February 23, 2009

Chapter 6: Introduction to Hypothesis Testing

One of the primary uses of statistics is to use data to *infer* something about a population or a probability model. Populations and probability models are defined by parameters. For instance, let μ_1 and μ_2 be the mean cholesterol levels of people who eat two different diets. We may be interested if there is a difference on average between cholesterol levels due to diet. In order to test this, we can collect data on cholesterol levels from people who eat the two different diets and perform a *hypothesis test* to determine if the mean cholesterol levels differ depending on diet. This chapter introduces the basics of hypothesis testing: terminology, the logic of hypothesis testing, test statistics, and decision making.

1 Introduction

In order to introduce the topic of hypothesis testing, we will begin with an illustrative example:

Example: Body Temperatures. It is widely believed that the average body temperature for healthy adults is 98.6 degrees Fahrenheit. Is this true? Where did the 98.6 degree value come from? Do all healthy people have exactly the same body temperature? A study was conducted a few years go to examine this belief. The body temperatures of $n = 130$ healthy adults were measured (half male and half female). The average temperature from the sample was found to be $\bar{x} = 98.249$ with standard deviation $s = 0.7332$. Do these statistics contradict the belief that the average body temperature is 98.6? If the true average temperature is indeed 98.6 and we obtain a sample of $n = 130$ healthy adults, we would not expect the sample mean to come out exactly equal to 98.6. We observed $\bar{x} = 98.249$ – can this deviation from 98.6 be explained by chance or is it unlikely we would observe a value this different from 98.6? Two people debating this issue could come to different conclusions. What is needed is an objective method to determine if the data contradict the hypothesis that the average body temperature is 98.6.

In this example, the parameter of interest is μ, the mean temperature of healthy adults. We want to test a hypothesis about μ. The way hypothesis testing is done is that a hypothetical value is proposed for μ which we denote by μ_0. The **null hypothesis**, denoted H_0, specifies that $\mu = \mu_0$:

$$H_0 : \mu = \mu_0.$$

In the body temperature example, $\mu_0 = 98.6$ and the null hypothesis is

$$H_0 : \mu = 98.6.$$

Typically, the null hypothesis represents the "status quo." The purpose of many studies is to determine if the data leads us to *reject* the null hypothesis. The **alternative hypothesis**, denoted H_a, is set up to represent the research goal:

$$H_a : \mu \neq \mu_0.$$

In the body temperature example, we have

$$H_a : \mu \neq 98.6.$$

This is an example of a *two-sided* alternative because we will reject the null hypothesis if there is evidence that the true mean lies to either side (greater or less than) of the hypothesized mean value of 98.6

The way the alternative hypothesis is set up depends on the scientific objective at hand. Many examples of alternative hypotheses are *one-sided*. For example, if we want to determine if an environmental toxin such as PCB reduces the mean eggshell thickness of pelican birds, then we would set up our hypotheses as:

$$H_0 : \mu = \mu_0 \text{ versus } H_a : \mu < \mu_0,$$

where μ_0 is the mean thickness for birds not exposed to PCB.

On the other hand, if we want to test if a new drug increases the mean survival time for people suffering from a particular type of cancer, then we would set up our hypotheses as:

$$H_0 : \mu = \mu_0 \text{ versus } H_a : \mu > \mu_0,$$

where μ_0 is the mean survival time without the medication.

Once the data is collected and analyzed, a decision has to be made. Should we reject H_0 and accept H_a? Or is there insufficient evidence to reject H_0? When making decisions, there are four possible scenarios and two of them involve errors:

1. Accept H_0 when in fact H_0 is true (good decision).
2. Accept H_0 when in fact H_0 is false (an error).
3. Reject H_0 when in fact H_0 is true (an error).
4. Reject H_0 when H_0 is false (good decision).

The two possible errors above have names:

Definition. A **Type I Error** is rejecting the null hypothesis H_0 when H_0 is true. The probability of committing a type I error is denoted by α and is called the **significance level** of the test.

Definition. A **Type II Error** is accepting H_0 when H_0 is false. The probability of a type II error is denoted by β.

Definition. The **power** of a statistical test is $1 - \beta$, which is the probability of rejecting H_0 when H_0 is false.

When testing hypotheses, we would like the test to have high power which means the ability to conclude the null hypothesis is false when it really is false with high probability. We would also like the probability of a type I error, α, of our tests to be small. Unfortunately, making the probability of a type I error smaller makes the test less powerful; making the test more powerful leads to a higher type I error. Therefore, a compromise is needed between these competing goals when performing hypothesis testing.

Generally, tests are set up so as to minimize the probability of committing a type I error. Typical values for the significance level α, the probability of a type I error, used in practice are 0.05, 0.01, or 0.10. We do not want to reject a null hypothesis that is true. In the body temperature, committing a type I error means that one would conclude the average body temperature differs from 98.6 when in fact the average body temperature is 98.6. Most thermometers for humans are marked at 98.6. Imagine throwing all these thermometers out because a scientific study says they are marked wrong and then realizing later that they were actually marked correctly.

A useful analogy for hypothesis testing is a court of law. The defendant is assumed innocent till proven guilty. Thus, the null hypothesis is that the defendant is innocent and the alternative hypothesis is that the defendant is guilty:

$$H_0: \text{ Innocent}$$

versus

$$H_a: \text{ Guilty.}$$

The trial starts and evidence is presented. In the statistical setting, the data is the evidence. Does the data allow us to reject H_0 and conclude H_a? Convicting an innocent man is committing a type I error: rejecting H_0 when it is true. We certainly do not want to convict innocent people, so we set up hypothesis tests to minimize the probability of committing this error. On the other hand, we do not want to let a guilty defendant go free (i.e. commit a type II error). Note that there are two reasons a defendant will not be convicted in practice:

1. The defendant is innocent (H_0 is true).
2. The defendant is guilty (H_0 is false), but we lack enough evidence to convict (a type II error). In statistics, lack of evidence corresponds to lack of data. If we do not have much data (i.e. the sample size is too small), then we will lack the evidence needed to reject H_0 when it is false.

We do not necessarily say a defendant is innocent (accept the null hypothesis) if we fail to convict because the failure could be due to insufficient evidence (reasonable doubt remains). Similarly, in hypothesis testing, if we do not reject the null hypothesis, we generally refrain from saying that we accept the null hypothesis (guard against a type II error); instead we may say that we "fail to reject H_0."

Students often find the logic of hypothesis testing difficult to understand at first. The court of law analogy should help with the understanding.

2 t-Test

Now, for the mechanics of performing a hypothesis test, the main idea is to compare the observed sample mean $\bar{x}$ to the hypothesized value μ_0 and see if the difference $\bar{x} - \mu_0$ is big or small. However, in order to have a framework to decide if the difference is big or not, we need to standardize this difference so that we will have a scale for comparisons. The standardized difference is our *test statistic* t:

$$\textbf{t-Test Statistic:} \qquad t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}. \tag{1}$$

Fact: If H_0 is true, then test statistic t in (1) follows a t-distribution on $n-1$ degrees of freedom, provided the data are from a normal distribution.

Thus, the t-distribution is our reference distribution for making a decision. If our observed test statistic t looks like it came from the t-distribution, then that would be consistent with the null hypothesis. However, if t does not look like it came from a t-distribution (because it is too big), then the standard difference $t = (\bar{x} - \mu_0)/(s/\sqrt{n})$ is too big to be explained by chance and therefore we would reject the null hypothesis.

We need a cut-off value to decide if our test statistic t is "too big."

Definition. The **Rejection Region** is the range of values of the test statistic t for which H_0 is rejected.

To determine the cut-off value, one chooses a significance level α (e.g. $\alpha = 0.05$). Recall that α is the probability of making a type I error (i.e. rejecting H_0 when H_0 is true). In the body temperature example, our alternative hypothesis is two-sided which means we will reject H_0 if t is too big in the positive or negative direction. Let us denote our cut-off value by c. Then we need to choose c such that

$$\begin{aligned} \alpha &= P(\text{Rejecting } H_0 \text{ when } H_0 \text{ is true}) \\ &= P(t > c \text{ or } t < -c, \text{ when } \mu = \mu_0) \end{aligned}$$

This implies that c must equal the $1 - \alpha/2$ percentile of the t-distribution on $n - 1$ degrees of freedom.

Summarizing, we have the following testing procedure:

Two-Sided test for the Mean of a Normal Distribution

$$H_0 : \mu = \mu_0 \quad \text{versus} \quad H_a : \mu \neq \mu_0.$$

Test Statistic: $t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$.

Decision:

$$\begin{array}{lll} \text{Rejection Region:} & \text{Reject } H_0 & \text{if } t > t_{n-1,1-\alpha/2} \text{ or } t < -t_{n-1,1-\alpha/2}. \\ & \text{Fail to Reject } H_0 & \text{otherwise.} \end{array}$$

Notice that we had to split the significance level α in two for the two-sided test. For one-sided tests, we do not need to split the significance level because we will only reject H_0 if the test statistic lies to one side of the hypothesized mean. Using the same reasoning, we get the following results for one-sided tests:

$$H_0 : \mu = \mu_0 \quad \text{versus} \quad H_a : \mu < \mu_0$$

Rejection Region: Reject H_0 if $t < -t_{n-1,1-\alpha}$.

Otherwise, fail to reject H_0.

$$H_0 : \mu = \mu_0 \quad \text{versus} \quad H_a : \mu > \mu_0$$

Rejection Region: Reject H_0 if $t > t_{n-1,1-\alpha}$.

Otherwise, fail to reject H_0.

Body Temperature Example continued. We can apply the hypothesis testing procedure to the body temperature example. Let us test the hypothesis using a significance level of $\alpha = 0.05$. Recall that the sample size is $n = 130$. Because it is a two-sided alternative, we shall reject the null hypothesis if our test statistic exceeds

$$t_{n-1,1-\alpha/2} = t_{129,0.975}$$

in absolute value. Using SAS's tinv function, we find that

$$t_{129,0.974} = 1.97852 \approx 1.98.$$

Thus, if the test statistic t exceeds 1.98 in magnitude, we shall reject the null hypothesis that the average body temperature is $\mu_0 = 98.6$. If $|t|$ does not exceed 1.98, then we will not reject H_0 and conclude that there is insufficient evidence that the mean body temperature differs from 98.6.

The sample mean and standard deviation for the body temperature data are $\bar{x} = 98.25$ and $s = 0.73$ respectively. The test statistic is

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{98.25 - 98.6}{0.73/\sqrt{130}} = -5.455.$$

The test statistic $t = -5.455 < -1.98$ falls in the rejection region, so we reject H_0 using a significance level $\alpha = 0.05$. Here is the logic: if H_0 is true and the average body temperature is indeed 98.6, then our test statistic t should look like it came from a t-distribution on 129 degrees of freedom; but the observed test statistic is $t = -5.455$. It is very unlikely that this value of t would be produced by a t-distribution on 129 degrees of freedom. Therefore, the assumption that H_0 is true does not seem plausible. Hence we reject H_0 and accept H_a.

Conclusion: At the 5% significance level we reject H_0 and conclude that the average body temperature for healthy adults *differs* from 98.6.

Here are two important notes about the hypothesis test:

1. When stating your conclusion, state the significance level at which you conducted the test so the reader will know the strength of the statistical evidence against the null hypothesis. We rejected H_0 using $\alpha = 0.05$. Therefore, with this testing procedure, there is only a 5% probability of incorrectly rejecting the null hypothesis.

2. Your conclusion should be stated to be consistent with your hypotheses. Recall that we rejected H_0 and accepted the alternative hypothesis that $\mu \neq 98.6$. Therefore, our conclusion is that the mean body temperature *differs* from 98.6. We cannot change our mind after seeing the data and decide we want the alternative hypothesis to be $H_a : \mu < 98.6$ since the sample mean came out to be lower than 98.6. This will inflate the type I error probability.

3 p-Values

The method of testing described above is known as the **critical value method** since we specified a significance level α which in turn determined a cut-off value or *critical value* (in this case $t_{n-1,1-\alpha/2}$) for deciding whether or not to reject the null hypothesis. To use this method, one needs to decide upon the significance level α and this decision will depend on the type of application. Typically, the value of $\alpha = 0.05$ is used. For better assurance against committing a type I error, one could use $\alpha = 0.01$. Using a smaller α will make it more difficult to reject the null hypothesis, decreasing the power of the test.

A problem with the critical value method is the following: Recall that the critical value in the body temperature example was 1.9785. Suppose the test statistic came out to be $t = -1.94$ (instead of -5.455). Then the conclusion of the test would be to *not* reject H_0 at $\alpha = 0.05$. However, a value of $t = -1.94$ is right on the border of the rejection region and the evidence is still fairly strong for rejecting H_0, but not quite at the $\alpha = 0.05$ level. If we test at $\alpha = 0.05$ all we can state is that we fail to reject H_0 at $\alpha = 0.05$ without letting on that the actual strength of evidence is quite strong.

Also note that the actual test statistic value is $t = -5.455$. Could we have rejected H_0 using $\alpha = 0.025$ or $\alpha = 0.01$? What is the smallest significance level at which we could still reject H_0? The answer to this question is the *p-value.* The p-value is the smallest significance level at which we would reject the null hypothesis. For this reason, the p-value is sometimes called the *observed significance level.*

Let T denote a t-random variable on $n-1$ degrees of freedom. We can define the p-value as following:

Definition: The **p-value** is the probability that the test statistic takes the value we observed or more extreme away from the null hypothesis if the null hypothesis is true. For a *two-sided* alternative hypothesis, the p-value is computed as:

$$p\text{-value:} = P(T > |t| \text{ or } T < -|t|, \text{ when } H_0 \text{ is true}) \text{ (Two-Sided } p\text{-value)}.$$

The p-value answers the following question: If the null hypothesis is true, how likely is it that our observed test statistic takes the value we observed or more extreme? If this probability is small, then we reject the null hypothesis. If the p-value is not small, then we do not reject the null hypothesis.

Interpreting p-values. Here are some rough guidelines for interpreting p-values which can be used in any testing scenario (not just for testing hypotheses about the mean). Let p denote the p-value of a test:

- If $p \leq 0.01$, very strong evidence against the null hypothesis.
- If $0.01 < p \leq 0.05$, strong evidence against the null hypothesis.
- If $0.05 < p \leq 0.10$, the evidence against H_0 is moderate.
- If $0.10 \leq p < 0.20$ the evidence against H_0 is fairly weak.
- If $p > 0.20$, there is no evidence against H_0.

In the body temperature example, the p-value is $P(T < -5.455 \text{ or } T > 5.455)$. Because the t-distribution is symmetric about zero, we can write the p-value as

$$p = 2P(T > |t|) = 2P(T > 5.455).$$

We can use SAS to compute this probability for us using the "probt" function as follows:

```
data;
p = 2*probt(-5.455, 129);
proc print;
run;
```

The probt function computes cumulative probabilities for the t-distribution. You have to specify the degrees of freedom (in this case $129 = 130 - 1$). Because we are performing a two-sided test, we need to multiply the probability by 2 to get the correct p-value.

Alternatively, SAS's analyst can perform one-sample t-tests for us automatically:

$$\text{Solutions} \rightarrow \text{Analysis} \rightarrow \text{Analyst}.$$

In the window that opens, choose "Open by SAS name" under the file menu. Double-click on the "Work" file which contains all the SAS data sets in operation. Open the SAS data set of interest. At the menu at the top, click

$$\text{Statistics} \rightarrow \text{Hypothesis Tests} \rightarrow \text{One Sample } t\text{-test for a Mean}.$$

In the window that opens, click on the variable of interest. Be sure to specify the hypothesized mean value μ_0 under "NULL". Also specify the correct alternative hypothesis ($\neq, <, >$). The SAS analyst output for the body temperature example is below:

N	Mean	Std. Dev.	Std. Error
130	98.25	0.73	0.06

Hypothesis Test

Null hypothesis: Mean of temp = 98.6
Alternative: Mean of temp ^= 98.6

t Statistic	Df	Prob > t
-5.455	129	<.0001

SAS automatically computes the t-test statistic and the corresponding two-tailed p-value of $p < 0.0001$. Note that the p-value is very small and therefore we have very strong evidence against the null hypothesis.

Using the p-value when stating the results of a hypothesis test is quite popular because it allows the reader to see the exact strength of the evidence for or against the null hypothesis.

Here are some additional notes on hypothesis tests:

1. If we are testing a hypothesis using the critical value method and a significance level of α is being used, then we will reject the null hypothesis if the p-value $< \alpha$.
2. A statistically significant result does not necessarily mean the result is of *scientific* significance. If very large sample sizes are used, statistical tests will be very powerful and able to detect minor

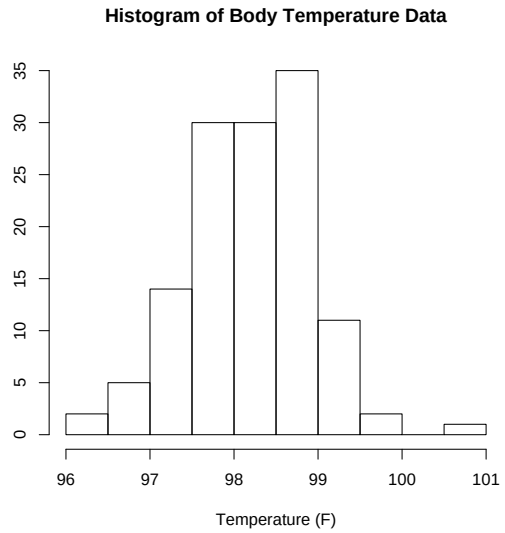

Figure 1: Histogram of the Body Temperature Data.

differences from the null hypothesis. A minor deviation from the null hypothesis may not be of scientific interest. For this reason, some prefer to state results using confidence intervals that give a range of plausible values for the parameter of interest.

3. The t-testing procedure requires that the underlying distribution is normal. The t-testing procedure is fairly robust to departures from normality and will produce approximately valid results for non-normal distributions provided the distribution does not deviate too strongly from normality. For larger sample sizes, the normality assumption can be relaxed more due to the central limit theorem effect. However, if the distribution does deviate strongly from normality, then an alternative testing procedure might be considered, such as a nonparametric test based on ranks or using the bootstrap testing procedure. If the distribution is strongly skewed, then perhaps interest should lie with some other aspect of the distribution other than the mean. Figure 1 shows a histogram of the body temperature data. The body temperature distribution in Figure 1 looks consistent with a normal distribution indicating that the t-test procedure should be reliable.

4. **One-Sided Test.**

   If we want to test $H_0 : \mu = \mu_0$ versus $H_a : \mu < \mu_0$, then the p-value $= P(T < t)$.

   If we want to test $H_0 : \mu = \mu_0$ versus $H_a : \mu > \mu_0$, then the p-value $= P(T > t)$.

4 Power and Sample Size

In the body temperature example, we had a sample size of $n = 130$. How did the investigators decide upon this sample size? If a scientist is to embark on a research project, one of the questions that needs to be answered is: *how many subjects*? One of the common requests of statisticians in biostatistical settings is to determine appropriate sample sizes for a study. Or, if the investigator has a particular sample size in

mind, maybe due to budget or time constraints, it would be useful to know the power of the test. Recall that the power of a test is the probability of rejecting the null hypothesis when it is false. High power is desirable and larger sample sizes lead to higher power.

The goal of a study that employs hypothesis testing is to determine if the null hypothesis can be rejected and the alternative hypothesis can be accepted. It would be unfortunate if time and energy is put into a study and the results of the study do not allow the rejection of the null hypothesis when the null hypothesis is false. This is analogous to putting a guilty person on trial but having to acquit due to lack of evidence. Therefore, it is important to power a study adequately in the planning stages.

To illustrate a power computation, suppose we are testing $H_0 : \mu = \mu_0$ versus the one-tailed alternative $H_a : \mu < \mu_0$. Suppose the true mean is $\mu_1 < \mu_0$ (and therefore the null hypothesis is false). Suppose also for the sake of argument that we know the population standard deviation σ. The critical value of the test will be based on a standard normal distribution instead of the t-distribution if σ is known and the test statistic is $z = (\bar{x} - \mu_0)/(\sigma/\sqrt{n})$. If we are testing at a significance level α, then the power of the test is

$$\begin{aligned}
\text{Power} &= P(\text{Rejecting } H_0 | H_0 \text{ is false}) \\
&= P(\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} < -z_{1-\alpha} | \mu = \mu_1) \\
&= P(\frac{\bar{X}}{\sigma/\sqrt{n}} < \frac{\mu_0}{\sigma/\sqrt{n}} - z_{1-\alpha} | \mu = \mu_1) \\
&= P(\frac{\bar{X} - \mu_1}{\sigma/\sqrt{n}} < \frac{\mu_0 - \mu_1}{\sigma/\sqrt{n}} - z_{1-\alpha} | \mu = \mu_1) \\
&= P(Z < \frac{\mu_0 - \mu_1}{\sigma/\sqrt{n}} - z_{1-\alpha}) \\
&= \Phi(\frac{\mu_0 - \mu_1}{\sigma/\sqrt{n}} - z_{1-\alpha}).
\end{aligned}$$

If β denotes the probability of a type II error (failing to reject H_0 when H_0 is false), then

$$\text{Power} = 1 - \beta.$$

From the preceding computation, we have

$$1 - \beta = \Phi(\frac{\mu_0 - \mu_1}{\sigma/\sqrt{n}} - z_{1-\alpha}),$$

which implies

$$z_{1-\beta} = \frac{\mu_0 - \mu_1}{\sigma/\sqrt{n}} - z_{1-\alpha}.$$

We can solve this equation for n to determine the required sample size for a given power $1 - \beta$:

$$n = \frac{\sigma^2 (z_{1-\beta} + z_{1-\alpha})^2}{(\mu_0 - \mu_1)^2}. \tag{2}$$

This sample size formula works for one-sided alternatives of the form $H_a : \mu > \mu_1$ also. Similar power computations can be carried out for two-sided alternatives. Of course, in practice, σ has to be estimated and the power computations require probability computations using the *non-central t-distribution.* However, the above computations allows us to make some observations about power and sample size:

1. Increasing the sample size n increases the power. Higher power requires larger sample sizes.
2. As the difference $\mu_0 - \mu_1$ grows bigger, the power increases. That is, if the true mean μ_1 lies far from the hypothesized mean μ_0, then it will be more likely to detect this difference.
3. As the significance level α gets smaller, the power decreases. In order to maintain a given power, larger sample sizes are required for smaller significance levels. This follows because as α gets smaller, the standard normal percentile $z_{1-\alpha}$ gets bigger.
4. If σ decreases, then the power increases. Similarly, for a given power, the required sample sizes decreases as σ decreases. If the investigator an can collect data in a way that minimizes the variance of the individual observations, then power will be increased.

Typically when computing an adequate sample size, one needs the following:

- Specify the desired power. Typical values range around 0.80 to 0.90.
- Specify a "guess" of the population standard deviation σ since data will not be available at the planning stages to estimate σ (unless a pilot study is conducted).
- Specify the size of the difference (between μ_0 and the true mean) you would like to be able to detect. For instance, in the body temperature example, we may not be interested in detecting if the true mean differs from 98.6 by a tenth of a degree since that is a small difference. However, if the true average body temperature differs from 98.6 by 0.3 degrees, then we might want to have the power to detect this. The **effect size** is the *standardized* difference in the parameters that the test can detect. Some software programs require inputting the effect size.

In the old days, statistics books came with numerous tables for estimating power and sample size. Nowadays there are software packages that do power and sample size computations. The following website (introduced in the last chapter) is useful for simple power and sample size computations:

`http://www.stat.uiowa.edu/~rlenth/Power/`

In SAS's analyst, choose

Statistic → Sample Size → One Sample t-test,

and then fill the dialogue box. For the output below, I plugged in $\sigma = 0.7$, a significance level $\alpha = 0.05$ and an alternative hypothesis value of $\mu = 98.4$

```
Null hypothesis:   Mean of temp =  98.6
Alternative:       Mean of temp ^= 98.6

       t Statistic      Df      Prob > t
       ---------------------------------
         -5.455        129       <.0001

                        One-Sample t-Test
```

```
      Null Mean = 98.6   Alternate Mean = 98.4
Standard Deviation = .7   Alpha = 0.05   2-Sided Test

                Power     N

                0.700     78
                0.750     87
                0.800     99
                0.850    112
                0.900    131
```

The output shows that with a sample size of $n = 78$, one has a power of 0.70 of rejecting the null hypothesis if the true mean differs from 98.6 by 0.2 degrees. The power jumps to 0.90 when the sample size is $n = 131$.

Confidence Intervals and Two-Sided Tests

We conclude this chapter by noting the following equivalence between a two-sided test with significance level α and a $100 \times (1 - \alpha)$ confidence interval:

Suppose we are testing $H_0 : \mu = \mu_0$ versus $H_a : \mu \neq \mu_0$ at a significance level α. Then we will reject H_0 if and only if the μ_0 *does not* lie in the $100 \times (1 - \alpha)$ confidence interval for μ.

5 Problems

1. An experiment is carried out on $n = 15$ rats where each rat is given a unit dose an experimental drug. Interest lies in measuring the response time to a neurological stimulus. The mean response time for rats not injected with the drug is 1.2 seconds. The experimenter wants to determine if the mean response time for rats injected with the drug differs from 1.2 seconds. Let μ denote the mean response time for rats injected with a unit dose of the experimental drug. Do the following parts:

   a) Set up an appropriate null and alternative hypothesis for this problem in terms of μ.

   b) In the context of this problem, what is a type I error?

   c) In the context of this problem, what is a type II error?

   d) Suppose the mean response time for the $n = 15$ rats was found to be $\bar{x} = 1.05$ seconds with standard deviation $s = 0.5$ seconds. Compute the t-test statistic for this problem.

   e) The hypothesis test is to be conducted using a significance level $\alpha = 0.05$. Draw a picture of the t-density and mark the rejection region for this test.

   f) Based on the results from parts (d) and (e), what is the conclusion of your test? Write a sentence summarizing your decision.

2. Which of the following numbers corresponds to a p-value showing strong evidence against the null hypothesis in the significance test $H_0 : \mu = \mu_0$ versus $H_a : \mu \neq \mu_0$? (circle the correct answer)

   a) 0.757 b) 0.013 c) 15.25 d) -0.001 e) 0.231 f)-22.331

4. Before performing an experiment, a power analysis is conducted to determine the necessary sample size. The hypothesis concerns the mean of a single population. Suppose the required sample size is found to be $n = 50$ using $\alpha = 0.05$ for a power of 80% to detect a difference of $\delta = 1$. What happens to the required sample size if:

   a) The power is raised from 80% to 90%? (Circle one) n goes UP or DOWN.

   b) The minimum detectable difference goes from $\delta = 1$ to $\delta = 1.5$? (Circle one) n goes UP or DOWN.

   c) α goes from 0.05 to 0.01. (Circle one) n goes UP or DOWN.

5. A study is to be done to determine if the cognitive ability of children living near a lead smelter is negatively impacted by increased exposure to lead. Suppose the average IQ for children in the United States is 100. From a pilot study, the standard deviation was estimated to be $s = 14.4$. Use a statistical power/sample size software to answer the following questions.

   a) Should a one-sided or two-sided hypothesis test be used?

   b) If 80% power is desired to detect a difference of 5 IQ points using a significance level $\alpha = 0.05$, what is the required sample size?

   c) Re-do part (b) if 90% power is desired. What happens to the required sample size?

   d) Re-do part (b) is α changes from 0.05 to 0.01. What happens to the required sample size?

   e) Re-do part (b) if one wants 80% power to detect a difference of 10 IQ points. What happens to the required sample size?

   f) Suppose a two-tailed test is desired instead of a one-tailed test. Re-do part (b). What happens to the required sample size?

   g) Suppose you can only budget $n = 10$ IQ tests on children and you want to be able to detect a difference of 5 points in mean IQ using $\alpha = 0.05$. Would it be worthwhile to perform the 20 IQ evaluations? Compute the power and explain.

   h) A study was actually done and the average IQ from $n = 124$ children living near a lead smelter was found to be $\bar{x} = 91.1$ with standard deviation $s = 14.4$. Compute the p-value for this test and state a conclusion based on the p-value.

6. The treatment for patients with prostate cancer depends on whether or not the cancer has spread to surrounding lymph nodes. A surgical procedure (laparectomy) into the abdominal cavity can determine the extent of this nodal involvement. It is hypothesized that prostate cancer patients whose cancer has spread to surrounding lymph nodes will have elevated levels of their serum acid phosphatase. The mean level of serum acid phosphatase in prostate cancer patients where the cancer has not spread to surrounding lymph nodes is 0.645. Twenty prostate cancer patients whose cancer had spread to surrounding lymph nodes were evaluated. The serum acid levels for these $n = 20$ patients is given in the SAS program below. Use this data to perform an appropriate hypothesis test to determine if the mean serum acid level for prostate cancer patients whose cancer has spread to surrounding lymph nodes is greater than 0.645.

   Define μ and write down H_0 and H_1 in terms of μ. Compute the t-test statistic and the p-value for this test. Write a one paragraph report explaining the results of the statistical test. Give an introductory sentence or two, the statistical results, and the interpretation of the results.

```
/*********************************************************
 Data below gives the serum acid phosphatase levels
   in 20 prostate cancer patients whose cancer has spread
   to surrounding lymph nodes.
*********************************************************/
data prostate;
input acid;
datalines;
 0.56
 0.67
 0.99
 1.36
 0.82
 0.48
 0.51
 0.49
 0.84
 0.81
 0.76
 0.70
 0.78
 0.70
 0.67
 0.82
 0.67
 0.72
 0.89
 1.26
;
run;
proc print;
run;
proc means;
run;
data new;
set prostate;
d = acid - 0.645;
run;
proc means;
run;
```

7. Atrazine is a chemical used in herbicides. Over the last decade, the average atrazine concentration per liter of water in Lake Michigan was 10 ng/L. Efforts have been made to reduce the use of atrazine. A year after these efforts began, a study was conducted to determine if the average atrazine concentration had decreased in Lake Michigan. A random sample of $n = 100$ liters of Lake Michigan water were tested for atrazine a year after the efforts began. The sample mean was found to be $\bar{x} = 9.1$ ng/L with standard deviation $s = 2.3$ ng/L. Do the following parts:

a) Let μ denote the average atrazine (ng/L) concentration in Lake Michigan a year after the effort to reduce the use of atrazine. Set up the appropriate null and alternative hypotheses *in terms of* μ.

b) In the context of this problem, what is a type I error?

c) In the context of this problem, what is a type II error?

d) Using a level of significance $\alpha = 0.05$, test the hypothesis stated in part (a). Be sure to compute the test statistic and see if it falls in the critical region. Write a sentence discussing the result of your test.

e) Form a 99% confidence interval for μ.

References.

Mackowiak, P. A., Wasserman, S. S., and Levine, M. M. (1992), "A Critical Appraisal of 98.6 Degrees F, the Upper Limit of the Normal Body Temperature, and Other Legacies of Carl Reinhold August Wunderlich," *Journal of the American Medical Association*, **268**, 1578-1580.