Notes 4: Hypothesis Testing: Hypothesis Testing, One Sample Z test, and Hypothesis Testing Errors

1. Coin Toss and Hypothesis Testing Logic – Is this result real; what is the probability of such a result?

(a) Hypothesis Testing and Probabilities – All Starts with H_0

Hypothesis testing is based upon probabilities and judging those probabilities against a known or theoretical standard. The standard against which probabilities are calculated is stated in the null hypothesis. Consider, for instance this hypothesis:

> ***H_0: a fair coin has a 50:50 chance of heads and tails ($\mu = .5$)***
> ***H_1: coin is not fair; it does not have a 50:50 chance of heads and tails ($\mu \neq .5$)***

The null hypothesis above (H_0) states that a coin, if fair, should land on heads 50% of the time and tails 50% of the time.

(b) Compare Empirical Results Against What Was Expected in H_0

How can we test whether a coin appears to be fair (50:50 chance of heads/tails)?

> ***We can empirically test that stated in the null hypothesis (H_0) by flipping a coin (taking a sample of coin tosses) and then compare our sample coin flip results to what is expected assuming the coin is fair (i.e., comparing our results to what was expected in the null hypothesis)***.

(c) Reject or Fail to Reject H_0 Based Upon Empirical Results

If the results we obtain in a sample are consistent with the null hypothesis (e.g., coin appears fair, the probabilities of heads/tails from our experimental coin tosses are similar to what is expected in the null hypothesis), then we will ***fail to reject*** the null and state that the sample data appear to be consistent with the null, thus the coin appears to be fair.

If, however, our sample results are odd, rare, or unexpected – that is, the results obtained are those that occur only with low probability – then we will ***reject*** the null hypothesis of fairness and conclude instead that the sample results from the coin are not consistent with the null hypothesis, therefore the null hypothesis is untenable and we reject it in favor of the alternative hypothesis.

(d) Empirical Probabilities Compared Against What Standard?

How does one judge whether empirical results obtain in a sample are consistent or inconsistent with the null hypothesis? What standard is used to judge whether results are rare if the null hypothesis is true?

An empirical example with a coin toss.

2. One Sample z test

(a) Hypotheses for one sample z test

In all hypothesis testing, the null is assumed true and it is the null that is tested. For a one sample z test, the null hypothesis will state the a sample mean will be equal to a population standard (or population mean):

Written:

There will be no difference in verbal SAT scores between GSU undergraduates and undergraduates nationwide.

Symbolized:

H_0: $\mu = 500$

where μ represents the mean (i.e., 500) of the population from which the sample was thought to be selected.

The alternative (H_a or H_1) in this example would be just the opposite of the null, i.e.,

Written:

Verbal SAT scores for undergraduates at GSU will differ from the nationwide average SAT score.

Symbolized:

H_1: $\mu \neq 500$

Notice that the above hypothesis, H_1: $\mu \neq 500$, does not specify whether GSU students will score higher or lower than the national average, it only indicates that the scores will not be the same (i.e., the scores will be different).

At this point, before collecting data, we may not know whether GSU students will score higher or lower than the national average, so we must consider possible results in both directions. Using this logic is the basis for a non-directional hypothesis and leads to non-directional statistical tests.

Some notes:

- Statistical hypotheses are symbolized by H_0 (statistical hypotheses are commonly referred to as null hypotheses)
- Research hypotheses are symbolized by H_1 or H_a where the '1' and 'a' subscripts denote the alternate or alternative hypothesis
- Note that hypotheses form pairs, the null, H_0, and the alternative, H_1
- The *null hypothesis is presumed to be true*, and through inferential techniques researchers will make a decision to either reject the null (and thereby conclude that the null is not tenable) or fail to reject the null (and thereby conclude that the null is tenable)

(b) Calculating Probabilities for z test

Z formula:

To calculate probabilities of sample data given the null hypothesis stated above, one first converts the obtained sample mean to a z statistic using the following formula which assumes both μ and σ are known:

$$Z_{\overline{X}} = Z_M = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}} = \frac{\overline{X} - \mu}{\sigma_{\overline{X}}}$$

The Z test is appropriate when one wishes to determine whether an obtained sample mean deviates, by an improbable amount, from some pre-specified value (which is usually the population mean).

Thus, for a particular sample mean, the above formula may be used to determine how far the sample mean is from the *population mean* in standard error (deviation) units.

Example Calculation:

For the example, suppose one takes a random sample of size 6 from all Georgia high school students who took the SAT and found the mean for the verbal section of the SAT to be $\overline{X}$ = 446.67. Verbal SAT scores have a national mean of μ = 500 and standard deviation of σ = 100.

The Z_M score is:

$$Z_M = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}} = \frac{446.67 - 500}{100 / \sqrt{6}} = \frac{-53.33}{\left(100 / 2.45\right)} = \frac{-53.33}{40.82} = -1.31.$$

This sample mean is -1.31 *standard errors* below the population mean, μ.

Finding Probabilities using Z:

As noted above the null hypothesis is assumed true. Data obtained in the sample are then compared to the null to determine the likelihood of obtaining data like that obtained in the sample if, in fact, the null is true.

Recall the null for this example:

Written:

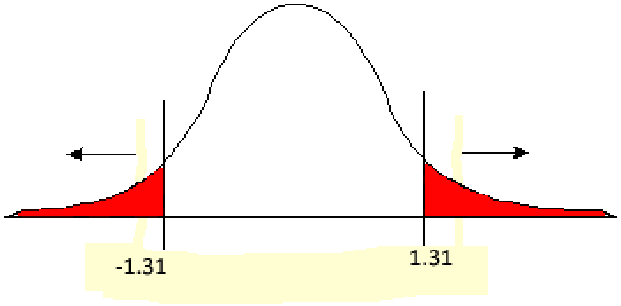
> There will be no difference in verbal SAT scores between GSU undergraduates and undergraduates nationwide.

Symbolized:

H_0: $\mu = 500$

Assuming this is true, what is the probability of randomly sampling 6 Georgia high school students who would have a sample mean that is -1.31 standard errors below the population mean (or lower), or 1.31 standard errors above the population mean (or greater)?

Graphically, we are looking for the area denoted in the distribution below:



The reason we seek to calculate probabilities for both upper (1.31 and above) and lower (-1.31 and below) tails is because we have a non-directional alternative hypothesis

H_1: $\mu \neq 500$

which states that Georgia students have a verbal SAT that is different from the population mean so it could be higher or lower if due to random chance.

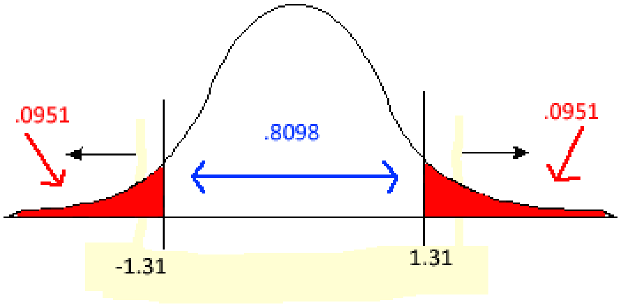
The z-table can be used to find the appropriate probabilities. Find these two probabilities:

$$p\left(Z_{\overline{X}} \leq -1.31\right)$$

and

$$p\left(Z_{\overline{X}} \geq 1.31\right)$$

In both cases the probability is .0951 at each end, which means that about .8098 of the Z scores will fall between -1.31 and 1.31 and about .1902 proportion of the Z scores lie below -1.31 and above 1.31. This is illustrated below.



Assuming the null hypothesis is true and Georgia students have the same verbal SAT scores, on average, as students nationwide, then we can expect a sample mean with a Z score of -1.31 or less, or 1.31 or greater 19.02% of the time. This probability of .1902 is called a ***p-value***.

> ***p-value for a Z test***: Assuming the null hypothesis is true, the p-value represents the probability of obtaining a sample mean that deviates from the population by this amount, or more, in a random sample.

That is, the chance, or probability, of a sample mean being this small, or smaller, i.e., $p\left(Z_{\overline{X}} \leq -1.31\right)$, or this large or larger, i.e., $p\left(Z_{\overline{X}} \geq 1.31\right)$ occurring in a given *random* sample is about .1902, or about 19.02% of the time if the *true* population mean of SAT scores is 500. Note that $p\left(Z_{\overline{X}} \leq -1.31\right)$ represents the probability of the event "less than or equal to -1.31" occurring and $p\left(Z_{\overline{X}} \geq 1.31\right)$ represents the probability of the event “greater than or equal to 1.31.”

Stated somewhat differently:

> If one sampled from a population with μ = 500 and σ = 100 a large number of times, by chance alone, roughly 19.02% of the random samples (of size 6, i.e., n = 6) selected would have a mean less than, or equal to 446.67, or a mean greater than, or equal to 553.33. Similarly, one would expect 80.98% of the samples would have a mean greater than 446.67 and less than 553.33.

Questions:

(a) How was the 553.33 derived? (to find sample mean given a Z score for the sample mean, use this formula:

$$\overline{X} = \mu + (Z_{\overline{X}} \times \sigma_{\overline{X}})$$

This is just the Z formula for the sample mean solved for $\overline{X}$.

(b) Why would one be interested in determining the rarity of an event—why focus on the extremes?

(c) Why the interest in both directions—above and below the population mean? That is, why find

$$p\left(Z_{\overline{X}} \geq \left|{}_{obtained}Z_{\overline{X}}\right|\right)$$

For the current example $p\left(Z_{\overline{X}} \geq \left|{}_{obtained}Z_{\overline{X}}\right|\right)$ = $p\left(Z_{\overline{X}} \geq \left|1.31\right|\right)$

Side Note—Sample size and Z_M *:*

Recall that the formula for Z_M scores is:

$$Z_{\overline{X}} = Z_M = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}} = \frac{\overline{X} - \mu}{\sigma_{\overline{X}}}$$

Suppose the population mean, μ, is 100 and the sample mean, $\overline{X}$, is 105. Also, assume the population standard deviation, σ, is 15. What effect upon Z_M does alteration of n have?

As the sample size increases, Z_M increases (in absolute value), as illustrated below.

(1) If n = 5, the standard error is $15/\sqrt{5}$ = 6.7, and Z_M is 0.74.
(2) If n = 10, the standard error is $15/\sqrt{10}$ = 4.7, and Z_M is 1.06.
(3) If n = 15, the standard error is $15/\sqrt{15}$ = 3.9, and Z_M is 1.28.
(4) If n = 20, the standard error is $15/\sqrt{20}$ = 3.4, and Z_M is 1.47.

As sample size increases, more information about the population is included, so one may state with more confidence whether a given sample mean is unusual relative to the population mean. In addition, as the sample size increases, Z_M becomes larger and $p\left(Z_{\overline{X}} \geq \left|obtained \quad Z_{\overline{X}}\right|\right)$ becomes smaller, which indicates, again, that obtaining a sample mean this far from the population mean becomes rarer (is more unlikely).

(c) Decision Regarding H_0—Reject or Fail to Reject

If the evidence from the Z test suggests that the null hypothesis is untenable, then H_0 is *rejected* in favor of the alternative hypothesis, H_1.

For example, if the statistical evidence indicates that it is unlikely the sample was drawn from a population with a mean of 500, then one might conclude that GSU students have a verbal SAT average which is different from 500.

If, however, the statistical evidence suggests that the null hypothesis is not false, then one would *fail to reject* the null hypothesis, i.e., *fail to reject* H_0. Failing to reject the null is simply stating that there is not enough evidence, based on the calculated probabilities, to reject the null. So in the example above, if H_0 is not rejected, then one would conclude that GSU SAT scores are similar to SAT scores from students nationwide.

Alpha:

What is considered small probability? If the probability for a given Z_M is small, say less than .01 or .05, then H_0 is rejected; if the probability is larger than .01 or .05, then H_0 is not rejected.

The probability that one selects as the cut-off for rejecting H_0 (e.g., .10, .05, or .01) is called the significance level and is denoted by the symbol α.

Note that the researcher (e.g., you) sets the significance level. The researcher decides what is and is not a small probability. Note:

- As mentioned above, the small probability (the significance level) is symbolized by α, and sometimes researchers will refer to the significance level as the "alpha level."
- The value of α is determined by the researcher, but traditionally significance levels are set at $\alpha = .10$, $\alpha = .05$, or $\alpha = .01$.
- One must choose the value of α before the experiment or hypothesis test is performed.
- If the calculated probability based upon Z_M is less than or equal to the alpha level, say .05, then one "rejects H_0 at the 5 percent level of statistical significance," or one states that "the result of the test was statistically significant at the .05 level," or states simply that $p < .05$.

Decision Rule:

To help decide whether the null hypothesis should be rejected, the following decision can be used:

If $p \leq \alpha$, the reject H_0; otherwise, fail to reject H_0

where p = $p\left(Z_{\bar{X}} \geq \left|obtained \quad Z_{\bar{X}}\right|\right)$, which is the p-value.

With the current example using Georgia verbal SAT scores, the p-value is .1902

$p\left(Z_{\bar{X}} \geq \left|obtained \quad Z_{\bar{X}}\right|\right) = .1902$.

If alpha is set to .05 then we have the following:

If $.1902 \leq .05$, the reject H_0; otherwise, fail to reject H_0

Since .1902 is greater than .05, we fail to reject H_0 and conclude that the sample data do not appear to contradict the null hypothesis there for there is no statistical evidence that Georgia mean verbal SAT scores differ from students nationwide.

Side Note—Confidence level

Confidence level is expressed as 1– α, but often in percentage form. Commonly α = .05, so one's confidence level for a given statistical test or confidence interval would be 1 – .05 = .95 or 95% confidence, i.e., one is 95% confidence the population value is within the stated confidence interval.

(d) Summary of Testing H_0

Testing the null hypothesis requires four steps:

- determine and specify both null and alternative hypotheses (in both written and symbolic form)
- specify the degree of risk of a Type 1 error (set the α error) one is willing to make (Type 1 error discussed shortly)
- find the p-value that corresponds to these data for this H_0; that is, assuming H_0 to be true, determine probability of obtaining a statistic that differs from the parameter in H_0 by an amount, in absolute value, as large or larger than that which was observed in the sample given the sample variability observed
- make decision about H_0—reject or fail to reject

Example: A Z test with α = .05:

Suppose one randomly sampled 256 GSU students and the obtained SAT mean, for both math and verbal combined, was M = 1025. If μ = 1000, and σ = 200, what is the probability of observing, in a random sample, a sample mean that deviates from 1000 by this amount, i.e., $p\left(Z_{\overline{X}} \geq \left|obtained \quad Z_{\overline{X}}\right|\right)$?

To calculate the probability, first find the Z value for the sample mean:

$$Z_{\overline{X}} = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} = \frac{1025 - 1000}{200/\sqrt{256}} = \frac{1025 - 1000}{200/16} = \frac{25}{12.5} = 2.00$$

The sample mean ($\overline{X}$ = 1025) is 2.00 standard errors above the population mean. Since the sampling distribution of the sample means is roughly normally distributed (due to the central limit theorem), one can find the probability of obtaining a sample mean this deviant (either above or below) from the population mean for a sample of size 256:

$$p\left(Z_{\overline{X}} \geq |2.00|\right)$$

The probability that $Z_{\overline{X}}$ is greater than or equal to 2.00 is

$$p\left(Z_{\overline{X}} \geq 2.00\right) = .0228,$$

and the probability that $Z_{\overline{X}}$ is less than or equal to –2.00 is

$$p\left(Z_{\overline{X}} \leq -2.00\right) = .0228,$$

so the probability that $Z_{\overline{X}} \geq |2.00|$ is

$$p\left(Z_{\overline{X}} \geq |2.00|\right) = 2(.0228) = .0456.$$

The probability of observing a sample mean that deviates this far from 1000 is .0456, or roughly 4.5 times out of 100.

Given this small probability, would you say this mean difference of 25 points is the result of (a) sampling fluctuation (random chance difference), or (b) some underlying difference between students at GSU and students nationwide who take the SAT?

If performing a hypothesis test and $\alpha = .05$, then one would conclude that such a small probability is unlikely, so the null hypothesis is untenable and the mean difference observed does not appear to be due to chance.

Since the p-value [$p\left(Z_{\bar{X}} \geq |2.00|\right) = .0456$] is less than α, H_0 is rejected in favor of H_1.

The null and alternative hypotheses in this example might be:

Null

GSU students have an average SAT, or
H_0: $\mu = 1000$

Alternative

GSU students do not have an average SAT, or
H_1: $\mu \neq 1000$

How would the conclusions and interpretation change if p = .1753?

Additional Examples:

(i) Hypothesis: GSU students have an average IQ. Note that the population parameters for IQ follows: μ = 100 and σ = 15. The sample of GSU students included 25 students (n = 25) with $\bar{X}$ = 109 (set α = .05.)

(a) What are the null and alternative hypotheses in both written and symbolic form?
(b) What is the calculated Z score?
(c) Is the GSU average statistically different from 100?
(d) So what conclusion to you draw?

(ii) Hypothesis: GSU's SAT average differs from students nationwide. (Note: μ = 1000, σ = 100, n = 40, $\bar{X}$ = 971, α = .10.)

(a) What are the null and alternative hypotheses in both written and symbolic form?
(b) What is the calculated Z score?
(c) Is the GSU average statistically different from 1000?
(d) So what conclusion to you draw?
(e) Does your conclusion change if alpha = .05 or .01?

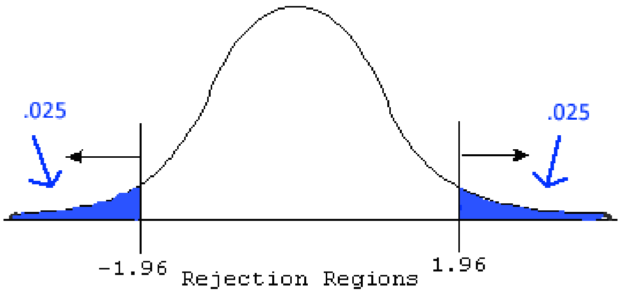
(iii) Hypothesis: The dropout rate across US states differs from Georgia's 2001-2002 dropout rate. (Note: $\mu = 6.5$ [Georgia's dropout rate], $\sigma = 1.76$, $n = 47$, $\overline{X} = 4.40$ [average dropout rate across states], $\alpha = .05$.)

(a) What are the null and alternative hypotheses in both written and symbolic form?
(b) What is the calculated Z score?
(c) Is the national dropout rate different from Georgia's?
(d) So what conclusion to you draw?
(e) Does your conclusion change if alpha = .01?

3. Short Cuts to P-values: Critical Values, Rejection Regions, and Decision Rules

Critical Values:

If alpha = .05, then to distribute alpha to both tails of the z distribution for a non-directional hypothesis, simply divide alpha in half: $\alpha / 2$. So alpha is set at .05, then .05/2 = .025, thus .025 would be in the lower tail, and .025 would be in the upper tail as illustrated below.



By distributing alpha into the tails of the z distribution, another method for hypothesis testing is developed. Rather than finding probabilities for $Z_{\overline{X}}$, critical Z values can be used. For example, with an alpha of .05, a two-tailed test (.05/2 = .025) would result in critical Z, or Z_{crit} of 1.96 and –1.96—more succinctly, ±1.96.

Question – From where were the values ±1.96 derived?

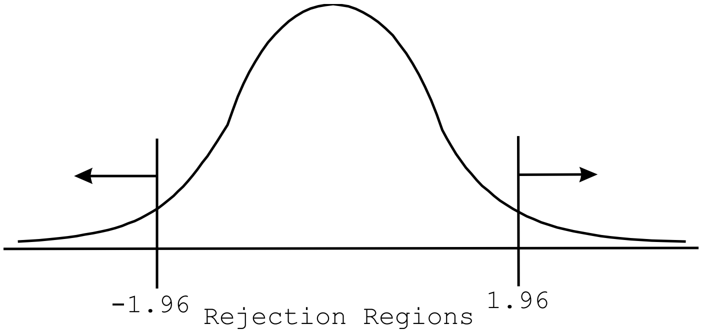
In hypothesis testing, if the calculated Z score for the sample mean is greater than 1.96 or less than –1.96, then H_0 is rejected and the alternative, H_1, is accepted.

Using the z table, find Z_{crit} values for the following:

- What are the critical values for a two-tailed test and $\alpha = .01$?
- What are the critical values for a two-tailed test and $\alpha = .10$?

Rejection Regions:

The regions in the z distribution that critical values establish are called rejection regions because one would reject H_0 if the test statistic, $Z_{\overline{X}}$, fell into one of these regions. The rejection region is expressed in terms of a statistic, z, not a probability. So, for example, a two-tailed test with an α of .05 would have the following rejection regions: $Z_{\overline{X}} \leq -1.96$ and $Z_{\overline{X}} \geq 1.96$. See figure below.



It is important to remember rejection regions are expressed in terms of $Z_{\overline{X}}$ (a test statistic), not probabilities.

Decision Rules for $Z_{\overline{X}}$ scores:

Decision rules are precise statements that indicate when a test statistic, such as $Z_{\overline{X}}$, would lead to a reject or fail to reject H_0 decision. For example, for a two-tailed test (non-directional H_1) using $Z_{\overline{X}}$, the decision rule is:

If $Z_{\overline{X}} \leq -\mathbf{Z}_{\mathbf{crit}}$ or $Z_{\overline{X}} \geq \mathbf{Z}_{\mathbf{crit}}$, then reject $\mathbf{H_0}$; otherwise, fail to reject $\mathbf{H_0}$

where $\mathbf{Z}_{\mathbf{crit}}$ is the $\alpha/2$ critical value from the standard normal distribution, the z table, and $Z_{\overline{X}}$ is the obtained or calculated Z value for the sample mean.

Example: A Z test with Critical Values:

Using the example offer earlier, suppose one randomly sampled 256 GSU students and the obtained SAT mean, for both math and verbal combined, was M = 1025. If $\mu = 1000$, and $\sigma = 200$, is there any evidence that GSU students differ, statistically, from students nationwide at the .05 level of significance?

The calculated Z:

$$Z_{\overline{X}} = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} = \frac{1025-1000}{200/\sqrt{256}} = \frac{1025-1000}{200/16} = \frac{25}{12.5} = 2.00$$

The sample mean ($\overline{X}$ = 1025) is 2.00 standard errors above the population mean.

With an alpha of .05, the critical Z values are ±1.96 so the decision rule:

If $Z_{\overline{X}} \leq -Z_{crit}$ or $Z_{\overline{X}} \geq Z_{crit}$, then reject H_0; otherwise, fail to reject H_0

with relevant numbers:

If $2.00 \leq -1.96$ or $2.00 \geq 1.96$, then reject H_0; otherwise, fail to reject H_0

Since 2 is greater than 1.96, the null hypothesis is rejected and one would conclude that GSU students have a higher SAT combined score than the national average.

Additional Examples:

(i) Hypothesis: GSU students have an average IQ. Note that the population parameters for IQ follows: μ = 100 and σ = 15. The sample of GSU students included 25 students (n = 25) with $\overline{X}$ = 93 (set α = .05.)

(a) What are the null and alternative hypotheses in both written and symbolic form?
(b) What is the calculated Z score?
(c) Is the GSU average statistically different from 100?
(d) So what conclusion to you draw?

(ii) Hypothesis: GSU's SAT average differs from students nationwide. (Note: μ = 1000, σ = 100, n = 40, $\overline{X}$ = 1030, α = .10.)

(a) What are the null and alternative hypotheses in both written and symbolic form?
(b) What is the calculated Z score?
(c) Is the GSU average statistically different from 1000?
(d) So what conclusion to you draw?
(e) Does your conclusion change if alpha = .05 or .01?

(iii) Hypothesis: The dropout rate across US states differs from Georgia's 2001-2002 dropout rate. (Note: μ = 6.5 [Georgia's dropout rate], σ = 1.76, n = 47, $\overline{X}$ = 4.40 [average dropout rate across states], α = .05.)

(a) What are the null and alternative hypotheses in both written and symbolic form?
(b) What is the calculated Z score?
(c) Is the national dropout rate different from Georgia's?
(d) So what conclusion to you draw?
(e) Does your conclusion change if alpha = .01?

4. Assumptions of the Z test

For valid application of the Z test, several assumptions are needed. *Assumptions* are conditions placed on a test statistic, such as $Z_{\bar{X}}$, that are necessary for its valid use in hypothesis testing. Two general assumptions for the Z test are:

> *Normality* – Assume that the sample was taken from a population which is normally distributed; the Z test is usually robust to this assumption due to the central limit theorem. Normality is needed to calculate correct p-values.
>
> *Independence* – Assume that each respondent's score is unrelated to the next respondent's score; one person's answer does not depend upon someone else's answer; if true random sampling is used to select observations, then independence can usually be assumed.

5. Errors in Hypothesis Testing

In hypothesis testing, two decisions can be made, either reject H_0 or fail to reject H_0. Two errors can also be made in deciding whether to reject or fail to reject H_0. The table below specifies each of these errors.

		Population Situation Regarding H_0	
		H_0 True	H_0 False
One's Decision	Reject H_0	Mistake (α) Type I error	Correct $(1 - \beta)$
	Fail to Reject H_0	Correct $(1 - \alpha)$	Mistake (β) Type II error

Case 1: Reject H_0 when H_0 is true.

This is an error because the null was rejected and it should not have been rejected. This is a Type I error (rejecting H_0 when H_0 is true). The probability of making this type of error is equal to α, the alpha level that researchers set, which is traditionally set at .05 or .01 (and sometimes .10).

Case 2: Fail to reject H_0 when H_0 is false (H_1 is true).

This is also an error, and is known as a Type II error (failing to reject H_0 when H_0 is false). This error occurs when one does not reject the null hypothesis when it should have been rejected because there really are differences (thus the alternative hypothesis is actually true). The probability of making this type of error is equal to β; unlike α, the researcher cannot directly set the level of α, but must manipulate other factors which influence α like sample size and/or the alpha level.

Case 3: Reject H_0 when H_0 is false (H_1 is true).

This is a correct decision because H_0 is not true so we adopt the alternative hypothesis, H_1. The probability of this occurring is 1 - β , and this probability is called power.

Case 4: Fail to reject H_0 when H_0 is true.

This is also a correct decision because H_0 was not rejected, and no differences actually exist. The probability of this occurring is 1 - α.

The researcher only has direct control of the α error level. The researcher cannot directly manipulate the α error level; however, several factors can increase or decrease α . These factors include sample size, the alpha level, type of hypothesis, and the amount of variability in the study. These factors are discussed in more detail below.

6. Power (and Factors that Impact Upon It)

Power Described

Errors in hypothesis testing include the Type I error (rejecting H_0 when H_0 is true) and the Type II error (failing to reject H_0 when H_0 is false). The probability of a Type I error is α and is set directly by the researcher. The probability of a Type II error is β and is controlled indirectly by factors which influence the ***power*** to the test.

The ***power*** of a test is the probability of rejecting a false H_0, p(rejecting false H_0); the probability of detecting differences if they actually exist. Power is influenced by (a) effect size, (b) n, (c) control of the variability in studies, (d) choice of hypotheses, and (e) α-level.

Factors Affecting Power

Effect Size—For a Z test (or one sample t test, discussed later) the size of the difference between the true value of μ and that value tested in H_0 ($\overline{X}$) is referred to as the *effect size (ES).* For example, suppose one wanted to test the difference in IQ of this statistics class vs. the national average. The average IQ in the statistics class is 130. One simple measure of effect size is 130 – 100 = 30. If, however, the average IQ in the statistics class was 105, then the effect size would be 105 – 100 = 5.

In general, the larger the effect size the more power the test has for detecting differences. If there are large differences, it will be easier to find them (i.e., easier to reject H_0). But if there are small differences, it will be more difficult to find them (i.e., more difficult to reject H_0).

(Explain why larger ESs provide more power.)

Sample Size – In general, the larger the sample size (n), the more powerful the test. Why does increasing n increase power? Recall that the formula for the standard error of the mean is $\sigma/\sqrt{n}$, so it is easy to see that as n increases, the standard error decreases. Since the standard error is the denominator in the z-score formula for the sample mean, $Z_{\overline{X}} = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$, as the sample size increases, so will $Z_{\overline{X}}$ for given sample mean.

So how does a change in $Z_{\overline{X}}$ affect power? Recall the decision rule:

If $Z_{\overline{X}} \leq -Z_{crit}$ or $Z_{\overline{X}} \geq Z_{crit}$, then reject H_0; otherwise, fail to reject H_0

So as the sample size increases, $Z_{\overline{X}}$ will become larger (in absolute value), and this increases the probability of rejecting H_0, therefore power is increased.

(Explain why increases in n increases power.)

Variability in Studies—Smaller variability yields larger power. As the population variance, σ^2, decreases, power increases. Using the same logic as above, note the formula for the standard error for the sample mean is $\sigma/\sqrt{n}$, so it is easy to see that as σ decreases, the standard error will decrease. Since the standard error is the denominator in the z-score formula for the sample mean, $Z_{\overline{X}} = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$, as σ decreases, the z-score for sample means will increase in absolute value.

So how does this affect power? Recall the decision rule:

If $Z_{\overline{X}} \leq -Z_{crit}$ or $Z_{\overline{X}} \geq Z_{crit}$, then reject H_0; otherwise, fail to reject H_0

So as the σ decreases, the absolute value of $Z_{\overline{X}}$ becomes larger, and this increases the probability of rejecting H_0, so power is increased.

(Explain why decreases in variability increases power.)

Choice of Hypotheses—Directional hypotheses give more power than non-directional hypotheses if the prediction of direction is correct, but directional hypotheses provide zero power if the prediction of direction is incorrect.

This relationship can be shown as follows. If the researcher sets $\alpha = .05$, then for a two-tailed (non-directional) test the critical values are 1.96 and –1.96. For a one-tailed (upper-tailed) test, however, the critical value is 1.64 for $\alpha = .05$.

So, if one calculates a z-score for a sample mean and it equals, say, 1.78 (i.e., $Z_{\overline{X}} = 1.78$), then which test is more powerful, the two-tail or one-tailed test?

With the two-tailed test, H_0 would not be rejected since 1.78 does not fall within the rejection region (i.e, 1.78 is not greater than 1.96 or less than -1.96). However, with the one-tailed test H_0 is rejected because the obtained z score, 1.78, lies within the rejection region (i.e, 1.78 > 1.64). So directional hypotheses are more powerful because their critical values are smaller than the corresponding critical values of non-directional tests.

(Explain why directional tests are more powerful than non-directional tests. Are directional tests always more powerful; if not, under what circumstances are they less powerful?)

Alpha (α)—The larger the α, the greater the power. That is, the greater the probability of rejecting a false H_0, the greater the chance of finding a difference (accepting H_1).

As α becomes larger, say from .01 to .05, one should easily see that it will be easier to reject H_0, and since it is easier to reject H_0, power is increased. For example, the critical value for a one-tailed test with α = .01 is 2.32, but increasing α to .05 results in a critical value of 1.64. Since the critical Z values, Z_{crit}, are smaller with larger α's, smaller calculated $Z_{\overline{X}}$ values are needed to reject H_0. In short, larger α's result in more power.

(Why does increasing alpha provide increased power?)

Which Factors to Alter?

To increase power, the easiest factors for the researcher to manipulate are n and α, but α is usually set at .10, .05, and .01 by tradition. In some circumstances one may also be able to choose directional tests.