

Chapter 2

Descriptive Statistics

The Mean

"When she told me I was average, she was just being mean".

The mean is probably the most often used parameter or statistic used to describe the central tendency of a population or sample. When we are discussing a population of scores, the mean of the population is denoted with the Greek letter μ . When we are discussing the mean of a sample, we utilize the letter X with a bar above it. The sample mean is obtained as

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n} \quad (2.1)$$

The population mean for a finite population of values may be written in a similar form as

$$\mu = \frac{\sum_{i=1}^N X_i}{N} \quad (2.2)$$

When the population contains an infinite number of values which are continuous, that is, can be any real value, then the population mean is the sum of the X values times the proportion of those values. The sum of values which can be an arbitrarily small in differences from one another is written using the integral symbol instead of the Greek sigma symbol. We would write the mean of a set of scores that range in size from minus infinity to plus infinity as

$$\mu = \int_{-\infty}^{+\infty} Xp(X)dx \quad (2.3)$$

where $p(X)$ is the proportion of any given X value in the population. The tall curve which resembles a script S is a symbol used in calculus to mean the “sum of” just like the symbol Σ that we saw previously. We use Σ to represent “countable” values, that is values which are discrete. The “integral” symbol on the other hand is used to represent the sum of values which can range continuously, that is, take on infinitely small differences from one-another.

A similar formula can be written for the sample mean, that is,

$$\bar{X} = \sum_{i=1}^n X_i p(X_i) \quad (2.4)$$

where $p(X)$ is the proportion of any given X_i value in the sample.

If a sample of n values is randomly selected from a population of values, the sample mean is said to be an unbiased estimate of the population mean. This simply means that if you were to repeatedly draw random samples of size n from the population, the average of all sample means would be equal to the population mean. Of course we rarely draw more than one or two samples from a population. The sample mean we obtain therefore will typically *not* equal the population mean but will in fact differ from the population mean by some specific amount. Since we usually don’t know what the population mean is, we therefore don’t know how far our sample mean is from the population mean. If we have, in fact, used random sampling though, we do know something about the shape of the distribution of sample means; they tend to be *normally* distributed. (See the discussion of the Normal Distribution in the section on Distributions). In fact, we can estimate how far the sample mean will be from the population mean some (P) percent of the time. The estimate of sampling errors of the mean will be further discussed in the section on testing hypotheses about the difference between sample means.

Now let us examine the calculation of a sample mean. Assume you have randomly selected a set of five scores from a very large population of scores and obtained the following:

$$X_1 = 3$$

$$X_2 = 7$$

$$X_3 = 2$$

$$X_4 = 8$$

$$X_5 = 5$$

The sample mean is simply the sum (2.3) of the X scores divided by the number of the scores, that is

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n} = \sum_{i=1}^5 (X_1 + X_2 + X_3 + X_4 + X_5)/5 = (3 + 7 + 2 + 8 + 5)/5 = 5.0 \quad (2.5)$$

We might also note that the proportion of each value of X is the same, that is, one out of five. The mean could also be obtained by

$$\begin{aligned} \bar{X} &= \sum_{i=1}^n X_i p(X_i) \\ &= 3 (1/5) + 7 (1/5) + 2 (1/5) + 8 (1/5) + 5 (1/5) = 5.0 \end{aligned} \quad (2.6)$$

The sample mean is used to indicate that value which is “most typical” of a set of scores, or which describes the center of the scores. In fact, in physics, the mean is the center-of-gravity (sometimes called the first moment of inertia) of a solid object and corresponds to the fulcrum, the point at where the object is balanced.

Unfortunately, when the population of scores from which we are sampling is not symmetrically distributed about the population mean, the arithmetic average is often not very descriptive of the “central” score or most representative score. For example, the population of working adults earn an annual salary of \$21,000.00. These salaries however are not symmetrically distributed. Most people earn a rather modest income while there are a few who earn millions. The mean of such salaries would therefore not be very descriptive of the typical wage earner. The mean value would be much higher than most people earn. A better index of the “typical” wage earner would probably be the *median*, the value which corresponds to the salary earned by 50% or fewer people.

Examine the two sets of scores below. Notice that the first nine values are the same in both sets but that the tenth scores are quite different. Obtain the mean of each set and compare them. Also examine the score below which 50% of the remaining scores fall. Notice that it is the same in both sets and better represents the “typical” score.

SET A: (1, 2, 3, 4, 5, 6, 7, 8, 9, 10)

Mean = ?

Median = ?

SET B: (1, 2, 3, 4, 5, 6, 7, 8, 9, 1000)

Mean = ?

Median = ?

Did you know that the great majority of people have more than the average number of legs? It's obvious really; amongst the 57 million people in Britain there are probably 5,000 people who have got only one leg. Therefore the average number of legs is: $((5000 * 1) + (56,995,000 * 2)) / 57,000,000 = 1.9999123$ Since most people have two legs...

Variance and Standard Deviation

A set of scores are seldom all exactly the same if they represent measures of some attribute that varies from person to person or object to object. Some sets of scores are much more variable than others. If the attribute measures are very similar for the group of subjects, then they are less variable than for another group in which the subjects vary a great deal. For example, suppose we measured the reading ability of a sample of 20 students in the third grade. Their scores would probably be much less variable than if we drew a sample of 20 subjects from across the grades 1 through 12!

There are several ways to describe the variability of a set of scores. A very simple method is to subtract the smallest score from the largest score. This is called the *exclusive range*. If we think the values obtained from our measurement process are really point estimates of a continuous variable, we may add 1 to the exclusive range and obtain the *inclusive range*. This range includes the range of possible values. Consider the set of scores below:

5, 6, 6, 7, 7, 7, 8, 8, 9

If the values represent discrete scores (not simply the closest value that the precision of our instrument gives) then we would use the exclusive range and report that the range is $(9-5) = 4$. If, on the other hand, we felt that the scores are really point estimates in the middle of intervals of width 1.0 (for example the score seven is actually an observation someplace between 6.5 and 7.5) then we would report the range as $(9-5) + 1 = 5$ or $(9.5-4.5) = 5$.

While the range is useful in describing roughly how the scores vary, it does not tell us much about how MOST of the scores vary around, say, the mean. If we are interested in how much the scores in our set of data tend to differ from the mean score, we could simply average the distance that each score is from the mean. The mean deviation, unfortunately is always 0.0! To see why, consider the above set of scores again:

$$\text{Mean} = (5 + 6 + 6 + 7 + 7 + 7 + 8 + 8 + 9) / 9 = 63 / 9 = 7.0$$

Now the deviation of each score from the mean is obtained by subtracting the mean from each score:

$$5 - 7 = -2$$

$$6 - 7 = -1$$

$$6 - 7 = -1$$

$$7 - 7 = 0$$

$$7 - 7 = 0$$

$$7 - 7 = 0$$

$$8 - 7 = +1$$

$$8 - 7 = +1$$

$$\underline{9 - 7 = +2}$$

$$\text{Total} = 0.0$$

Since the sum of deviations around the mean always totals zero, then the obvious thing to do is either take the average of the absolute value of the deviations OR take the average of the squared deviations. We usually average the squared deviations from the mean because this index has some very important application in other areas of statistics.

The average of squared deviations about the mean is called the *variance* of the scores. For example, the variance, which we will denote as S^2 , of the above set of scores would be:

$$S^2 = \frac{(-2)^2 + (-1)^2 + (-1)^2 + 0^2 + 0^2 + 0^2 + 1^2 + 1^2 + 2^2}{9} = 1.3333$$

approximately. (2.7)

Thus we can describe the score variability of the above scores by saying that the average squared deviation from the mean is about 1.3 score points.

We may also convert the average squared value to the scale of our original measurements by simply taking the square root of the variance, e.g. $S = \sqrt{1.3} = 1.1547$ (approximately). This index of variability is called the *standard deviation* of the scores. It is probably the most commonly used index to describe score variability!

Estimating Population Parameters: Mean and Standard Deviation

We have already seen that the mean of a sample of scores randomly drawn from a population of scores is an estimate of the population's mean. What we have to do is to imagine that we repeatedly draw samples of size n from our population (always placing the previous sample back into the population) and calculate a sample mean each time. The average of all (infinite number) of these sample means is the population mean. In algebraic symbols we would write:

$$\mu = \frac{\sum_{i=1}^k \bar{X}_i}{k} \text{ as } k \rightarrow \infty \quad (2.8)$$

Notice that we have let $\bar{X}$ represent the sample mean and μ represent the population mean. We say that the sample mean is an *unbiased* estimate of the population mean because the average of the sample statistic calculated in the same way that we would calculate the population mean leads to the population mean. We calculate the sample mean by dividing the sum of the scores by the number of scores. If we have a finite population, we could calculate the population mean in exactly the same way.

The sample variance calculated as the average of squared deviations about the sample mean is, however, a *biased* estimator of the population variance (and therefore the standard deviation also a biased estimate of the population standard deviation). In other words, if we calculate the average of a very large (infinite) number of sample variances this average will NOT equal the population variance. If, however, we multiply each sample variance by the constant $n/(n-1)$ then the average of these “corrected” sample variances will, in fact, equal the population variance! Notice that if n , our sample size, is large, then the bias $n/(n-1)$ is quite small. For example a sample size of 100 gives a correction factor of about 1.010101. The bias is therefore approximately one hundredth of the population variance. The reason that the average of squared deviations about the sample means is a biased estimate of the population variance is because we have a slightly different mean (the sample mean) in each sample.

If we had knowledge of the population mean μ and always subtracted μ from our sample values X , we would not have a biased statistic. Sometimes statisticians find it more convenient to use the biased estimate of the population variance than the unbiased estimate. To make sure we know which one is being used, we will use different symbols for the biased and unbiased estimates. The biased estimate will be represented here by a S^2 and the unbiased by a s^2 . The reason for use of the square symbol is because the square root of the variance is the standard deviation. In other words we use S for the biased standard deviation and s for the unbiased standard deviation. The Greek symbol sigma σ is used to represent the population standard deviation and σ^2 represents the population variance. With these definitions in mind then, we can write:

$$\sigma^2 = \frac{\sum_{j=1}^K s_j^2}{k} \text{ as } k \rightarrow \infty \quad (2.9)$$

or

$$\sigma^2 = \frac{\sum_j \frac{n}{n-1} S_j^2}{k} \text{ as } k \rightarrow \infty \quad (2.10)$$

where n is the sample size, k the number of samples, S^2 is the biased sample variance and s^2 is the unbiased sample variance.

You may have already observed that multiplying the biased sample variance by $n/(n-1)$ gives a more direct way to calculate the unbiased variance, that is:

$$s^2 = (n / (n - 1)) * S^2$$

or

$$s^2 = \frac{n}{n-1} \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n} = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1} \quad (2.11)$$

In other words, we may directly calculate the unbiased estimate of population variance by dividing the sum of square deviations about the mean by the sample size minus 1 instead of just the sample size.

The numerator term of the variance is usually just called the “sum of squares” as sort of an abbreviation for the sum of squared deviations about the mean. When you study the Analysis of Variance, you will see a much more extensive use of the sum of squares. In fact, it is even further abbreviated to SS . The unbiased variance may therefore be written simply as

$$s^2 = \frac{SS_x}{n-1}$$

The Standard Error of the Mean

In the previous discussion of unbiased estimators of population parameters, we discussed repeatedly drawing samples of size n from a population with replacement of the scores after drawing each sample. We noted that the sample mean would likely vary from sample to sample due simply to the variability of the scores randomly selected in each sample. The question may therefore be asked “How variable ARE the sample means?”. Since we have already seen that the variance (and standard deviation) are useful indexes of score variability, why not use the same method for describing variability of sample means? In this case, of course, we are asking how much do the sample means tend to vary, on the average, around the population mean. To find our answer we could draw, say, several hundred samples of a given size and calculate the average of the sample means to estimate Since we have already seen that the variance (and standard deviation) are useful indexes of score variability, why not use the same method for describing variability of sample means? In this case, of course, we are asking how much do the sample

means tend to vary, on the average, around the population mean. To find our answer we could draw, say, several hundred samples of a given size and calculate the average of the sample means to estimate: and then get the squared difference of each sample mean from this estimate. The average of these squared deviations would give us an approximate answer. Of course, because we did not draw ALL possible samples, we would still potentially have some error in our estimate. Statisticians have provided mathematical proofs of a more simple, and unbiased, estimate of how much the sample mean is expected to vary. To estimate the variance of sample means we simply draw ONE sample, calculate the unbiased estimate of X score variability in the population then divide that by the sample size! In symbols

$$s_{\bar{X}}^2 = \frac{s_X^2}{n} \quad (2.12)$$

The square root of this estimate of variance of sample means is the estimate of the standard deviation of sample means. We usually refer to this as the *standard error of the mean*. The standard error of the mean represents an estimate of how much the means obtained from samples of size n will tend to vary from sample to sample. As an example, let us assume we have drawn a sample of seven scores from a population of scores and obtained:

1, 3, 4, 6, 6, 2, 5

First, we obtain the sample mean and variance as:

$$\bar{X} = \frac{\sum_{i=1}^7 X_i}{7} = 3.857 \text{ (approximately)} \quad (2.13)$$

$$s^2 = \frac{\sum_{i=1}^7 (X_i - \bar{X})^2}{7 - 1} = \frac{127}{6} = 3.81 \quad (2.14)$$

Then the variance of sample means is simply

$$s_{\bar{X}}^2 = \frac{s_X^2}{n} = \frac{3.81}{7} = 0.544 \quad (2.15)$$

and the standard error of the mean is estimated as

$$s_{\bar{X}} = \sqrt{s_{\bar{X}}^2} = 0.74 \quad (2.16)$$

You may have noticed by now, that as long as we are estimating population parameters with sample statistics like the sample mean and sample standard deviation, that it is theoretically possible to obtain estimates of the variability of ANY sample statistic. In principle this is true, however, there are relatively few that have immediate practical use. We will only be using the expected variability of a few sample statistics. As we introduce them, we will tell you what the estimate is of the variance or standard deviation of the statistic. The standard error of the mean, which we just examined, will be used in the z and t-test statistic for testing hypotheses about single means. More on that later.

Testing Hypotheses for Differences Between or Among Means

The Nature of Scientific Investigation

People have been trying to understand the things they observe for as long as history has been recorded. Understanding observed phenomenon implies an ability to describe and predict the phenomenon. For example, ancient man sought to understand the relationship between the sun and the earth. When man is able to predict an occurrence or change in something he observes, it affords him a sense of safety and control over events. Religion, astrology, mysticism and other efforts have been used to understand what we observe. The scientific procedures adopted in the last several hundred years have made a large impact on human understanding. The scientific process utilizes inductive and deductive logic and the symbols of logic, mathematics. The process involves:

- (a) Making systematic observations (**description**)
- (b) Stating possible relationships between or differences among objects observed (**hypotheses**)
- (c) Making observations under controlled or natural occurrences of the variations of the objects hypothesized to be related or different (**experimentation**)
- (d) Applying an accepted decision rule for stating the truth or falsity of the speculations (**hypothesis testing**)
- (e) Verifying the relationship, if observed (**prediction**)
- (f) Applying knowledge of the relationship when verified (**control**)
- (g) Conceptualizing the relationship in the context of other possible relationships (**theory**).

The rules for deciding the truth or falsity of a statement utilizes the assumptions developed concerning the chance occurrence of an event (observed relationship or difference). These decision rules are particularly acceptable because the user of the rules can ascertain, with some precision, the likelihood of making an error, whichever decision is made!

As an example of this process, consider a teacher who observes characteristics of children who mark false answers true in a true-false test as different from children who mark true answers as false. Perhaps the hypothetical teacher happens to notice that the proportion of left-handed children is greater in the first group than the second. Our teacher has made a systematic observation at this point. Next, the teacher might make a **scientific statement** such as “Being left-handed increases the likelihood of responding falsely to true-false test items.” Another way of making this statement however could be “The proportion of left-handed children selecting false options of true statements in a true-false test does not differ from that of right handed children beyond that expected by sampling variability alone.” This latter statement may be termed a **null hypothesis** because it states an absence (null) of a difference for the groups observed. The null hypothesis is the statement generally accepted for testing because the alternatives are innumerable. For example (1) no difference exists or (2) some difference exists. The scientific statement which states the principle of interest would be difficult to test because the possible differences are innumerable. For example, “increases” in the example above is not specific enough. Included in the set of possible “increases” are 0.0001, 0.003, 0.012, 0.12, 0.4, etc. After stating the null hypothesis, our scientist-teacher would make **controlled observations**. For example, the number of “false” options chosen by left and right handed children would be observed after controlling for the total number of items missed by each group. This might be done by **matching** left handed children with right handed children on the total test scores. The teacher may also need to insure that the number of boys and girls are also matched in each group to control for the possibility that sex is the variable related to option choices rather than handedness. We could continue to list other ways to control our observations in order to rule out variables other than the hypothesized ones possibly affecting our **decision**.

Once the teacher has made the controlled observations, decision rules are used to **accept or reject** the null hypothesis. We will discover these rules involve the chances of rejecting a true null hypothesis (**Type I error**) as well as the chances of accepting a false null hypothesis (**Type II error**).

Because of the chances of making errors in applying our decision rules, results should be verified through the observation of additional samples of subjects.

Decision Risks

Many research decisions have different losses which may be attached to outcomes of an experiment. The figure below summarizes the possible outcomes in testing a null hypothesis. Each outcome has a certain probability of occurrence. These probabilities (chances) of occurrence are symbolized by Greek letters in each outcome cell.

Possible Outcomes of an Experiment			
Experimenter conclusion based on observed data	True State of Nature		
	H_0 True H_0 False		
	accept	$1 - \alpha$	β
	H_0		Type II error
	reject	Type I Error	
	H_0	α	$1 - \beta$

In the above figure α (alpha) is the chance of obtaining a sample which leads to rejection of the null hypothesis when in the population from which the sample is drawn the null hypothesis is actually true. On the other hand, we also have the chance of drawing a sample that leads us to accept a null hypothesis when, in fact, in the population we should reject it. This latter error has β (Beta) chances of occurring. Greek symbols have been used rather than numbers because the experimenter may control the types of error! For example, by selecting large samples, by reducing the standard deviation of the observed variable (for example by improving the precision of measurement), or by decreasing the size of the discrepancy (difference) we desire to be sensitive to, we can control both Type I and Type II error.

Typically, the chances of getting a Type I error is arbitrarily set by the researcher. For example, the value of alpha may be set to .05. Having set the value of α , the researcher can establish the sample size needed to control Type II error which is also arbitrarily chosen (e.g. $\beta = 0.2$). In other cases, the experimenter is limited to the sample size available. In this case the experimenter must also determine the smallest difference or effect size (alternate hypothesis) to which he or she wishes to be sensitive.

How does a researcher decide on α , β and a minimum discrepancy? By assessing or estimating the loss or consequences in making each type of error! For example, in testing two possible cancer treatments, consider that treatment 1 costs \$1,000 while treatment 2 costs \$100. Consider the null hypothesis

H_0 : no difference between treatments (i.e. equally effective)

and consider the alternative

H_1 : treatment 1 is more effective than treatment 2.

If we reject H_0 : and thereby accept H_1 : we will pay more for cancer treatment. We would probably be glad to do this if treatment 1 were, in fact, more effective. But if we have made a Type I error, our losses are 10 to 1 in dollars lost. On the other hand, consider the loss if we should accept H_0 : when, in fact, H_1 : is correct. In this case lives will be lost that might have been saved. What is one life worth? Most people would probably place more than \$1,000 value on a life. If so, you would probably choose a smaller β value than for α . The size of both these values

are dependent on the size of risk you are willing to take. In the above example, a $\beta = 0.001$ would not be unreasonable.

Part of our decision concerning α and β also is based on the cost for obtaining each observation. Sometimes **destructive** observation is required. For example, in testing the effectiveness of a manufacturer's military missiles, the sample drawn would be destroyed by the testing. In these cases, the cost of additional observations may be as large as the losses associated with Type I or Type II error!

Finally, the size of the discrepancy selected as "meaningful" will affect costs and error rates. For example, is an IQ difference of five points between persons of Group A versus Group B a "practical" difference? How much more quickly can a child of 105 IQ learn over a child of 100 IQ? The larger the difference selected, the smaller is the sample needed to be sensitive to true population differences of that size. Thus, cost of data collection may be conserved by selecting realistic differences for the alternative hypothesis. If sample size is held constant while the discrepancy is increased, the chance of a Type II error is reduced, thus reducing the chances of a loss due to this type of error. We will examine the relationships between Type I and Type II error, the discrepancy chosen for an alternative hypothesis, and the sample size and variable's standard deviation in the following sections.

Hypotheses Related to a Single Mean

In order to illustrate the principles of hypothesis testing, we will select an example that is rather simple. Consider a hypothetical situation of the teacher who has administered a standardized achievement test in algebra to high school students completing their first course in algebra. Assume that extensive "norms" exist for the test showing that the population of previously tested students obtained a mean score equal to 50 and a standard deviation equal to 10. Further assume the teacher has 25 students in the class and that the class test mean was 55 and the standard deviation was 9. The teacher feels that his particular method of instruction is superior to those used by typical instructors and results in superior student performance. He wishes to provide evidence for his claim through use of the standardized algebra test. However, other algebra teachers in his school claim his teaching is really no better than theirs but requires half again as much time and effort. They would like to see evidence to substantiate their claim of no difference. What must our teachers do? The following steps are recommended by their school research consultant:

1. Agree among themselves how large a difference between the past population mean and the mean of the sampled population is a practical increment in algebra test performance.
2. Agree upon the size of Type I error they are willing to accept considering the consequences.
3. Because sample size is already fixed ($n = 25$), they cannot increase it to control Type II error. They can however estimate what it will be for the alternative

hypothesis that the sampled population mean does differ by a value as large or larger than that agreed upon in (2) above.

4. Use the results obtained by the classroom teacher to accept or reject the null hypothesis assuming that the sample means of the kind obtained by the teacher are normally distributed and unbiased estimates of the population mean. This is equivalent to saying we assume the teacher's class is a randomly selected sample from a population of possible students taught by the instructor's method. We also assume that the effect of the instructor is independent for each student, that is, that the students do not interact in such a way that the score of one student is somehow dependent on the score obtained by another student.

By assuming that sample means are normally distributed, we may use the probability distribution of the normally distributed z to test our hypothesis. Based on a theorem known as the "Central Limit Theorem", it can be demonstrated that sample means obtained from scores that are NOT normally distributed themselves DO tend to be normally distributed! The larger the sample sizes, the closer the distribution of sample means approaches the normal distribution. You may remember that our z score transformation is

$$z = \frac{X - \bar{X}}{S_x} = \frac{d}{S_x} \quad (2.17)$$

when determining an individual's z score in a sample. Now consider our possible sample means in the above experiment to be individual scores that deviates (d) from a population mean (μ) and have a standard deviation equal to

$$S_{\bar{X}} = \frac{S_x}{\sqrt{n}} \quad (2.18)$$

That is, the sample means vary inversely with the square root of the sample size. The standard deviation of sample means is also called the standard error of the mean. We can now transform our sample mean (55) into a z score where $\mu = 50$ and the standard error is $S_e = S_x/\sqrt{n} = 10/5 = 2$. Our result would be:

$$z_0 = \frac{\bar{X} - \mu_0}{S_e} = \frac{55 - 50}{2} = 2.5 \quad (2.19)$$

Note we have used a small zero subscript by the population mean to indicate this is the null hypothesis mean.

Before we make any inference about our teacher's student performance, let us assume that the teachers agreed among themselves to set the risk of a Type I error rather low, at 0.05, because of the inherent loss of greater effort and time on their part if the hypothesis is rejected (assuming they adopt the superior teaching method). Let us also assume that the teachers have agreed that a class that achieves an average mean at least 2 standard deviations of the sample means above the previous

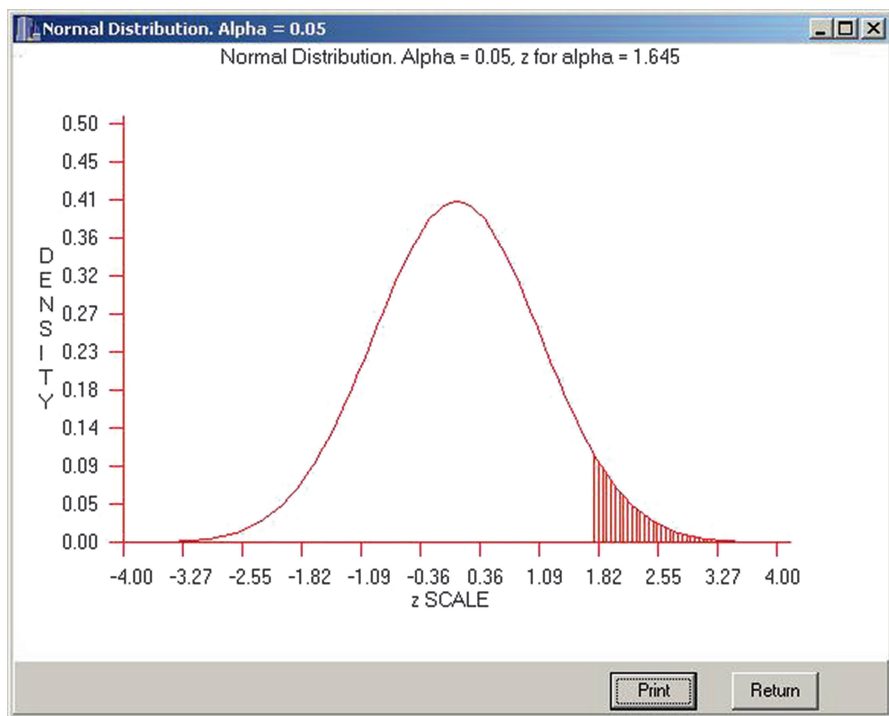


Fig. 2.1 Distribution of sample means

population mean is a realistic or practical increment in algebra learning. This means that the teachers want a difference of at least 4 points from the mean of 50 since the standard error of the means is 2.

Now examine the figure. In this figure the distribution of sample means is shown (since the statistic of interest is the sample mean.) A small caret (^) may be shown at the scale point where our specific sample statistic (the mean) falls in the theoretical distribution that has a mean of 50 and standard error of 2. Also shown, by shading is the area corresponding to the extreme 0.05 area of the distribution (Fig. 2.1).

Examination of the previous figure indicates that the sample mean obtained deviates from the hypothesized mean by a considerable amount (5 points). If we were obtaining samples from a population in which the mean was 50 and the standard error of the means was 2, we would expect to obtain a sample this deviant only 0.006 of the time! That is, only 0.006 of normally distributed z scores are as large or larger than the $z = 2.5$ that we obtained! Because our sample mean is *SO* deviant for the hypothesized population, we reject the hypothesized population mean and instead accept the alternative that the population from which we did sample has a mean greater than 50. If our statistic had not exceeded the z score corresponding to our Type I error rate, we would have accepted the null hypothesis.

Using a table of the normally distributed z score you can observe that the critical value for our decision is a $z_{\alpha} = 1.645$.

To summarize our example, we have thus far:

1. Stated our hypothesis. In terms of our critical z score corresponding to μ , we may write the hypothesis as

$$H_0 : z < z_{\mu} \quad (2.20)$$

2. Stated our alternate hypothesis which is

$$H_1 : z > z_{\mu}$$

3. Obtained sample data and found that $z > z_{\mu}$ which leads us to reject H_0 : in favor of H_1 :

Determining Type II Error and Power of the Test

In the example described above, the teachers had agreed that a deviation as large as 2 times the standard deviation of the means would be a “practical” teaching gain. The question may be asked, “What is the probability of accepting the null hypothesis when the true population mean is, in fact, 2 standard deviations (standard error) units above the hypothesized mean?” The figure below illustrates the theoretical distributions for both the null hypothesis and a specific alternate hypothesis, i.e. $H_1 = 54$ (Fig. 2.2).

The area to the left of the α value of 1.645 (frequently referred to as the region of rejection) under the null distribution (left-most curve) is the area of “acceptance” of the null hypothesis—any sample mean obtained that falls in this region would lead to acceptance of the null hypothesis. Of course, any sample mean obtained that is larger than the $z = 1.645$ would lead to rejection (the shaded portion of the null distribution). Now we may ask, “If we consider the alternative distribution (i.e. $\mu = 54$), what is the z value in that distribution which corresponds to the z value for μ under the null distribution?” To determine this value, we will first transform the z score for α under the null distribution back to the raw score X to which it corresponds. Solving the z score formula for X we obtain

$$\bar{X} = z_{\mu} S_{\bar{X}} + \mu_0 \quad (2.21)$$

or

$$\bar{X} = 1.645 (2) + 50 = 53.29$$

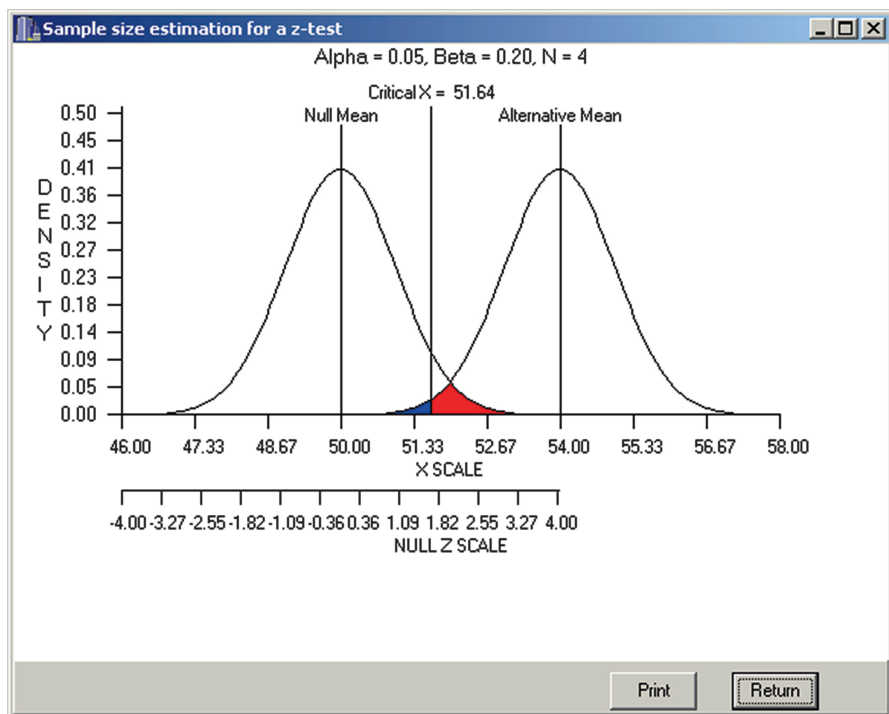


Fig. 2.2 Sample size estimation

Now that we have the raw score mean for the critical value of alpha, we can calculate the corresponding z score under the alternate distribution, that is

$$z_1 = \frac{\bar{X} - \mu_1}{S_{\bar{X}}} = \frac{53.29 - 54}{2} = -.355 \quad (2.22)$$

We may now ask, “What is the probability of obtaining a unit normal z score less than or equal to -0.355 ?” Using a table of the normal distribution or a program to obtain the cumulative probability of the z distribution we observe that the probability is $\beta = 0.359$. In other words, the probability of obtaining a z score of -0.355 or less is 0.359 under the normal distribution. We conclude then that the Type II error of our test, that is, the probability of incorrectly accepting the null hypothesis when, in fact, the true population mean is 54 is 0.359. Note that this nearly 36% chance of an error is considerably larger than the 5% chance of making the Type I error!

The sensitivity of our statistical test to detect true differences from the null hypothesized value is called the Power of our test. It is obtained simply as $1 - \beta$. For the situation of detecting a difference as large as 4 (two standard deviations of the sample mean) in our previous example, the power of the test was

$1 - 0.359 = 0.641$. We may, of course, determine the power of the test for many other alternative hypotheses. For example, we may wish to know the power of our test to be sensitive to a discrepancy as large as 6 X score units of the mean. The figure below illustrates the power curves for different Type I error rates and differences from the null hypothesis.

Again, our procedure for obtaining the power would be

- (a) Obtain the raw X-score mean corresponding to the critical value of α (region of rejection) under the null hypothesis. That is

$$\begin{aligned}\bar{X} &= z_{\alpha} S_{\bar{X}} + \mu_0 \\ &= 1.645 (2) + 50 = 53.29\end{aligned}\quad (2.23)$$

- (b) Obtain the z_1 score equivalent to the critical raw score for the alternate hypothesized distribution, e.g.

$$\begin{aligned}z_1 &= (\bar{X} - \mu_1) / S_{\bar{X}} \\ &= (53.29 - 56) / 2 \\ &= -2.71 / 2 \\ &= -1.355\end{aligned}\quad (2.24)$$

- (c) Determine the probability of obtaining a more extreme value than that obtained in (b) under the unit-normal distribution, e.g.

$$\begin{aligned}P(z < z_1 \mid \text{ND} : \mu = 0, \sigma = 1) \\ = P(z < -1.355 \mid \text{ND} : \mu = 0, \sigma = 1) = .0869\end{aligned}\quad (2.25)$$

- (d) Obtain the power as $1 - \beta = 1.0 - .0869 = .9131$ (2.26)

One may repeat the above procedure for any number of alternative hypotheses and plot the results in a figure such as that shown above. The above plot was made using the OpenStat option labeled “Generate Power Curves” in the Utilities menu.

As the critical difference increases, the power of the test to detect the difference increases. Minimum power is obtained when the critical difference is equal to zero. At that point power is equal to α , the Type I error rate. A different “power curve” may be constructed for every possible value of α . If larger values of α are selected, for example 0.20 instead of 0.05, then the test is more powerful for detecting true alternative distributions given the same meaningful effect size, standard deviation and sample size.

The Fig. 2.3 below shows the power curves for our example when selecting the following values of α : 0.01, 0.05, and 0.10.

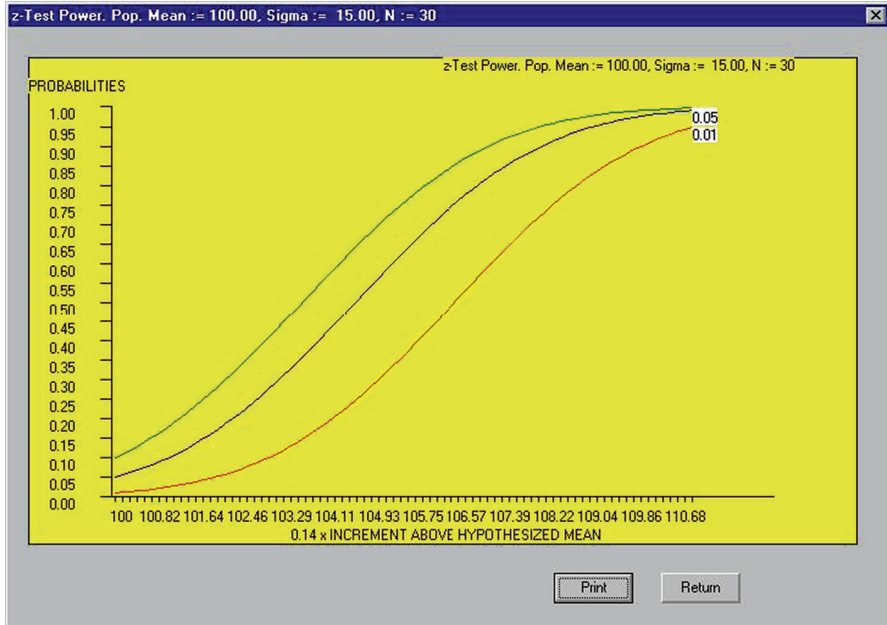


Fig. 2.3 Power curves

Sample Size Requirements for the Test of One Mean

The translation of a raw score mean into a standard score was obtained by

$$z = \frac{\bar{X} - \mu}{S_{\bar{X}}} \quad (2.27)$$

Likewise, the above formula may be rewritten for translating a z score into the raw score mean by:

$$\bar{X} = S_{\bar{X}}z + \mu \quad (2.28)$$

Now consider the distribution of an infinite number of sample means where each mean is based on the same number of randomly selected cases. Even if the original scores are not from a normally distributed population, if the means are obtained from reasonably large samples ($N > 30$), the means will tend to be normally distributed. This phenomenon is known as the *Central Limit Theorem* and permits us to use the normal distribution model in testing a wide range of hypotheses concerning sample means.

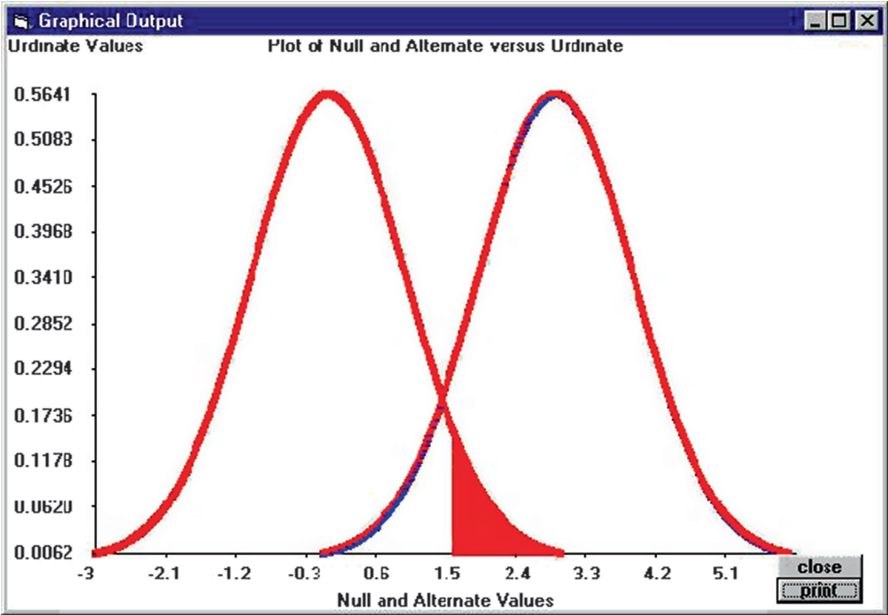


Fig. 2.4 Null and alternate hypotheses for sample means

The extreme “tails” of the distribution of sample means are sometimes referred to as “critical regions.” Critical regions are defined as those areas of the distribution which are extreme, that is unlikely to occur often by chance, and which represent situations where you would reject the distribution as representing the true population should you obtain a sample in that region. The size of the region indicates the proportion of times sample values would result in rejection of the null hypothesis by chance alone—that is, result in a “Type I” error. For the situation of our last example, the full region “R” of say 0.05 may be split equally between both tails of the distribution, that is, 0.025 or R/2 is in each tail. For normally distributed statistics a 0.025 extreme region corresponds to a z score of either -1.96 for the lower tail or $+1.96$ for the upper tail. The critical sample mean values that correspond to these regions of rejection are therefore

$$\bar{X}_c = \pm \sigma_{\bar{X}} z_{\alpha/2} + \mu_0 \tag{2.29}$$

In addition to the possibility of a critical score ($\bar{X}_c$) being obtained by chance part of the time (α) there also exists the probability (β) of accepting the null hypothesis when in fact the sample value is obtained from a population with a mean different from that hypothesized. Carefully examine the Fig. 2.4 above.

This figure represents two population distributions of means for a variable. The distribution on the left represents the null hypothesized distribution. The distribution on the right represents an alternate hypothesis, that is, the hypothesis that a

sample mean obtained is representative of a population in which the mean differs from the null distribution mean by a given difference D . The area of this latter distribution to the left of the shaded alpha area of the left curve and designated as β represents the chance occurrence of a sample falling within the region of acceptance of the null hypothesis, even when drawn from the alternate hypothesized distribution. The score value corresponding to the critical mean value for this alternate distribution is:

$$\bar{X}_c = \sigma_{\bar{X}} z_{\beta} + \mu_1 \quad (2.30)$$

Since formulas (2.29) and (2.30) presented above are both equal to the same critical value for the mean, they are equal to each other! Hence, we may solve for N , the sample size required in the following manner:

$$\sigma_{\bar{X}} z_{\alpha} + \mu_0 = \sigma_{\bar{X}} z_{\beta} + \mu_1 \quad (2.31)$$

where $\mu_1 = \mu_0 - D$

$$\text{and } \sigma_{\bar{X}} = \sigma_x / \sqrt{N} \quad (2.32)$$

Therefore,

$$(\sigma_x / \sqrt{N}) z_{\alpha} + \mu_0 = (\sigma_x / \sqrt{N}) z_{\beta} + \mu_1 \quad (2.33)$$

$$\text{or } \mu_1 - \mu_0 = (\sigma_x / \sqrt{N}) z_{\alpha} - (\sigma_x / \sqrt{N}) z_{\beta} \quad (2.34)$$

$$\text{or } D = \sigma_x / \sqrt{N} (z_{\alpha} - z_{\beta}) \quad (2.35)$$

$$\text{or } \sqrt{N} = (\sigma_x / D)(z_{\alpha} - z_{\beta}) \quad (2.36)$$

Note: z_{β} is a negative value in the above drawing because we are showing an alternative hypothesis above the null hypothesis. For an alternative hypothesis below the null, the result would yield an equivalent formula.

By squaring both sides of the above equation, we have an expression for the sample size N required to maintain both the selected α rate and β rate of errors, that is

$$N = \frac{\sigma_x^2}{D^2} (z_{\alpha} + z_{\beta})^2 \quad (2.37)$$

To demonstrate this formula (2.37) let us use the previous example of the teacher's experiment concerning a potentially superior teaching method. Assume that the teachers have agreed that it is important to contain both Type I error (α) and Type II error (β) to the same value of 0.05. We may now determine the number of students that would be required to teach under the new teaching method and test.

Statistics and Measurement Concepts with OpenStat

Miller, W.

2013, XIV, 237 p., Softcover

ISBN: 978-1-4614-5742-8