17

The Normal Curve and Hypothesis Testing

The Normal Curve Defined
Level of Significance
Sampling Distributions
Hypothesis Testing

In the last chapter we explained the elementary relationship of means, standard deviations, and z-scores. In this chapter we extend this relationship to include the **Normal Curve**, which allows us to convert z-score differences into probabilities. On the basis of laws of probability, we can make inferences from sample statistics to population parameters and make decisions about differences in scores. Using z-scores and the Normal Curve, we can convert differences in scores to probabilities.

The chapter is divided into the following sections:

The Normal Curve Defined. What is the nature of the Normal Curve? How does the Normal Curve and its associated Distribution table, link z-score with area under the curve? How does area under the curve relate to the concept of probability?

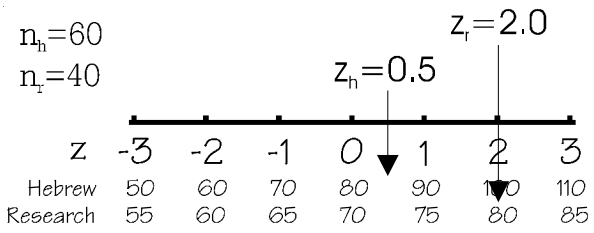
Level of Significance. What do the terms "level of significance" and "region of rejection" mean? What is alpha (α)? What is a critical value?

The Sampling Distribution. What is a sampling distribution? How does it differ from a frequency distribution?

Hypothesis Testing. How do we statistically test a hypothesis?

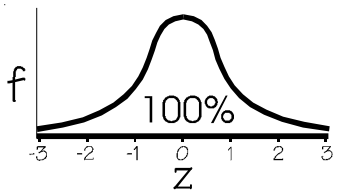
The Normal Curve

On page 16-11 we presented a z-scale with the z-scores for John's Research and Hebrew test scores. It looked like this:

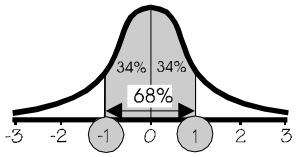


Recall that the mean of the z-scale equals zero and extends, practically speaking, 3 points in either direction. Each point on the z-scale equals one standard deviation away from the mean. A score of 100 in John's Hebrew class equals 2 standard deviations above the mean (μ=80, σ=10, z=+2.0). A score of 55 in John's Research class equals 3 standard deviations below the mean (μ=70, σ=5, z=-3.0).

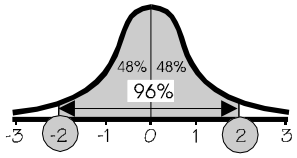
The z-scale assumes that the distribution of standardized scores forms a bell-shaped curve, called a Normal Curve. The normal curve is plotted on a set of X-Y axes, where the X-axis represents, in this case, "z-scores" and the Y-axis "frequency of z-scores." It looks like the diagram at left. The area between the "bell" and the baseline is a fixed area, which equals 100 percent of the scores in the distribution. We will use this area to determine the probabilities associated with statistical tests. ***There is an exact and unchanging relationship between the z-scores along the x-axis and the area under the curve.***



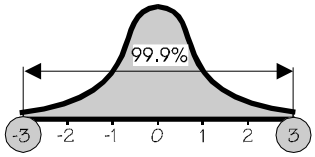
The area under the curve between z = ±1 (read "z equals plus or minus 1") standard deviation is 68% of the scores (p=0.68).



The area between ±2 standard deviations is 96%, or 0.96 of the curve.



The "tails" of the distribution theoretically extend to infinity, but 99.9% of the scores fall between z = ±3.00.



Now, let's use the normal curve in a practical way with John's classes. We can use the information in the diagram above to answer questions about John's classes.

Example 1: How many Hebrew students scored between 70 and 90?

For the Hebrew class, a score of 70 equals a z-score of -1 and a 90 equals a z-score of +1. The area under the normal curve between -1 and +1 is 68%. Therefore, the proportion of students in Hebrew scoring between 70 and 90 is **0.68**.

How many students is that?

Multiply the proportion (p=0.68) times the number of students in the class (60). The answer is 40.8. Rounding to the nearest whole student we would say that ***41 Hebrew students fall between 70 and 90 on this test.***

Example 2: How many research students scored between 60 and 80?

For the research class a score of 60 equals a z-score of -2; an 80 equals a z-score of +2. The area under the curve between -2 and +2 is 96%. Therefore, the proportion of the students in Research scoring between 60 and 80 is **0.96.**

How many students is that? (0.96)(40)=38.4. Rounding off to the nearest whole student we would say that ***38 research students fall between 60 and 80 on this test.***

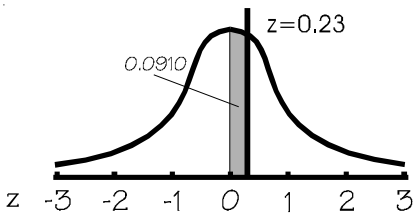
The Normal Curve Table

The Normal Curve distribution table allows us to determine areas under the Normal Curve between a mean and a z-score. Look up this table and use it to follow along the following description. You will find this table on page 1 in the Tables Appendix at the back of the book (Appendix A3-1).

The left column of the Normal Curve Table is labelled "Standard score z." Under this heading are z-scores in the form "x.x" beginning with 0.0 at the top and ending with 4.0 at the bottom.

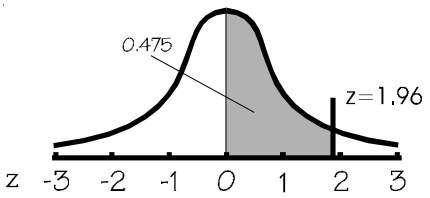
Across the top of the chart are the hundredths (0.0x) digits of z-scores, the numbers 0.00 through 0.09. To find the area under the normal curve between a mean (z = 0) and z = 0.23, look down the left column to 0.2 and then over to the column headed by .03. Where the ***0.2 row*** and ***.03 column*** you'll find the area under the Normal Curve between $z_1 = 0$ and $z_2 = 0.23$. This area (shown in gray) is **0.0910** or 9.1%.

	.00	.01	.02	.03	.04 ...
0.0				|	
0.1				|	
0.2	———	———	———	-.0910	
0.3					



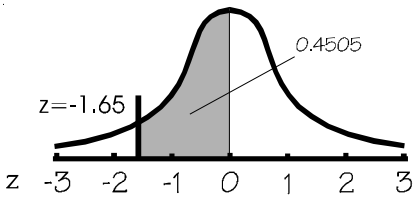
What is the area under the curve between the mean and z=+1.96? Look down the left column to the row labelled 1.9 and then across to the column labelled .06. Where these cross in the chart you will find the answer: 0.4750. That means that 47.5% of the scores in the group fall between the mean and 1.96 standard deviations away from the mean.

	.03	.04	.05	.06	.07 ...
1.7				|	
1.8				|	
1.9	———	———	———	-.4750	
2.0					



What is the area under the curve between the mean and z=-1.65? The normal curve is symmetrical, which means that the negative half mirrors the positive half. We can find the area under the curve for negative z-scores as easily as we can for positive ones. Look down the column for the row labelled 1.6 and then across to the column labelled 0.05. Where these cross you will find the answer: 0.4505. Forty-five percent (45%) of the scores of a group falls between the mean and -1.65 standard deviations from the mean.

	.02	.03	.04	.05	.06 ...
1.4				|	
1.5				|	
1.6	———	———	———	-.4505	
1.7					



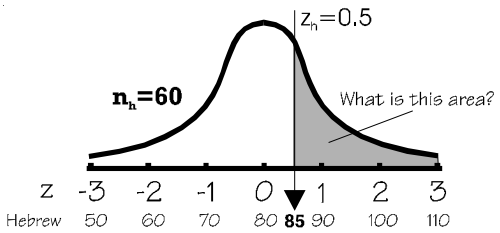
The Normal Curve Table in Action

Let's continue to use John's exam scores to further illustrate the use of the Normal Curve. We know John scored 85 in Hebrew. ***How many students scored higher than***

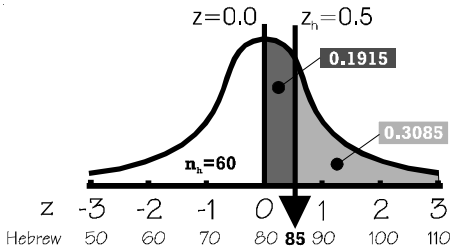
John?

Our first step is to compute the z-score for the raw score of 85, which we have already done. We know that the standard score for John's Hebrew score of 85 is z_h = 0.500 (diagram on 16-11 and 17-1).

The second step is to draw a picture of a normal curve with the area we're interested in. Notice that I've lightly shaded the area to the right of the line labelled **z = 0.5**. This is because we want to determine how many students scored higher than John. Since higher scores move to the right, the shaded area, which is equal to the proportion of students, is what I need. But just how much area is this?

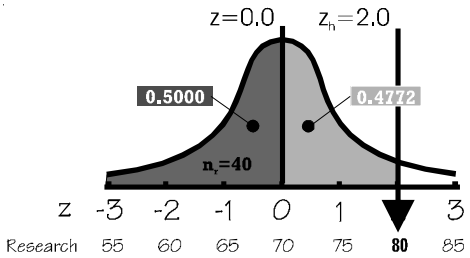


Look at the Normal Curve Table for the proportion linked to a z-score of 0.5. Down the left column to "0.5." Over to the first column headed ".00." The area related to z=0.5 is **0.1915.** I have shaded this area darker in the diagram below. Our ***lightly shaded area*** is on the other side of z=0.5! The area under the entire Normal Curve represents 100% of the scores. Therefore, the area under half the curve, from the mean outward, represents 50% (0.5000) of the scores. So, the lightly shaded area in the diagram is equal to 0.5000 minus 0.1915, or **0.3085.**



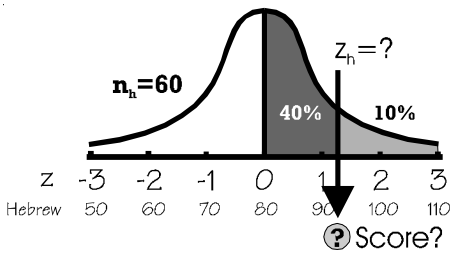
So we know that 30.85% of the students in John's Hebrew class scored higher than he did. ***How many students is that?*** Multiplying .3085 (proportion) times 60 (students in class) gives us 18.51, or *19 students*. ***Nineteen of 60 students scored higher than John on the Hebrew exam.***

Here's another. John scored 80 in Research. ***How many students scored lower than this?*** We've already computed John's z-score in Research as +2.00. The area under the curve between the mean and z = 2.00 is 0.4772. Find 0.4772 in the Table.



Since John also score higher than all the students in the lower half of the curve, we must *add the 0.5000 from the negative half of the curve* to the 0.4772 value of the positive half to get our answer. So, **97.72%** of the students in John's research class scored lower than he. How many students is this? It is (40 * .9772 =39.09) ***39 students***.

Here's an example which takes another perspective. We've used the Normal Curve table to translate z-scores into proportions. We can also translate proportions into z-scores. Take this question: ***What score did a student in John's Hebrew class have to make in order to be in the top 10% of the class?*** We start with an area (0.10) and work back to a z-score, then compute the raw score (X) using the mean and standard deviation for the group. Draw a picture of the problem -- like the one below.



We have "cut off" the top 10% of the curve. What proportion do I use in the Normal Curve table? We know we want the upper 10%. We also know that the table works from the mean out. So, the z-score that cuts off the *upper 10%* must be the same z-score that cuts off *40% of the scores between itself and the mean* (50% - 10% = 40%).

The proportion we look for in the table is **0.4000**. Search the proportion values in the table and find the one closest to .4000. The closest one in our table is "*0.3997*."

Look along this row to the left. The z-score value for this row is "***1.2***." Look up the column from 0.3997 to the top. The z-score hundredth value is ".08." *The z-score which cuts off the upper 10% of the distribution is **1.28.***

.	.05	.06	.07	**.08**	.09 ...
1.0					
1.1					
1.2	———	———	———	***.3997***	
1.3					

The z-score formula introduced in Chapter 16 yields a z-score from a raw score when we know the mean and standard deviation of a group of scores (left formula below).

This z-score formula can be transformed into a formula that computes X from z. Multiply both sides of the z-score formula by s and add $\overline{X}$. This produces the formula below right. Do you see how the two equations below are the same? One solves for z and the other for X.

$$z = \frac{X-\overline{X}}{s} \qquad X = \overline{X} + zs$$

Substituting the values of z=1.28, $\overline{X}$=80, and s=10 into the equation above right we get the following:

```
X = 80 + (1.28 * 10)
  = 80 + 12.80
X = 92.80
```

A student had to make ***92.8 or higher*** to be in the upper 10% of the Hebrew class.

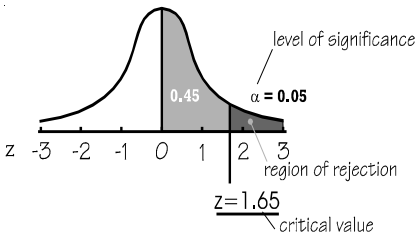
These examples may seem contrived, but they demonstrate basic skills and concepts you'll need whenever you use parametric inferential statistics. **Learn them well,** become fluent in their use, because you'll soon be using them in more complex, but more meaningful, procedures.

Level of Significance

John's Hebrew score was different from the class mean, but was the difference greater than we might expect by chance. Or as a statistician would ask it, was the score ***significantly*** different? John's research score was different from the class mean, but was it *significantly* different?

Criticial Values

We determine whether a difference is significant by using a criterion, or **critical value**, for testing z-scores. The critical value cuts off a portion of the area under the normal curve, called the **region of rejection**. The proportion of the normal curve in the region of rejection is called the **level of significance**. Level of significance is symbolized by the Greek letter alpha (α).



In this example, the **critical value** of 1.65 cuts off 5% of the normal curve. The **level of significance** shown above is α = 0.05. Any z-score greater than 1.65 falls into the **region of rejection** and is declared "significantly different" from the mean. Convention calls for the level of significance to be set at either ***0.05 or 0.01***.

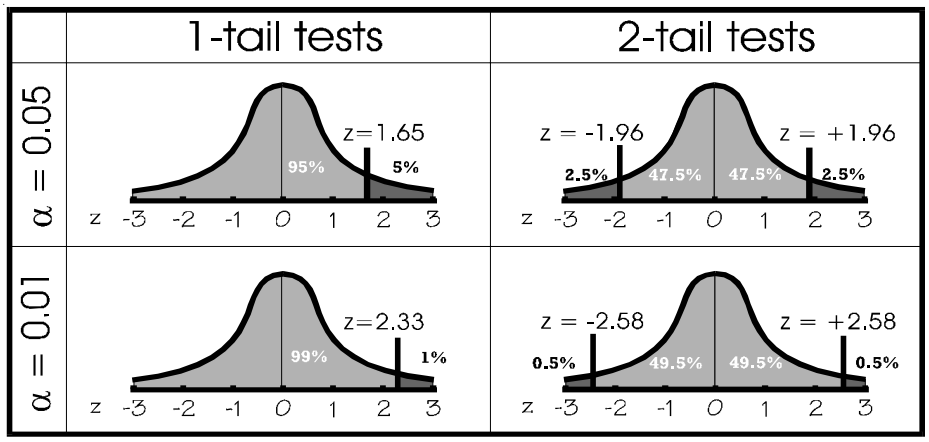
One- and Two-Tailed Tests

When all of α is in one tail of the normal curve, the test is called a "one-tailed test." When we statistically test a *directional hypothesis,* we use a *one-tail statistical test.* (Refer back to Chapter 4, if necessary, to review "directional hypothesis")

We can also divide the region of rejection between the tails of the normal curve in order to test non-directional hypotheses. To do this, place half of the level of significance (α/2) in each of the two tails. When statistically testing a *non-directional hypothesis,* use a *two-tailed test.*

The chart below summarizes the four conditions. Notice the effect of 1- or 2-tailed tests and α=.01 or .05 on the critical values used to test hypotheses. ***Memorize the conditions for each of the four conventional critical values: 1.65, 1.96. 2.33 and 2.58.***

Notice that the one-tail critical values (1.65, 2.33) are smaller than the two-tail values (1.96, 2.58). Having chosen a *directional hypothesis* (demonstrating greater confidence in your study), you can show "significance" with a *smaller z-score* (easier to obtain) than is possible with a non-directional study.



So now we return to our question at the beginning of this section. **Did John score "significantly higher" than his class averages in research and Hebrew?** Since this is a directional hypothesis, we'll use a 1-tail test, with α =0.05. Under these conditions, John had to score **1.65** standard deviations above the mean in order for his score to be considered "significantly different."

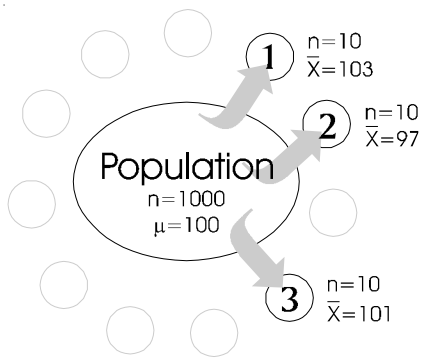
In Research, John scored **2.00** standard deviations above the mean. Since 2.00 is greater than 1.65, we can say with 95% confidence that ***John scored significantly higher in research than the class average***. In Hebrew, John scored **0.5** standard deviations above the mean. Since 0.5 is less than 1.65, we conclude that ***John did not score significantly higher in Hebrew than the class average.***

Our discussion to this point has focused on single scores (e.g., John's exam grades) within a *frequency distribution of scores*. While this has provided an elementary scenario for building statistical concepts, we seldom have interest in comparing single scores with means. We have much more interest in testing differences between a sample of scores and a given population, or between two or more samples of scores. Among the example Problem Statements in Chapter 4, you saw "Group 1 versus Group 2" types of problems. This requires an emphasis on ***group means*** rather than subject scores, on ***sampling distributions*** rather than frequency distributions.

--- --- Warning! This transition from scores to means is the most confusing element of the course --- ---

Sampling Distributions

A distribution of means is called a ***sampling distribution,*** which is necessary in making decisions about differences between group **means**. Just as naturally occurring ***scores*** fall into a normal curve distribution, so do the ***means*** of samples of scores drawn from a population. *The normal curve of **scores** forms a frequency distribution; the normal curve of **means** forms a sampling distribution.*



Look at the diagram at right. Here we see three samples drawn from a population. All three sample means are different, since each group of ten scores is a distinct subset of the whole. The variability among these sample means is called ***sampling error***. Even though we are drawing equal-sized groups from the same population, the means differ from one

another and from the population mean. Differences between means must be large enough to overcome this "natural" variation to be declared significant.

If we were to draw 100 samples of 10 scores each from a population of 1000 scores, we would have 100 different mean scores. *These 100 sample **means** would cluster around the population mean in a **sampling distribution**,* just as ***scores*** cluster around the *sample mean* in a ***frequency distribution***. If we were to compute the "*mean of the means*" we would find it would equal the *population mean.*

The two characteristics which define a normal frequency distribution are the mean and standard deviation. These same characteristics define a sampling distribution. The ***mean*** (μ) of a sampling distribution is the population mean, if it is known. If it is unknown, then the best estimate of the mean is one of the sample means ($\overline{X}$).

The *standard deviation* of the sampling distribution, called the ***standard error of the mean*** ($\sigma_{\overline{X}}$), is equal to the standard deviation of the population (σ) divided by the square root of the number of subjects in the sample ($\sqrt{n}$). Or, as in the formula below left,

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}} \qquad s_{\overline{X}} = \frac{s}{\sqrt{n}}$$

If the population standard deviation (σ) is unknown (which is usually the case), we must estimate it. In this case, the formula for standard error of the mean ($s_{\overline{X}}$) is based on the estimated standard deviation (s), as in the formula above right.

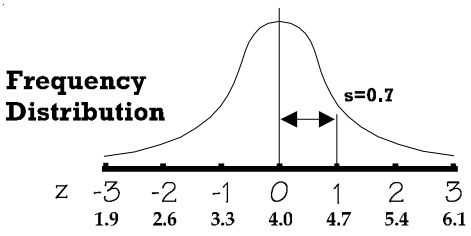
The Distinction Illustrated

Let's illustrate these concepts with the following scenario: a staff believes the education space in the church needs renovating. They want to measure "attitude toward building renovation" among the membership. They develop a "building renovation attitude scale" which has a range of 1 (low) to 7 (high). Because of several meetings already conducted, their hypothesis is that *church members have a negative attitude toward building*. They set $\alpha = 0.05$, and decide to use a 1-tail test since they are certain the scores will reveal a negative attitude. Here is the seven-point attitude scale used in the study.

1 2 3 4 5 6 7

Negative Neutral Positive

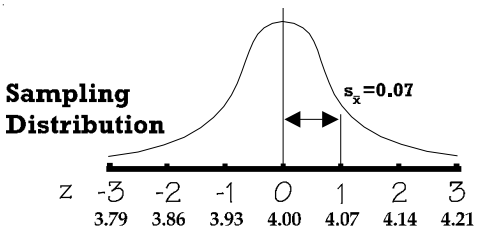
On a seven-point scale, the value of "**4**" is neutral. It represents the condition of **neutral attitude**. The research[1] hypothesis for their study was $\mathbf{H_a: \mu < 4.00}$ (church members will score significantly less than 4.00). The *null hypothesis* for the study was $\mathbf{H_0: \mu = 4.00}$.



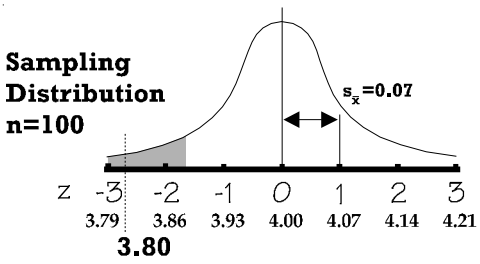
They randomly selected 100 church members and asked them to complete the attitude survey . After collecting and scoring the 100 forms, they computed the sample's mean and standard deviation, which equalled **3.80** and **0.7**. A *frequency distribution of scores* looked like the diagram at left. Notice the neutral value of *4.00* at z=0, s=0.7, and the computed *score-labels* for each of the z-scores (1.9 - 6.1).

But the staff wanted to know if the ***sample's*** attitude ***of 3.8 was significantly lower than 4.0?*** They developed the sampling distribution, with the neutral value of

[1]Research hypotheses are also called "alternative" hypotheses -- hence the reference "H_a"



4.00 at z=0, $s_{\bar{x}}$=0.07, and the computed raw mean-labels for each of the z-scores (3.79 - 4.21). The *x-axis of the sampling distribution reflects means, not scores.* Notice also in the diagram at left that the *much smaller differences* between the *mean-labels* than between the score-labels. This is because the standard error of the mean ($s_{\bar{x}}$=0.07) is much smaller than standard deviation of the sample (s=0.7).



Using a 1-tail test with α=0.05, the critical value needed to reject H_0 : μ=4.00 is **-1.65**. The area cut off by this critical value is shown shaded gray at left. Since the sample mean (3.8) falls into this area (beyond the dotted line), we declare that the ***3.80 is significantly lower than 4.00.*** Translating into English, we can say that ***the church at large does have a negative attitude toward building renovation.***

Let's now look at a slightly modified form of the z-score formula to compute the exact z-score for $\bar{X}$, as well as the exact probability of getting such a mean, all using the same procedure and Normal Curve table that we used before.

Using the z-Formula for Testing Group Means

Our statistical question is, ***Is 3.8 significantly lower than 4.0?*** We can see from the diagram above that the line representing 3.80 falls in the region of rejection. But we can also determine the exact number of ***standards errors*** the mean of 3.80 falls away from the hypothesized "neutral attitude" mean of 4.0. We do this in the same way we determined the number of ***standard deviations*** John's individual scores fell away from his class means. We use a z-Score formula much like the one on page 17-5, but adjusted for use with **sampling distributions**. We replace the standard deviation (s) with the standard error of the mean ($s_{\bar{x}}$), $\bar{X}$ with μ, and X with $\bar{X}$. It looks like this:

$$z = \frac{\bar{X}-\mu}{s_{\bar{x}}}$$

The above formula converts a ***mean*** into a ***z-score,*** given μ and $s_{\bar{x}}$. *Sampling distribution* z-scores are tested for significance just as we did with frequency distribution z-scores. Substituting 4.0, 3.8 and 0.07 for μ, $\bar{X}$ and $s_{\bar{x}}$ we have

$$z = \frac{\bar{X}-\mu}{s_{\bar{x}}} = \frac{3.8 - 4.0}{0.07} = -2.85$$

The mean of 3.80 is ***2.85 standard errors below the hypothesized mean of 4.00***. In order to be significant (1-tail, α = 0.05), the z-score must be 1.65 standard errors or more from the mean. Since -2.85 is farther from the mean than -1.65, we reject the null hypothesis and accept the alternative: "The congregation has a negative attitude toward renovation." In hypothesis testing, the null hypothesis is either *retained* (no significant difference) or it is *rejected* (significant difference). There are no partial decisions.

Moving back to the sampling distribution diagram, make a note that the dotted

line representing $\overline{X}$=3.8 is 2.85 standard errors below μ=4.0. Our finding is the same (3.8 falls into the shaded area), but we obtain a specific z-score, a more accurate measurement, by using the formula.

Summary

In this chapter we have introduced you to the process of testing hypotheses of parametric differences by way of the Normal Curve. We have differentiated between frequency and sampling distributions, and introduced the formula for computing the standardized score z.

Because our hypothesis decisions (reject or retain H_0) are based on probabilities (necessary since we work with sampling error and inferences), our results are always subject to errors. Such is science: hypotheses, data gathering, speculations, probabilities of findings. Our goal through proper research design and statistical analysis is to minimize errors and maximize "true findings." We will take up the topic of error rates and power in the next chapter.

Example

Dr. Robert DeVargas studied the change in moral judgment in students (3% sample, N=360) who used the *Lessons in Character* curriculum adopted by the Fort Worth I. S. D. for the 1996-1997 school year.[2] While much of his statistical analysis is far beyond the scope of this chapter, notice in his writing below the use of "level of significance" as a benchmark for his findings.

> Analysis of the fifth grade test data proceeded with the following steps:
>
> 1. An Analysis of Covariance (ANCOVA) was performed upon the post-test means of the treatment and control groups using the pre-test scores as a covariate variable. The mean score of the control group post-test score was 2.19 (n=30). The mean score of the treatment group post-test was 2.18 (n=31). The ANCOVA procedure produced an F value of 0.163 giving a significance of p=0.688. The critical value $F_{cv}(1, 60, \alpha=.05)$ =4.00.
>
> ...
>
> [3b] The treatment group's mean pre-test score equaled 2.0535 and the mean post-test score was 2.1835; the mean difference was 0.13 (n=31, SD=0.316). The standard error of the difference was 0.057 giving t = 2.29. The critical value t_{cv}(df=30, 1-tail, $\alpha=.05$) = ±1.697.[3]
>
> ...
>
> The step 3b analysis performed on the pre- and post-test means of the treatment group calculated the t value to be 2.29. Comparison to the critical value. . .reveals that. . .there exists a significant difference between the pre- and post-test scores. . . .it can be stated that the treatment [of moral judgement curriculum] made a significant difference in the level of moral judgement between the pre- and post-test scores of the treatment group.[4]

[2]Robert DeVargas, "A Study of Lessons in Character: The Effect of Moral Judgement Curriculum Upon Moral Judgement," (Ph.D. diss., Southwestern Baptist Theological Seminary, 1998)

[3]Ibid., 75 [4]Ibid., 78

Vocabulary

Alpha (α)	probability of rejecting true null hypothesis. Level of significance. 1%, 5%
Critical value	value beyond which the null hypothesis is rejected
Frequency distribution	categorization of scores into classes and class counts
Level of significance	probability of rejecting a true null. Symbolized by α
Normal curve	symmetrical mesokurtic (bell-shaped) distribution of scores
One-tail test	hypothesis test which places α in only one tail of distribution
Region of rejection	area under the normal curve beyond the critical value
Reject null	the decision of "statistically significant difference"
Retain null	the decision of "no statistically significant difference"
Sampling error	random differences among the means of randomly selected groups
Sampling distribution	normal curve distribution of sample **means** within a population
Standard error of the mean	the 'standard deviation' of a sampling distribution of means
Two-tail test	hypothesis test which places $\alpha/2$ in both tails of distribution

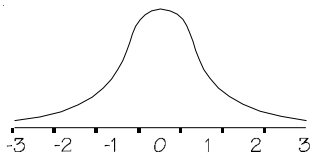
Study Questions

1. Define the following terms. Think of how you would explain them to someone in the class.

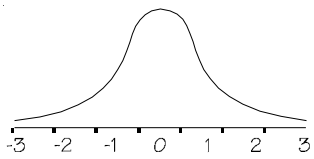
α	p
inferential statistics	null or statistical hypothesis (H_0)
level of significance	1- and 2-tail tests
sampling distribution	standard error of the mean
research hypothesis (H_a)	directional hypothesis
Normal Curve table	non-directional hypothesis

2. Determine the area under the normal curve between the following z-scores. Draw out the problems with the following diagrams.

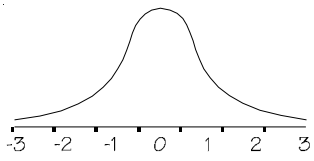
A. z = 0 and z = 2.3



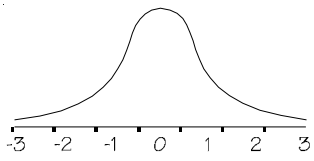
B. z = 0 and z = -1.7



C. z = 0.5 and z = 1.8



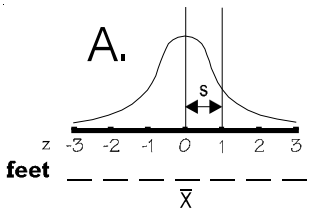
D. z = -1.2 and z = .75

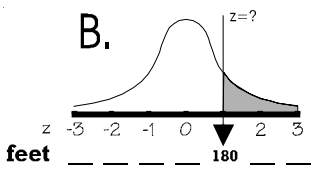


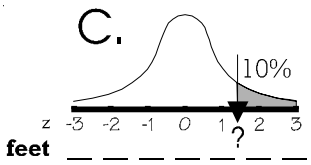
3. You go to an associational picnic. The men decide to have a contest to see who can throw a softball the farthest. You record the distances the ball is thrown, and compute the mean (= 164 feet) and the standard deviation (s = 16 feet). There were 100 men who threw the ball. Answer the following questions:

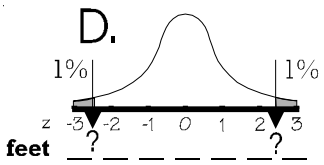
   A. With this information, sketch a normal curve and label it with z-scores and raw scores, mean and standard deviation.
   B. How many men threw the ball 180 feet or more? 120 feet or less?
   C. How far did one have to throw the ball to be in the top 10%?
   D. What distances are so extreme that only 1% of the men threw this far?
   E. Sketch a sampling distribution based on mean, s, n given above.
   F. A sister association joins the fellowship and challenges your association with a ball-toss of their own. Their average distance was 170 feet. Did they throw significantly better than your association?

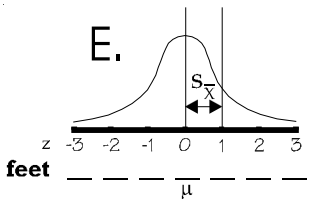
The following diagrams will help you work through problem 3.

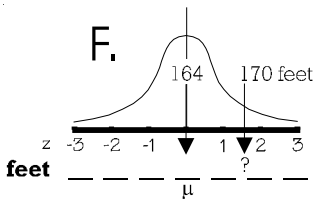












Sample Test Questions

1. The power of a statistical test refers to its ability to
   A. reject a true null
   B. retain a true null
   C. reject a false null
   D. retain a false null

2. A score of 85 would be converted to a z-score of ___, given a sample mean of 90 and a standard deviation of 10.
   A. +5.00
   B. +0.50
   C. -0.50
   D. -5.00

3. The area under the normal curve between z = ±2.00 is approximately
   A. 10%
   B. 50%
   C. 68%
   D. 96%

4. The point on the z-scale that "cuts off" the region of rejection is the _____ and the area under the curve to the right of this line is given by _____.
   A. critical value, p
   B. level of significance, 1-α
   C. critical value, α
   D. level of significance, 1-p