The Logic of Statistical Hypothesis Testing

Best Practice

Authored by: Jennifer Kensler, PhD STAT T&E COE

Reviewed by: Laura Freeman, PhD IDA

The goal of the STAT T&E COE is to assist in developing rigorous, defensible test strategies to more effectively quantify and characterize system performance and provide information that reduces risk. This and other COE products are available at www.AFIT.edu/STAT.

STAT T&E Center of Excellence
2950 Hobson Way – Wright-Patterson AFB, OH 45433

Table of Contents

Executive Summary

This best practice provides an introduction to statistical hypothesis testing, which uses observed data to draw conclusions about a claim regarding a larger population or populations. Statistical hypothesis tests are the building blocks upon which many statistical analysis methods rely and therefore it is important to understand the basics of hypothesis testing. The hypothesis test must be carefully constructed so that it accurately reflects the question the tester wants to answer. This includes clearly stating the hypotheses and understanding the assumptions that the hypothesis test makes. This best practice provides an overview of the logic behind hypothesis testing to introduce key concepts and terminology. It highlights the importance of understanding and correctly interpreting the results of a hypothesis test as well as common errors and misunderstandings. A simple example of a one sample t-test ***illustrates the concepts*** presented in the context of Department of Defense (DoD) testing. Many statistical analyses, including more complex analyses, utilize hypothesis testing. The analysis of data from a designed experiment simultaneously performs multiple hypothesis tests.

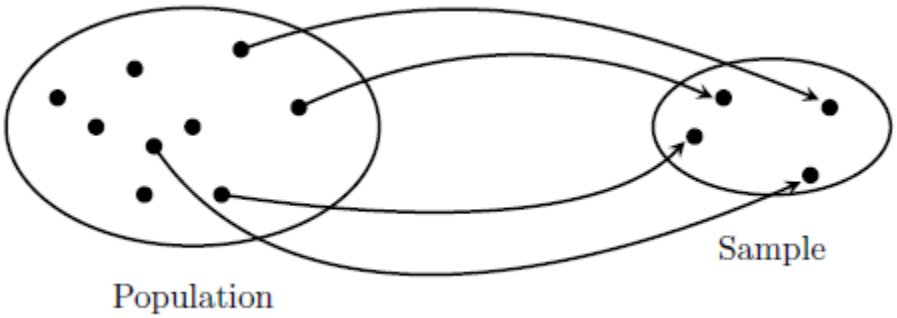
Keywords: Statistical Inference, Risk, Significance Level, P-Value, Sample Size

Introduction

Statistical Inference

Often one is interested in drawing conclusions about a population, but examining the entire population is usually impractical or impossible. Statistical inference involves using information obtained from a sample to draw conclusions about a larger population. Statistical inference is key to having rigorous and adequate DoD tests because we are often interested in future performance of a system under similar (or different) conditions. Since we do not know what the future holds we are dependent on statistical inference to make statements about future performance. Therefore, the sample must be representative of the population; this objective is usually achieved by obtaining a randomly selected sample. Figure 1 illustrates that a sample is a subset of the population. In hypothesis testing, one form of statistical inference, a claim about a population is evaluated using data observed from a sample of the population. The data one observes will be different depending on which individuals of the population the sample captures. Hypothesis testing addresses this random sampling "error" (i.e. variation) and allows one to evaluate claims regarding the values of a single parameter, several parameters, or the form of an entire probability distribution of a population.

Figure 1



An Illustrative Example

Consider the target location error for a rocket. The Air Force wants to determine if the mean target location error for the Roka rocket is less than 10 feet. In this case the population includes all Roka rockets and the mean target location error for the entire population is denoted by μ (read "mew"). On the other hand, the observed mean target location error of the sample is denoted by $\bar{y}$ (read "y-bar"). The mean target location error for the sample will be used to conclude whether the mean population target location error is less than 10 feet.

DoD Testing

Throughout the acquisition lifecycle many questions must be answered regarding the performance and suitability of a system. Does a missile hit its target at least 80% of the time? Is the mean miles between system failure at least 3,000? Statistical hypothesis testing is a vehicle for answering these questions. Care must be taken in setting up the hypothesis test to ensure that the analysis performed addresses the test objective. Too often DoD testing includes "implied" hypothesis tests in which the actual hypotheses are never explicitly stated! This ambiguity means that the statistical analysis may be answering a different question than the tester intended.

Foundations of Hypothesis Testing

The Null and Alternative Hypotheses

In statistical hypothesis testing there are two mutually exclusive hypotheses. The null hypothesis, denoted H_0 (read "H-naught"), and the alternative hypothesis, denoted H_A (read "H-a"). The null hypothesis is the default position: it represents the status quo, conventional thinking, or historical performance. The alternative hypothesis is the claim to be tested; it reflects what the tester is trying to show.

To illustrate the concept of null and alternative hypotheses consider a criminal trial. In the criminal justice system the defendant is presumed innocent until proven guilty. In the language of hypothesis testing the null hypothesis is $H_0: Defendant\ is\ Innocent$ and the alternative hypothesis is $H_A: Defendant\ is\ Guilty$. Since the defendant is considered innocent until proven guilty, the burden of

proof is on the prosecution who must show guilt beyond reasonable doubt in order to obtain a conviction. Likewise, in hypothesis testing the burden of proof is on the alternative hypothesis. The null hypothesis is not rejected unless there is strong evidence to support the alternative hypothesis. Thus, it is important to clearly state the hypotheses. A null hypothesis of H_0: $Defendant\ is\ Innocent$ has quite different implications for a trial than a null hypothesis of H_0: $Defendant\ is\ Guilty$!

Type I and Type II Errors

In the criminal trial analogy there are two possible ways that a mistake can be made. The jury can find an innocent person guilty or they can find a guilty person not guilty. Notice that the jury never finds the defendant innocent, but instead can declare the defendant not guilty. Similarly in hypothesis testing we do not accept the null hypothesis, but instead fail to reject the null hypothesis. Just because there is not enough evidence to reject the null hypothesis does not mean that the null hypothesis is true! Figure 2 summarizes the possible outcomes from a hypothesis test.

Figure 2

		Decision (Verdict)	
		Fail to Reject H_0 (Not Guilty)	**Reject H_0** (Guilty)
Truth	**H_0 is True** (Defendant is Innocent)	Correct	Type I Error
	H_A is True (Defendant is Guilty)	Type II Error	Correct

In hypothesis testing a type I error means rejecting the null hypothesis when the null hypothesis is true. The probability of a type I error is called α (alpha). A type II error means failing to reject the null hypothesis when the alternative hypothesis is true. The probability of a type II error is called β (beta). Type I and type II errors represent two ways an incorrect conclusion can be made, thus both α and β should be minimized. Unfortunately, α and β work in opposition to one another. If one is concerned with falsely convicting an innocent person the standard required for conviction can be raised, leading to a lower type I error rate α. However, making it harder to obtain a conviction has also made it harder to convict a guilty person. Thus, lowering the type I error rate has increased the type II error rate! Likewise lowering the type II error rate (making it easier to convict a guilty person) will increase the type I error rate (more innocent people will be found guilty).

Testers must carefully consider the relative consequences of making type I and type II errors in setting up their hypothesis test, so that the risk of making type I and type II errors reflects the severity of the consequences of these errors. One might ask how can both α and β be decreased? The (usually impractical) answer is to collect more information! In the criminal trial the risk of type I and type II errors

can be reduced by continuing the investigation and finding previously missed witnesses and clues—in other words collecting more data.

Working with limited resources is a major challenge in DoD testing. The amount of data collected will almost certainly be driven by resource constraints and be less than ideal. In any case, it is vitally important that everyone enter the test aware of the risks. The risk of making type I and type II errors must be fully understood and acceptable in light of the potential consequences.

Performing a Hypothesis Test

Setting Up the Hypothesis Test

For the sake of simplicity this best practice examines the case of a hypothesis test about a population mean. Table 1 shows the three forms of the null and alternative hypotheses where μ_0 is the value of the population mean under the null hypothesis. These three forms of the hypothesis are general and apply to tests about other parameters.

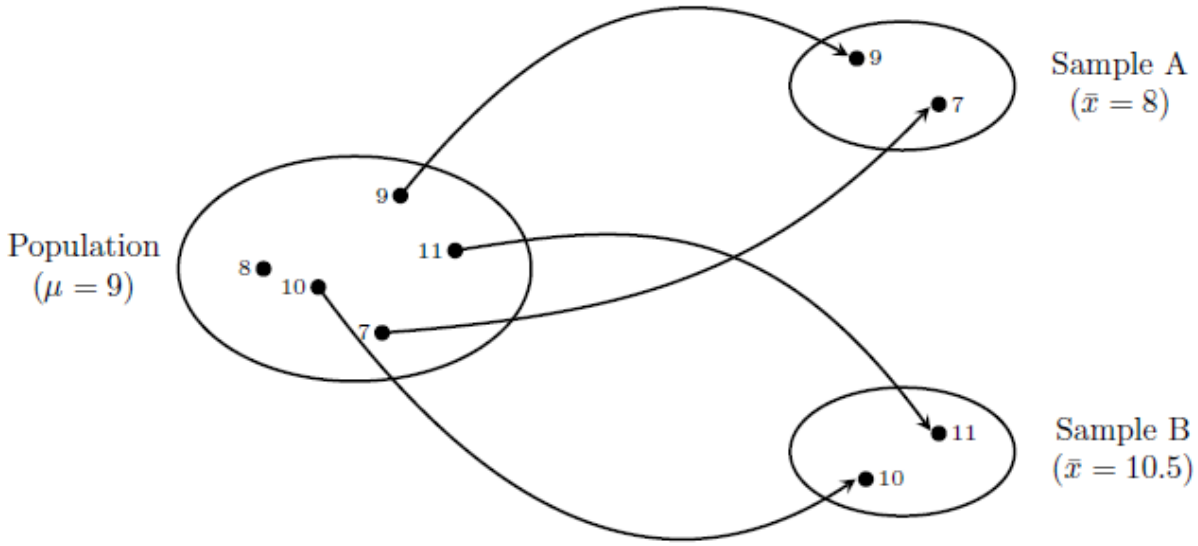
Table 1

Two-Tailed	Left-Tailed	Right-Tailed
$H_0: \mu = \mu_0$ $H_A: \mu \neq \mu_0$	$H_0: \mu = \mu_0$ $H_A: \mu < \mu_0$	$H_0: \mu = \mu_0$ $H_A: \mu > \mu_0$

Recall the target location error example where we are interested in testing whether the population mean target location error for the Roka rocket is less than 10 feet. A hypothesis test can answer the question of whether, based on a sample, there is evidence that the entire population of rockets has a mean target location error less than 10 feet. In this case a one sample t-test can be used to answer this question. The mathematical formulas are included in the Appendix; here the example motivates the logic behind the t-test. The hypotheses are $H_0: \mu = 10$ vs. $H_A: \mu < 10$.

A perfect decision rule would always reject the null hypothesis when the population mean target location error is less than 10 feet and always fail to reject the null hypothesis when it is not. A complication arises because we do not see the entire population but only a sample. Figure 3 illustrates that a single population could lead to different samples and hence different conclusions. The correct decision would be to reject the null hypothesis that the population mean is equal to 10 feet in favor of the alternative hypothesis that the population mean is less than 10 feet. The decision based on Sample B would be to incorrectly fail to reject the claim that the population mean is equal to 10 feet. Although a perfect decision rule does not exist, statistics quantifies the risk of drawing incorrect conclusions. Typically, the hypotheses are set up such that rejecting the null hypothesis when the null hypothesis is true reflects the worst possible outcome.

Figure 3



Part of setting up the hypothesis test includes characterizing the type I and type II error rates. The significance level α, which is the probability of a type I error, must be set **prior to the collection of data.** Common values for α include 0.01, 0.05, 0.1 and 0.2 and should be chosen based on the consequences of rejecting the null hypothesis when it is true. In our example a type I error means concluding that the target location error is less than 10 feet when it really is 10 feet.

The type II error depends on δ (delta), the operationally significant difference to be detected. What difference in mean target location error from 10 feet does the tester want to be able to detect? Does a mean target location error of 9.999998 represent a meaningful improvement from 10? No. What about 9.9? How about 9.5? What difference represents an operationally significant improvement in mean target location error? Consultation with stakeholders reveals that a difference of 0.75 feet is of practical importance.

Based on previous tests, the team estimates that the standard deviation (a measure of spread or variation in data) of the target location error is around 1.5 feet. The team uses statistical software to scope the trade space for different values of α, β, and sample size with $\delta = 0.75$ and a standard deviation of 1.5. After considering the consequences of type I and type II errors the team decides to set $\alpha = 0.1$ and $\beta = 0.2$ and to use a sample size of 19.

A value of $\alpha = 0.1$ means that if the population mean target location error is truly 10 feet there is 10% probability of observing test results that would lead the tester to conclude that the mean target location error is less than 10 feet. On the other hand power is the complement of the type II error rate; it is the probability of rejecting the null hypothesis when it is false, $1 - \beta$. Having 1 – 0.20 = 0.80 power to detect a difference of 0.75 feet means that if the population mean target location error is really 9.25 (10 – 0.75) there is an 80% chance of observing data that will lead the tester to conclude that the mean target

location error is less than 10 feet. Note that the probability of type I and type II errors are used for planning purposes and are determined prior to the collection of data.

Analyzing the Data

Suppose 19 rockets are launched and we obtain a sample mean target location error of 8.6 feet. Figure 4 shows a histogram of the data.

Figure 4

Before performing the t-test, the assumptions of the t-test must be checked. The t-test assumes that the observations are independent and come from a simple random sample, and that either the observations follow the normal distribution or the sample size is sufficiently large[1]. These assumptions are reasonable for the target location error experiment. Caution: failing to check these assumptions can lead to incorrect conclusions! The violation of these assumptions may necessitate the use of other statistical techniques to appropriately analyze the data.

In order to decide whether these data are evidence that the population mean target location error is less than 10 feet the p-value is compared to the significance level $\alpha = 0.10$. A p-value is the probability under the null hypothesis of obtaining results as extreme or more extreme that what was actually observed. In this example it is the probability, assuming the population mean target location error is 10 feet, of obtaining a sample mean of 8.6 feet or less. It is important to understand that the p-value ***is not*** the probability that the null hypothesis is true! A small p-value means that under the null hypothesis results as extreme as those observed are unlikely. If the p-value is less than the significance level α then reject the null hypothesis; otherwise fail to reject the null hypothesis.

[1] The sufficiently large option requires the sample size to be large enough to invoke the central limit theorem. The sample size needed to invoke the central limit theorem depends on the underlying distribution of the population. Although a common rule of thumb is 30, in many cases a sample size as small as 10 may be adequate to invoke the central limit theorem.

Statistical software shows that the p-value from the hypothesis test is 0.0002. Meaning that if the mean target location error really was 10 feet there is only a 0.02% chance of obtaining a sample mean as small as or smaller than 8.6 feet. Since 0.0002 is less than $\alpha = 0.10$ reject the null hypothesis and conclude that there is evidence at the $\alpha = 0.10$ significance level that the mean target location error of the population of rockets is less than 10 feet.

Types of Hypothesis Tests

The example in this tutorial uses a one sample t-test to compare a population mean to a requirement. In practice the situation is rarely this simple, but there are many types of hypothesis tests appropriate for more complex situations. The t-test can be extended to the case of multiple samples. For example, the two-sample t-test examines whether the population means of two groups are the same and the one-way Analysis of Variance (ANOVA) compares the population means of three or more groups. Extending this comparison further the general *k* factor ANOVA determines whether each factor or factor interaction affects the population mean by performing multiple hypothesis tests simultaneously!

In addition to the mean hypothesis tests are often performed on other parameters including proportions and variances. While there are countless types of hypothesis tests, Table 2 lists some common hypothesis test about means, proportions and variances. Regardless of the kind of hypothesis test being performed the fundamental process is the same as that illustrated with the one sample t-test example.

Table 2

Mean	Proportion	Variance
One Sample t-test	One Sample z-test for Proportions	One sample χ^2 test
Two Sample t-test	Two Sample z-test for Proportions	Two Sample F-test
Paired t-test	χ^2 Homogeneity of Proportions	
ANOVA		

Hypothesis Testing and Confidence Intervals

Although the scope of this best practice does not include confidence intervals it acknowledges that there is a relationship between hypothesis testing and confidence intervals. Namely it is possible (and statistically valid) to perform a hypothesis test at the α significance level using a confidence interval with a $100(1-\alpha)\%$ confidence level. Although it is common in DoD to frame hypothesis tests in terms of confidence intervals (i.e. discussing a 90% confidence interval as opposed to testing a hypothesis at the 10% significance level), this best practice does not include this approach for several reasons. 1) ***An explicit statement of the hypotheses is necessary*** regardless of whether a traditional hypothesis test is being performed or a confidence interval is being used to perform the test. However, in the confidence interval approach this explicit statement is almost always overlooked. 2) In the confidence interval

approach the ***critical upfront planning for power*** is often omitted. 3) "Confidence" has a very specific statistical meaning, but is often misconstrued in a more general colloquial sense and used incorrectly. 4) The confidence interval approach to hypothesis testing confounds confidence intervals and hypothesis testing, which while related have different objectives. Confidence intervals are used to estimate a parameter, while hypothesis tests are used to test a claim about a parameter.

Conclusions

Hypothesis tests use information from a sample to evaluate a claim about the larger population. Upfront planning is vitally important to ensure the hypothesis test adequately addresses the tester's objectives. Hypotheses must be clearly stated and assumptions understood. The trade space for the probabilities of type I and type II errors must be examined and the values chosen must reflect the severity of these errors. Finally, the results of the hypothesis test must be thoroughly understood and correctly interpreted.

References

Sullivan, M. (2004). *Statistics Informed Decisions Using Data.* Upper Saddle River, New Jersey: Pearson Education.

Appendix

The One Sample T-Test

The test statistic for the one sample t-test is

$$t_{obs} = \frac{\bar{y} - \mu_0}{s/\sqrt{n}}$$

where $\bar{y}$ is the sample mean, μ_0 is the population mean under the null hypothesis, n is the number of observations, and the sample standard deviation is $s = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}}$. For the Roka rocket target location error example the observed test statistic is

$$t_{obs} = \frac{8.6 - 10}{1.4/\sqrt{19}}$$

$$= -4.4.$$

The calculation of the p-value depends on the form of the alternative hypothesis and is given by

$$\text{p-value} = \begin{cases} 2P(T > |t_{obs}|) & \text{if } H_A: \mu \neq \mu_0 \\ P(T < t_{obs}) & \text{if } H_A: \mu < \mu_0 \\ P(T > t_{obs}) & \text{if } H_A: \mu > \mu_0 \end{cases}$$

where T follows the t-distribution with $n - 1$ degrees of freedom. For the Roka rocket target location error example the p-value is $P(T < -4.4) = 0.0002$.