dispersion

Measures of dispersion are measures that describe how far the data spread from the center.

The statistical values that describe the magnitude of variation around the mean

- **variance and**
- **standard deviation**

The variance defined as:

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{N - 1}$$

where: s^2 = variance of the data

N = sample size, number of observations

The standard deviation is the square root of the variance. It can be seen from the equation that what you are measuring is the distance of each data point from the mean. This equation can also be rewritten (for ease of use) as:

$$s^2 = \frac{1}{N}\sum_{i=1}^{N} x_i^2 - \left(\frac{N}{N-1}\right)\bar{x}^2$$

For grouped data, the standard deviation is found:

$$s = \sqrt{\frac{\sum fm^2 - N(\bar{x})^2}{N-1}}$$

Where all variables are as previously defined. The standard deviation (STD) may also be estimated as:

$$s_{est} = \frac{P_{85} - P_{15}}{2}$$

Where:
P_{85}:85th percentile value of the distribution (i.e., 85% of all data is at this value or less).
P_{15}:15th percentile value of the distribution (i.e., 15% of all data is at this value or less).

The ith percentile is defined as that value below which x% of the outcomes fall. P_{85} is the 85th percentile, often used in traffic speed studies; it is the speed that encompasses 85% of vehicles.

The median is the 50th percentile speed, or the median speed.

The coefficient of variation is the ratio of the standard deviation to the mean and is an indicator of the spread of outcomes relative to the mean.

The distribution or the underlying shape of the data is of great interest. Is it normal? Exponential? But the engineer is also interested in anomalies in the shape of the distribution (e.g., skewness or bimodality).
Skewness is defined as the (mean - mode)/s.d

- If a distribution is negatively skewed, it means that the data are concentrated to the left of the most frequent value (i.e., the mode).
- When a distribution is positively skewed, the data are concentrated to the right of the mode.

The engineer should look for the underlying reasons for skewness in a distribution. For instance, a negatively skewed speed distribution may indicate a problem such as sight distance or pavement condition that is inhibiting drivers from selecting higher travel speeds.

The Normal Distribution and Its Applications

One of the most common statistical distributions is the normal distribution, known by its characteristic bell-shaped curve (Fig. 1). The normal distribution is a continuous distribution. Probability is indicated by the area under the probability density function f(x) between specified values, such as P (40 < x< 50).

The equation for the normal distribution function is:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\left[\frac{(x-\mu)^2}{2\sigma^2}\right]}$$

Where:
x :a normally distributed statistic.
μ: true mean of the distribution.
σ: True standard deviation of the distribution.

The probability of any occurrence between values x_1 and x_2 is given by the area under the distribution function between the two values. The area may be found by integration between the two limits. Likewise, the mean μ, and the variance, σ^2, can be found through integration.
The normal distribution is the most common distribution because any process that is the sum of many parts tends to be normally distributed.

- SPEED,
- TRAVEL TIME,
- DELAY

Are all commonly described using the normal distribution? The function is completely defined by two parameters: the mean and the variance.

All other values in Equation above, including π, are constants. The notation for a normal distribution is x: N [μ, σ^2], which means that the variable x is normally distributed with a mean of μ and a variance of σ^2.

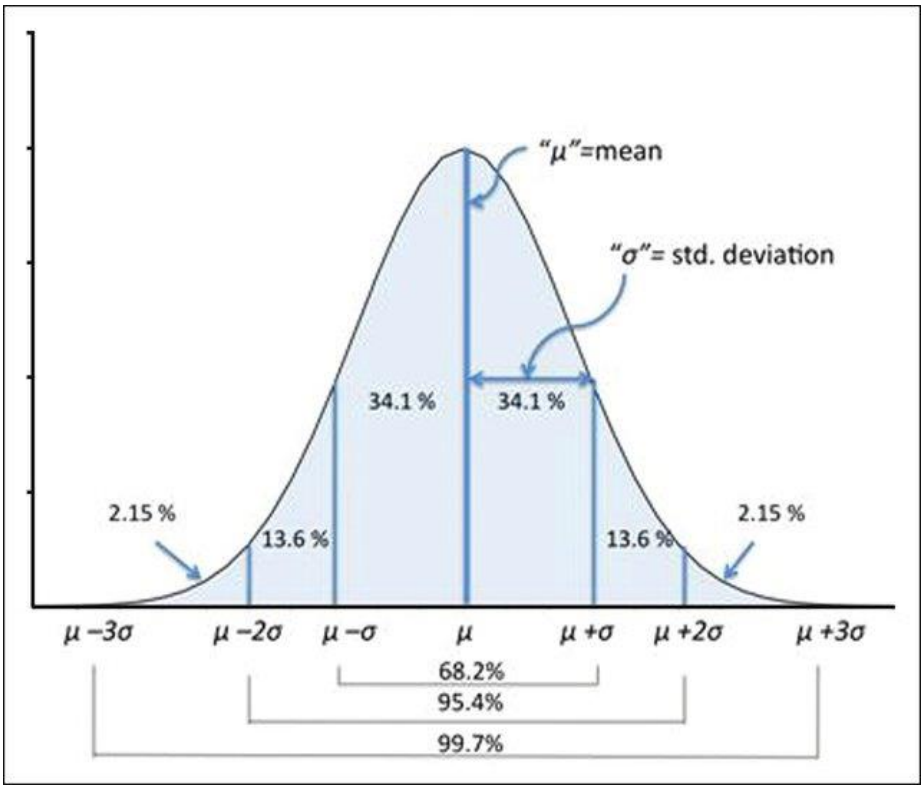


Fig. 1: The Normal Distribution.

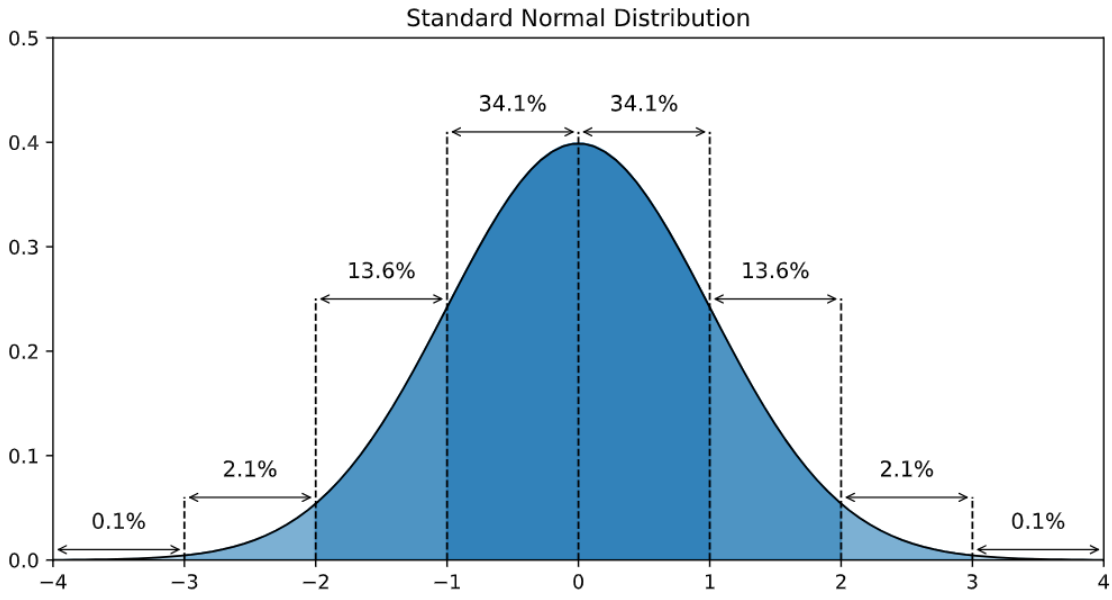
The Standard Normal Distribution

For the normal distribution, the integration cannot be done in closed form due to the complexity of the equation for f(x); thus tables for a "standard normal" distribution, with zero mean (μ = 0) and unit variance ($\sigma^2 = 1$), are constructed. Table 1 presents tabulated values of the standard normal distribution.
The standard normal is denoted z: Z [0, 1]. Any value of x on any normal distribution, denoted x: N [0, 1], can be converted to an equivalent value of z on the standard normal distribution.

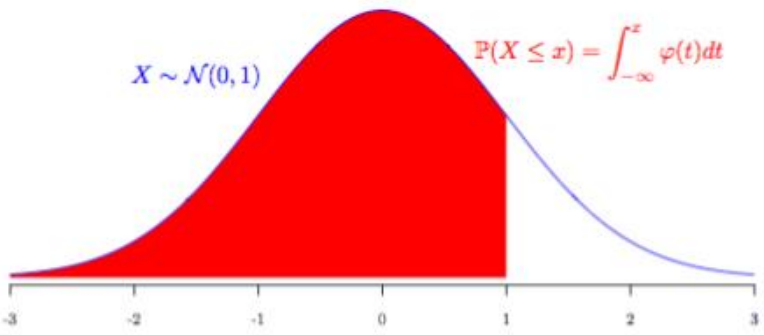
This can also be done in reverse when needed. The translation of an arbitrary normal distribution of values of z to equivalent values of z on the standard normal distribution is accomplished as:

$$z = \frac{x - \mu}{\sigma}$$

Where:
z = equivalent statistic on the standard normal distribution, z: N [0, l]
x = statistic on any arbitrary normal distribution,
x: N[μ, σ^2] other variables as previously defined.



The Standard Normal Distribution



	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990

Figure 2 shows the translation for a distribution of spot speeds that has a mean of 55 mi/h and standard deviation of 7 mi/h to equivalent values of z.

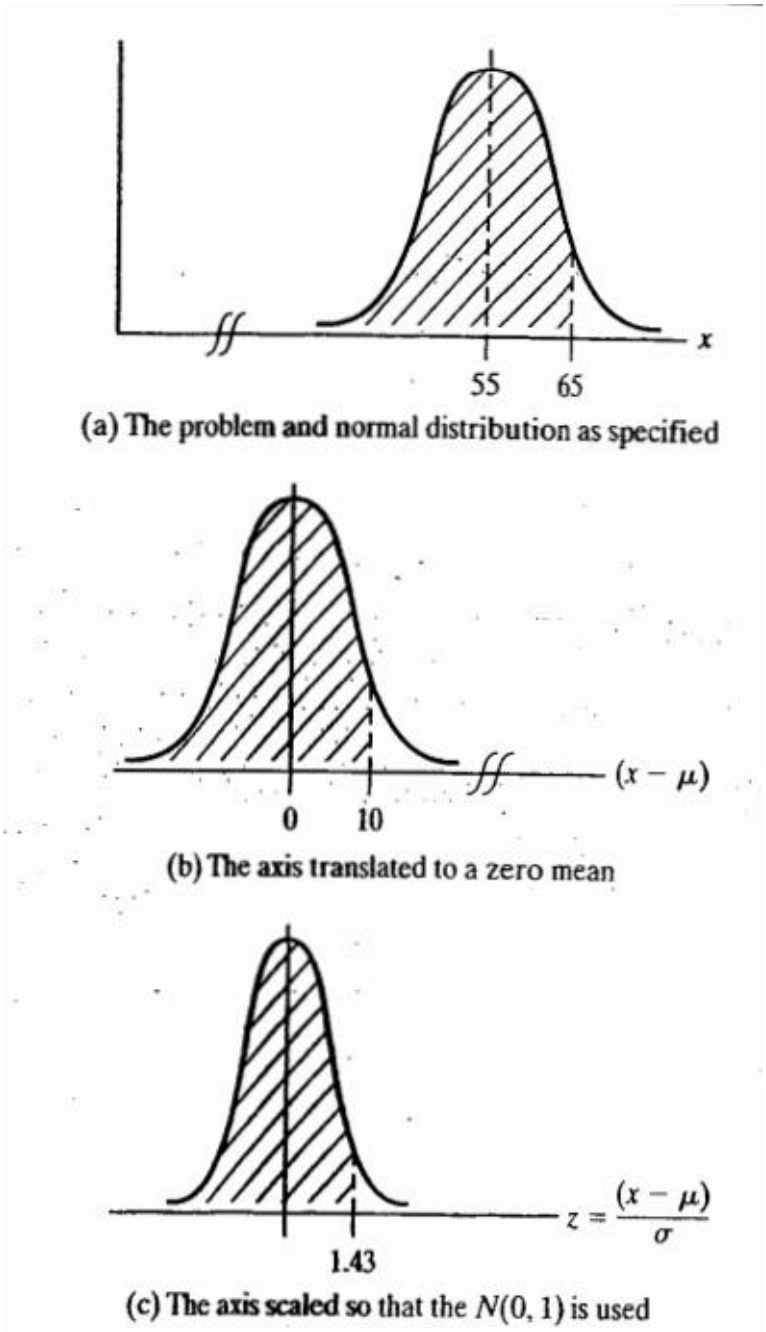


Figure 2: Shifting the Normal Distribution to the Standard Normal Distribution.

Consider the following example: For the spot speed distribution of Figure 2, x: **N [55, 49]**, what is the probability) that the next observed speed will be 65 mi/h or less? Translate and scale the x-axis as shown in Figure 2. The equivalent question for the standard normal distribution, z: N [0, 1], is found using as below:

Determine the probability that the next value of z will be less than:

$$z = \frac{65 - 55}{7} = 1.43$$

Entering Table of standard normal distribution on the vertical scale at 1.4 and on the horizontal scale at 0.03, the probability of having a value of z less than 1.43 is 0.9236, or 92.36%.

Another type of application frequently occurs: For the case just stated, what is the probability that the speed of the next vehicle is between 55 and 65 mi/h?

The probability that the speed is less than 65 mi/h has already been computed. We can now find the probability that the speed is less than 55 mi/h, which is equivalent to z = (55 - 55)/7 = 0.00, so that the probability is 0.50, or 50%, exactly.
The probability of being between 55 and 65 mi/h is just the difference of the two probabilities: (0.9236 - 0.5000) = 0.4236, or 42.36%.

For the case just stated, find the probability that the next vehicle ' s speed is less than 50 mi/h. Translating to the z-axis, we wish to find the probability of a value being less than z = (50 - 55)/7 = -0.71.

Negative values of z are not given in the Table of standard distribution, but by symmetry it can be seen that the desired shaded area is the same size as the area greater than + 0.71". Still, we can only find the shaded area less than + 0.71 (it is: 0.7611).
However, knowing that the total probability under the curve is 1.00, the remaining area (i.e., the desired quantity) is therefore (1.0000 - 0.7611) = 0.2389, or 23.89%.

From these illustrations, three important procedures have been presented

- The conversion of values from any arbitrary normal distribution to the standard normal distribution,
- The use of the standard normal distribution to determine the probability of occurrences, and
- The use of Table of standard normal distribution to find probabilities less than both positive and negative values of z, and between specified values of z.

Important Characteristics of the Normal Distribution Function

The preceding exercises allow us to compute relevant areas under the normal curve. Some numbers occur frequently in practice, and it is useful to have those in mind. For instance, what is the probability that the next observation will be within one standard deviation of the mean, given that the distribution is normal? That is, what is the probability that x is in the range ($\mu \pm 1.00\ \sigma$)? By a similar process to those just illustrated, we can find that this probability is 68. 3%.

The following ranges have frequent use in statistical analysis involving normal distributions:

- 68.3% of the observations are within $\mu \pm 1.00\ \sigma$.
- 95.0% of the observations are within $\mu \pm 1.96\ \sigma$.
- 95. 5% of the observations are within $\mu \pm 2\ \sigma$.
- 99. 7% of the observations are within $\mu \pm 3\ \sigma$.

The total probability under the normal curve is 1.00, and the normal curve is symmetric around the mean. It is also useful to note that the normal distribution is asymptotic to the x-axis and extends to values of $\pm \infty$. These critical characteristics will prove to be useful throughout the text.

Confidence Bounds

What would happen if we asked everyone in class (70 people) to collect 50 samples of speed data and to compute their own estimate of the mean? How many estimates would there be? What distribution would they have? There would be 70 estimates and the histogram of these 70 means would look normally distributed. Thus the “estimate of the mean" is itself a random variable that is normally distributed.

Usually we compute only one estimate of the mean (or any other quantity), but in this class exercise we are confronted with the reality that there is a range of outcomes. We may therefore, ask how good our estimate of the mean is. How confident are we that our estimate is correct? Consider that

- The estimate of the mean quickly tends to be normally distributed.
- The expected value (the true mean) of this distribution is the unknown fixed mean of the original distribution.
- The standard deviation of this new distribution of means is the standard deviation of the original distribution divided by the square root of the number of samples, N. (This assumes independent samples and infinite population.)

Standard error

The standard deviation of this distribution of the means is called the standard error of the mean (E)

$$E = \sigma/\sqrt{N}$$

Where the sample standard deviation, s, is used to estimate σ. and all variables are as previously defined. The same characteristics of any normal distribution apply to this distribution of means as well.

In other words, the single value of the estimate of the mean,$\bar{x}_n$ approximates the true mean population, μ, as follows:

$$\mu = \bar{x} \pm \text{E, with 68.3\% confidence}$$
$$\mu = \bar{x} \pm 1.96\ \text{E, with 95\% confidence}$$
$$\mu = \bar{x} \pm 3.00\ \text{E, with 99.7\% confidence}$$

The ± term (E, 1.96E, or 3.00E, depending on the confidence level) in the preceding equation is also called the tolerance and is given the symbol e.

Consider the following: 54 speeds are observed, and the mean is computed as 47.8 mi/h, with a standard deviation of 7. 80 mi/h. What are the 95% confidence bounds?

$$P[47.8 - 1.96 * (7.80/\sqrt{54})] \leq \mu$$
$$\leq [47.8 + 1.96 * (7.80/\sqrt{54})] = 0.95 \quad \text{or}$$
$$P(45.7 \leq \mu \leq 49.9) = 0.95$$

Thus it is said there is a 95% probability that the true mean lies between 45.7 and 49.9 mi/h. Further, although not proven here, any random variable consisting of sample means tends to be normally distributed for reasonably large, regardless of the original distribution of individual values.

Sample Size Computations

We can rewrite the equation for confidence bounds to solve for N, given that we want to achieve a specified tolerance and confidence. Resolving the 95% confidence bound equation or/V gives:

$$N \geq \frac{1.96^2 s^2}{e^2}$$

Where 1.962 is used only for 95% confidence. If 99.7% confidence is desired, then the 1.96 would be replaced by 3^2.

Consider another example: With 99.7% and 95% confidence, estimate the true mean of the speed on a highway, plus or minus 1 mi/h. We know from previous work that the standard deviation is 7.2 mi/h. How many samples do we need to collect?

$$N = \frac{3^2 * 7.2^2}{1^2} \approx 467 \text{ samples for 99.7\% confidence,}$$

$$N = \frac{1.96^2 * 7.2^2}{1^2} \approx 200 \text{ samples for 95\% confidence}$$

Consider further that a spot speed study is needed at a location with unknown speed characteristics. A tolerance of ± 0.2 mph and a confidence of 95% is desired. What sample size is required? Because the speed characteristics are unknown, a standard deviation of 5 mi/h (a most common result in speed studies) is assumed. Then for 95% confidence,

$N = (1.96^2 * 5^2)/0.2^2 = 2{,}401$ samples.

This number is unreasonably high. It would be too expensive to collect such a large amount of data. Thus the choices are to either reduce the confidence or increase the tolerance.
A 95% confidence level is considered the minimum that is acceptable; thus, in this case, the tolerance would be increased. With a tolerance of 0. 5 mi/h:

$N = (1.96^2 * 5^2)/0.5^2 = 384$ samples (vehicles number)

Thus the increase of just 0.3 mi/h in tolerance resulted in a decrease of 2,017 samples required.

Note that the sample size required depends on **s**, which was assumed at the beginning. After the study is completed and the mean and standard deviation are computed, should be rechecked. If N is greater (i.e., the actual s is greater than the assumed s), then more samples may need to be taken.

Another example: An arterial is to be studied, and it is desired to estimate the mean travel time to a tolerance of ±5 seconds with 95% confidence. Based on prior knowledge and experience, it is estimated that the standard deviation of the travel times is about 15 seconds. How many samples are required?

Based on an application of Equation, $N = 1.96^2(15^2)/(5^2) = 34.6$, rounded to 35 samples

As the data is collected, the s computed is 22 seconds, not 15 seconds. If the sample size is kept at N = 35, the confidence bounds will be $\pm 1.96(22^2)/\sqrt{35}$ or about ±7.3 seconds.

If the confidence bounds must be kept at ±5 seconds, then the sample size must be increased so that;
$N > 1.96^2 (22^2/5^2) = 74.4$ or 75 samples. Additional data will have to be collected to meet the desired tolerance and confidence level.