

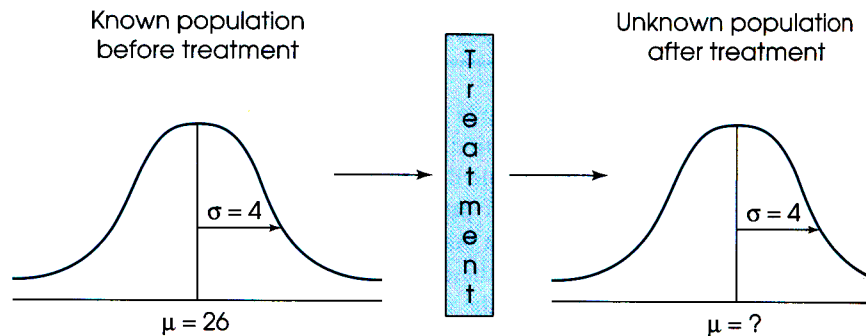
## Chapter 8: Introduction to Hypothesis Testing

We're now at the point where we can discuss the logic of hypothesis testing. This procedure will underlie the statistical analyses that we'll use for the remainder of the semester, so it's important that you understand the logic.

A *hypothesis test* is a statistical method that uses sample data to evaluate a hypothesis about a population parameter.

So, the procedure is to first state a hypothesis about a population parameter, such as  $\mu$ . The next step is to collect sample data that would address the hypothesis. And then to determine the extent to which the sample data are consistent with the hypothesis about the parameter.

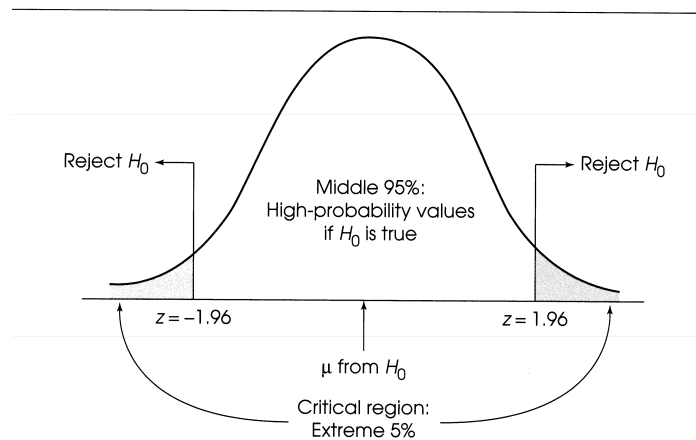
Here's an illustration of the logic of null hypothesis significance testing (NHST). Suppose that a population of 2-year-old children has a mean weight  $\mu = 26$  pounds and  $\sigma = 4$ . If you were to administer a treatment (handling) to every member of the population, what would happen to the scores? The effect of treatment is modeled as an additive constant (adding either plus or minus constant). If you recall the impact of an additive constant on variability, you'll realize that the standard deviation would stay the same. If the treatment has no effect, then the treatment constant would be zero, and the population mean would be unchanged. Schematically, the situation is illustrated below:



So, the question becomes, “Was the treatment effect zero, or was it greater than zero?” To test that question, we would typically construct a testable statistical hypothesis, called the null hypothesis ( $H_0$ ). In this case,  $H_0: \mu = 26$ . But, of course, we cannot treat and measure every member of the population. Instead, we will take a sample (e.g.,  $n = 16$ ) and give them extra handling. If the mean weight of the sample is near 26 pounds, then the handling treatment will likely be considered to be ineffective. To the extent that the mean weight of the sample is much larger (or smaller) than 26, then we would be inclined to think that the handling treatment was effective. The crucial question is, “How much must the mean weight differ from 26 pounds to convince us that the treatment was effective?”

The conventional way of determining the extent to which the treatment was effective is by establishing a level of significance or alpha level. What an alpha level represents is one's willingness to say that a sample mean was *not* drawn from the population with  $H_0$  true, when in

fact it was. In other words, what level of error are you willing to tolerate? (This particular kind of error is called a Type I Error, but more on that a bit later on.) Another way of thinking about alpha level is that it determines the weird or unusual scores that would occur in the distribution when  $H_0$  is true. The typical alpha level is .05 ( $\alpha = .05$ ), which means that the “definition” of weird is a score that would occur so infrequently as to lie in the lower or upper 2.5% of the sampling distribution. Alternatively, you could think that you are willing to conclude that if your sample mean falls in the lower or upper 2.5% of the distribution when  $H_0$  is true, you would be better off concluding that  $H_0$  is false. As you may recall, the  $z$ -scores that determine the lower and upper .025 of a normal distribution are  $-1.96$  and  $+1.96$ , as illustrated in the figure below.



After you’ve determined your null and alternative hypotheses, and your significance level (typically .05), you’re ready to collect your sample and determine the sample mean. With the sample mean, you’re now in a position to compute a  $z$ -score:

$$z = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}} = \frac{\text{obtained difference}}{\text{difference due to chance}}$$

Finally, given your  $z$ -score, you can make a decision about the null hypothesis. If the sample mean leads to a  $z$ -score that would fall in the critical region, you would decide to reject  $H_0$ . If the sample mean leads to a  $z$ -score that doesn’t fall in the critical region, you would fail to reject, or retain,  $H_0$ .

### *Uncertainty and Errors in Hypothesis Testing*

One of the reasons that we never talk about “proving” anything in science is that we recognize the ubiquity of uncertainty. Because we can never know the “truth” independent of our research, we can never be certain of our conclusions. Thus, when you decide to reject  $H_0$ , you need to be aware that  $H_0$  could really be false, in which case you have made a correct rejection. It’s also possible, however, that rejected  $H_0$  and it’s really true. If so, you’ve made an error. We call that error a Type I error. You should recognize that the significance level you choose is an expression of tolerance for a Type I error.

When you decide to retain  $H_0$ , it may well be that  $H_0$  is true, which is a correct retention. It's also possible, however, that  $H_0$  is false. In that case retaining  $H_0$  would be an error—a Type II error. You can summarize the four possibilities in a table:

		Actual Situation	
		$H_0$ True	$H_0$ False
Experimenter's Decision	Reject $H_0$	Type I Error	Correct Rejection
	Retain $H_0$	Correct Retention	Type II Error

Note that if you make the decision to reject  $H_0$ , you cannot make a Type II Error. ☺

*Example of Hypothesis Testing with a z-Score*

1. Question: Does prenatal alcohol have an impact on birth weight?
2. What are population characteristics normally?:  $\mu = 18$  grams,  $\sigma = 4$
3. State the null and alternative hypotheses, as well as the  $\alpha$ -level:

$$H_0: \mu = 18$$

$$H_1: \mu \neq 18$$

$$\alpha = .05, \text{ so if } |z| \geq 1.96, \text{ reject } H_0$$

4. Collect the data and compute the test statistic:

With a sample of  $n = 16$  and a sample mean ( $\bar{X}$ ) = 15, you could compute your z-score:

$$z = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}} = \frac{15 - 18}{1} = -3.0$$

5. Make a decision:

Because  $|z_{\text{Obtained}}| \geq 1.96$ , I would reject  $H_0$  and conclude that the impact of prenatal alcohol is to reduce birth weight. (Note that I could be making a Type I error.)

*Assumptions underlying hypothesis tests with z-scores*

You will actually see very few hypothesis tests using z-scores. That's because doing so requires that one knows  $\sigma$ , which is a highly unusual circumstance. Beyond that major assumption, there are other assumptions of note:

1. Participants are randomly sampled. (And that almost never happens in psychological research!)
2. Observations are independent of one another. Random sampling will typically produce observations that are independent of one another.
3. The value of  $\sigma$  will not be affected by the treatment. Even in the rare circumstances under which we might actually know  $\sigma$ , it's also crucial that any treatment we use has no impact on  $\sigma$ , otherwise that statistic will be thrown off.
4. The sampling distribution of the mean is normal. Of course, with increasingly large sample size (approaching infinity), the sampling distribution will become normal.

### Effect Size

Some people have expressed reservations about null hypothesis significance testing, which your text's authors detail (and you should read and think about). A crucial question that is not addressed by a significance test is the size of the treatment effect. That effect can be assessed by a number of measures, one of which is Cohen's  $d$ . You compute this measure of effect size as follows:

$$d = \frac{\text{mean difference}}{\text{standard deviation}}$$

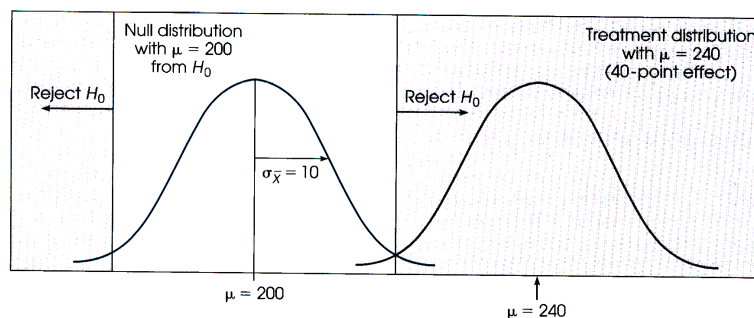
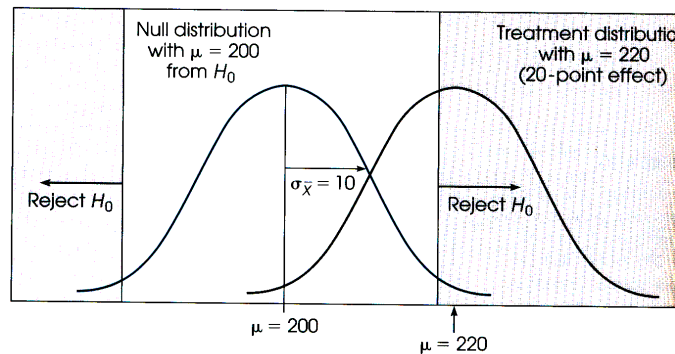
A value of  $d$  between 0 and 0.2 is considered a small effect, between 0.2 and 0.8 is considered a medium effect, and a value of  $d$  greater than 0.8 is considered a large effect.

### Statistical Power

Throughout this course, you will be exposed to the notion of power. The *power* of a test is the probability that the test will correctly reject a false null hypothesis. It represents the opposite of a Type II error, also called a  $\beta$  error. Thus, power is described as  $1 - \beta$ .

Several factors can influence power, but for now, you should think of the impact of treatment effect on power. In the example of the impact of prenatal alcohol on birth weight, if the alcohol had a greater impact, the mean weight of the sample would be smaller still (leading to an even larger z-score).

As you can see in the figures below, if the sample were drawn from a population with  $\mu = 240$ , power would be greater than if the sample were drawn from a population with  $\mu = 220$ . (Of course, in the real world, you'd never know the  $\mu$  of the population from which your sample was drawn.)



Problems:

1. In earlier notes, we used normally distributed gestation times ( $\mu = 268$ ,  $\sigma = 16$ ) to address various questions. In a similar fashion, we could test hypotheses using that distribution. For example, suppose that you had a sample of 16 women and their mean gestation period was 274 days. How likely is it that your sample was randomly drawn from the population of gestation periods? Would your conclusion change if you sample size was  $n = 64$ ? What would the effect size ( $d$ ) be in both cases?

$H_0$ : $H_1$ : $\alpha = .05$	
$n = 16$	$n = 64$
Compute statistics:          Decision:   Conclusion:	Compute statistics:          Decision:   Conclusion:

2. Given that IQ scores are normally distributed with  $\mu = 100$  and  $\sigma = 15$ , how likely is it that a sample of  $n = 25$  students with  $M = 109$  was randomly drawn from that population? What is the effect size ( $d$ ) in this case?

3. What is the relationship between effect size and power? If you have a great deal of power, do you need to worry about effect size? If you have a small effect size, do you need to worry about power? Provide examples to make your points.

4. Can you relate Null Hypothesis Significance Testing (NHST) to Signal Detection Theory (SDT)? What concept in NHST might you relate to  $d'$  in SDT?

5. What's the major limitation to conducting hypothesis tests with  $z$ -scores? How might you surmount that difficulty?