Excerpt from:
ANALYSIS OF MEASUREMENT ERROR

JOHANNES J. VOLWERK , PhD
Science Instructor, ret. Sheldon High School, Eugene, Oregon.

VIII. HYPOTHESES TESTING

The procedures described in the previous sections apply to situations were the objective is to obtain the best possible estimates of the value of a measured quantity and the precision of that measurement. This type of measurement includes many of the typical experiments conducted in teaching laboratories both in the physical and life sciences. However, some measurements performed in a high school or college general biology course require a different statistical approach. For example, a student may wish to investigate the effect of an environmental variable such as temperature, light level, or humidity on the growth rate of a certain plant. A possible experimental approach would be to allow two samples, each containing a small number of plants, to grow under the same conditions, except for the variable to be examined, and to compare the average growth rates of the two samples. The problem then is to evaluate if any observed difference in the sample means is statistically significant, or is simply due to the sampling error. Another typical biology experiment involves testing the "goodness of fit" of observed genetic ratios. For example, a student may wish to examine if crossing pink four-o'clocks produces the phenotype frequency distribution of red, white, and pink flowers expected from Mendel's laws.

Evaluating these types of observations leads us into the realm of what the statistician refers to as hypotheses testing, claims testing, or testing of significance. In essence, the first example involves testing a claim about population means, while the second example involves testing a claim about sample distribution across multiple classes. As we will see, the different nature of these experiments and the type of data that is obtained requires that different statistical procedures be applied in these tests.

Before we can proceed with some specific examples it is necessary to briefly outline the general statistical approach to hypothesis testing. In the language of the statistician **a hypothesis is a statement that something is true.** Note that this definition of the term hypothesis is somewhat different from the way hypothesis is used in science, i.e., an idea or concept that aims to explain an observation or set of observations. The following statements are examples of hypotheses that can be tested by various statistical procedures:

1. The average IQ of biology teachers is greater than that of the general population.
2. The average IQ of SHS students is the same as that of the general population of high school students.
3. Snow tires of brand X have a shorter skid distance than those of brand Y.
4. Douglas fir seedlings grow faster at 20° C than at 10° C.
5. In Oregon, red, white, and black cars of make X sell in the same ratio that is observed nationally.

Suppose you are a backgammon player and you have a special pair of dice that you claim will produce a pair of sixes at least 50% of the time. Any reasonable person wishing to test your claim would begin by examining each die for any gross irregularities such as having sixes on more than one side. If the dice appear normal the next reasonable step would be to roll the dice several times to see what would happen. Let's suppose that double sixes turn up 6 times out of 10 rolls. On the basis of these sample data, most people would conclude that your claim is correct. But is it? Perhaps the dice are normal and 6 double sixes out of 10 rolls is simply a very

unusual manifestation for a pair of dice that usually exhibit normal behavior. However, since the probability of getting 6 double sixes in 10 rolls is so incredibly small – the chance of getting double sixes in any roll is 1/36; the chance of rolling six double sixes in 10 rolls is roughly one in one billion– it is more reasonable to conclude that the dice do favor sixes and have been "doctored" in some way. This is the way statisticians think when they are testing hypotheses:

If an event can easily occur, it is attributed to chance, but if the event appears to be unusual, that significant departure is attributed to the presence of different characteristics.

What do we mean by unusual or unlikely? We can arbitrarily select 5% (or a probability, p, of 0.05) as the level that separates an unusual event from a chance occurrence. That is, we define unusual to mean that the event has a 5% chance (or less) of occurring. Clearly the chance of a normal pair of dice producing 6 double sixes out of 10 rolls is much less than 0.05 and thus an unusual event. We conclude that our data support the hypothesis that the dice are abnormal, or, more precisely, that based on the available information we must reject the hypothesis that the dice are normal. The significance level we choose depends on the severity of the consequences resulting from arriving at an erroneous conclusion, i.e., rejecting a true hypothesis or failing to reject a false hypothesis. In most science experiments involving hypothesis testing a significance level of 5% is reasonable, and commonly employed.

In general, a hypothesis test involves procedures that allow us to make inferences about whole populations by analyzing (small) samples. In this decision-making process we begin by formulating a hypothesis (which may just be a guess) about the population. After gathering sample data we try to determine whether the data support the hypothesis or whether they are statistically significant. The overall process involves the following steps:

1. ***Formulate the null hypothesis and the alternative hypothesis.*** The hypothesis to be tested is called the null hypothesis denoted by H_0. It is a statement of a zero or null difference that is directly tested so that the final conclusion will either be the rejection of H_0 or the failure to reject H_0. Written in symbolic form H_0 should contain the condition of equality, i.e., the symbol, $=$, $\leq$, or $\geq$. The alternative hypothesis, denoted by H_1, is the statement that must be true if the null hypothesis is false. Written in symbolic form H_1 will contain the symbol, $\neq$, $>$, or $<$. *Either H_0 or H_1 may correspond to the original claim, but it is always H_0 that is being tested directly*.
2. ***Select the level of significance appropriate for the hypothesis being tested***. In the examples discussed below the level of significance, denoted by p, is always set at 5% ($p = 0.05$).
3. ***Determine what statistic is relevant to the test and what is its sampling distribution.*** The relevant test statistic is the quantity that is measured or derived from the sampling data. For example, for hypotheses 1 and 2 above the relevant statistic is the sample mean, $\bar{x}$, which is compared to the known population mean, μ, with known standard deviation, σ. For hypotheses 3 and 4 two sample means are compared in which case the population mean(s) may not be known. For hypothesis 5 the proportions of red, white, and black cars of make Y sold over a given period in Eugene are compared to the known national averages.
4. ***Determine the test statistic from the sample data and determine the critical region and the critical value(s)***. The critical region encompasses the set of all values of the test statistic that would cause us to reject the null hypothesis (the shaded area in the graph below). The critical region is bounded by the critical value(s) that separate the critical region from the value(s) of the test statistic that would not

lead to rejection of the null hypothesis. The critical values depend on the assigned level of significance (here always 5%), the nature of the null hypothesis, and the relevant sampling distribution (both discussed below).

5. ***Reject the null hypothesis if the test statistic is in the critical region. Fail to reject the null hypothesis if the test statistic is not in the critical region.***
6. ***Restate the previous decision in simple non-technical terms.***

To illustrate the process of hypothesis testing we will work through an example in which we test hypothesis 1 stated above: *The mean IQ of biology teachers is greater than that of the general population.* Let's assume that we have the following data available:

- The mean IQ of the general population in the US equals 100 ($\mu = 100$) with a standard deviation of 12 ($\sigma = 12$).
- The mean IQ of a random sample of 36 biology teachers equals 104 ($\bar{x} = 104$).

Step 1:
In this example the original claim does not contain the condition of equality. The original claim thus corresponds to the alternative hypothesis, H_1. The null hypothesis, H_0, which is the one directly tested, then states that the mean IQ of all biology teachers is equal to or smaller than that of the general population. In symbolic form:

H_0: $\mu \leq 100$ (claim to be tested)

H_1: $\mu > 100$ (original claim)

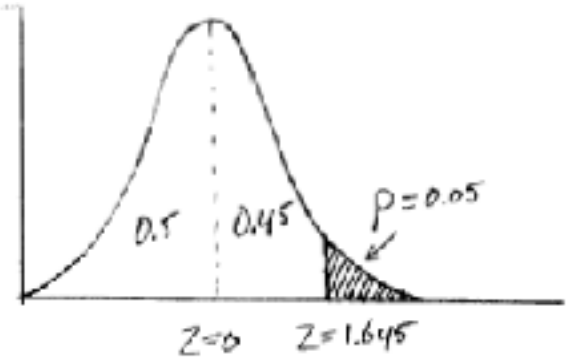
Step 2:
We set the level of significance at 5% (p = 0.05). The test of the claim now centers on the probability of getting a sample mean of 104. If the probability is 5% or less, we conclude that we must reject the null hypothesis and that the data support the original claim. If the probability higher than 5%, we conclude that the data fail to reject the null hypothesis and that the data do not support the original claim.

Step 3:
The relevant test statistic is the sample mean, $\bar{x}$. Because we know that the IQ values for the general population are normally distributed, and because the sample size is reasonably large (n = 36), we may assume that the sample means can be approximated by a normal distribution. To compare the sample mean to the population mean it is useful to calculate the number of standard errors that the sample mean is above or below the population mean. This is called the *standard score* or *z score*, z, where

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

By substituting the values for μ, $\bar{x}$, σ, and n, we calculate that a sample mean of 104 corresponds to z = 2.00.

Step 4:
Next we need to determine the critical value, that is, the cutoff separating unusual results (those with a probability of 5% or less) from chance fluctuations. Critical values are found in a data table (see Table I in the Appendix) that lists the z score values and their corresponding probabilities calculated for the standard normal curve. The procedure is best illustrated using the standard normal distribution graph shown below. The fraction of the total area under the curve bounded by the z score values represents the probability of observing an outcome in the range between these z values. For example, the probability of any outcome equals the total area under the curve or 1, the probability of an outcome with a z score between z = 0 and z = ∞ equals half the area under the curve or 0.5, etc.



We determine that z = 1.645 is the cutoff separating unusual results from chance fluctuations by observing that if the shaded area in the right tail of the curve (the critical region) represents 5% or 0.05 of the total area, then the leftmost limit of that region and z = 0 must encompass 45% or 0.45 of the total. In Table I we find that a probability of 0.45 corresponds to a z value of 1.645 for the leftmost limit of the critical region.

Step 5:
Since the observed sample mean has a z score of 2.00 which is in the critical region well above the critical value z = 1.645, we must reject the null hypothesis.

Step 6:
Stated in non-technical terms we decide that the results from our random sample of 36 biology teachers support the hypothesis that the mean IQ of biology teachers is higher than that of the general population. This does not prove that the original hypothesis is correct, just that, at our chosen level of significance (p = 0.05), the present data support it.

Although the level of significance is commonly set at 5%, it is important to realize that this choice can affect the outcome of the hypothesis testing process. In the example above choosing a 1% level of significance (p = 0.01) leads to a different conclusion. From Table I we see that a probability of 0.49 corresponds to z = 2.33. Now the z score (z = 2.00) of our sample mean is significantly less than the cutoff value separating chance fluctuations from unusual outcomes and we fail to reject the null hypothesis. At the 1% level of significance the present data do **not** support the claim that biology teachers have higher mean IQ.

As was stated above, assigning critical values depends on the nature of the claim to be tested. In our example the purpose was to decide if the observed sample mean was significantly higher than the population mean, that is, would fall in the critical region in the right tail of the standard normal curve. This is an example of a so-

called right-tailed test. A right-tailed test is performed when the alternative hypothesis, H_1, contains the symbol >. Similarly, a left-tailed test is performed when H_1 contains the symbol <. Testing the claim 'boxers have an IQ lower than 100' would be an example of a left-tailed test. A two-tailed test involves claims for which the alternative hypothesis contains the symbol $\neq$. For example, the claim 'prisoners have a mean IQ equal to 100' requires a two-tailed test because the sample mean may either be significantly higher or significantly lower than 100. In a two-tailed test the critical region is split equally between the right and left tail of the curve. At the 5% significance level the critical values are z = -1.96 and z = 1.96. Sample means with z scores lower than –1.96 or higher than 1.96 would cause us to reject the claim that prisoners have a mean IQ equal to 100.

The student t test

When it comes to applying the hypothesis testing procedure to the results of science experiments, determining the relevant sampling distribution becomes a crucial issue. Unfortunately, in many cases we do not have the luxury of all the information that was available in the example above. In many laboratory experiments the population mean or population standard deviation of a measured quantity will not be known nor can we always be certain that the sample distribution is normal. In addition, we usually do not have the advantage of large sample sizes.

Early in the twentieth century William S. Gosset developed specific distributions for small sample sizes while working in an Irish brewery. Because the brewery did not allow publication of research results, he published his work under the name "Student." The sampling distribution applicable to small samples became known as the "student t distribution" and the procedure for testing claims about small samples as the "student t test."

What is a large and what is a small sample? Statisticians generally agree that samples of size $n > 30$ are large enough so that the distribution of their means approaches a normal distribution. In that case we can use the sample standard deviation s as an approximation for the unknown population standard deviation σ since large random samples tend to be representative of the populations from which they originate. If $n < 30$ we use the student t distribution which has the same general shape and symmetry as the normal distribution but reflects the greater variability that is expected for small samples. The width of the student t distribution depends on the sample size with the distribution becoming narrower as the sample size increases. It can be shown mathematically that for $n > 30$ the difference between the normal and the student t distribution is negligibly small. In order to use the student t test, we must be able to reasonably assume that the parent population is essentially normal. Even if the population is not exactly normal we can expect good results using the student t distribution as long as the distribution is basically symmetrical.

For many experiments in the student biology laboratory we do not know the means and standard deviations of the populations we are studying. Yet we would like to draw conclusions or make generalizations about population means using data obtained from small samples. To illustrate the point and to show an example of the application of the student t test to a typical biology inquiry, we test the hypothesis and analyze the data from the following experiment:

A biology student is interested in studying the effect of light on the rate of seed germination and performs the following investigation. Two randomly selected sets of 10 corn seeds are allowed to germinate under identical conditions except that one set of 10 is exposed to direct light while the second set is kept in the dark. For each seed the root length is measured daily over a period of four days. The following data are obtained (there are actual data obtained by a student at Sheldon HS):

Corn seeds grown in the light for 4 days

Seed #	1	2	3	4	5	6	7	8	9	10
Root length (mm)	45	75	63	20	55	50	45	80	45	60

Corn seeds grown in the dark for 4 days

Seed #	1	2	3	4	5	6	7	8	9	10
Root length (mm)	60	65	20	60	32	0	60	20	50	62

From the data we calculate the following:

- Mean root length for seeds grown in the light

$$\bar{x}_\ell = \frac{\sum x_\ell}{n_\ell} = 53.8 \text{ mm}$$

with standard deviation

$$s_\ell = \sqrt{\frac{\sum (x_\ell - \bar{x}_\ell)^2}{n_\ell - 1}} = 17.2 \text{ mm}$$

where $n_\ell = 10$ is the sample size.

- Mean root length for seeds grown in the dark

$$\bar{x}_d = 40.9 \text{ mm}$$

with standard deviation

$$s_d = 22.3 \text{ mm}$$

and sample size $n_d = 10$.

Our task now is to determine whether the observed difference in the mean root length for seeds grown in the dark and in the light is statistically significant given the sample standard deviations.

Step 1:
We formulate the null and alternative hypotheses. Since we look for a statistically significant difference, we test the claim that the population mean root length of seeds grown in the light and in the dark is the same. In symbolic form:

$$H_0 : \mu_\ell = \mu_d$$
$$H_1 : \mu_\ell \neq \mu_d$$

This involves a two-tailed test since μ_ℓ can be either significantly larger or significantly smaller than μ_d (H_1 contains the symbol $\neq$).

Step 2:
As usual we set the level of significance at 5% (p = 0.05).

Step 3:
Because the sample sizes are small (n = 10), we use the student t distribution. Furthermore, both the population means and population standard deviations are unknown. It may be surprising that the population means and standard deviations are unknown: for sure this is not the first time this experiment has been done! The problem is one of controlled variables. As you may recall, the outcome of a science experiment is meaningful only when all factors that affect the results are tightly controlled except the one that is being investigated. In this experiment such factors may include, the type of corn, the age of the seeds, the way they were stored, as well as the specific details of the growing protocol. Since it is impossible to control all of these factors every time the experiment is done by different people and at different locations, we must conclude that useful population parameters, which can only be derived from large data sets collected under uniform circumstances, simply are not available.

Step 4:
The specific mathematical form of the test statistic t that must be calculated when two sample means are compared using the student t test depends on two more factors: (1) whether the samples are dependent or independent; and (2) whether the population standard deviations are (approximately) equal or not. Dependent means that there exists a relationship between the two samples. For example, blood pressure readings for a group of subjects before and after treatment are dependent because each pair of before and after readings belongs to the same subject. However, in our example of the germinating corn seeds the samples clearly are independent. Since the population standard deviations (σ_ℓ and σ_d) for the two samples are not known we don't know if they are equal or not. A statistical test of variance called the F test that uses the sample standard deviations to determine if $\sigma_\ell \cong \sigma_d$ can be applied, but its use in the t test for comparing two sample means is controversial. Here we will assume that the standard deviations are comparable. This assumption seems reasonable as the two samples consist of corn seeds randomly selected from a single batch that are grown under identical conditions except for the presence or absence of light. The required test statistic for comparison of two independent samples with $\sigma_\ell \cong \sigma_d$ is

$$t = \frac{(\bar{x}_\ell - \bar{x}_d) - (\mu_\ell - \mu_d)}{\sqrt{\frac{1}{n_\ell} + \frac{1}{n_d}}\sqrt{\frac{(n_\ell - 1)s_\ell^2 + (n_d - 1)s_d^2}{n_\ell + n_d - 2}}}$$

where $\mu_\ell - \mu_d = 0$ since we test the claim $\mu_\ell = \mu_d$. Substituting the values for $\bar{x}_\ell$, s_ℓ, $\bar{x}_d$, s_d, n_ℓ, and n_d we find

$$t = 1.452$$

We determine the critical values and critical regions corresponding to a 5% level of significance for the t distribution from tabulated data (see Table II in the appendix) in a manner similar to what was done in the previous example for the normal distribution. However, because the student t distribution applies to small samples the sample size itself is a factor that needs to be considered. As the sample size increases we obtain better estimates of the sample mean and standard deviation. This means that in the graph of the distribution the critical values shift in from the wings and the critical regions increase in size. Sample size is taken into account by considering what is called the degrees of freedom. Degrees of freedom (df) correspond to the number of values that may vary after certain restrictions are imposed on all values. In tests on a mean the number of degrees of freedom is simply the sample size minus 1 (df = n-1). In our case we have two samples with sizes n_ℓ and n_d so that

$$Df = n_\ell + n_d - 2 = 10 + 10 - 2 = 18$$

In Table II we find for a two-tailed test with p = 0.05 and df = 18 the critical values 2.101 and –2.101.

Step 5:
Since our calculated test statistic (t = 1.452) is neither larger than 2.101 nor smaller than –2.101 we fail to reject the null hypothesis.

Step 6:
Stated in non-technical terms we conclude that the present data do not support the notion that the presence or absence of light significantly affects the germination rate of corn seeds. The observed difference between the sample means most likely must be attributed to sampling error.

TABLE II[1)]

The t distribution

	p	
Degrees of Freedom	0.05 (two tails)	0.05 (one tail)
1	12.706	6.314
2	4.303	2.920
3	3.182	2.353
4	2.776	2.132
5	2.571	2.015
6	2.447	1.943
7	2.365	1.895
8	2.306	1.860
9	2.262	1.833
10	2.228	1.812
11	2.201	1.796
12	2.197	1.782
13	2.160	1.771
14	2.145	1.761
15	2.132	1.753
16	2.120	1.746
17	2.110	1.740
18	2.101	1.734
19	2.093	1.729
20	2.086	1.725
21	2.080	1.721
22	2.074	1.717
23	2.069	1.714
24	2.064	1.711
25	2.060	1.708
26	2.056	1.706
27	2.052	1.703
28	2.048	1.701
29	2.045	1.699
Large	1.960	1.645

1) The listed values represent the critical t values at the 5% level of significance ($p = 0.05$) as a function of the number of degrees of freedom for a one- and two-tailed test. For example, when $df = 10$ the critical regions for a two-tailed test include all values of the test statistic $t > 2.228$ or $t < -2.228$. For a one-tailed test the critical regions are $t > 1.812$ (right-tailed) and $t < -1.812$ (left-tailed). A more complete list of t values can be found at www.statsoft.com/textbook/sttable.html#z.