CHAPTER 1

An Overview of Statistical Applications

1.1 Introduction

Civil engineering is considered to be one of the oldest engineering disciplines. It deals with planning, analysis, design, construction and maintenance of the physical and naturally built environment. The subject is grouped into several specialty areas namely, Structural Engineering, Construction Engineering and Management, Transportation Engineering, Water Resource Engineering, Surveying, Environmental Engineering, Geotechnical Engineering. Transportation engineering involves the collection of huge amount of data for performing all types of traffic and transportation studies. The analysis is carried out based on the collected and observed data. The statistical aspects are also an important element in transportation engineering specifically traffic engineering. Statistics facilitates to resolve how much data will be obligatory, as well as what consequential inferences can confidently be finished based on that observed and collected data. Generally statistics is required whenever it is not possible to directly measure all of the values required. If the traffic engineer needs to know the average speed of all vehicles on a particular section of roadway, not all vehicles could be observed. Even if all speeds of vehicles could be measured over a specified time period, speeds of vehicles arriving before or after the study period or on a different day than the sample day would be unknown. In effect, no matter how many speeds are measures, there are always more that are not known. For all practical and statistical purposes, the number of vehicles using a particular section of road way over time is infinite. Therefore, the traffic engineering

often observes and measures the characteristics of a finite sample of vehicles in a population that is effectively infinite. The mathematics of statistics is used to estimate characteristics that cannot be established with absolute certainty and to assess the degree of certainty that exists.

1.2 Probability Functions and Statistics

Before exploring some of the more complex statistical applications in traffic engineering, some basic principles of probability and statistics that are relevant to transportation and traffic engineering subject are reviewed.

1.2.1 Discrete versus Continuous Functions

Discrete functions are made up of discrete variables that is, they can assume only specific whole values and not any value in between. Continuous functions, made up of continuous variables, on the other hand, can assume any value between two given values. For example, Let N = the number of cars in a family. N can equal 1, 2, 3, etc., but not 1.5, 1.6, 2.3. Therefore it is a discrete variable. Let H = the height of an individual car. H can equal 1.5 m, 1.75 m, 2.25 m, 2.50 m and 2.75 m, etc., and therefore, is a continuous variable. Examples of discrete probability functions are the Bernoulli, binomial, and Poisson distributions, which will be discussed in the Chapter 4. Some examples of continuous distribution are the normal, exponential, and chi-square distributions.

1.2.2 Distributions Describing Randomness

Some events are very predictable, or should be predictable. If you add mass to a spring or a force to a beam, you can expect it to deflect a predictable amount. If you depress the gas pedal a certain amount and you are on level terrain, you expect to be able to predict the speed of the vehicle. On the other hand, some events may be totally random. The emission of the next particle from a radioactive sample is said to be completely random. Some events may have very complex mechanisms and appear to be random for all practical purposes. In some cases, the underlying mechanism cannot be perceived, while in other cases we cannot afford the time or money necessary for the investigation. consider the question of who turns north and who turns south after crossing a bridge. Most of the time, we simply say there is a probability p that a vehicle will turn north, and we treat the outcome as a random event. However, if we studied who was driving each car and where each driver worked, we might expect to make the estimate a very predictable event, for each and every car. In fact, if we kept a record of their license plates and their past decisions, we could make very predictable estimates. The events to a large extent are not random. Obviously, it is not worth that trouble because the random assumption serves us well adequate. In fact, a number of things are modeled as random for all practical purposes, given the investment we can afford. Most of the time, these judgments are just fine and are very reasonable but, as with every engineering judgment, they can sometimes cause errors.

1.2.3 Data Organization

In the data collection process, the data is collected for use in traffic studies, the raw data can be looked at as individual pieces of data or grouped into classes of data for easier comprehension. Most of the data will fit into a common distribution. Some of the common distributions found in traffic engineering are the normal distribution, the exponential distribution, the chi-square distribution, the Bernoulli distribution, the binomial distribution, and the Poisson distribution. As part of the process for determining which distribution fits the data, one often summarizes the raw data into classes and creates a frequency distribution table. This makes the data more easily readable and understood. You could put the data into an array format, which means listing the data points in order from either lowest to highest or highest to lowest. This will give you some feeling for the character of the data but it is still not very helpful, particularly when you have a large number of data points.

1.2.4 Common Statistical Estimators

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

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:

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \qquad \text{.....(1.1)}$$

where $\bar{x}$ = arithmetic average or mean of observed values

x_i = i^{th} individual value of statistic

N = sample size, number of values x_i

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:

$$\bar{x} = \frac{\sum_j f_j \, m_j}{N} \qquad \text{.....(1.2)}$$

where f_j = number of observation in group j

m_j = middle value of variable in group j

N = total sample size or number of observations

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, half are lower. For non-grouped data, it is the middle value; for example, for the set of numbers (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 curve.

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 is 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 will be the same.

Measures of Dispersion

Measures of dispersion are measures that describe how far the data spread from the center. The variance and standard deviation are statistical values that describe the magnitude of variation around the mean, with the variance defined as:

$$S^2 = \frac{\sum (x_i - \bar{x})^2}{N-1} \qquad \text{.....(1.3)}$$

where S^2 = variance of the data

N = sample size, number of observations

All other variables are previously defined.

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 as:

$$S^2 = \frac{1}{N}\sum_{i=1}^{N} x_i^2 - \left(\frac{N}{N-1}\right)\bar{x}^2 \qquad \text{.....(1.4)}$$

For grouped data, the standard deviation is found from:

$$S = \sqrt{\frac{\sum f_m^2 - N(\bar{x})^2}{N-1}} \qquad \text{.....(1.5)}$$

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

$$S_{est} = \frac{P_{85} - P_{15}}{2} \qquad \text{.....(1.6)}$$

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 x^{th} 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. P_{50} is the 50th percentile speed or the median.

1.2 Applications of Normal Distribution

One of the most common statistical distribution is the normal distribution, known by its characteristic bell shaped curve. 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:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\left[\frac{(x-\mu)^2}{2\sigma^2}\right]} \qquad \text{.....(1.7)}$$

where x = normally distributed statistic

μ = true mean of the distribution

σ = true standard deviation of the distribution

π = 3.14

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, α^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, and 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 (1.6) including π, are constants. The notation for a normal distribution is x: $N[\mu, \alpha^2]$, which means that the variable x is normally distributed with a mean of μ and a variance of σ^2.

1.2.1 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 ($\mu = 0$) and unit variance ($\sigma^2 = 1$), are constructed. The standard normal is denoted z: N[0, 1]. Any value of x on any normal distribution, denoted x: N[μ, σ^2], 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 x to equivalent values of z on the standard normal distribution is accomplished as:

$$z = \frac{x - \mu}{\sigma} \qquad \text{.....(1.8)}$$

where z = equivalent statistic on the standard normal distribution, z: N[0, 1]

x = statistic on any arbitrary normal distribution, x: N[μ, σ^2] other variables as previously defined.

1.2.2 Characteristics of the Normal Distribution Function

The forgoing exercise allow one 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 ($\mu \pm 1.00\sigma$)? For example the tentative value can find that this probability is 68.3% by adopting some standard method.

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 $\mu \pm 2.00\sigma$
- 99.7% of the observations are within $\mu \pm 3.00\sigma$

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.

1.3 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 is our estimate of the mean. How confident are we that our estimate is correct? Consider that:

1. The estimate of the mean quickly tends to be normally distributed.
2. The expected value (the true mean) of this distribution is the unknown fixed mean of the original distribution.
3. 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.)

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

$$E = \sigma / \sqrt{N} \qquad \text{.....(1.9)}$$

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 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 population, μ as follows:

$\mu = \bar{x} \pm E$, with 68.3% confidence

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

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

The $\pm$ term (E, 1.96E, or 3,00E, depending upon the confidence level) in the above 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 kmph, with a standard deviation of 7.80 kmph. What are the 95% confidence bounds?

$$P[47.8 - 1.96 \times \left(7.80/\sqrt{54}\right) \le \mu$$

$$\le \left[47.8 + 1.96 \times 7.80/\sqrt{54}\right] = 0.95 \text{ or}$$

$$P\left(45.7 \le \mu \le 49.9\right) = 0.95$$

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

1.4 Determination of Sample Size

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 for N gives:

$$N \geq \frac{1.96^2 x^2}{e^2} \qquad \text{.....(1.10)}$$

where 1.96^2 is used only for 95% confidence. If 99.7% confidence is desired, then the 1.96^2 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 mph. We know from previous work that the standard deviation is 7.2 kmph. How many samples do we need to collect?

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

and

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

Consider further that a spot speed study is needed at a location with unknown speed characteristics. A tolerance of ± 0.2 kmph and a confidence of 95% is desired. What sample size is required? Since the speed characteristics are unknown, a standard deviation of 5 km/h (a most common result in speed studies) is assumed. Then for 95% confidence, $N = (1.96^2 \times 5^2)/0.2^2 = 2{,}401$ samples. This number is unreasonably high. 1t 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:

$$N = \frac{1.96^2 \times 5^2}{0.5^2} = 384 \text{ vehicles}$$

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 is dependent on s, which was assumed at the beginning. After the study is completed and the mean and standard deviation me

computed, N 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 $\pm$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 13.10, $N = 1.96^2(15^2)/(5^2) = 34.6$, which is 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)/$\sqrt{35}$ or about ±7.3 seconds. If the confidence bounds must be kept at $\pm$ 5 seconds, then the sample size must be increased so that $N \geq 1.96^2(22^2)/(5^2) = 74.4$ or 75 samples. Additional data will have to be collected to meet the desired tolerance and confidence level.

1.5 Random Variables Summation

One of the most common occurrences in probability and statistics is the summation of random variables, often in the form $Y = a_1X_1 + a_2X_2$ or in the more general form:

$$Y = \sum a_i X_i \qquad \text{.....(1.11)}$$

where the summation is over i. usually from 1 to n.

It is relatively straightforward to prove that the expected value (or mean) μ_r of the random variable Y is given by:

$$\mu_Y = \sum a_i \mu_{Xi} \qquad \text{.....(1.12)}$$

and that if the random variables x_i are independent of each other, the variance σ_Y^2 of the random variable Y is given by:

$$\sigma_Y^2 = \sum a_i^2 \sigma_{xi}^2 \qquad \text{.....(1.13)}$$

The fact that the coefficients, a_i. are multiplied has great practical significance for us in all our statistical work.

1.5.1 The Central Limit Theorem

One of the most impressive and useful theorems in probability is that the slim of n similarly distributed random variables tends to the normal distribution, no matter what the initial, underlying distribution is. That is, the random variable $Y = \Sigma X_i$, where the X_i have the same distribution, tends to the normal distribution.

The words "tends to" can be read as "tends to look like" the normal distribution. In mathematical terms, the actual distribution of the random variable Y approaches the normal distribution asymptotically.

Sum of Travel Times

Consider a trip made up of 15 components, all with the same underlying distribution, each with a mean of 10 minutes and standard deviation of 3.5 minutes. The underlying distribution is unknown. What can you say about the total travel time?

While there might be an odd situation to contradict this, n = 15 should be quite sufficient to say that the distribution of total travel times tends to look normal. From the standard Equation, the mean of the distribution of total travel times is found by adding 15 terms ($a_i \mu_i$) where $a_i = 1$ and $\mu_i = 10$ minutes, or

$$\mu_y = 15 \times (1 \times 10) = 150 \text{ minutes}$$

The variance of the distribution of total travel times is found from Equation (1.13) by adding 15 terms $\left(a_i^2 \sigma_i^2\right)$ where a_i is again 1, and σ_i is 3.5 minutes. Then:

$$\sigma_y^2 = 15 \times (1 \times 3.5^2) = 183.75 \text{ minutes}^2$$

The standard deviation, σ_y is, therefore, 13.6 minutes.

If the total travel times are taken to be normally distributed, 95% of all observations of total travel time will lie between the mean (150 minutes) ± 1.96 standard deviations (13.6 minutes), or:

$$X_y = 150 \pm 1.96\ (13.6)$$

Thus, 95% of all total travel times would be expected to fall within the range of 123 to 177 minutes (values rounded to the nearest minute).

Hourly Volumes

Five-minute counts are taken, and they tend to look rather smoothly distributed but with some skewness (asymmetry). Based on many observations, the mean tends to be 45 vehicles in the five-minute count, with a standard deviation of seven vehicles. What can be said of the hourly volume?

The hourly volume is the sum of 12 five-minute distributions, which should logically be basically the same if traffic levels are stable. Thus, the hourly volume will tend to look normal, and will have a mean computed using Equation (1.12), with $a_i = 1$, $\mu_i = 45$ vehicles, and n = 12, or $12 \times (1 \times 45) = 540$ veh/h. The variance is computed using Equation (1.13), with $a_i = 1$, $\sigma_i = 7$, and n = 12, or $12 \times (1^2 \times 7^2) = 588$ (veh/h)^2. The standard deviation is 24.2 veh/h. Based on the assumption of normality, 95% of hourly volumes would be between $540 \pm 1.96(24.2) = 540 \pm 47$ veh/h (rounded to the nearest whole vehicle).

Note that the summation has had an interesting effete. The σ/μ ratio for the five-minute count distribution was 7/45 = 0.156, but for the hourly volumes it was 47/540 = 0.087. This is due to the summation, which tends to remove extremes by canceling "highs" with "lows" and thereby introduces stability. The mean of the sum grows in proportion n, but the standard deviation grows in proportion to the square root of n.

Sum of Normal Distributions

Although not proven here, it is true that the sum of any two normal distributions is itself normally distributed. By extension, if one normal is formed by n_1 summations of one underlying distribution and another normal is formed by n_2 summations of another underlying distribution, the sum of the total also tends to the normal.

Thus, in the foregoing travel-time example, not all of the elements had to have exactly the same distribution as long as sub-groupings each tended to the normal.

1.6 The Binomial Distributions

1.6.1 Bernoulli and the Binomial Distribution

The Bernoulli distribution is the simplest discrete distribution, consisting or only two possible outcomes: yes or no, heads or tails, one or zero, etc. The first occurs with probability p, and therefore the second occurs with probability (1 – p = q). This is modeled as:

$$P(X = 1) = p$$

$$P(X = 0) = I – p = q$$

In traffic engineering, it represents any basic choice – to park or not to park; to take this route or that; to take auto or transit (for one individual). It is obviously more useful to look at more than one individual, however, which leads us to the binomial distribution. The binomial distribution can be thought of in two common ways:

1. In N outcomes of the Bernoulli distribution, make a record of the number of events that have the outcome “1”, and report that number as the outcome X.
2. The binomial distribution is characterized by the following properties:
    - There are N events, each with the same probability p of a positive outcome and (l – p) of a negative outcome.
    - The outcomes are independent of each other.
    - The quantity of interest is the total number X of positive outcomes, which may logically vary between 0 and N.
    - N is a finite number.

The two ways are equivalent, for most purposes.

Consider a situation in which people may choose “transit” or “auto” where each person has the same probability p = 0.25 of choosing transit, and each person's decision is independent of that of all other persons. Defining "transit" as the positive choice for the purpose of this example and choosing N = 8. note that:

1. Each person is characterized by the Bernoulli distribution, with p = 0.25.
2. There are $2^8 = 256$ possible combinations of choices, and some of the combinations not only yield the same value of X but also have the same probability of occurring. For instance, the value of X = 2 occurs for both

TTAAAAAA

and

TATAAAAA

and several other combinations, each with probability of $p^2(1 - P)^6$, for a total of 2^8 such combinations.

Stated without proof is the result that the probability P(X = x) is given by

$$P(X = x) = \frac{N!}{(N - x)!x!} p^x (1 - p)^{N-x} \qquad \text{…..(1.14)}$$

with a mean of Np and a variance of Npq where q = 1– p. The derivation may be found in any standard probability text.

There is an important concept that the reader should master, in order to use statistics effectively throughout this text. Even though on average two out of eight people will choose transit, there is absolutely no guarantee what the next eight randomly selected people will choose, even if they follow the rules (same p, independent decisions, etc). In fact, the number could range anywhere from X = 0 to X = 8. And we can expect that tile result X = 1 will occur 10.0% of the time, X = 4 will occur 8.7% of the time, and X = 2 will occur only 31.1% of the time.

This is the crux of the variability in survey results. If there were 200 people in the senior class and each student surveyed eight people from the subject population, we would get different results. Likewise, if we average our results, the result would probably be close to 2.00, but would almost surely not be identical to it.

1.6.2 Asking People Questions Survey Results

Consider that the commuting population has two choices X = 0 for auto and X = 1 for public transit. The probability p is generally unknown, and it is usually of great interest. Assuming the probability is the same for all people (to our ability to discern, at least). then each person is characterized by the Bernoulli distribution.

If we ask n = 50 people for their value of X, the resulting distribution of the random variable Y is binomial and may tend to look like the normal. Applying some "quick facts"

and noting that the expected value (that is the mean) is 12.5 (50 × 0.25), the variance is 9.375 (50 × 0.25 × 0.75), and the standard deviation is 3.06, one can expect 95%, of the results 10 fall in the range 12.5 ± 6.0 or between 6.5 and 18.5.

If n = 200 had been selected. then the mean of Y would have been 50 when p = 0.25 and the standard deviation would have been 6.1, so that 95% of the results would have fallen in the range of 38 to 62.

1.6.3 The Binomial and the Normal Distributions

The central limit theorem informs us that the sum of Bernoulli distributions (i.e., the binomial distribution) tends to the normal distribution. The only question is: How Fast? A number of practitioners in different fields use a rule of thumb that says "for large n and small p" the normal approximation can be used without restriction. This is incorrect and can lead to serious errors.

The most notable case in which the error occurs is when rare events are being described, such as auto accidents per million miles traveled or aircraft accidents. Which is an exact rendering of the actual binomial distribution for $p = 0.7(10)^{-6}$ and two values of n namely $n = 10^6$ and $n = 2(10)^{-6}$, respectively. Certainly p is small and n is large in these cases, and, just as clearly, (they do not have the characteristic symmetric shape of the normal distribution.

It can be shown that in order for there to be some chance of symmetry- that is, in order for the normal distribution to approximate the binomial distribution the condition that $np/(1-p) \geq 9$ is necessary.

1.7 The Poisson Distribution

The Poisson distribution is known in traffic engineering as the "counting" distribution. It has the clear physical meaning or a number of events X occurring in a specified counting interval of duration T and is a one-parameter distribution with:

$$P(X = x) = e^{-m}\frac{m^x}{x!} \qquad \text{.....(1.15)}$$

with mean $\mu = m$ and variance $\sigma^2 = m$.

The fact that one parameter m specifies both the mean and the variance is a limitation, in that if we encounter field data where the variance and mean are clearly different, the Poisson does not apply.

The Poisson distribution often applies to observations per unit of time, such as the arrivals per five minute period at a toll booth. When headway times are exponentially distributed with mean $\mu = 1/\lambda$, the number of arrivals in an interval of duration T is Poisson distributed with mean $\mu = m = \lambda T$.

Applying the Poisson distribution is done the same way we applied the Binomial distribution earlier. For example, say there is an average of five accidents per day on the Florida freeways.

1.8 Testing of Hypothesis

Very often traffic engineers must make a decision based on sample information. For example, is a traffic control effective or not? To test this, we formulate a hypothesis, H_0, called the null hypothesis and then try to disprove it. The null hypothesis is formulated so that there is no difference or no change, and then the opposite hypothesis is called the alternative hypothesis, H_1.

When testing a hypothesis, it is possible to make two types of errors: (1) We could reject a hypothesis that should be accepted (e.g., say an effective control is not effective). This is called a Type I error. The probability of making a Type I error is given the variable name, α. (2) We could accept a false hypothesis (e.g., say an ineffective control is effective). This is called a Type II error. A Type II error is given the variable name β.

Consider this example: An auto inspection program is going to be applied to 100,000 vehicles, of which 10,000 are "unsafe" and the rest are "safe, “Of course, we do not know which cars are safe and which are unsafe.

We have a test procedure, but it is not perfect, due to the mechanic and test equipment used. We know that 15% of the unsafe vehicles are determined to be safe, and 5% of the safe vehicles are determined to be unsafe.

We would define: H_0: The vehicle being tested is "safe," and H_1: the vehicle being tested is "unsafe." The Type I error, rejecting a true null hypothesis (false negative), is labeling a safe vehicle as "unsafe." The probability of this is called the level of significance, α, and in this case $\alpha = 0.05$. The Type II error. failing to reject a false null hypothesis (false positive), is labeling an unsafe vehicle as "safe." The probability of this, β is 0.15. 1ngeneral, for a given test procedure, one can reduce Type I error only by living with a higher Type II error, or vice versa.

1.8.1 Before-and-After Tests with Two Distinct Choices

In a number of situations, there are two clear and distinct choices, and the hypotheses seem almost self-defining:

- Auto inspection (acceptable, not acceptable)
- Disease (have the disease, don't)
- Speed reduction of 5 mph (it happened, it didn't)
- Accident reduction of 10% (it happened, it didn't)
- Mode shift by five percentage points (it happened, it didn't)

Of course, there is the distinction between the real truth (reality, unknown to us) and the decision we make. as already discussed and related to Type I and Type II errors. That

is, we can decide that some cars in good working order need repairing and we can decide that some unsafe cars do not need repairing.

There is also the distinction that people may not want to reduce the issue to a binary choice or might not the able to do so. For instance, if an engineer expects a 10% decrease in the accident rate, should we test "H_0: no change" against "H_1: 10% decrease" and not allow the possibility of a 5% change? Such cases are addressed in the next section. For the present section, we will concentrate on binary choices.

Application: Travel Time Decrease

Consider a situation in which the existing travel time on a given route is known to average 60 minutes, and experience has shown the standard deviation to be about 8 minutes. An "improvement" is recommended that is expected to reduce the true mean travel time to 55 minutes.

This is a rather standard problem, with what is now a fairly standard solution. The logical development of the solution follows. The first question we might ask ourselves is whether we can consider the mean and standard deviation of the initial distribution to be truly known or whether they must be estimated. Actually, we will avoid this question simply by focusing on whether the after situation has a true mean of 60 minutes or 55 minutes. Note that we do not know the shape of the travel time distribution, but the central limit theorem tells us: a new random variable Y, formed by averaging several travel time observations, will tend to the normal distribution if enough observations are taken. The shape of Y for two different hypotheses, which we now

form:

H_0: The true mean of Y is 60 minutes

H: The true mean of Y is 55 minutes

A logical decision rule: if the actual observation Y falls to the right of a certain point, Y*, then accept H_0; if the observation falls to the left of that point, then accept H_1.

Note that:

1. The n travel time observations are all used to produce the one estimate of Y.
2. If the point Y* is fixed, then the only way the Type I and Type II errors can be changed is to increase n, so that the shapes of the two distributions become narrower because the standard deviation of Y involves the square root of n in its denominator.
3. If the point Y* is moved, the probabilities of Type I and Type II errors vary, with one increasing while the other decreases.

To complete the definition of the test procedure, the point Y* must be selected and the Type I and Type II errors determined. It is common to require that the Type I error (also known as the level of significance, α) be set at 0.05, so that there is only a 5% chance of

rejecting a true null hypothesis. In the case of two alternative hypotheses, it is common to set both the Type I and Type II errors to 0.05, unless there is very good reason to imbalance them (both represent risks, and the two risks-repairing some cars needlessly versus having unsafe cars on the road, for instance may not be equal).

Y* will be set at 57.5 min in order to equalize the two probabilities. The only way these errors can be equal is if the value of Y* is set at exactly half the distance between 55 and 60 min. The symmetry of the assumed normal distribution requires that the decision point be equally distant from both 55 and 60 min, assuming that the standard deviation of both distributions (before and after) remains 8 min.

To ensure that both errors are not only equal but have an equal value of 0.05, Y* must be 1.645 standard deviations away from 60 minutes, based on the standard normal table. Therefore, $n \geq (1.645^2)\ (8^2)/2.5^2$ or 2^8 observations, where 8 = the standard deviation, 2.5 = the tolerance (57.5 mph is 2.5 mph away from both 55 and 60 mph), and 1.645 corresponds to the z statistic on the standard normal distribution for a beta value of 0.05 (which corresponds to a probability of $z \leq 95\%$).

The test has now been established with a decision point of 57.5 min. If the "after" study results in an average travel lime of under 57.5 min, we will accept the hypothesis that the true average travel time has been reduced to 55 min. If the result of the "after" study is an average travel time of more than 57.5 min, the null hypothesis- that the true average travel time has stayed at 60 min-is accepted.

Was all this analysis necessary to make the commonsense judgment to set the decision time at 57.5 min half way between the existing average travel time of 60 min and the desired average travel time of 55 min? The answer is in two forms: the analysis provides the logical basis for making such a decision. This is useful. The analysis also provided the minimum sample size required for the "after" study to restrict both alpha and beta errors to 0.05. This is the most critical result of the analysis.

Application: Focus on the Travel Time Difference

The preceding illustration assumed that we would focus on whether the underlying true mean or the "after" situation was either 60 minutes or 55 minutes. What are some of the practical objections that people could raise?

Certainly one objection is that we implicitly accepted at face value that the "before' condition truly had an underlying true mean or 60 minutes. Suppose, to overcome that, we focus on the difference between before and after observations.

The n_1 "before" observations can be averaged to yield a random variable Y_1 with a certain mean μ_1 and a variance of σ_1^2/n_1. Likewise, the n_2 "after" observations can be averaged to yield a random variable Y_2 with a (different?) certain mean μ_2 and a variance of σ_2^2/n_2. Another random variable can be formed as $Y = (Y_2 - Y_1)$, which is itself

normally distributed and that has an underlying mean of ($\mu_2 - \mu_1$) and variance $\sigma^2 = \sigma_2^2/n_2 + \sigma_1^2/n_1$. This is often referred to as the normal approximation.

What is the difference between this example and the previous illustration? The focus is directly on the difference and does not implicitly assume that we know the initial mean. As a result, "before" samples are required. Also, there is more uncertainty, as reflected in the larger variance. There are a number of practical observations stemming from using this result: it is common that the "before" and "after" variances are equal, in which case the total number of observations can be minimized if $n_1 = n_2$. If the variances are not known, the estimators s_i^2 are used in their place.

If the "before" data was already taken in the past and n_1 is therefore fixed, it may not be possible to reduce the total variance enough (by just using n_2) to achieve a desired level of significance, such as $\alpha = 0.05$. Comparing with the previous problem, note that if both variances are 8^2 and $n_1 = n_2$ is specified, then $n_1 \geq 2 \times (1.645^2)\ (8^2)/(2.5^2)$ or 55 and $n_2 \geq 55$. The total required is 110 observations. The fourfold increase is a direct result of focusing on the difference of –5 mph rather than the two binary choices (60 or 55 minutes).

1.8.2 Before-and-After Tests with Generalized Alternative Hypothesis

It is also common to encounter situations in which the engineer states the situation as "there was a decrease" or "there was a change" versus "there was not," but does not state or claim the magnitude of the change. In these cases, it is standard practice to set up a null hypothesis of "there was no change" ($\mu_1 = \mu_2$) and an alternative hypothesis of "there was a change" ($\mu_1 \neq \mu_2$). In such cases, a level of significance of 0.05 is generally used.

The null hypotheses for two possible cases both have a null hypothesis of "no change." The first case implicitly considers that if there were a change, it would be negative-that is, either there was no change or there was a decrease in the mean. The second case does not have any sense (or suspicion) about the direction of the change, if it exists, Note that:

1. The first is used when physical reasoning leads one to suspect that if there were a change, it would be a decrease. In such cases, the Type I error probability is concentrated where the error is most likely to occur, in one tail.
2. The second is used when physical reasoning leads one to simply assert "there was a change" without any sense of its direction. In such cases, the Type I error probability is spread equally in the two tails.

In using the second case often we might hope that there was no change, and really not want to reject the null hypothesis. That is, not rejecting the null hypothesis in this case is a measure of success. There are, however, other cases in which we wish to prove that there is a difference. The same logic can he used, but in such cases, rejecting the null hypothesis is "success."

An Application: Travel Time Differences

Let's assume, we have made some improvements and suspect that there is a decrease in the true underlying mean travel time. Using information from the previous illustration, let us specify that we wish a level of significance a = 0.05. The decision point depends upon the variances and the n_i. If the variances are as stated in the prior illustration and $n_1 = n_2 = 55$, then the decision point Y*= –2.5 mph, as before.

Let us now go one step further. The data is collected, and Y = –3.11 results. The decision is clear: reject the null hypothesis of "there is no decrease." But what risk did we take?

Consider the following:

- Under the stated terms, had the null hypothesis been valid, we were taking a 5% chance of rejecting the truth. The odds favor (by 19 to 1, in case you are inclined to wager with us) not rejecting the truth in this case.
- At the same time, there is no stated risk of accepting a false hypothesis H_1, for the simple reason that no such hypothesis was stated.
- The null hypothesis was rejected because the value of Y was higher than the decision value of 2.5 mph. Since the actual value of – 3.11 is considerably higher than the decision value, one could ask about the confidence level associated with the rejection. The point Y = – 3.11 is 2.033 standard deviations away from the zero point, as can be seen from:

$$\sigma_Y = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

$$= \sqrt{\frac{8^2}{55} + \frac{8^2}{55}} = 1.53$$

and z = 3.11/1.53 = 2.033 standard deviations. Entering the standard normal distribution table with z = 2.033 yields a probability of 0.9790. This means that if we had been willing to take only a 2% chance of rejecting a valid H_0. we still would have rejected the null hypothesis: we are 98% confident that our rejection of the null hypothesis is correct. This test is called the Normal Approximation and is only valid when $n_1 \geq 30$ and $n_2 \geq 30$.

Since this reasoning is a little tricky, let us state it again: If the null hypothesis had been valid. you were initially willing to take a 5%, chance of rejecting it. The data indicated a rejection. Had you been willing to take only a 2% chance, you still would have rejected it.

One-Sided Versus Two-Sided Tests

The material just discussed appears in the statistics literature as "one-sided" tests, for the obvious reason that the probability is concentrated in one tail (we were considering only a

decrease in the mean). If there is no rationale for this, a "two-sided" test should be executed, with the probability split between the tails. As a practical matter, this means that one does not use the probability tables with a significance level of 0.05. but rather with 0.05/2 = 0.025

1.8.3 Other Useful Statistical Tests

The t-Test

For small sample sizes (N < 30), the normal approximation is no longer valid. It can be shown that if x_1 and x_2 are from the same population, the statistic t is distributed according to the tabulated, distribution, where:

$$t = \frac{x_1 - x_2}{S_p\sqrt{1/n_1 + 1/n_2}} \qquad \text{…..(1.16)}$$

and S_p is a pooled standard deviation. which equals:

$$S_p = \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}} \qquad \text{…..(1.17)}$$

The t distribution depends upon the degrees of freedom, f, which refers to the number of independent pieces of data that form the distribution. For the t distribution, the non-independent pieces of data are the two means, x_1 and x_2. Thus:

$$f = n_1 + n_2 - 2 \qquad \text{…..(1.18)}$$

Once the t statistic is determined, the tabulated values, yield the probability of a t value being greater than the computed value. In order to limit the probability of a Type I error to 0.05, the difference in the means will be considered significant only if the probability is less than or equal to 0.05 that is, if the calculated t value falls in the 5% area of the tail, or in other words, if there is less than a five percent chance that such a difference could be found in the same population. If the probability is greater than 5% that such a difference in means could be found in the same population, then the difference would be considered not significant.

The F-Test

In using the t-test, and in other areas as well, there is an implicit assumption made that the $\sigma_1 = \sigma_2$. This may be tested with the F distribution, where:

$$F = \frac{S_1^2}{S_2^2} \qquad \text{….(1.19)}$$

(by definition the larger s is always on top)

It can be proven that this F value is distributed according to the F distribution. The F distribution is tabulated according to the degrees of freedom in each sample, thus $f_1 = n_1 - 1$ and $f_2 = n_2 - 1$. Since the f distribution in the shaded area in the tail, like the t distribution, the decision rules are as follows:

- If [Prob F ≥ F] ≤ 0.05, then the difference is significant.
- If [Prob F ≥ F] > 0.05, then the difference is not significant.

The f distribution is tabulated for various probabilities, as follows, based on the given degrees of freedom: thus:

when p = 0.10, F = 1.94
p = 0.05, F = 2.35
p = 0.025, F = 2.77

The F values are increasing, and the probability [F ≥ 1.56] must be greater than 0.10 given this trend; thus, the difference in the standard deviations is not significant. The assumption that the standard deviations are equal, therefore, is valid.

Chi-Square Test: Hypotheses or an Underlying Distribution f(x)

One of the early problems stated was a desire to "determine" the underlying distribution. such as in a speed study. The most common rest to accomplish this is the Chi-square (χ^2) goodness-of-fit test.

In actual fact, the underlying distribution will not be determined. Rather, a hypothesis such as "H_0: The underlying distribution is normal" will be made, and we test that it is not rejected, so we may then act as if the distribution were in fact normal.

The procedure is best illustrated by an example. Consider data on the height of 100 people. To simplify the example, we will test the hypothesis that this data is uniformly distributed (i.e., there are equal numbers of observations in each group).

In order to test this hypothesis, the goodness-or-fit test is done by following these steps:

1. Compute the theoretical frequencies, f_i, for each group. Since a uniform distribution is assumed and there are 10 categories with 100 total observations, $f_i = 100/10 = 10$ for all groups.
2. Compute the quantity:

$$\chi^2 = \sum_{i=1}^{N} \frac{(n_i - f_i)^2}{f_i} \qquad \text{…..(1.20)}$$

3. As shown in any standard statistical text, the quantity χ^2 is chi-squared distributed, and we expect low values if our hypothesis is correct. (If the observed samples exactly equal the expected, then the quantity is zero.) Therefore, refer to a table of the chi-square distribution and look up the number that we would not exceed more than 5% of the time (i.e., $\alpha = 0.05$). To do this, we must also have the number of degrees of freedom, designated df, and defined as df = N – 1 – g, where N is the number of categories and g is the number of things we estimated from the data in defining the hypothesized distribution. For the uniform distribution, only the sample size was needed to estimate the theoretical frequencies, thus "0"

parameters were used in computing χ^2 (g = 0). Therefore, for this case, df = 10 – 1 – 0 = 9.

4. Now entered with $\alpha = 0.05$ and df = 9. A decision value of $\chi^2 = 16.92$ is found. As the value obtained is 43.20 > 16.92, the hypothesis that the underlying distribution is uniform must be rejected.

In this case, a rough perusal of the data would have led one to believe that the data were certainly not uniform, and the analysis has confirmed this. The distribution actually appears to be closer to normal, and this hypothesis can be tested. Determining the theoretical frequencies for a normal distribution is much more complicated, however. An example applying the χ^2 test to a normal hypothesis is discussed in detail in chapter 9. Note that in conducting goodness-of-fit tests, it is possible to show that several different distributions could represent the data. How could this happen? Remember that the test does not prove what the underlying distribution is in fact. At best, it does not oppose the desire to assume a certain distribution. Thus, it is possible that different hypothesized distributions for the same set of data may be found to be statistically acceptable. A final point on this hypothesis testing: the test is not directly on the hypothesized distribution, but rather on the expected versus the observed number of samples. The computations involve the expected and observed number of samples; the actual distribution is important only to the extent that it influences the probability of category, That is, the actual test is between two histograms. It is therefore important that the categories be defined in such a way that the "theoretic histogram" truly reflects the essential features and detail of the hypothesized distribution. That is one reason why categories of different size are sometimes used: the categories should each match the fast-changing details of the underlying distribution.

1.9 Summary

In transportation engineering the specifically traffic studies cannot be done without using some basic statistics. While using statistics, the engineer is often faced with the need to act despite the lack of certainty, which is why statistical analysis is employed. This chapter is meant to review basic statistics for the reader to be able to understand and perform everyday traffic studies before entering to the main and detailed statistics applications to traffic and transportation engineering.

There are a number of very important statistical techniques presented in this book in further chapters that are useful for traffic engineers and transportation engineers.