CHAPTER 12

INFERENTIAL STATISTICS:

PARAMETER ESTIMATION

AND

HYPOTHESIS TESTING

Descriptive statistics play an important role in data analysis when either samples or population are studied. Nevertheless, the descriptive treatment of data focuses only upon the entity (or entities) being confronted. Frequently the analyst desires to go beyond the data at hand and generalize to some broader phenomenon. This implicit inferential process is the bedrock of statistical induction. In brief, the purpose of statistical inference is to study a sample of elements drawn from the population and from the knowledge of the sample alone make an intelligent generalization about the universe of elements from which the sample was selected. In this section we will look at the underlying logic of parameter estimation, one branch of statistical inference. In the next section the other branch of inferential statistics will be pursued, namely, hypothesis testing.

To understand the rationale of parameter estimation (and hypothesis testing) a knowledge of sampling theory and probability theory is indispensable. Suppose our task is to estimate the proportion of college students at University A that participate in intercollegiate athletics. We decide to use a sampling fraction (proportion of the population sampled) of 5%. If the university enrolled 20,000, it would be necessary to select 1,000 students. How are the sample elements, i.e., students, chosen? We decide to conduct the survey in one large building of the campus complex. Unbeknown to us we choose the physical education building. Suppose 750 of the 1,000 interviewed individuals indicate they participate in intercollegiate athletics. In proportion terms, .75 (or 75%) of the respondents responded affirmatively to the sports participation query. In inferential statistics the crux of the matter is not to simply determine the sample outcome, instead, the sample is instrumental in telling us about the entire set of elements. In other words, we use the few, i.e. , sample, to predict the many, i.e., population.

Parameters and Statistics.

A "good" sample is representative of the population from which it is chosen. On the basis of our sampling findings we would expect, assuming the mini-version

were representative, that somewhere in the neighborhood of .75 (or 75%) of all university students engaged in intercollegiate athletics. It quickly becomes apparent that this projected population figure is out of line. While we are not certain we're very, very sure that the true proportion--the parameter we're attempting to estimate--is not near the .75 (or 75%) sample statistic. The purpose of parameter estimation is to estimate a population parameter--a numerical indicator of the entire population's characteristic(s)--on the basis of a sample statistic--a numerical characteristic determined by the outcomes of a sampling experiment. Since it is important to differentiate population and sample characteristics, it is conventional to let Greek letters, e.g., θ (theta), σ (sigma), μ (mu), represent population values and Roman (or Latin) letters, e.g., t, s, $\bar{X}$, sample values.

Probability and Non-probability Sampling.

Since our hypothetical estimation problem involved estimating the proportion of all students who participated in collegiate athletics on the basis of a subset of students, we may raise the question, "how good an estimate is p (sample proportion) of π (population proportion)? The sample estimate is not a good estimator because of the sampling design employed.

As we've seen there are two general bodies of sampling techniques: 1) probability, and 2) non-probability approaches. Within each of these broad categories are further subdivisions, e.g., simple random, stratified, and cluster sampling are components of probability sampling although differences exist among them in terms of how the sample is selected, how much information one has about the universe, etc; purposive, quota, and accidental samples are components of the non-probability approach. For purposes of expositing the sampling statistics logic the specific varieties of each need not concern us, rather the critical difference between the two is sufficient. The virtue of probability sampling is that each potential sample element in the universe has an equal or known probability of being selected. This is abbreviated "EPSEM" meaning "equal probability of selection method". In contradistinction, the non-probability sampling approach can not make such

a claim. Under the latter case, the investigator simply samples elements that are convenient to do so until the sample reaches the designated or desired size.

In comparing probability and non-probability sampling techniques it is not so much that the latter is automatically unrepresentative as it is that one remains ignorant of its representativeness. More often than not, however, such a sample ends up being biased because probability theory can not come to its rescue. The hypothetical sample alluded to above is a non-probability sample. The sample proportion is biased because the sampling procedure was faulty. If one attempts to estimate the proportion of an entire university engaged in a particular activity, like sports, one does not draw the entire sample, however large, solely from the physical education building.

Sources of Error in Sampling

There are two important sources of error in sampling.[2] First, what is termed systematic bias or error accrues in investigations when the sampling design is faulty (as the case is here). Second, is what is termed random, chance, or probability error, what statisticians simply call random sampling error. Even in randomly drawn samples this second type of error occurs. Systematic bias can be virtually eliminated by employing appropriate sampling techniques, which generally means some form of probability sampling. Sampling error can not be eliminated even in random sampling but it can be measured. It is the fact that we can quantitatively assess the magnitude of sampling error that makes probability samples so important in inferential statistics. Let us turn to how we might select a probability sample and measure the degree of sampling error.

Using the simple random sample (SRS) as our model we know that, by definition, it is a sample that gives each element in the universe an equal probability of selection. The mechanics of such a procedure might be described as follows. Each element in the universe is given a unique number from 1 to N. For example, if a student roster is available (assuming the roster is exhaustive and accurate) we might number each student from 00001 to 20,000.

Then we select students until the desired sample size is reached. The particular elements selected may be obtained by consulting a table of random digits and using a field of five columns (since 20,000 would take up that many columns) select those persons whose number appears as we move up and down the column headings. If our sampling fraction were 5%, after we randomly selected 1,000 persons we would have to locate them. Note that random sampling requires more work (particularly in trying to track down students) but has so many advantages that the time, energy, and cost usually outweigh the deficits. The use of a table of random numbers provides an operational measure of a random sample, that is, some elements (like those housed in the pe department) do not have a greater chance of being selected; rather, each member of the entire student body has the same probability of being included in the sample.

Measuring Sampling Error: Central Tendency, Dispersion, and Form of the Sampling Distribution.

How is sampling error measured and how is it used to the advantage of the researcher? The answer to these queries is not only ingenius but necessitates a discussion of the concept sampling distribution and the central tendency, dispersion and form of the sampling distribution. The concept of sampling distribution is a theoretical construct but the empirical generation of one should impregnate its statistical import. Recall that the purpose of statistical estimation is to estimate the population characteristic of interest on the basis of a sample. We do not know the parameter values but attempt to provide reasonably good estimators of them. For illustrative purpose we will assume that we do know the population parameters and reveal how probability sampling allows us to estimate what these values are. Our example will necessarily be simplistic for didactic purpose but the same logic applies regardless of the complexity of the matter. Our hypothetical universe contains five individuals with salaries of $10,000, $12,000, $14,000, $16,000 and $18,000 whose mean annual salary we want to estimate on the basis of a sample of size two ($n=2$).

The population parameter, μ, is \$14,000 and the population parameter, σ, is \$2828.43. Usually only the central tendency and dispersion of the sampling distribution is necessary to know in parameter estimation.

Now, if we take all possible samples of size two from this hypothetical universe, we can compute the arithmetic mean and standard deviation of this distribution. A sampling distribution is the distribution of a sample statistic that would occur if all possible samples of a given size were taken from a fixed universe. A sample statistic, as used in the description above, is a summary statistical index, like the mean, and does not directly refer to individual raw score values. To conceptualize this distribution we will construct a matrix of sample statistics for this example. This matrix contains each and every possible combination of samples of size two from a universe of five and is presented in Table 12.1

TABLE 12.1

SAMPLING DISTRIBUTION OF ARITHMETIC MEANS FOR SAMPLES OF SIZE TWO FROM A UNIVERSE OF FIVE ELEMENTS in \$1,000's

	10	12	14	16	18
10	10	11	12	13	14
12	11	12	13	14	15
14	12	13	14	15	16
16	13	14	15	16	17
18	14	15	16	17	18

The concept of sampling distribution implies sampling with replacement, i.e., each selected element is thrown back into the population from which it was drawn after it has been selected. This contrasts with sampling without replacement which means that the particular element chosen is not returned for potential re-selection. Ordinarily if the sample is large, like our sample of 20,000 students, it makes little difference which sampling type is employed.

In fact, sampling without replacement is often used when the universe is large. Speaking in probabilistic terms this does not alter the probability of elements being chosen since, for example, the probability of selecting the first person in a universe of 20,000 is 1/20,000 or .00005. If the element is not returned to the population pool the probability of the second element is 1/19,999 or .0000500025; the third is 1/19,998 or .000050005; et cetera. Practically speaking, with a universe of this size it doesn't make a great deal of difference since by the time the 1,000th person is selected the probability has become 1/19,000 or .0000526316.

Notice that the entries in Table 12.1 are sample statistics, all possible sample arithmetic means of samples of size two. A sampling distribution is a special kind of frequency distribution and earlier we said that the salient statistical features of a univariate frequency distribution are central tendency, variability, and form. Computing statistical indices for each of these properties of the sampling distribution will provide us with a very meaningful statistical lesson regarding parameter estimation.

The Mean of the Sampling Distribution

Conceptually, the mean of the sampling distribution, symbolized $\mu_{\bar{X}}$, might be thought of as a mean of means. Here it is the arithmetic average of all possible sample means of size two from a population of 25. To compute $\mu_{\bar{X}}$ we may sum the twenty-five entries (which are means) in Table 12.1 and divide by 25. Doing this we have:

$$\mu_{\bar{X}} = \Sigma \bar{X}_i / N = 350/25 = 14$$

If we compare the mean of the sampling distribution of means with the previously calculated population mean we have:

$$\mu_{\bar{X}} = \mu = \$14$$

This identity suggests that the sample mean is an unbiased estimate of the population mean because the mean of the sampling distribution equals μ.

Notice that it is not true to say that any single sample mean is equal to μ because it obviously is not. Probability theory apropos parameter estimation entails a large number of occurrences, not a single one. Moreover, it is our knowledge of the sampling distribution of a statistic, the mean in this case, not a specific mean or the universe's characteristics (which we generally do not know anyway) that provides the basis for the legitimacy of our inference.

The Standard Deviation of the Sampling Distribution.

Conceptually, the standard deviation of the sampling distribution, symbolized $\sigma_{\bar{X}}$, indicates the degree of variability of sample means around the parameter mean. It is a special type of standard deviation, the standard deviation of the distribution of sample means. To calculate $\sigma_{\bar{X}}$ we have:

$$\sigma_{\bar{X}} = \sqrt{\frac{\Sigma \bar{X}_i^2 - (\Sigma \bar{X}_i)^2/N}{N}} = \sqrt{\frac{5000 - (350)^2/25}{25}} = 2$$

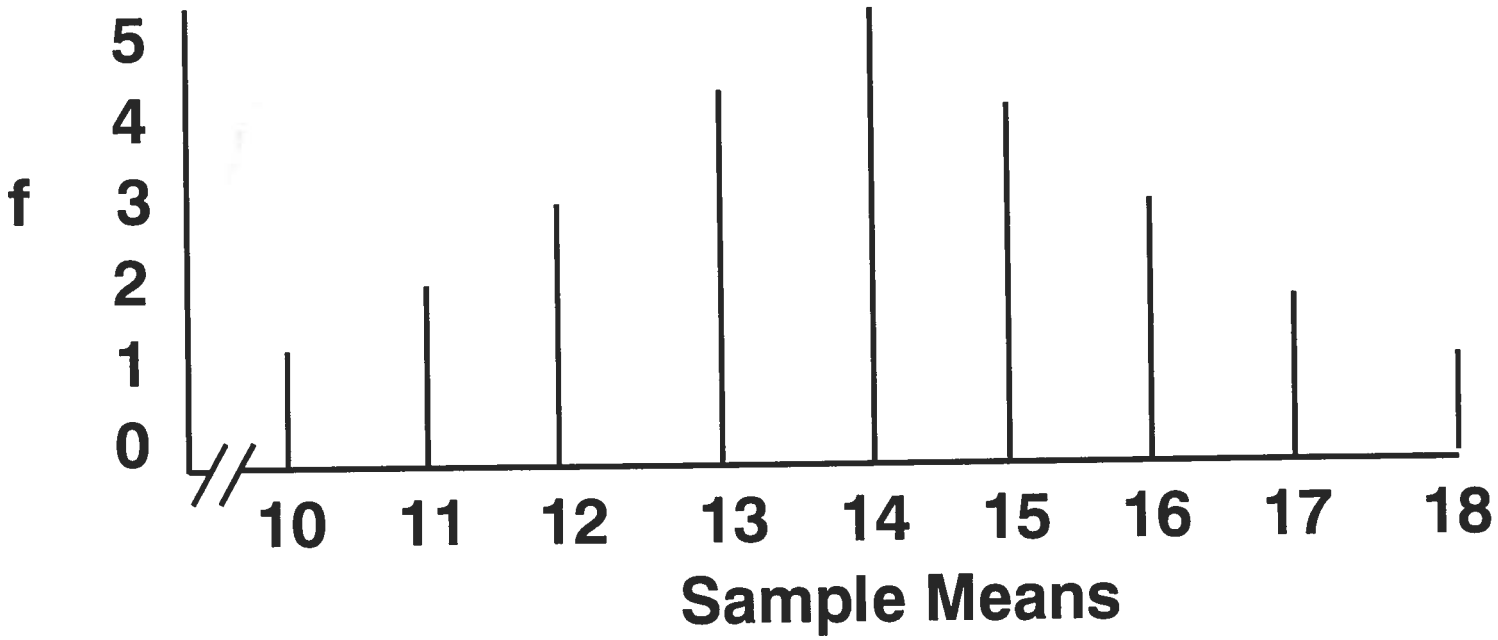
If we now compare $\sigma_{\bar{X}}$ with σ we notice a slight discrepancy, namely, $\sigma_{\bar{X}}$ is an underestimate of the actual standard deviation of the population. Mathematically this virtually always occurs and leads statisticians to contend that the sample standard deviation is a biased estimate of the population standard deviation because the standard deviation of the sampling distribution of means is less than the population standard deviation. Because of this the denominator in the ultimate estimation formula will be n-1 to correct for this known discrepancy.

The Form of the Sampling Distribution

Form or shape implies two statistical indices: 1) skewness and 2) kurtosis. The construction of a histogram (or polygon) enables us to visually capture the shape of the distribution and describe these properties. Let us construct a bar graph (see Figure 12.1) for the quantities in Table 12.1 Notice that the form of their distribution takes on the appearance of the normal curve.

FIGURE 12.1

FREQUENCY GRAPH OF SAMPLE MEANS FROM TABLE 12.1



Statistical Theorems

Having discussed the mathematical properties of a sampling distribution, a pivotal concept in inductive statistics, let us directly raise the query, "How does a knowledge of the central tendency, dispersion, and form of the sampling distribution provide a sound basis for estimating population parameters?" Or saying it differently, "What criteria are used in judging the accuracy of a sample estimate when the population values may never be known?" A couple of fundamental statistical theorems interwoven with the development of the previous concepts will provide the statistical answer.

For the ensuing discussion we are assuming random sampling. Conceptualizing the distribution of sample statistics as a probability distribution it is sensible and rational to assume that the probability of selecting a sample of size two with a mean of $10,000 to be 1/25 or .04. In other words, the probability in selecting at random, a couple of individual observations with a mean of $10,000 would be .04. Common sense as well as empirical verification combines to suggest that the larger the sample size, the more individual elements included in the sample, the closer the correspondence between the sample estimator and the population parameter. In more formalized terms this observation is known as the law of large numbers and states: as the sample size n increases, the sample statistics, e.g., $\bar{X}$ and s, will approximate more and more closely the population parameters, μ and σ.

The Central Limit Theorem

Our previous discussion of central tendency, variability, and form of the sampling distribution is succinctly exposited in the central limit theorem (CLT) which describes what happens to the mean, standard deviation, and form of the sampling distribution as n, the sample size, becomes larger and larger (what the mathematician calls "as the sample size is increased without limit").[3] The CLT has three ramifications:[4]

1) the mean of the sampling distribution, $\mu_{\bar{X}}$, coincides with the mean of the population, μ. That is, $E(\bar{X})=\mu$.

2) the standard deviation of the sampling distribution, $\sigma_{\bar{X}}$, is computed $\sigma/\sqrt{n}$. This expression says two things. First, the standard deviation of the sampling distribution, $\sigma_{\bar{X}}$, is proportional to the population standard deviation. Secondly, the standard error is also a function of n, sample size, and as n becomes larger the corresponding standard error becomes smaller. But, $E(s)\neq\sigma$, making it a biased estimate.

3) the shape of the sampling distribution tends to normality, (see Figure 12.1) i.e., takes on the appearance of the normal curve, as n is increased regardless of the shape of the population from which the sample was derived.

In a nutshell the central limit theorem states that as n becomes large the shape of the sampling distribution tends to normality with mean, $\mu_{\bar{X}}$, equal to μ and standard deviation, $\sigma_{\bar{X}}$, equal to $\sigma/\sqrt{n}$.

Point Estimates and Interval Estimates

There are two kinds of parameter estimates. When a single value is used to estimate the corresponding population parameter it is said to be a point estimate. When a range of values within which you estimate the population parameter to lie is used it is called an interval estimate. For all practical purposes estimating population proportions, i.e., the proportion in the population that does something, votes a certain way, ad infinitum, and population means, i.e., the mean income, height, age, etc. of the population, are the only two types of parameter estimations with which we need to deal. In this section the logic, computation, and interpretation of point estimates and confidence intervals for a proportion and a mean will be highlighted using the previous discussion as a backdrop.

Point Estimates are popular in the mass media basically because of (apparent) ease of interpretation. To illustrate, at present a point estimate for the proportion of unemployed workers is .07 (or 7%: .07x100) while a point estimate for the mean income of Americans is $53,000. These single value estimates of the population of unemployed workers and the population mean income are fairly easy to digest but provide no estimate of probable error (sampling error in statistical vernacular). Returning to our earlier examples of estimating the proportion of students participating in intercollegiate athletics and the (arithmetic) mean income of five individuals, the proportion of .75 (750 out of 1000 students indicated affirmation of the query) and the mean of $11,000 (resulting from randomly selecting (two) individuals with annual earnings of $10,000 and $12,000) are single-value estimates. Both these statistical indices, .75 and $11,000, are point estimates--single values--computed from a random sample of elements in the respective universes. Statisticians prefer interval estimates to point estimates because the former indicate the degree of accuracy or the range within which the actual parameter values are likely to fall. Here's how interval estimates, called confidence intervals, are constructed.

Interval estimates, like point estimates, are attempts to estimate an unknown population parameter we'll call θ, theta. A random sample is selected from the universe and an estimator of θ, called $\hat{\theta}$, (theta hat) a sample statistic, is computed. This randomly selected sample value, $\hat{\theta}$, may be thought of as a random variable which is an estimator of θ. Like any random variable, the sample estimate has a probability distribution, described in terms of central tendency, dispersion, and form. Statisticians distinguish between unbiased and biased estimators. For example, both $\bar{X}$ and Md are unbiased estimates of the population central tendency, i.e., $E(\bar{X})$ and $E(Md)=\mu$. However, the mean is preferable because it is more efficient than the median, that is, there is less variability of sample means about the population mean than is the case with the median. In statistical notation, $s^2_{(\bar{X})} < s^2_{(Md)}$.

On the other hand, the sample variance is a biased estimator of the population variance, because E (s^2) $\neq \sigma^2$; instead, the E(s^2) = $\sqrt{n/n-1(\sigma^2)}$. Therefore, in calculating the variance a modified formula is used (note the correction factor of n-1) in the denominator . An interval estimate for a parameter may be written as follows:

$$1 - \alpha \text{ confidence limits for } \theta \text{ are } \hat{\theta} \pm k \text{ SE } (\hat{\theta})$$

This equation means that the 95 and 99% confidence limits (1-.05=.95 and 1-.99 = .01) for a parameter, θ, are determined by taking the estimate of the parameter, labeled theta hat ($\hat{\theta}$), adding to and subtracting from the point estimate a constant (k) according to the desired confidence level. The standard scores z=1.96 and a z=2.58 dissect the normal curve so that 95% and 99% of its area are contained within plus and minus values of that magnitude, times the standard error (SE) of the sample estimator ($\hat{\theta}$). A couple of illustrations will demonstrate how the formula works.

Estimating a Population Mean

Suppose a social welfare agency of a large metropolitan community wishes to estimate the family income of persons living in the inner-city. While it would be possible to survey the entire geographical area it would not be practical to do so. Hence, a random sample of 36 individuals is drawn and each of the persons is consulted and requested to supply the agency their gross yearly income. The arithmetic mean is computed to be $6,380 and the standard deviation, $1,100. The agency is attempting to estimate μ, the population yearly income and the sample mean is used to make such an inference. To construct the 95% confidence limits the formula below is made specific for estimating the parameter μ, therefore:

$$\bar{X} = k\ (\sigma/\sqrt{n})$$

Since we don't know σ, the population standard deviation, it is estimated from the sample standard deviation employing the corrected formula (for known bias):

(s)$\sqrt{n-1}$. Substituting the present values into the formula we have:

$= s/\sqrt{n-1}$ $\quad$ $\$6380 \pm 2.58\ (185.94)$

$= 1100/\sqrt{35}$ $\quad$ $\$6380 \pm 479.72$

$= 185.94$ $\quad$ $\$5900.28$ to $\$6859.72$

Before interpreting the 99% confidence interval let us compute the 95% confidence limits. The only change is in k, in which case it becomes 1.96 for 1-.05 =.95. Substituting we have:

$$\$6380 + 1.96\ (185.94)$$

$$\$6380 \pm 364.44$$

$$\$6015.56 \text{ to } \$6744.44$$

The 95% confidence limits for these data extend from $6,015.56 to $6,744.44.

What do these confidence levels mean? How are they interpreted? For one thing, the parameter one is trying to estimate is or is not within the constructed interval of values. It can't be 95%(or 99%) in and 5% (or 1%) out! Recalling that statistical principles apply to long run events as well as large numbers of events--not single occurrences--a confidence interval is interpreted to mean that if repeated random samples were selected and a range of values constructed around each point estimate, we would expect about 95% (or 99%, depending on the confidence level selected) of the confidence intervals to contain the parameter value we're trying to estimate.

Notice as one's confidence increases from 95% to 99% the interval becomes larger because of the affect of k upon SE. In other words, one can be increasingly confident but at the expense of a wider band of values within which the parameter θ may actually lie.

Our confidence or "faith" resides not in any particular interval, but in the procedure which is based upon statistical/mathematical principles like the central limit theorem and the law of large numbers.[5]

Estimating a Population Proportion.

Suppose our concern is to estimate the proportion of potential voters who will vote for a particular Presidential candidate. Call this population parameter π. As is the case with any parameter estimate we draw a random sample, compute a statistic purportedly estimating the population value and construct an interval around it. A sample of 1,000 is taken and 550 of the sample indicate a preference for candidate X. To convert the frequency into a proportion we divide f by N or 550/1,000 = .55. The specific manifestation of the generic confidence interval formula becomes:

$$p \pm k\,(\sigma_p) = .55 \pm 1.645\left(\sqrt{pq/n}\right) = .55 \pm 1.645\left(\sqrt{(.55)(.45)/1000}\right)$$

According to the parameter estimate about 90% of the time the expectation would be that the population parameter lies between .524 and .576. Pragmatically speaking, the election should favor X although a "sure" winner would not be completely possible to prognisticate.

Summarizing the inferential technique known as parameter estimation it can be said that the purpose is to determine what samples, particular sample statistics which are estimators of population values, can tell us about populations, particularly population parameters. There are point estimates and interval estimates which may be made for either population means or population proportions. The logic of statistical estimation is achieved by understanding and interrelating the concepts of sampling distribution and its properties (central tendency, dispersion, and form), and the central limit theorem and law of large numbers. These same concepts provide the rationale for the other branch of inferential statistics, namely, hypothesis testing.

Hypothesis Testing

The purpose of hypothesis testing is to use statistical knowledge and reasoning to make decisions in the face of uncertainty. This contrasts with the other side of inferential statistics, parameter estimation, in which the purpose is to estimate population parameters.

In testing hypothesis we use what are termed tests of significance to help us make a decision as to whether or not chance factors--random sampling error--could have produced the obtained results. The basic logic behind hypothesis testing is something that most of us have been doing our entire lives, although it is much more formalized than that which we use in everyday affairs.[6] Let us begin by providing a link between your intuitive use of decision making under uncertain conditions and that of hypothesis testing.

Hypothesis Testing and Intuition

The fundamental concern in hypothesis testing is to make a decision concerning the likelihood that an event will occur given certain assumptions about the existing state of affairs. Informally we utilize this logic practically every day. For example, you ask a charming individual for a date and much to your dismay the person refuses you. The essential question is: "How likely is it that this individual would refuse you if you were liked?" With a sample of one (one request denied) we can't make too much of the evidence but suppose a second, third, . . . sixth time the person doesn't accept your solicitation for a date. You ask yourself, "How likely is it that I'd be turned down six times assuming the other party was interested? As the law of large numbers implies, your answer to the question could probably be answered with a bit more certitude, although, of course, you can never be absolutely sure. Probably you would want to reject your original assumption that the person cared for you, reasoning that it is highly unlikely that someone would deny you a social occasion six consecutive times if they were truly interested in your company.

Statisticians in the course of hypothesis testing formalize this kind of thinking and provide precise numerical cut off points for accepting or rejecting one's assumption about the initial state of affairs. Consider the following example.[7]

A large industrial firm has a job training program to train new employees. Past data have indicated that it takes an average (arithemetic mean) of 24 days to accomplish this on-the-job training program with a variability index (standard deviation) of 1.8 days. The company is contemplating purchasing new equipment which is quite costly but wantsto be sure that the new materials significantly reduce the length of the training program. A random sample of 30 new employees serves as an experimental group for determining the effectiveness of the new appratuses. This new group of recruits reaches maximum productivity in 19.6 days. The board of trustees wants to know if 19.6 days (mean) is significantly different from 20 days. In other words, the board decides that unless the mean number of training days is 20 or less the equipment would not be financially pragmatic to buy. In hypothesis testing the following steps are taken:

1. The hypothesis to be tested must be phrased in such a manner that evidence can be brought to bear enabling it to be accepted or rejected. The reason the hypothesis cannot be directly substantiated is that it is much easier to disprove an hypothesis than to prove it. For example, if a coin is tossed 100 times and 55 heads are flipped, do we have sufficient evidence to reject the assumption that the coin is unbiased. The answer is probably not, even though theoretically we'd expect 50 heads and 50 tails in 100 coin flips. What if 80 heads occurred on 100 tosses? Is the coin fair? While it is possible to obtain such an outcome it is unlikely if the coin were, in fact, unbiased. In short, it is easier to suggest that probably the coin isn't "honest" than to conclude that it is "honest".

Statistically speaking, we achieve this goal by formulating what is called the null hypothesis, symbolized H_0, an hypothesis one ordinarily wants to reject or "nullify". For the present example, H_0 would be expressed as:

$$H_0: \theta = \theta_0$$

$$H_1: \mu = 20$$

Notice that the null hypothesis is made with regard to parameters. In other words a particular sample is not our concern, rather our interest is with the probability that the sample reflects the actual state of affairs in the larger population from which it has been selected.

The counter hypothesis, the hypothesis against which we compare the null, is called the alternative hypothesis, symbolized H_1 (or H_A). The alternative hypothesis can take on three possible forms. Generically the three possibilities are: 1) $\theta \neq \theta_0$ (this is called a two tailed hypothesis or two tailed test of significance); 2) $\theta < \theta_0$ (this is called a left-tail test in which you propose that the parameter is less than a certain value); and 3) $\theta > \theta_0$ (this is called a right-tail test in which you propose to test whether or not the parameter is greater than a certain value). Sometimes these latter two conditions are called one tailed (or directional) hypothesis, in contrast to two tailed (or non-directional) hypothesis. In order to decide which of the three alternative hypotheses is most sensible the consequences of our decision must be considered. The figure below summarizes the consequences.

If H_1 is:	action taken when H_0 is rejected
$\mu > 20$	not buy (ok)
$\mu < 20$	buy (mistake)

If the alternative hypothesis is, in fact, $\mu > 20$, and we rejected the null, the company would decide not to purchase the new equipment. Such a decision would be ok because the criterion set by the company, namely, that only if the new materials reduced the training period to 20 days would the expenditures be justified. On the other hand if H_1 were $\mu < 20$, and we rejected H_0, the decision would be to buy the equipment but this would be a mistake from the criterion set. Therefore the decision is to use a one-tailed test, particularly a left tail test.

2) The second step is to decide upon a level of significance or p value at which the data will be tested. In rhetorical form the level of significance issue raises the question, "How big a risk is one willing to take in rejecting H_0 when it is correct?" Again in visual form we may consider the following chart which contrasts the conditions of H_0 with the decision to accept or reject it.

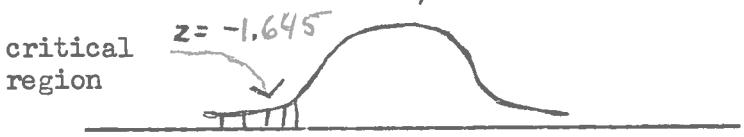
Actual State of Affairs regarding H_0	Decision Regarding H_0: Accept	Reject
true	ok	type I error (α)
false	type II error (β)	ok

As can be seen from the chart, if H_0 is true and we accept it, or when H_0 is false and we reject it the decision is correct, ("ok"). But, if H_0 is true and rejected we make an error (called a type I or alpha error) or if H_0 is false and we accept it we also make an error (called a type II or beta error). Since type II errors are more complex and because statisticians prefer to avoid type I errors it is the type I error we wish to establish as the major decision criterion.

Common values of alpha are .05, .01. and .001. These levels of significance are really policy, not statistical, matters per se. The chosen level depends upon the consequences of making the errors themselves. Substantively an alpha of .05 means the researcher is willing to take the risk of rejecting H_0 when H_0 is true 5% of the time. If this risk is too great then either the .01 or .001 levels of significance may be employed since the probability of type I errors is reduced to 1% of .1%, respectively. Alpha and beta errors are inversely related, you can reduce one of them only at the expense of increasing the other. For example, in the present illustration the consequences of a type I error means that the company will waste money in purchasing the equipment; the consequences of making a type II error mean that productivity will be lost. Couching alpha and beta errors in dollars and cents terms serves to underscore the practical policy making nature of these decision criteria.

3) The third step is to select a test statistic θ (which is an estimate of the parameter). It is imperative that the distribution (what we've previously called the sampling distribution of the test statistic) of θ be known when the null hypothesis is true because it is only when this condition is met that any possibility of rejecting H_0 exists. Then we proceed in this fashion. The sampling distribution of the test statistics θ is divided into two parts: 1) the region of acceptance (of H_0), and 2) the region of rejection (of H_0). If θ falls in the critical region, the region of rejection, H_0 is automatically rejected. The statistician reasons that if the null hypothesis were true and a value of this magnitude occurs, it is highly improbable, although possible, that the initial statement of the null was correct. As we indicated in step two there is a remote possibility, specified by the level of significance, that our reasoning is erroneous and H_0 is, in fact, correct. However, we play the odds and 95% vs. 5% is probably a sufficient betting strategy.

4) The fourth step involves the decision criteria. The diagram below is the sampling distribution of the test statistic. Five percent (.05 of the sampling distribution's area) of the curve is blocked off on the left. If the test statistics value falls in that area H_0 will be rejected. In other words the null hypothesis will be rejected if: $\bar{X} - \mu/\sigma/\sqrt{n} < z = .05$



5) The final procedure is to compute the test statistic's value. The relevant information for performing the test is:

$H_0 : \mu = 20$ $\qquad \alpha = .05$

$H_1 : \mu < 20$ $\qquad z_\alpha = -1.645$

Therefore,

$$\frac{19.6 - 20.0}{1.8/\sqrt{30}} = -1.33$$

Does a z=-1.33 lie in the region of rejection? It does not and we are obliged to retain the null hypothesis. This means that we have not accumulated sufficient information to believe that the new training equipment achieves the policy level decision necessary warranting the company to purchase it. This test is called a single sample test of significance in which a sample statistic is compared with known population parameters. Often the parameters are not known nor is a single sample test practical; under these circumstances a two sample test of the differences between means may be employed.

Suppose we have two groups of students, social science majors and physical science majors, who take a current events test. A random sample of 50 (n_1) social science majors is selected with the social science students achieving a mean ($\overline{X_1}$) = 500 and a standard deviation (s_1) = 21. A random sample of 50 (n_2 = 50) physical science majors is drawn and their mean ($\overline{X_2}$) is 490 with a standard deviation (S_2) of 19. The question becomes: "Is there a significant difference between the two mean scores on the current events test?"

The null hypothesis is that the mean of social science students and physical science students is equal. The alternative hypothesis is that the means are different. Recalling that H_0 and H_1 are always specified in terms of parameter values we have:

$$H_0: \mu_1 = \mu_2$$

$$H_0: \mu_1 \neq \mu_2$$

The .05 level of significance is set. The test statistic is that of the differences between means symbolized, $\bar{X}_1 - \bar{X}_2$. The standard error of the difference between means is computed to be $\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$.

The sampling distribution, the normal curve, is divided into two regions of rejection since the alternative hypothesis simply predicts a difference without indicating the direction (or tail) of the difference. Substituting the present values we have:

(Table A in Appendix)

$$\frac{500 - 490}{\sqrt{\frac{21^2}{50} + \frac{19^2}{50}}} \quad \frac{= 10}{4} = 2.5$$

Does a z= 2.5 lie in the critical region? It does and therefore we reject the null hypothesis and conclude that there is a statistically significant difference between the mean current events scores of social and physical science majors.

The Chi Square Test.

In the previous section we discussed two interval-ratio level tests of significance, the: 1) z single sample test, and 2) z two sample test. In this section we will consider two nominal level tests, the: 1) chi square single sample test, and 2) chi square two sample test. Chi square is one of the most versatile and widely used tests of statistical significance when data are nominal in nature.

The Chi Square Single Sample Test. The chi square single sample test is also known as a test of the goodness-of-fit. The researcher collects empirical data and then determines how close the data "fit" some predetermined theoretical or hypothesized distribution. When this chi square test is used the research analyst is attempting to determine whether an observed data distribution differs from a theoretical one. To illustrate, if a coin were completely "honest" one would theoretically anticipate exactly half heads and half tails. Therefore, if a penny were tossed 30 times we would expect 15 heads and 15 tails on the basis of chance. Although it is improbable that we would obtain exactly 15 heads and 15 tails the chi square goodness-of-fit test would enable us to determine if the results actually obtained would frequently occur by chance if the coin were "fair". If we obtained 16 heads and 14 tails, or 14 heads and 16 tails we would certainly not be disturbed about the honesty of the coin. However, if we obtained 25 heads and 5 tails, or vice-versa, we'd surely begin to doubt the honesty of the coin (ruling out some deception upon the coin flipper's part). The application of this test of significance would permit us to determine the particular outcomes that would seriously lead us to doubt the coin's honesty.

Consider the following example from sport sociology. The World Series is a best of seven series of baseball games. The first team to win four games is declared the winner. Suppose there have been seventy-six Series played. Thirteen of these have been won in four games, 17 have been won in five games, and 15 and 31 have been won in six and seven games, respectively. Statistically speaking, we want to know if the observed number of games the World Series has lasted differs from chance. If the Series was as likely to go 4 as 5 as 6 as 7 games we would expect 19 games to have lasted 4 games, 19 to have lasted 5 games and 19 to have lasted 6 and 7 games. To determine the expected frequencies we would divide N by K where N = the total number of World Series contests and k = the number of possible games the series could last. Having presented the logic for deriving the number of expected frequencies for each Series if one outcome was as likely as another we present the chi square formula:

$$x^2 = \Sigma \frac{(O_f - E_f)^2}{E_f}$$

Where: O_f = observed frequency

E_f = expected frequency

Σ = summation operator

The observed frequencies (O_f) are those actually obtained through the data collection procedure.[8] The expected frequencies (E_f) are those generated under the assumption that the Series is as likely to be of the duration 4, 5, 6, and 7 games. Translating these abstract statements into the null and alternative hypothesis we have:

H_0: the World Series is as likely to last 4 as 5 as 6 as 7 games

H_1: the World Series is not as likely to last 4 as 5 as 6 as 7 games

$\alpha = .05$

The following working table (Table 12.2) is helpful for computing the chi square goodness-of-fit statistic.

TABLE 12.2

Working Table for Chi Square Goodness of Fit Test

# of Games	O_f	E_f	$O_f - E_f$	$(O_f - E_f)$	$(O_f - E_f)^2/E_f$
4	13	19	6	36	1.89
5	17	19	2	4	.21
6	15	19	4	16	.84
7	31	19	12	144	7.58
					$\Sigma = 10.52 = x^2$

The chi square value for this data set equals 10.52. What does it mean? We consult the chi square sampling distribution (Table B in the Appendix of the text) to assess its interpretation for the matter at hand. Along the far left hand column of the table are listed df (degrees of freedom) values from 1 to 30 and along the top are listed various probability levels (different levels of significance). For the chi square goodness-of-fit test the degrees of freedom are found by: k-1 where k = the number of categories and 1 is a constant. For the present data there are four categories (4, 5, 6, 7), hence df = 4 - 1 = 3. There exist three degrees of freedom in the present case. At the outset we decided to test the null hypothesis at the .05 level of significance. The value in the body of the table where df = 3 and probability = .05 intersect is the critical value. The critical value (critical since of the computed value equals or exceeds it the null hypothesis is rejected) is 7.815. Since the computed value is larger than the critical value, (that is, $10.52 > 7.815$) the null hypothesis is rejected. The rationale for the decision

is this: if, in fact, the null hypothesis were correct only 5 times in 100 would a chi square value of 7.815 or larger be obtained. Since 95 times out of 100 you would not get a value of that magnitude the researcher decides to reject the null and accept the alternative. Substantively speaking, the World Series is significantly more likely to last a certain number of games than others. By percentaging the various number of games the Series has lasted we can see that 41% of the time the Series went 7 games, 22% of the time it lasted 5 games, and 20% and 17% it went 6 and 4 games, respectively.

Thus we say that the outcome is statistically significant. To say that the outcome is statistically significant is to imply that it is implausible that the null hypothesis is true at the preselected level of significance. Using statistical symbols we'd write: $p < .05$. This means that the null hypothesis would be expected to be true less than 5 times in 100.

The Chi Square Two Sample Test. The chi square two sample test assumes that two randomly and independently drawn samples have been selected and the researcher wishes to test whether the two variables are independent of each other in the populations from which they have been selected. The data for such a test typically appear in a contingency table. Suppose we selected a random sample of 100 Republicans and 100 Democrats. Each individual is asked whether they approve or disapprove of President Obama's handling of the Iranian affair. When the data are cross-tabulated the following contingency table (Table 12.3) emerges.

TABLE 12.3

Attitudes Toward Obama's Handling of the Iranian Crisis by Political Preference

		Republican	Democrat	
Attitude	Approve	40 (a)	75 (b)	115
	Disapprove	60 (c)	25 (d)	85
		100	100	200

The null hypothesis may be expressed as follows: Attitudes toward handling of the Iranian affair are independent of political preference. The alternative hypothesis would be: Attitudes toward handling of the Iranian affair are related to political preference. The level of significance is set at .01.

The expected frequencies for the chi square two sample test are generated using the row and column marginal totals. In Table 12.3 there are two row marginal totals, 115 and 85, and two column marginal totals, 100 and 100. To generate the expected frequencies under the truth of the null hypothesis we multiply the row marginal total common to the cell (for which we want to determine the expected frequency) by the column marginal total common to the cell (for which we want to determine the expected frequency) and divide by N. For cell a we multiply 115 by 100 and divide by 200. This process produces an E_f for cell a of 57.5. This procedure is employed for all cells in the contingency table. More simply, the expected frequencies in any size contingency table are generated as follows:

$$E_f = \frac{\text{(row marginal total) (column marginal total)}}{N}$$

Table 12.4 contains a working format for calculating the chi square test of independence's statistical value.

TABLE 12.4

Working Table for Computing Chi Square

Cell	O_f	E_f	$O_f - E_f$	$(O_f - E_f)^2$	$(O_f - E_f)^2/E_f$
a	40	57.5	-17.5	306.25	5.33
b	75	57.5	17.5	306.25	5.33
c	60	42.5	17.5	306.25	7.21
d	25	42.5	-17.5	306.25	7.21
					$\Sigma = 25.08 = x^2$

We again ask, what does a chi square value of 25.08 mean? We consult the chi square sampling distribution (Table in the Appendix), determine the degrees of freedom for the data, and find the critical value at the intersection of df and the predetermined level of significance. For a chi square test of independence the degrees of freedom are determined by: $(r - 1)(c - 1)$ where r = number of rows, c = number of columns, and 1 (inside each set of parentheses) is a constant. Since we have two rows and two columns, making our table a 2 x 2 one we have 1 degree of freedom $(2 - 1)(2 - 1) = 1$. At the intersection of 1 df and the .01 level of significance is found the critical value of 6.635. This value means that if the null hypothesis were true only 1 time in 100 would a value of 6.635 or larger be obtained. Since our computed value is considerably larger ($25.08 > 6.635$) we reject the null hypothesis at the .01 level of significance and claim that, indeed, there is a relationship between attitudes toward handling of the Iranian crisis and political preference. Substantively, the data in Table 12.3 tell us that Democrats (75%) are more approving than are Republicans (40%). Such a finding is said to be statistically significant and is symbolized $p < .01$.

The same procedure for determining the chi square test of independence's statistical value is used regardless of the size of the table. A table could be 7 x 5, 3 x 4, 6 x 8, etc. Nevertheless the same working format and process for generating the expected frequencies would be used. Then the chi square sampling distribution would be consulted according to df and level of significance to determine the critical value. If the computed value is equal to or exceeds the critical value at the chosen level of significance the null hypothesis is rejected.

The One-Sample Runs Test. The one-sample runs test is geared to help the researcher determine if a sample of observations is randomly distributed

or not. Computationally, it entails a statistical procedure based on the order or sequence of the empirical data. A run in the present context is defined as a series of particular observations which are followed or preceded by different observations or by no specific observations at all. Consider the following illustration.

Once again we will consider the outcomes of the World Series.

Of the seventy-six games played 46 (60 ½%) have been won by the American League representative and 30 (39 ½%) by the National League pennant winner. To utilize the one-sample runs test the observations must be of a dichotomous nature. A systematic tabulation of winners produced the following sequence (A = American League team won; N = National League team won).[9] We will determine whether or not the distribution of winners occurred in a random sequence. The data are as follows:

1	2	3	4	5	6	7	8	9	10	11	12
AAA	NNN	AAAA	N	AAAA	N	A	NN	AA	NN	AAAA	N

13	14	15	16	17	18	19	20	21	22	23	24	25
A	NN	AAAAA	N	A	N	A	N	A	N	AAAAAAA	NN	A

26	27	28	29	30	31	32	33	34	35	36	37	38	39	40
N	A	NN	AA	NNN	A	N	A	N	A	N	AAA	NN	N	N

In this example there are 40 runs. The total number of runs (designated r) permits us to determine whether or not a sample of observations (World Series victor) is random.

The null hypothesis is that the sequence of World Series winners occurs in a random order. The alternative hypothesis is that the sequence is not random. Let us choose the .05 level as the significance criterion. The sample size is 76 (the number of World Series).

Table C in the Appendix contains the critical values of r for the runs test

under the truth of the null hypothesis. Notice that the runs table (actually two different tables) includes the number of runs that are too few as well as too many to reject H_0. One or the other, but not both, would be used for one-tailed tests (depending on the direction specified in the alternative hypothesis). Both tables would be used for a two-tailed (non-directional) alternative hypothesis.

To calculate r let N_1 = the number of American League teams who have won the Series ($N_1 = 46$) and N_2 = the number of National League teams who have done the same ($N_2 = 30$). To employ the one-sample runs test we determine the sequence in which the N_1 and N_2 items occur and count r, the number of runs. In our example there are 40 runs where $N_1 = 20$ and $N_2 = 20$. These are indicated in the listings above. In the runs tables we locate the value at the intersection of N_1 and N_2 and discover that an r of fourteen or less or twenty-eight or more is sufficient to reject the null hypothesis at the .05 level.

Notice that if the observed number of runs is equal to or smaller than 14 or equal to or larger than 28 we can reject the null hypothesis that the World Series winners are randomly arranged. Since our r = 40 we can reject H_0 and consequently conclude that the sequence of World Series winners is not random insofar as American and National League pennant winners are concerned. In notation form, $p < .05$.

Summary

In this chapter we have briefly discussed the two branches of inferential statistics: 1) parameter estimation and 2) hypothesis testing. Parameter estimation entails selecting a random sample of population elements and using the sample statistic(s) to estimate the population parameter(s). Hypothesis testing involves procedures for either accepting or rejecting

the null hypothesis in comparison to some alternative hypothesis.

Both subdivisions of inferential statistics are based upon one's knowledge of the sampling distribution (the distribution of all possible sample statistics that would occur if one were to choose an indefinite number of random samples from a fixed universe) of the statistic in question. The nature of an empirical sampling distribution was exemplified and special attention was paid to the correspondence between the mean and standard deviation of both the sampling distribution and population. Researchers do not have to generate the sampling distributions of most statistics since this has already been done. Instead, the analyst must know the relevance of sampling distributions and how to use them.

There are two varieties of parameter estimators: 1) point estimates, and 2) interval estimates. Point estimates are single values which have been determined through the selection of random samples. Interval estimates consist of a range of values within which it is probable that the parameter would lie if a large number of confidence intervals were constructed. We demonstrated the construction of two interval estimates, one for a proportion and a second for a mean.

Hypothesis testing makes use of statistical knowledge and reasoning to make decisions in the face of uncertainty. The logic of hypothesis testing was demonstrated with special attention focused on such statistical concepts as:. null and alternative hypotheses, level of significance, one vs. two tailed tests, type 1 and 2 errors, region of rejection, critical value, sampling distribution, and the test statistic value. There are many different specific tests of significance. Many statistics books give a fuller treatment of the available techniques. Since our scope is to be representative rather than exhaustive we discussed tests appropriate for the nominal, ordinal, and interval-ratio measurement levels. For interval-ratio level data we

illustrated the application of a z single sample test and a z two sample test. For ordinal level data we applied the one sample runs test and for nominal level data the application of the chi square single sample test (the goodness-of-fit type) and the chi square test for two independent samples. These tests should be sufficient to convey to the reader a feel and flavor for hypothesis testing procedures.

Important Concepts Discussed in This Chapter

Inferential Statistics
Parameter Estimation
Hypothesis Testing
Parameter
Statistic
Sampling Fraction
Probability Sampling
Non-probability Sampling
Systematic Bias or Error
Random, Chance or Probability Error
Sampling Distribution
    Central Tendency
    Dispersion
    Form
Sampling With Replacement
Sampling Without Replacement
Biased Estimate
Unbiased Estimate
Central Limit Theorem
Point Estimate
Interval Estimate
Estimating a Population Proportion
Null and Alternative Hypothesis
One-tailed test (directional test)
Two-tailed test (non-directional test)
Level of Significance
Type 1 and 2 Errors
Region of Rejection
Test Statistic Value
Critical Value
z single sample test
z two sample test
Chi Square One Sample Test (goodness-of-fit)
Chi Square Two Sample Test
Observed Frequency
Expected Frequency
Degree of Freedom
One Sample Runs Test
Run