Confidence Intervals and Hypothesis Testing: A Primer[1]

By Allan T. Mense, Ph.D., PE, CRE
Principal Engineering Fellow
Raytheon Missile Systems

Introduction

There appears to be a need to discuss confidence intervals and how to compute them under a variety of situations. These techniques are crucial to understanding how to use statistics to make decisions.

Confidence intervals[2] are used to bound the value of some population parameter of interest. The population parameters are designated (by convention) by Greek letters e.g. μ, σ^2, β, etc. Since we do not know these parameters and will probability never know them exactly, we choose to estimate them using statistics (data). The estimators for these population parameters are designates by Latin letters e.g. $\bar{x}, s^2, \hat{b}$, etc. The estimators are themselves random variables as they take on values dependent on the sampled data.

We would like these estimators to have certain useful properties such as minimum variance, i.e. each estimator has some variance and covariance, also it is useful if the estimator is unbiased in that it correctly estimates the parameter we are seeking and not some value above or below the parameter value. These requests seem natural but are not always necessary.

It seems relatively clear that once a quantity of interest, e.g. a sample mean $\bar{x}$ or sample variance, s^2, is computed from data, that one only has a point estimate of the population parameter e.g. μ or σ^2. It would be helpful to understand the accuracy of this estimate. To place an interval around the population value (lower value < population value< upper value) of the quantity of interest one might guess that one would make use of the point estimate already computed for the estimator and then calculate, using some soon to be disclosed formulas, the upper and lower values of the confidence interval. As an example we have the expression ($L < \mu < U$) but for any given set of data we have to estimate L and U, call the estimates l and u, and so the best we can say is that the probability of the population parameter being between the computed values is $Pr\{l < \mu < u\} = 1-\alpha$ where $1-\alpha$ is called the confidence level and α is called the significance level.

To perform this confidence interval calculation we need to go through the following steps.

1) What population parameter are we interested in evaluating, e.g. μ?

2) Find the estimator (preferably an unbiased estimator) for the population parameter, i.e. $\bar{x}$?

---

[1] Some of this material follows the text "Applied Statistics and Probability for Engineers, 3rd Edition," by Montgomery and Runger.

[2] There are several types of intervals; confidence intervals for parameters, tolerance intervals for bounding intervals containing some percentage of the population, and prediction intervals to bound future values of variables (see footnote 1)

3) Do we know the probability distribution for the estimator? If yes then we can proceed if no then call 911-statistician because we will have to either find the correct distribution for that estimator or create the distribution using numerical (Monte Carlo) techniques.

4) Once the distribution of the estimator is known, e.g. $f_n(\bar{x})$ or $F_n(\bar{x})$ which are respectively the pdf and CDF of the sample mean of sample size n, then we can calculate any probability of obtaining a given value $\bar{x}$. For example, one can find the value, $\bar{x}_{upper}$, that marks the value of $\bar{x}$ for which say 95% of the possible sample means of sample size n would be at or below.

One can also calculate the value of $\bar{x}$ marking the mean value for which say only ≤ 5% of mean values would occur, call this $\bar{x}_{lower}$, and from these two values of $\bar{x}$ one creates an interval, called a confidence interval, that contains may contain the actual population mean, μ, in 90% of the samples of size n you may take including the interval using the data you just took to calculate $\bar{x}$.

Here is an example. Suppose that $X_1, X_2, \ldots, X_n$ is a random sample from a normal distribution with unknown mean μ and *known* variance σ^2. We know from The Central Limit Theorem that the sample mean, $\bar{x}$ is distributed as a normal distribution whose mean is μ and whose variance is σ^2/n. Written as $N(\mu,\sigma^2/n)$. We may "standardize" the sample mean by subtracting the population mean μ and dividing the difference by σ^2/n. By the way this expression σ^2/n is often called the standard error of the mean.

The next few paragraphs can get a little complicated so hang in there and read them more than once! As stated previously a confidence interval estimate for μ is an interval of the form $l \le \mu \le u$, where the endpoints l and u are computed from the sample data. Because different samples will produce different values of l and u, these end-points are instantiations of the random variables L and U, respectively. One is looking for values of L and U such that the following probability statement is true: $Pr\{ L \le \mu \le U\}=1-\alpha$ where $0 \le \alpha \le 1$. The value of $1-\alpha$ is typically 0.90 or 0.95.

There is a probability of $1-\alpha$ of selecting a sample for which the confidence interval (CI) will contain the true value of μ. However, once we have selected the sample, so that $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$, and we have computed l and u, the resulting **confidence interval** either includes μ or it does not. More will be made of this statement in the next section. Thus, another sample of n data points would result in another value for the sample mean and another set of calculated values for l and u.
To set up the interval (l,u) it is easier to deal with the transformed variable z instead of dealing with the normal distribution for $\bar{x}$ so we will instead deal with the standardized or unit normal distribution for $z= (\bar{x}-\mu)/(\sigma/\sqrt{n})$ written as N(0,1)[3].

Using this z transformation one bounds Z by values that make the unit normal distribution equal to 95% and 5% respectively which will produce a 90% confidence interval. This is done in the following more complicated manner.

[3] When the variance is unknown we need to estimate it from the data we then use what is called a student t-distribution instead of the unit normal and the transformed variable of interest is $t=(\bar{x}-\mu)/(s/\sqrt{n})$.

Note that we are interested in the values of Z that fall between two symmetrically placed values $Z_{\alpha/2}$ and $Z_{1-\alpha/2}$ such that the probability between these two Z values gives $(1-\alpha/2)-\alpha/2 = 1-\alpha$. This is written many times as a probability statement

$$\Pr\{ Z_{\alpha/2} \le (\bar{x} - \mu)/(\sigma/\sqrt{n}) \le Z_{1-\alpha/2}\} = 1-\alpha.$$

Rearranging this expression one obtains $\Pr\{ \bar{x} - (\sigma/\sqrt{n}) Z_{\alpha/2} \ge \mu \ge \bar{x} - (\sigma/\sqrt{n}) Z_{1-\alpha/2}\} = 1-\alpha$, however $Z_{\alpha/2} = - Z_{1-\alpha/2}$ and therefore one has the final expression,

$$\Pr\{ \bar{x} + Z_{1-\alpha/2}\, \sigma/\sqrt{n} \ge \mu \ge \bar{x} - Z_{1-\alpha/2}\, \sigma/\sqrt{n}\} = 1-\alpha,$$

Not all confidence intervals can be written this way as will be seen later in fact the confidence interval for the mean one of the few CIs that can be written in this form. Z represents some multiple of the standard deviation of the mean and is chosen so that a fraction 1-a of the possible mean values fall within the bounds shown above, ($\bar{x} \pm Z_{1-\alpha/2}\, \sigma/\sqrt{n}$) .

As noted in footnote 2, for the case in which the variance of the data is not known one uses the t-distribution (which already has a mean = 0, variance =1/(1-(2/dof))) instead of the unit normal distribution N(0,1) and the interval is given by ($\bar{x} \pm s/\sqrt{n}\ t_{\alpha/2,\nu-1}$), The variable s is of course the sample standard deviation.

This situation now has an additional dependency on the sample size n through the t-distribution? The t-distribution looks a lot like the unit normal distribution but has a lower peak value at t=0 and "fatter" tails which means that for the same level of confidence that $t_{\alpha/2,\nu-1} > Z_{1-\alpha/2}$ for all values of n. Therefore the confidence interval is larger for the population mean when the variance is unknown and must be calculated from the sample data.

In any case to compute any confidence interval for a population parameter we need to know the distribution function of the estimator of that parameter which for the sample mean is the normal distribution or the t-distribution and for the sample variance it is the chi square distribution, usually designated as χ^2. Here is a numerical example

---

ASTM Standard E23 defines standard test methods for notched bar impact testing of metallic materials. The Charpy V-notch (CVN) technique measures impact energy and is often used to determine whether or not a material experiences a ductile-to-brittle transition with decreasing temperature. Ten measurements of impact energy (*J*) on specimens of A238 steel cut at 60ºC are as follows: 64.1, 64.7, 64.5, 64.6, 64.5, 64.3, 64.6, 64.8, 64.2, and 64.3. Assume that impact energy is normally distributed with $\sigma = 1J$. We want to find a 95% CI for μ, the mean impact energy. The required quantities are $z_{1-\alpha/2} = z_{0.975} = 1.96$, $n = 10$, $\sigma = 1$, and $\bar{x} = 64.46$. The resulting 95% CI is found as follows:

$$\bar{x} + Z_{1-\alpha/2}\, \sigma/\sqrt{n} \ge \mu \ge \bar{x} - Z_{1-\alpha/2}\, \sigma/\sqrt{n}$$

$$64.46 - 1.96 * 1 / (10)^{1/2} \le \mu \le 64.46 + 1.96 * 1 / (10)^{1/2}$$

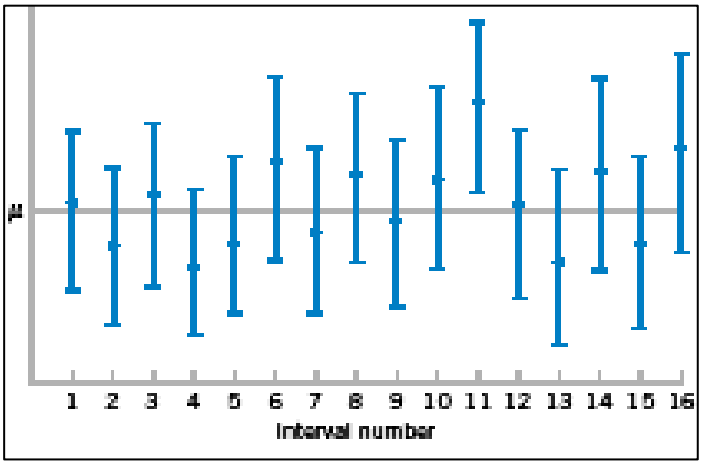
$63.84 \leq \mu \leq 65.08$

with a confidence level of 95%. That is, based on the sample data, a range of highly plausible values for mean impact energy for A238 steel at 60°C is $63.84J \leq \mu \leq 65.08J$.

Interpreting a Confidence Interval

How does one interpret a confidence interval? In the impact energy estimation shown above the 95% CI is 63.84 ≤μ≤ 65.08, so it is tempting to conclude that μ is within this interval with probability 0.95. However, with a little reflection, it's easy to see that this cannot be correct; the true value of μ is unknown and the statement 63.84 ≤μ≤ 65.08 is either correct (true with probability 1) or incorrect (false with probability 1). The correct interpretation lies in the realization that a confidence Interval (CI) is a random interval because in the probability statement defining the end-points of the interval, L and U are random variables calculated from data. Consequently, the correct interpretation of a $100(1-\alpha)\%$ CI depends on the relative frequency view of probability. Specifically, if an infinite number of random samples are collected and a $100(1-\alpha)\%$ confidence interval for μ is computed from each sample, using the prescription shown above, $100(1-\alpha)\%$ of these intervals will contain the true value of μ.
The situation is illustrated in figure below, which shows several $100(1-\alpha)\%$ confidence intervals for the mean μ of a normal distribution. The dots at the center of the intervals indicate the point estimate of μ(that is, $\bar{x}$). Notice that one of the intervals fails to contain the true value of μ. If this were a 95% confidence interval, in the long run only 5% of the intervals would fail to contain μ.

Now in practice, we obtain only one random sample and calculate one confidence interval. Since this interval either will or will not contain the true value of μ, it is not reasonable to attach a probability level to this specific event. The appropriate statement is the observed interval [l, u] brackets the true value of μ with confidence $100(1-\alpha)\%$. This statement has a frequency interpretation; that is, we don't know if the statement is true for this specific sample, but the method used to obtain the interval [l, u] yields correct statements $100(1-\alpha)\%$ of the time.



Confidence Level and Precision of Estimation

Notice in the above example that our choice of the 95% level of confidence was essentially arbitrary. What would have happened if we had chosen a higher level of confidence, say, 99%? In fact, doesn't it seem reasonable that we would want the higher level of confidence? At α=0.01, we find $z_{1-\alpha/2} = z_{0.995}$ = 2.58 while for α=0.05, $z_{1-\alpha/2} = z_{0.975}$ = 1.96. Thus, the **length** of the 95% confidence interval is 2(1.96 $\sigma/\sqrt{n}$) = 3.92 $\sigma/\sqrt{n}$, whereas the length of the 99% CI is 2(2.58 $\sigma/\sqrt{n}$) = 5.16 $\sigma/\sqrt{n}$. Thus, the 99%

CI is longer than the 95% CI. This is why we have a higher level of confidence in the 99% confidence interval. Generally, for a fixed sample size *n* and standard deviation σ, the higher the confidence level, the longer the resulting CI.
The length of a confidence interval is a measure of the **precision** of estimation. From the preceding discussion, we see that precision is inversely related to the confidence level. It is desirable to obtain a confidence interval that is short enough for decision-making purposes and that also has adequate confidence. One way to achieve this is by choosing the sample size *n* to be large enough to give a CI of specified length or precision with prescribed confidence. Remember the population from which we are drawing our sample is not changing its properties e.g. μ & σ it is only the distribution of the sample mean values whose properties change with sample size e.g. $\sigma/\sqrt{n}$.

Choice of Sample Size

The precision of the confidence interval in the above example is $2z_{1-\alpha/2}\,\sigma/\sqrt{n}$ This means that in using $\bar{x}$ to estimate μ, the error is less than or equal to E=| $\bar{x}-\mu$ |with confidence 100$(1-\alpha)$%. In situations where the sample size can be controlled, we can choose *n* so that we are 100$(1-\alpha)$ percent confident that the error in estimating μ is less than a specified bound on the error *E*. The appropriate sample size is found by choosing *n* such that $z_{1-\alpha/2}\,\sigma/\sqrt{n}$ =E. Solving this equation for n gives the following formula.

$$n=\left(z_{1-\alpha/2}\,\sigma/E\right)^2$$

If the right-hand side of this equation is not an integer, it must be rounded up. Choosing n in this manner will insure that the level of confidence does not fall below 100$(1-\alpha)$%. Notice that 2*E* is the length of the resulting confidence interval and YOU must set this value.
Let's work an example.

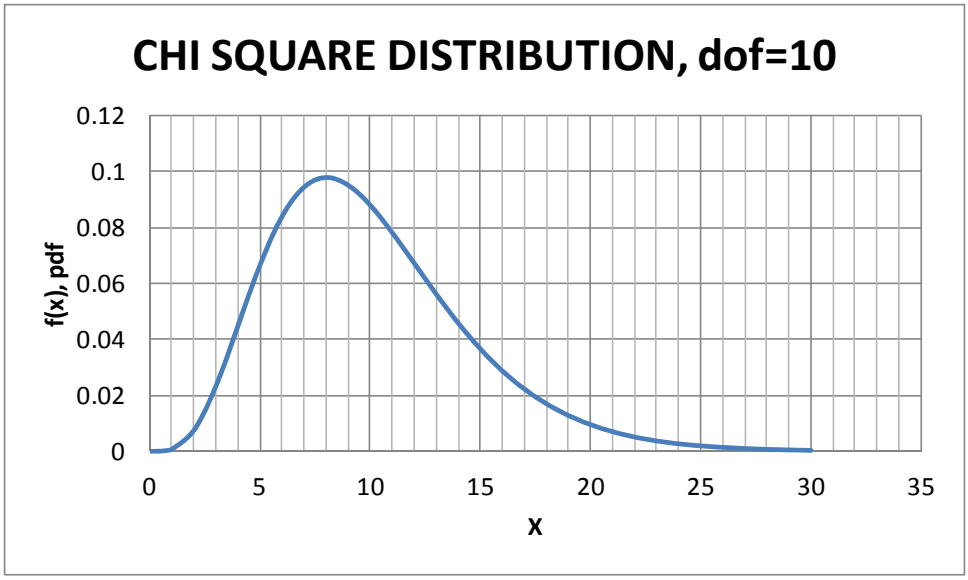
To illustrate the use of this procedure, consider the CVN test described in the previous example, and suppose that we wanted to determine how many specimens must be tested to ensure that the 95% CI on m for A238 steel cut at 60°C has a length of at most 1.0 J i.e. 2E=1 J. Since the bound on error in estimation E is one-half of the length of the CI, to determine n we use the above equation with E = 0.5, σ= 1, and $z_{1-\alpha/2}$=1.96. The required sample size is 16 because n must be an integer, n = (1.96*1/0.5)2 =15.37 so the required sample size is n =16. Notice the general relationship between sample size, desired length of the confidence interval 2E, confidence level 100$(1-\alpha)$, and standard deviation σ: As the desired length of the interval 2E decreases, the required sample size n increases for a fixed value of σ and specified confidence.

How about bounds for the variance?

Roughly the same procedure applies for the variance as for the mean but there are a couple of algebraic differences. First the problem is to determine the confidence interval for the population variance σ^2 using the sample variance s^2. It turns out that the sampling statistic for comparing is $(n-1)s^2/\sigma^2$. This ratio turns out to be distributed as a chi square variable that is if x is defined to equal $(n-1)s^2/\sigma^2$ then x obeys the distribution shown below.

$$f(x)=\begin{cases}\dfrac{x^{\frac{\nu}{2}-1}e^{-\frac{x}{2}}}{2^{\frac{\nu}{2}}\Gamma(\nu/2)}, & x>0\\ 0, & x\le 0\end{cases}$$

This distribution is not hard to derive but requires some math that has not been introduced in this note. If the data you are analyzing is normally distributed and you standardize the data so you are using the Z variable then the χ^2 distribution tells you how the square of that data is distributed and the square of Z can be related to the sample variance by $(n-1)s^2/\sigma^2 = \sum \frac{(x_i-\bar{x})^2}{\sigma^2} = \sum z_i^2$. The χ^2 distribution has the parameter ν = degrees of freedom and $\Gamma(\nu/2)$ is the complete gamma function which can be looked up in tables of found using excel[4]. This is the general form of the chi square distribution and the mean of the distribution occurs at $x=\nu$ and the mode at $\nu-2$.



Now let's use the chi square distribution to find the confidence interval for the variance. Again I write this statement in the form of a probability statement i.e.

$\Pr\{\chi^2(\alpha/2,n-1) < (n-1)s^2/\sigma^2 < \chi^2(1-\alpha/2,n-1)\}=1-\alpha.$

Note that the lower and upper values of χ^2 are different in this distribution. The upper bound is evaluated at $1-\alpha/2$ and the lower bound at $\alpha/2$. There is no symmetry so we cannot write limits as some mean +/- some increment as we did for the mean value. If I rearrange the probability statement just a bit we obtain **$\Pr\{(n-1)s^2/\chi^2(1-\alpha/2,n-1) < \sigma^2 < (n-1)s^2/\chi^2(\alpha/2,n-1)\}=1-\alpha$,** which is in a form of a confidence interval for the population variance, σ^2.

Example: Using $s^2 = 25$, $n=11$, $\alpha=0.05$ one has $\chi^2(.025,10)=3.247$, $\chi^2(.975,10)=20.483$. ***Note: In excel the values of alpha start at zero at the far right side where x = +infinity and then goes to 1 at x=0 so evaluating chi square at alpha/2 using excel requires you insert 1-alpha/2 instead e.g. CHIINV(0.025,10)=20.483. You also need to pay attention in excel about calculating 2-tailed or 1-tailed values. It's a bit of a mess so pay attention.*** The normal distribution is done in just the opposite manner so just be careful. This produces a 95% confidence interval with bounds given below.

[4] In excel one uses the following function to find $\Gamma(y)$=EXP(GAMMALN(y))

(10)(25)/20.483 < σ^2 < (10)(25)/3.247 or <u>12.2 < σ^2 < 77</u>.

This interval is very large because the number of data values (n=11) is very small.

Confidence Interval for Proportions:

In binary experiments such as Pass/Fail tests for parts or printed circuit boards etc. one performs n tests and then records the number of failures x in these n tests. Typically one estimates the fraction or proportion of failures in the population with the point estimator <p>=x/n. From this information and knowing that the distribution of this point estimator is the binomial distribution one can calculate confidence bounds by appropriately summing the binomial distribution. Remember we are trying to find two values P_L and P_U using the data that bound the population proportion, p, with some level of confidence. Since the sampling distribution for the estimator x/n is the binomial distribution we can find an interval by manual iteration of the following sums.

$$\sum_{k=x}^{n}\binom{n}{k}p_L^k(1-p_L)^{n-k}=\alpha/2, x=1,2,\ldots,n-1$$

$$\sum_{k=0}^{x}\binom{n}{k}p_U^k(1-p_U)^{n-k}=\alpha/2, x=1,2,\ldots,n-1$$

For x=n or all failures we only have 1-sided bound $p_L=(\alpha)^{1/n}$ since $p_U=1$ while for x=0 or no failures the 1-sided bound is $p_U=1-(\alpha)^{1/n}$ since $p_L=0$. (See Agresti & Coull (1998) for more discussion)

The explanation that goes along with the above exact equations is seen to be difficult to understand so most text books use what is called the normal approximation to the binomial distribution by setting $\mu=x/n$ and $\sigma^2/n = (x/n)(1-(x/n))/n$ in the normal distribution function. The confidence interval (called the normal approximation) obtained is given by

$$\Pr\{(x/n)-z_{1-\alpha/2}[(x/n)(1-(x/n))/n]^{1/2} < p < (x/n)+z_{1-\alpha/2}[(x/n)(1-(x/n))/n]^{1/2}\}=1-\alpha$$

However the <u>accuracy of this approximate interval is only accurate near the value <p>=x/n ~1/2</u> (e.g. use for election results) and gives physically incorrect results near x/n ~ 0 and x/n ~ 1.

There have been many other approximations proposed for the CI for a proportion and I list them here in order of increased accuracy. Not that you will ever use these but here they are for future reference.

Approximation #1: (Does not produce physically incorrect results near x=0 and x=n)

$$\Pr\left\{\frac{\left(\frac{x}{n}\right)+\left(\frac{Z_{1-\alpha/2}^2}{2n}\right)-\sqrt{\frac{Z_{1-\alpha/2}^2}{n}\left(\frac{x}{n}\right)\left(1-\frac{x}{n}\right)+\left(\frac{Z_{1-\alpha/2}^2}{2n}\right)^2}}{1+\frac{Z_{1-\alpha/2}^2}{n}} < p < \frac{\left(\frac{x}{n}\right)+\left(\frac{Z_{1-\alpha/2}^2}{2n}\right)+\sqrt{\frac{Z_{1-\alpha/2}^2}{n}\left(\frac{x}{n}\right)\left(1-\frac{x}{n}\right)+\left(\frac{Z_{1-\alpha/2}^2}{2n}\right)^2}}{1+\frac{Z_{1-\alpha/2}^2}{n}}\right\}=1-\alpha$$

Approximation #2: (Poisson approximation good for p small only)

$$\Pr\left\{\chi^2_{1-\alpha/2,2x}/2n < p < \chi^2_{1-\alpha/2,2(x+1)}/2n\right\} = 1-\alpha$$

Approximation #3: (arcsine transformation, fairly accurate for all p biased near p~0 and p~1)

$$\Pr\left\{\left(\sin\left(\arcsin\left(\sqrt{\frac{x}{n}}\right) - \frac{Z_{1-\alpha/2}}{2\sqrt{n}}\right)\right)^2 < p < \left(\sin\left(\arcsin\left(\sqrt{\frac{x}{n}}\right) + \frac{Z_{1-\alpha/2}}{2\sqrt{n}}\right)\right)^2\right\} = 1-\alpha$$

Exact non parametric interval: Most accurate, gives largest interval and used in SSE.xls (F is the inverse of the Fisher F-distribution,β is the inverse of the beta distribution)

$$\Pr\left\{\left(1+\frac{n-x+1}{xF^{Inv}_{1-\alpha/2,2x,2(n-x+1)}}\right)^{-1} < p < \left(1+\frac{n-x}{(x+1)F^{Inv}_{\alpha/2,2(x+1),2(n-x)}}\right)^{-1}\right\} = 1-\alpha$$

or

$$\Pr\left\{1-\beta^{Inv}_{1-\alpha/2,n-x+1,x} < p < 1-\beta^{Inv}_{\alpha/2,n-x,x+1}\right\} = 1-\alpha$$

For a more complete discussion see Leemis (2006).

Hypothesis Testing [5]

Now that we know how to calculate a confidence interval for a population parameter we can use that information to perform what is called hypothesis testing. I personally use confidence intervals whenever possible but a lot of decision analysis uses hypothesis testing and so it is important that one learns the basics.

In the first part of this paper I illustrated how to construct a confidence interval estimate of a parameter from sample data. However, many problems in engineering require that we decide whether to accept or reject a statement about some parameter. The statement is called a hypothesis, and the decision-making procedure about the hypothesis is called hypothesis testing. This is a useful aspect of statistical inference, since many types of decision-making problems, tests, or experiments in the engineering world can be formulated as hypothesis-testing problems.

Statistical hypothesis testing and confidence interval estimation of parameters are the fundamental methods used at the data analysis stage of a comparative experiment, in which the engineer is interested, for example, in comparing the mean of a population to a specified value. These simple comparative experiments are frequently encountered in practice and provide a good foundation for the more complex experimental design problems that are discussed in class. In this note I discuss comparative experiments involving a single population, and, as always, the focus is on testing hypotheses concerning the parameters of the population.

A **statistical hypothesis** is a statement about the parameters of one or more populations.

Since we use probability distributions to represent populations, a statistical hypothesis may also be thought of as a statement about the probability distribution of a random variable. The hypothesis will usually involve one or more parameters of this distribution. For example, suppose that we are interested in the burning rate of a solid propellant used to power aircrew escape systems. Now burning rate is a random variable that can be described by a probability distribution. Taking an example from Montgomery [1], suppose that our interest focuses on the mean burning rate (a parameter of this distribution). Specifically, we are interested in deciding whether or not the mean burning rate is 50 centimeters per second. We may express this formally as a two-sided hypothesis test.

H_0: $\mu = 50$ cm/s.
H_1: $\mu \neq 50$ cm/s.

The statement H_0: $\mu = 50$ centimeters per second in the above equation is called the **null hypothesis**, and the statement H_1: $\mu \neq 50$ centimeters per second is called the **alternative hypothesis**.
Since the alternative hypothesis specifies values of μ that could be either greater or less than 50 centimeters per second, it is called a two-sided alternative hypothesis.
We might use instead a one sided hypothesis test such as;

H_0: $\mu = 50$ cm/s.
H_1: $\mu > 50$ cm/s.

or possibly,

[5] This section mirrors the material in Montgomery & Runger, "Applied Statistics and Probability for Engineers, 3rd Edition," Wiley Publishing.

H_0: μ = 50 cm/s.
H_1: μ < 50 cm/s.
The test depends on what you want to establish.

It is important to remember that hypotheses are always statements about the population parameters under study, not statements about the sample.

The value of the population parameter specified in the null hypothesis (50 centimeters per second in the above example) is usually determined in one of three ways. First, it may result from past experience or knowledge of the process, or even from previous tests or experiments. The objective of hypothesis testing then is usually to determine whether the parameter value has changed. Second, this value may be determined from some theory or model regarding the process under study. Here the objective of hypothesis testing is to verify the theory or model. A third situation arises when the value of the population parameter results from external considerations, such as design or engineering specifications, or from contractual obligations. In this situation, the usual objective of hypothesis testing is conformance testing.

A procedure leading to a decision about a particular hypothesis is called a test of a hypothesis. Hypothesis-testing procedures rely on using the information from a random sample taken from the population of interest. If this information is consistent with the hypothesis, we will conclude that the hypothesis is true; however, if this information is inconsistent with the hypothesis, we will conclude that the hypothesis is false. We emphasize that the truth or falsity of a particular hypothesis can never be known with certainty, unless we can examine the entire population. This is usually impossible in most practical situations. Therefore, a hypothesis-testing procedure should be developed with the probability of reaching a wrong conclusion in mind.

The structure of hypothesis-testing problems is identical in all the applications that we will consider.

1. The null hypothesis is the status quo or the truth as we know it. It is the hypothesis we wish to test.
2. Rejection of the null hypothesis always leads to accepting the alternative hypothesis. This is the hypothesis we wish to prove or demonstrate and depends obviously on the sample data.
3. The null hypothesis will always be stated so that it specifies an exact value of the population parameter of interest e.g. H_0: μ = 50 cm/s. The alternate hypothesis will allow the parameter to take on several possible values e.g. H_1: μ > 50cm/s.

To illustrate the general concepts, consider the propellant burning rate problem introduced earlier. The null hypothesis is that the mean burning rate is 50 centimeters per second, and the alternate is that it is not equal to 50 centimeters per second. That is, we wish to test

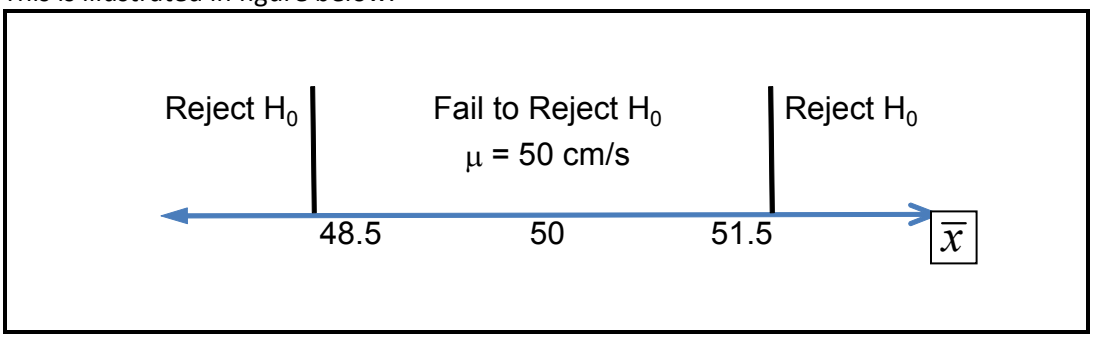
H_0: μ = 50 cm/s.
H_1: $\mu \neq$ 50 cm/s.

Testing the hypothesis involves taking a random sample, computing a test statistic from the sample data, and then using the test statistic to make a decision about the null hypothesis. The null hypothesis is that the mean burning rate is 50 centimeters per second, and the alternate is that it is not equal to 50 centimeters per second. Suppose that a sample of specimens is tested and that the sample mean burning rate is observed. The sample mean is an estimate of the true population mean. A value of the sample mean that falls close to the hypothesized value of centimeters per second is evidence that the true mean is really 50 centimeters per second; that is, such evidence supports the null hypothesis H_0. On

the other hand, a sample mean that is considerably different from 50 centimeters per second is evidence in support of the alternative hypothesis. Thus, the sample mean is the test statistic in this case. The sample mean can take on many different values.

Suppose that if we assume that $48.5 < \bar{x} < 51.5$ would be an acceptable range of values for the sample burn rate so measure mean burn rates within that range will imply that we cannot reject the null hypothesis. However, we will reject the null hypothesis H_0: $\mu = 50$ cm/s , in favor of H_1: $\mu \neq 50$ cm/s. if either $\bar{x} < 48.5$ or $\bar{x} > 51.5$.

This is illustrated in figure below.



The values of that are less than 48.5 and greater than 51.5 constitute the critical region for the test, while all values that are in the interval $48.5 < \bar{x} < 51.5$ form a region for which we will fail to reject the null hypothesis. By convention, this is usually called the acceptance region.

The boundaries between the critical regions and the acceptance region are called the critical values. In our example the critical values are 48.5 and 51.5. It is customary to state conclusions relative to the null hypothesis H_0. Therefore, we reject H_0 in favor of H_1 if the test statistic falls in the critical region and fail to reject H_0 otherwise.

Errors and Drawing Wrong Conclusions.

This decision procedure can lead to either of two *wrong* conclusions. For example, the true mean burning rate of the propellant could be equal to 50 centimeters per second. However, for the randomly selected propellant specimens that are tested, we could observe a value of the test statistic (mean burn rate) that falls into the critical region. We would then reject the null hypothesis H_0 in favor of the alternate when, in fact, H_0 is really true. This type of wrong conclusion is called a **type I error**.

Definition: Rejecting the null hypothesis H_0 when it is true is defined as a type I error. Measured by α.

Now suppose that the true mean burning rate is different from 50 centimeters per second, yet the sample mean falls in the acceptance region. In this case we would fail to reject H_0 when it is false. This type of wrong conclusion is called a **type II error**.

Definition: Failing to reject the null hypothesis when it is false is defined as a type II error.

	Truth	
Decision	H0 is True	H0 is False
Fail to Reject H0	no error	Type II error
Reject H0	Type I error	no error

Thus, in testing any statistical hypothesis, four different situations determine whether the final

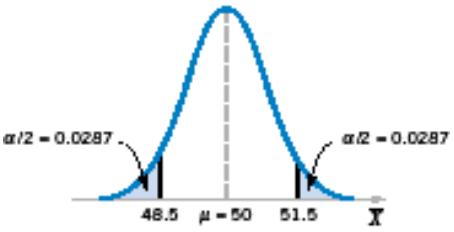
decision is correct or in error. These situations are presented in table at the left.

Because our decision is based on random variables, probabilities can be associated with the type I and type II errors in above table. The probability of making a type I error is denoted by the Greek letter α.

Pr(type I error) = Pr(reject H_0 when H_0 is true) =α

Sometimes the type I error probability is called the *significance level*, or the α*-error*, or the size of the test. In the propellant burning rate example, a type I error will occur when either $\bar{x}$ <48.5 or $\bar{x}$ >51.5 when the true mean burning rate is μ=50 centimeters per second.

Suppose that the standard deviation of burning rate is σ=2.5 centimeters per second and that the burning rate has a distribution for which the conditions of the central limit theorem apply, so the distribution of the sample mean is approximately normal with mean $\bar{x}$ =50 and standard deviation = $\sigma / \sqrt{n}$ = 2.5 / $\sqrt{10}$ =0.79. The probability of making a type I error (or the significance level of our test) is equal to the sum of the areas that have been shaded in the tails of the normal distribution in the figure below.



We may find this probability as

Pr(X < 48.5 when μ= 50) + Pr(X > 51.5 when μ= 50)

The z-values that correspond to the critical values 48.5 and 51.5 are z1 =(48.5-50)/0.79 =-1.90 and z2= (51.5-50)/0.79m=1.90.

Therefore α = Pr(Z <-1.90) +Pr(Z>1.90) = 0.028717 + 0.028717 = 0.057434.

This implies that 5.76% of all random samples would lead to rejection of the hypothesis when the true mean burning rate is really 50 centimeters per second.

From inspection of the above figure, notice that we can reduce α by widening the acceptance region. For example, if we make the critical values 48 and 52, the value of α is

α = Pr(Z < (48-50)/0.79) +Pr(Z>(52-50)/0.79) = 0.0057 + 0.0057 = 0.0114

We could also reduce α by increasing the sample size n. If n=16, s=2.5, $\sigma / \sqrt{n}$ =0.625, and using the original critical region , we find z1 = (48.5 – 50)/0.625 = -2.40 and z2=(51.5-50)/0.625 = +2.40 which produces an

α = Pr(Z<-2.4) + Pr(Z>2.4) = 0.0082 + 0.0082 =0.0164.

Now to complete our look at hypothesis testing we need to explore how to calculate the Type II error. <u>This has always proven to be confusing to students so some concentration at this point in the reading is required!</u>

In evaluating a hypothesis-testing procedure, it is also important to examine the probability of a type II error, which we will denote by β.

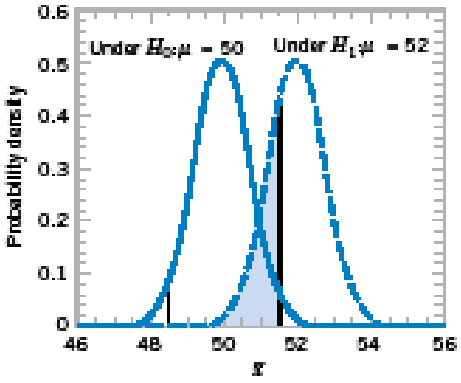
Definition: β = Pr(type II error) = Pr(fail to reject H_0 when H_0 is false)

To calculate β (sometimes called the **β-error**), we must have a specific alternative hypothesis; that is, we must have a particular alternate value of μ. For example, suppose that it is important to reject the null hypothesis H_0: μ= 50 whenever the mean burning rate μ is greater than 52 centimeters per second or less than 48 centimeters per second. We could calculate the probability of a type II error β for the values μ= 52 and μ= 48 and use this result to tell us something about how the test procedure would perform. Specifically, how will the test procedure work if we wish to detect, that is, reject H_0, for a mean value of μ= 52 or μ= 48? Because of symmetry, it is necessary only to evaluate one of the two cases—say, find the probability of accepting the null hypothesis H_0: μ= 50 centimeters per second when the true mean is μ= 52 centimeters per second.

The Figure below will help us calculate the probability of type II error β. The normal distribution on the left in figure below is the distribution of the test statistic when the null hypothesis H_0: μ= 50 is true (this is what is meant by the expression "under H_0: μ= 50"), and the normal distribution on the right is the distribution for the alternative hypothesis being true and the value of the mean is 52 (or "under H_1: μ=52"). Now a type II error will be committed if the sample mean falls between 48.5 and 51.5 (the critical region boundaries) when μ= 52 as opposed to μ=50. Make sure you understand why this is correct.

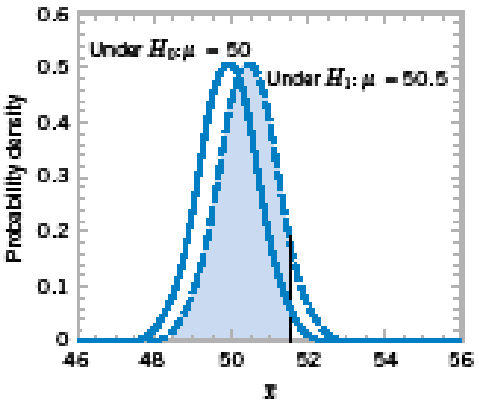
As seen in the figure below, this β is just the probability, shown by the shaded area under the normal distribution for μ=52. Therefore, referring to the figure below, we find that

$\beta = Pr(48.5 \le X \le 51.5 \text{ when } \mu = 52)$



The z-values corresponding to 48.5 and 51.5 when μ= 52 are z1= (48.5-52)/0.79=-4.43 and z2= (51.5-52)/0.79 = -0.63 so
β = Pr(-4.43 < Z < -0.63) = 0.2643.
The power of this test is defined as 1-β = 0.7357 and this is the probability of rejecting H_0: μ=50 when the true mean is μ=52.
In general we would be happy with experiments that produced confidence level, CL = 1-α = 0.9 and Power = 1-β = 0.80.

If instead of an alternate hypothesis of μ=52 we had instead hypothesized the true distribution to have a mean μ=50.5 we would find the β to be much higher and the power much lower since it would be more difficult to differentiate between the two distributions. If the true value of the mean is μ= 50.5 and the hypothesized value is H_0: μ= 50. The true value of μ is very close to 50, and the value for β =Pr(48.5<μ<51.5 when μ=50.5).



As shown in figure below, the z-values corresponding to 48.5 and 51.5 when μ=50.5 are
z1 = -2.53 and z2 = 1.27 which calculates to be β = Pr(Z<1.27) - Pr(Z<-2.53) = 0.8980 - 0.0057 = 0.8923 and therefore the power of this test is 1-β = 0.1077

Thus, the type II error probability is much higher for the case where the true mean is 50.5 centimeters per second than for the case where the mean is 52 centimeters per second.

In design of experiments we are often asked to compute the power of the test BEFORE we get any data. This of course is not possible however we can parameterize the power (or just β) with the ratio of $\delta = (\mu_0 - \mu_1)/\sigma$ and then explore the power versus this ratio varying δ from say 1 to 3.

Discussion.

This analysis of the calculation of β reveals four important points:
1. The size of the critical region, and consequently the probability of a type I error, can always be reduced by appropriate selection of the critical values.
2. Type I and type II errors are related. A decrease in the probability of one type of error always results in an increase in the probability of the other, provided that the sample size n does not change.
3. An increase in sample size, n, will generally reduce both α and β, provided that the critical values are held constant.
4. When the null hypothesis is false, β increases as the true value of the parameter approaches the value hypothesized in the null hypothesis ($\mu_1 \rightarrow \mu_0$). The value of β decreases as the difference between the true mean and the hypothesized value increases.

Generally, the analyst controls the type I error probability α when he or she selects the critical values. Thus, it is usually easy for the analyst to set the type I error probability at (or near) any desired value. Since the analyst can directly control the probability of wrongly rejecting H_0, we always think of rejection of the null hypothesis H_0 as a strong conclusion.
On the other hand, the probability of type II error β is not a constant, but depends on the true value of the parameter. It also depends on the sample size that we have selected. Because the type II error probability β is a function of both the sample size and the extent to which the null hypothesis H_0 is false, it is customary to think of the decision to accept H_0 as a weak conclusion, unless we know that β is acceptably small.
Therefore, rather than saying we "accept H_0", we prefer the terminology "fail to reject H_0". Failing to reject H_0 implies that we have not found sufficient evidence to reject H_0, that is, to make a strong statement.
Failing to reject H_0 does not necessarily mean that there is a high probability that H_0 is true. Remember the old adage "the absence of evidence is not evidence of absence" even though our judicial system works that way! It may simply mean that more data are required to reach a strong conclusion. This can have important implications for the formulation of hypotheses.

Finally in addressing specifically the concept of power $(1-\beta)$ of a statistical test Power is a very descriptive and concise measure of the sensitivity of a statistical test, where by sensitivity we mean the ability of the test to detect differences. In this case, the sensitivity of the test for detecting the difference between a mean burning rate of 50 centimeters per second and 52 centimeters per second is 0.7357. That is, if the true mean is really 52 centimeters per second, this test will correctly reject and "detect" this difference 73.57% of the time. If this value of power is judged to be too low, the analyst can increase either α or the sample size n.

This note started out to explain the concepts of confidence intervals and hypothesis testing without overloading the brain of the reader. Hopefully this goal was reasonably successful.

Reference:

1. Montgomery et al.,"Engineering Statistics, 5th Ed.", (2011),Wiley& Sons.
2. Agresti and Coull,(1998) The American Statistician, Vol. 52, pp 119-126
3. L. Leemis, "Lower System Reliability Bounds from Binary Failure Data Using Bootstrapping,"(2006), J. Quality Technology, Vol. 38, No. 1