MODULE 14

HYPOTHESIS TESTS

Contents

A STATISTIC IS AN IMPERFECT ESTIMATE of a parameter because of sampling variability. There are two calculations using the results of a single sample that recognize this imperfection and allow conclusions to be made about a parameter. First, a researcher may form an *a priori* hypothesis about a parameter and then use the information in the sample to make a judgment about the "correctness" of that hypothesis. Second, a researcher may form, from the information in the sample, a range of values that is likely to contain the parameter. The first method is called *hypothesis testing* and is the subject of this module. The second method consists of constructing a *confidence region*, which is introduced in Module 15. Specific applications of these two techniques are described in Modules 16-**??**.

14.1 Hypothesis Testing & The Scientific Method

In its simplest form, the scientific method has four steps:

1. Observe and describe a natural phenomenon.
2. Formulate a hypothesis to explain the phenomenon.
3. Use the hypothesis to predict new observations.
4. Experimentally test the predictions.

If the results of the experiment do not match the predictions, then the hypothesis is rejected and an alternative hypothesis is proposed. If the results of the experiment closely match the predictions, then belief in the hypothesis is gained, though the hypothesis will likely be subjected to further experimentation.

Statistical hypothesis testing is key to using the scientific method in many fields of study and, in fact, closely follows the scientific method in concept. Statistical hypothesis testing begins by formulating two competing statistical hypotheses from a research hypothesis. One of these hypotheses (the null) is used to predict the parameter of interest. Data is then collected and statistical methods are used to determine whether the observed statistic closely matches the prediction made from the null hypothesis or not. Probability (Module 12) is used to measure the degree of matching with sampling variability taken into account. This process and the theory underlying statistical hypothesis testing is explained in detail in this module.

14.2 Statistical Hypotheses

Hypotheses are classified into two types: (1) research hypothesis and (2) statistical hypotheses. A research hypothesis is a "wordy" statement about the question or phenomenon that the researcher is testing. Four example research hypotheses are:

1. A medical researcher is concerned that a new medicine may change patients' mean pulse rate (from the "known" mean pulse rate of 82 bpm for individuals in the study population not using the new medicine).
2. A chemist has invented an additive to car batteries that she thinks will extend the current 36 month average life of a battery.
3. An engineer wants to determine if a new type of insulation will reduce the average heating costs of a typical house (which are currently \$145 per month).
4. A researcher is concerned whether, on average, Alzheimer's caregivers at a particular facility are clinically depressed (as suggested by a mean Beck Depression Inventory (BDI) score greater than 25)

Research hypotheses are converted to statistical hypotheses that are mathematical and more easily subjected to statistical methods. There are two types of statistical hypotheses: (1) the null hypothesis and (2) the alternative hypothesis. The **null hypothesis**, abbreviated as H_0, is a specific statement of no difference between a parameter and a specific value or between two parameters. The H_0 ALWAYS contains an equals sign because it always represents "no difference." The **alternative hypothesis**, abbreviated as H_A, always states that there is some sort of difference between a parameter and a specific value or between two parameters. The type of difference comes from the research hypothesis and will require use of a less than ($<$), greater than ($>$), or not equals ($\neq$) sign. Null and alternative hypotheses that correspond to the four research hypotheses above are:

1. $H_A : \mu \neq 82$ and $H_0 : \mu = 82$ (where μ represents the mean pulse rate for individuals in the study population that take the new medicine; thus, the alternative hypothesis represents a change from the "normal" pulse rate).
2. $H_A : \mu > 36$ and $H_0 : \mu = 36$ (where μ represents the mean life of batteries with the new additive; thus, this alternative hypothesis represents an extension of the current battery life).
3. $H_A : \mu < 145$ and $H_0 : \mu = 145$ (where μ represents the mean monthly heating bill for houses that receive the new type of insulation; thus, this alternative hypothesis represents a decline in heating bills from the previous "normal" amount).
4. $H_A : \mu > 25$ and $H_0 : \mu = 25$ (where μ represents the mean BDI score; thus, this alternative hypothesis represents a mean score that indicates clinical depression).

The sign used in the alternative hypothesis comes directly from the wording of the research hypothesis (Table 14.1). An alternative hypothesis that contains the $\neq$ sign is called a **two-tailed alternative**, as the value can be "not equal" to another value in two ways; i.e., less than or greater than. Alternative hypotheses with the $<$ or the $>$ signs are called **one-tailed alternatives**. The null hypothesis is easily constructed from the alternative hypothesis by replacing the sign in the alternative hypothesis with an equals sign.

Table 14.1. Common words that indicate which sign to use in the alternative hypothesis.

$>$	$<$	$\neq$
is greater than	is less than	is not equal to
is more than	is below	is different from
is larger than	is lower than	has changed from
is longer than	is shorter than	is not the same as
is bigger than	is smaller than	
is better than	is reduced from	
is at least	is at most	
is not less than	is not more than	

14.2.1 Hypothesis Testing Concept

Statistical hypothesis testing begins by using the null hypothesis to predict what value one should expect for the mean in a sample. So, for the Square Lake example (from Module 1), if $H_0 : \mu = 105$ and $H_A : \mu < 105$, then one would expect, if the null hypothesis is true, that the observed sample mean would be 105. If the observed sample mean was NOT equal to 105 and sampling variability did not exist, then the prediction based on the null hypothesis would not be supported and one would conclude that the null hypothesis was incorrect. In other words, one would conclude that the population mean was not equal to 105.

Of course, sampling variability does exist and it complicates matters. The simple interpretation of not supporting H_0 because the observed sample mean did not equal the hypothesized population mean canNOT be made because, with sampling variability, one would not expect a statistic to exactly equal the parameter in the population from which the sample was extracted. For example, even if the null hypothesis was correct, one would not expect, with sampling variability, the observed sample mean to exactly equal 105; rather, one would expect the observed sample mean to be **reasonably** close to 105.

Thus, hypothesis testing is a process to determine if the difference between the observed statistic and the expected statistic based on the null hypothesis is "large" **relative to sampling variability**. For example, the standard error of $\bar{x}$ for samples of $n = 50$ in the Square Lake example is $\frac{\sigma}{\sqrt{n}} = \frac{31.5}{\sqrt{50}} = 4.45$. With this sampling variability, an observed sample mean of 103 would be considered reasonably close to 105 and one would have more belief in $H_0 : \mu = 105$ (Figure 14.1). However, an observed sample mean of 90 is further away from 105 than one would expect based on sampling variability alone and belief in $H_0 : \mu = 105$ would lessen (Figure 14.1).

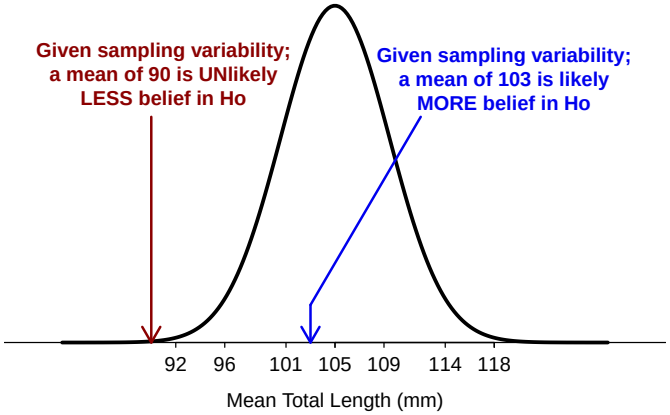

Figure 14.1. Sampling distribution of samples means of n=50 from the Square Lake population ASSUMING that μ=105.

While the above procedure is intuitively appealing, the conclusions are not as clear when the examples chosen (i.e., samples means of 103 and 90) are not as extremely close or distant from the null hypothesized value. For example, what would one concludeif the observed sample mean was 97?. A first step in creating a more objective decision criteria is to compute the "p-value." A p-value is the probability of the observed statistic or a value of the statistic more extreme assuming that the null hypothesis is true. The p-value is described in more detail below given its centrality to making conclusions about statistical hypotheses.

The meaning of the phrase "or more extreme" in the p-value definition is derived from the sign in H_A (Figure 14.2). If H_A is the "less than" situation, then "or more extreme" means "less than" or "shade to the left" for the probability calculation. The "greater than" situation is defined similarly but would result in shading to the "right." In the "not equals" situation, "or more extreme" means further into the tail AND the exact same size of tail on the other side of the distribution. It is clear from Figure 14.2 why "less than" and "greater than" are one-tailed alternatives and "not equals" is a two-tailed alternative.

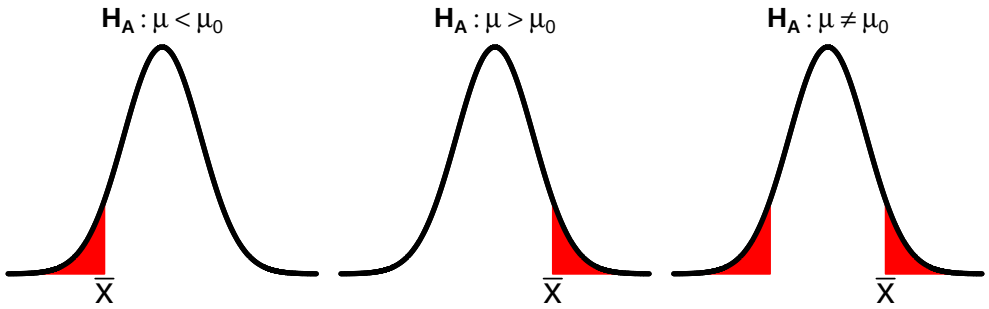

Figure 14.2. Depiction of "or more extreme" (red areas) in p-values for the three possible alternative hypotheses.

The "assuming that the null hypothesis is true" phrase is used to define a μ for the sampling distribution on which the p-value will be calculated. This sampling distribution is called the **null distribution** because it depends on the value of μ from the null hypothesis. One must remember that the null distribution represents the distribution of all possible sample means assuming that the null hypothesis is true; it does NOT represent the actual sample means.[1] The null distribution in the Square Lake example is thus $\bar{x} \sim N(105, 4.45)$ because $n = 50 > 30$ (so the Central Limit Theorem holds), $H_0 : \mu = 105$, and SE=$\frac{31.49}{\sqrt{50}}$=4.45.

The p-value is computed with a "forward" normal distribution calculation on the null sampling distribution. For example, suppose that a sample mean of 100 was observed with $n = 50$ from Square Lake (as it was in Table 2.2). The p-value in this case would be "the probability of observing $\bar{x} = 100$ or a smaller value assuming that $\mu = 105$." This probability is computed by finding the area to the left of 100 on a $N(105, 4.45)$ null distribution and is the exact same type of calculation as that made in Section 13.3. Thus, this p-value of $p = 0.1308$ is computed as below and shown in Figure 14.3.

```
> ( distrib(100,mean=105,sd=31.49/sqrt(50)) )
[1] 0.1307722
```

Interpreting the p-value requires critically thinking about the p-value definition and how it is calculated. Small p-values appear when the observed statistic is "far" from the null hypothesized value. In this case there is a small probability of seeing the observed statistic ASSUMING that H_0 is true. Thus, the assumption is likely wrong and H_0 is likely incorrect. In contrast, large p-values appear when the observed statistic is close to the null hypothesized value suggesting that the assumption about H_0 may be correct.

[1] Of course, unless the null hypothesis happens to be perfectly true.

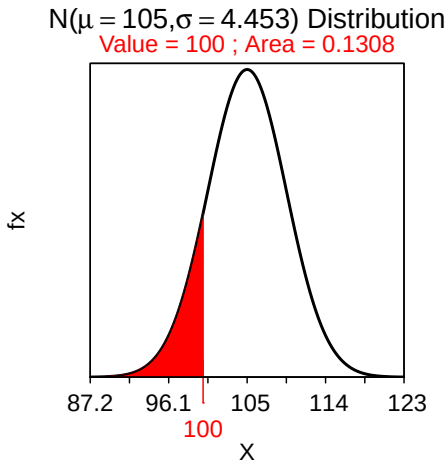

Figure 14.3. Depiction of the p-value for the Square Lake example where $\bar{x} = 100$ and $H_A : \mu < 105$.

The p-value serves as a numerical measure on which to base a conclusion about H_0. To do this objectively requires an objective definition of what it means to be a "small" or "large" p-value. Statisticians use a cut-off value, called the rejection criterion and symbolized with α, such that p-values less than α are considered small and would result in rejecting H_0 as a viable hypothesis. The value of α is typically small, usually set at 0.05, although $\alpha = 0.01$ and $\alpha = 0.10$ are also commonly used.

The choice of α is made by the person conducting the hypothesis test and is based on how much evidence a researcher demands before rejecting H_0. Smaller values of α require a larger difference between the observed statistic and the null hypothesized value and, thus, require "more evidence" of a difference for the H_0 to be rejected. For example, if rejection of the null hypothesis will be heavily scrutinized by regulatory agencies, then the researcher may want to be very sure before claiming a difference and should then set α at a smaller value, say $\alpha = 0.01$. The actual choice for α MUST be made before collecting any data and canNOT be changed once the data has been collected. In other words, once the data are in hand, a researcher cannot lower or raise α to achieve a desired outcome regarding H_0.

◇ The value of the rejection criterion (α) is set by the researcher BEFORE data is collected.

The null hypothesis in the Square Lake example is not rejected because the p-value (i.e., 0.1308) is larger than any of the common values of α. Thus, the conclusion in this example is that it is possible that the mean of the entire population is equal to 105 and it is not likely that the population mean is less than 105. In other words, observing a sample mean of 100 is likely to happen based on random sampling variability alone and it is unlikely that the null hypothesized value is incorrect.

14.3 Test Statistics and Effect Sizes

Instead of reporting the observed statistic and the resulting p-value, it may be of interest to know how "far" the observed statistic was from the hypothesized value of the parameter. This is easily calculated with

$$\text{Observed Statistic} - \text{Hypothesized Parameter}$$

where "Hypothesized Parameter" represents the specific value in H_0. However, the meaning of this difference is difficult to interpret without an understanding of the standard error of the statistic. For example, a difference of 10 between the observed statistic and the hypothesized parameter seems "very different" if the standard error is 3 but does not seem "different" if the standard error is 15 (Figure 14.4).

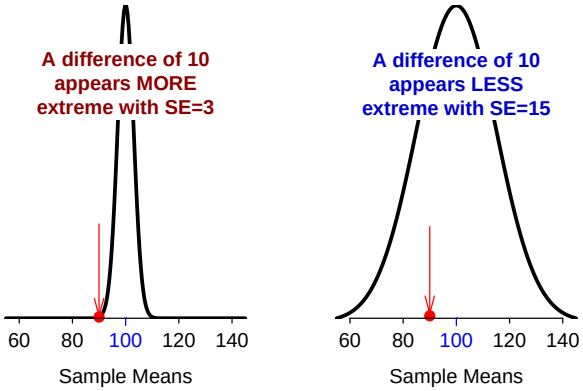

Figure 14.4. Sampling distribution of samples means with SE=3 (Left) and SE=15 (Right). A single observed sample mean of 90 (a difference of 10 from the hypothesized mean of 100) is shown by the red dot and arrow.

The difference between the observed statistic and the hypothesized parameter is standardized to a common scale by dividing by the standard error of the statistic. The result is called a *test statistic* and is generalized with

$$\text{Test Statistic} = \frac{\text{Observed Statistic} - \text{Hypothesized Parameter}}{SE_{\text{Statistic}}} \tag{14.3.1}$$

Thus, the test statistic (14.3.1) measures how many standard errors the observed statistic is away from the hypothesized parameter. A relatively large value is indicative of a difference that is likely not due to randomness (i.e., sampling variability) and suggests that the null hypothesis should be rejected.

The test statistic in the Square Lake Example is $\frac{100-105}{\frac{31.49}{\sqrt{50}}} = -1.12$. Thus, the observed mean total length of 100 mm is 1.12 standard errors below the null hypothesized mean of 105 mm. From our experience, a little over one SE from the mean is not "extreme" and, thus, it is not surprising that the null hypothesis was not rejected.

There are other forms for calculating test statistics, but all test statistics retain the general idea of scaling the difference between what was observed and what was expected from the null hypothesis in terms of sampling variability. Even though there is a one-to-one relationship between a test statistic and a p-value, a test statistic is often reported with a hypothesis test to give another feel for the magnitude of the difference between what was observed and what was predicted.

14.4 Hypothesis Testing Concept Summary

In summary, hypotheses are statistically examined with the following procedure.

1. Construct null and alternative hypotheses from the research hypothesis.
2. Construct an expected value of the statistic based on the null hypothesis (i.e., assume that the null hypothesis is true).
3. Calculate an observed statistic from the individuals in a sample.
4. Compare the difference between the observed statistic and the expected statistic based on the null hypothesis in relation to sampling variability (i.e., calculate a test statistic and p-value).
5. Use the p-value to determine if this difference is "large" or not.
    - If this difference is "large" (i.e., p-value$< \alpha$), then reject the null hypothesis.
    - If this difference is not "large" (i.e., p-value$> \alpha$), then "Do Not Reject" the null hypothesis.

Statisticians say "do not reject H_0" rather than "accept H_0 as true" when the p-value $> \alpha$ for two reasons. First, there are several other possible values, besides the specific value in the null hypothesis, that would lead to "do not reject" conclusions. For example, if a null hypothesized value of 105 was not rejected, then values of 104.99, 104.98, etc. would also likely not be rejected.[2] So, we don't say that we "accept" a particular hypothesized value when we know many other values would also be "accepted."

Second, the null hypothesis is almost always not true. Consider the null hypothesis of the Square Lake example (i.e., "that the mean length is 105 mm"). The mean length of fish in Square Lake is undoubtedly not exactly equal to 105. It may be 104.9, 105.01, or some other more disparate value. The point is that the specific value of the hypothesis is likely never true, especially for a continuous variable. The problem is that it takes large amounts of data to be able to distinguish means that are very close to the true population mean (i.e., it is difficult to distinguish between 104.9 and 105 when sampling variability is present). Very often we will not take a sample size large enough to distinguish these subtle differences. Thus, we will say that we "do not reject H_0" because there simply was not enough data to reject it.

14.5 Errors and Power

The goal of hypothesis testing is to make a decision about H_0. Unfortunately, because of sampling variability, there is always a risk of making an incorrect decision. Two types of incorrect decisions can be made (Table 14.2). A Type I error occurs when a true H_0 is falsely rejected. In other words, even if H_0 is true, there is a chance that a rare sample will occur and H_0 will be deemed incorrect. The probability of making a Type I error is set when α is chosen. A Type II error occurs when a false H_0 is not rejected. The probability of a Type II error is denoted by β.

Table 14.2. Types of decisions that can be made from a hypothesis test.

		Decision from Data	
		Reject	Not Reject
Truth About	H_0	Type I	Correct
Population	H_A	Correct	Type II

[2] In fact, for example, the values in a 95% confidence interval – see Module **??** – represent all possible hypothesized values that would not be rejected with a two-tailed H_A using $\alpha = 0.05$.

The decision in the Square Lake example above produced a Type II error because $H_0 : \mu = 105$ was not rejected even though we know that $\mu = 98.06$ (Table 2.1). Unfortunately, in real life, it will never be known exactly when a Type I or a Type II error has been made because the true μ is not known. However, it is known that a Type I error will be made $100\alpha\%$ of the time. The probability of a type II error (β), though, is never known because this probability depends on the true μ. Decisions can be made, however, that affect the magnitude of β (discussed below with power).

A concept that is very closely related to decision-making errors is the idea of **power**. Power is the probability of correctly rejecting a false H_0. In other words, it is the probability of detecting a difference from the hypothesized value if a difference really exists. Power is used to demonstrate how sensitive a hypothesis test is for identifying a difference. High power related to a H_0 that is not rejected implies that the H_0 really should not have been rejected. Conversely, low power related to a H_0 that was not rejected implies that the test was very unlikely to detect a difference, so not rejecting H_0 is not surprising nor particularly conclusive.

Power is equal to $1 - \beta$ and, thus, like β it cannot be computed directly. However, a researcher can make decisions that will positively affect power (Figure 14.5). For example, a researcher can increase power by increasing α or n. Increasing n is more beneficial because it does not result in an increase in Type I errors as would occur with increasing α.

In addition, power decreases as the difference between the hypothesized mean (μ_0) and the actual mean (μ_A) decreases (Figure 14.5). This means that the ability to detect increasingly smaller differences decreases. In addition, power decreases with an increasing amount of natural variability (i.e., σ; Figure 14.5). In other words, the ability to detect a difference decreases with increasing amounts of variability among individuals. A researcher cannot control the difference between μ_0 and μ_A or the value of σ. However, it is important to know that if a situation with a "large" amount of variability is encountered or the difference to be detected is small, the researcher will need to increase n to gain power.

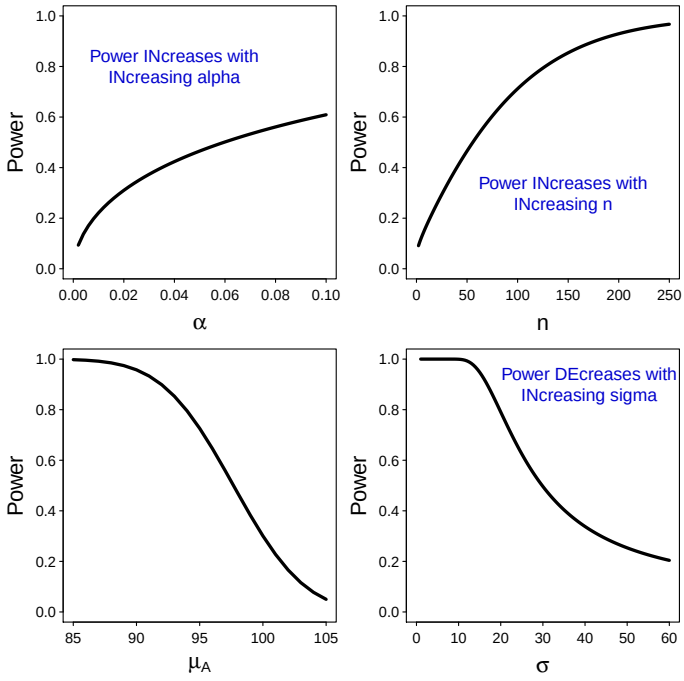

Figure 14.5. The relationship between one-tailed (lower) power and α, n, actual mean (μ_A), and σ. In all situations where the variable does not vary, $\mu_0 = 105$, $\mu_A = 98.06$, $\sigma = 31.49$, $n = 50$, and $\alpha = 0.05$.

Power cannot usually be calculated because the actual mean (μ_A) is not known. However, in the Square Lake example, μ_A is known and power can be calculated in four steps:

1. Draw the sampling distribution assuming the H_0 is true (called the null distribution).
   - The null distribution is $N(105, \frac{31.49}{\sqrt{50}})$ because $H_0 : \mu = 105$, $\sigma = 31.49$, and $n = 50$.
2. Find the rejection region borders (based on α and H_A) in terms of the value of the statistic (a "reverse" calculation on the null distribution).
   - The rejection region is delineated by the $\bar{x}$ that has $\alpha = 0.10$ to the left (because H_A is a "less than"). This reverse calculation on the null distribution gives $\bar{x}$=99.2928.

   ```
   > ( rejreg <- distrib(0.10,mean=105,sd=31.49/sqrt(50),type="q") )
   [1] 99.29279
   ```
3. Draw the sampling distribution corresponding to the "actual" parameter value (SE is the same as that for the null distribution).
   - The actual μ is 98.06. Thus, the actual sampling distribution is $N(98.06, \frac{31.49}{\sqrt{50}})$.
4. Compute the portion of the "actual" sampling distribution in the REJECTION region of the null distribution (i.e., a "forward" calculation on the actual distribution).
   - This computation is to find the area to the left of $\bar{x}$=99.2928 on $N(98.06, \frac{31.49}{\sqrt{50}})$. The area to the left of this Z is 0.6090.

   ```
   > ( distrib(rejreg,mean=98.06,sd=31.49/sqrt(50)) )
   [1] 0.6090419
   ```

Thus, the power to detect a $\mu_A = 98.06$ was 0.6090. This means that in only about 61% of the samples will the false $H_0 : \mu = 105$ be correctly rejected. Thus, it is not too surprising that H_0 was not rejected in this example. If n could be doubled to 100, however, the power to correctly reject $H_0 : \mu = 105$ would increase to approximately 0.82 (Figure 14.5).