XI. ADVANCED STATISTICS

MOST PEOPLE WOULD RATHER LIVE WITH A PROBLEM THEY CAN'T SOLVE, THAN ACCEPT A SOLUTION THEY CAN'T UNDERSTAND.

R. E. D. WOOLSEY AND H. S. SWANSON

Advanced Statistics

Advanced Statistics is presented in the following topic areas:

- Statistical decision making
- Analysis of variance (ANOVA)
- Relationships between variables
- Design and analysis of experiments

Statistical Decision Making

Statistical Decision Making is presented in the following topic areas:

- Point estimates
- Confidence intervals
- Hypothesis testing
- Paired-comparison tests
- Goodness-of-fit tests
- Contingency tables

Statistical Inference

The objective of statistical inference is to draw conclusions about population characteristics based on the information contained in a sample. The steps involved in statistical inference are:

- Precisely define the problem objective
- Formulate a null and an alternate hypothesis
- Decide if the problem will be evaluated by a one-tail or two-tail test
- Select a test distribution and a critical value for the test statistic
- Calculate a test statistic from the sample
- Make a inference by comparing the calculated and the critical values
- Report the findings

Point Estimate for Population Mean

In analyzing sample values to arrive at population probabilities, two major estimators are used: point estimates and confidence intervals.

A point estimate of the population mean, μ, is the sample mean, $\bar{X}$.

$$\mu \approx \bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

Example: Given the following tensile strength readings from 4 piano wire segments: 28.7, 27.9, 29.2, and 26.5 psi, calculate the point estimation of the population mean.

$$\mu \approx \bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{28.7 + 27.9 + 29.2 + 26.5}{4}$$

$$\mu \approx \bar{X} = 28.08 \text{ psi}$$

28.08 psi is the point estimate for the population mean.

Point Estimate for Population Variance

The sample variance, s^2, is the best point estimate of the population variance, σ^2.

The sample standard deviation, s, is the best point estimate of the population standard deviation, σ.

$$s^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n - 1} \qquad \sigma^2 = \frac{\sum_{i=1}^{n}(X_i - \mu)^2}{N}$$

$$s = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n - 1}} \qquad \sigma = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \mu)^2}{N}}$$

Confidence Interval for the Mean

Continuous Data - σ Known

The confidence interval of the population mean, μ, when the population standard deviation, σ, is known, is calculated using the sample mean, $\bar{X}$, the population standard deviation, σ, the sample size, n, and the normal distribution.

$$\bar{X} - Z_{\alpha/2}\frac{\sigma_X}{\sqrt{n}} \leq \mu \leq \bar{X} + Z_{\alpha/2}\frac{\sigma_X}{\sqrt{n}}$$

From sample data one can calculate the interval within which the population mean, μ, is predicted to fall. Confidence intervals are always estimated for population parameters. A confidence interval is a two-tail event and requires critical values based on an alpha/2 risk in each tail.

Continuous Data - σ Unknown

The confidence interval of the population mean, μ, when the population standard deviation, σ, is unknown, is calculated using the sample mean, $\overline{X}$, the sample standard deviation, s, the sample size, n, and the t distribution.

$$\overline{X} - t_{\alpha/2,\, n-1}\frac{s}{\sqrt{n}} \leq \mu \leq \overline{X} + t_{\alpha/2,\, n-1}\frac{s}{\sqrt{n}}$$

Example: The average of 25 samples is 18 with a sample standard deviation of 6. Calculate the 95% confidence interval for the population mean.

$$\overline{X} - t_{\alpha/2,\, n-1}\frac{s}{\sqrt{n}} \leq \mu \leq \overline{X} + t_{\alpha/2,\, n-1}\frac{s}{\sqrt{n}}$$

$$18 - 2.064\frac{6}{\sqrt{25}} \leq \mu \leq 18 + 2.064\frac{6}{\sqrt{25}}$$

$$15.52 \leq \mu \leq 20.48$$

Confidence Interval for Proportion

For large sample sizes, with np and n(1-p) greater than or equal to 5, the binomial distribution can be approximated by the normal distribution to calculate a confidence interval for population proportion.

$$p_s - Z_{\alpha/2}\sqrt{\frac{p_s(1 - p_s)}{n}} \leq p \leq p_s + Z_{\alpha/2}\sqrt{\frac{p_s(1 - p_s)}{n}}$$

p_s = sample proportion
p = population proportion
n = sample size

Confidence Interval for Variance

The confidence interval or interval estimate for the population variance, σ^2, is given by:

$$\frac{(n - 1)s_X^2}{X^2_{\alpha/2,\, n-1}} \leq \sigma^2 \leq \frac{(n - 1)s_X^2}{X^2_{1-\alpha/2,\, n-1}}$$

s^2 = sample variance
n = sample size
n - 1 = degrees of freedom

Confidence Interval for Standard Deviation

The confidence interval for the population standard deviation, σ, is given by:

$$\sqrt{\frac{(n - 1)s_X^2}{X^2_{\alpha/2,\, n-1}}} \leq \sigma \leq \sqrt{\frac{(n - 1)s_X^2}{X^2_{1-\alpha/2,\, n-1}}}$$

Hypothesis Testing

Hypothesis testing is a type of statistical inference in which a null hypothesis and alternative hypothesis are stated. The null hypothesis is a statement about the value of a population parameter such as the mean, and must contain the condition of equality.

The alternative hypothesis is a statement that must be true if the null hypothesis is false.

A null hypothesis can only be rejected, or fail to be rejected, it cannot be accepted because of a lack of evidence to reject it.

As an example of hypothesis tests for a population mean, there are only three possible forms, where μ is the population mean and μ_0 is a specified value:

$H_0: \mu = \mu_0$	$H_0: \mu \leq \mu_0$	$H_0: \mu \geq \mu_0$
$H_1: \mu \neq \mu_0$	$H_1: \mu > \mu_0$	$H_1: \mu < \mu_0$

Hypothesis Testing (Continued)

The steps of hypothesis testing are:

- State the null and alternative hypothesis
- Specify the level of significance, α
- Determine the critical values separating the reject and nonrejection areas
- Determine the sampling distribution and test statistic
- Determine if the test statistic is in the reject or nonrejection area
- Conclude if the null hypothesis is rejected or failed to be rejected
- State the statistical decision in terms of the original problem

Types of Errors

When formulating a conclusion regarding a population based on observations from a small sample, two types of errors are possible:

- Type I error: This error results when the null hypothesis is rejected when it is, in fact, true. The probability of making a type I error is called α (alpha) or producer's risk.

- Type II error: This error results when the null hypothesis is not rejected when it should be rejected. This error is called the consumer's risk and is denoted by the symbol β (beta).

The degree of risk (α) is normally chosen by the concerned parties (α is often taken as 5%) in arriving at the critical value of the test statistic. Increasing the sample size can reduce both the α and β risks.

Types of Errors (Continued)

The types of errors are shown in the Figure below:

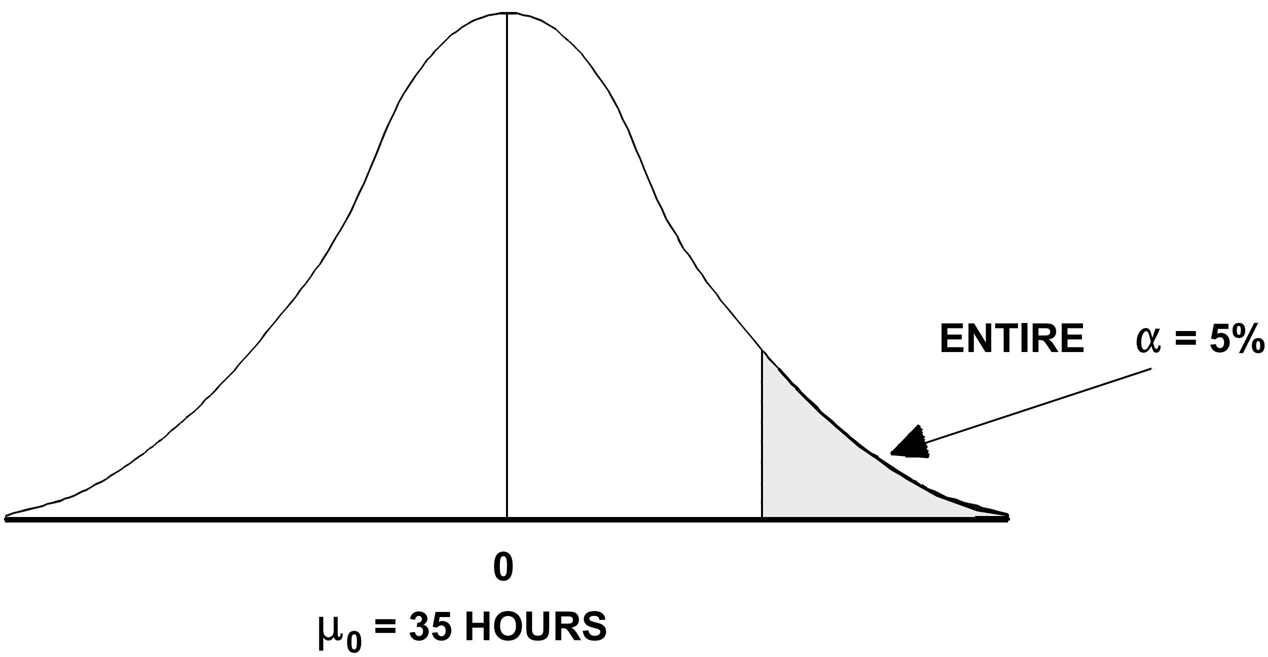
		Null Hypothesis: True	Null Hypothesis: False
The Decision Made	Fail to Reject H_0	$p = 1 - \alpha$ Correct Decision	$p = \beta$ Type II Error
	Reject H_0	$p = \alpha$ Type I Error	$p = 1- \beta$ Correct Decision

Error Matrix

One-Tail Test

If a null hypothesis is established to test whether a sample value is smaller or larger than a population value, then the entire α risk is placed on one end of a distribution curve. This constitutes a one tail-test.

H_0: new $\leq$ to present **H_1: new > present**

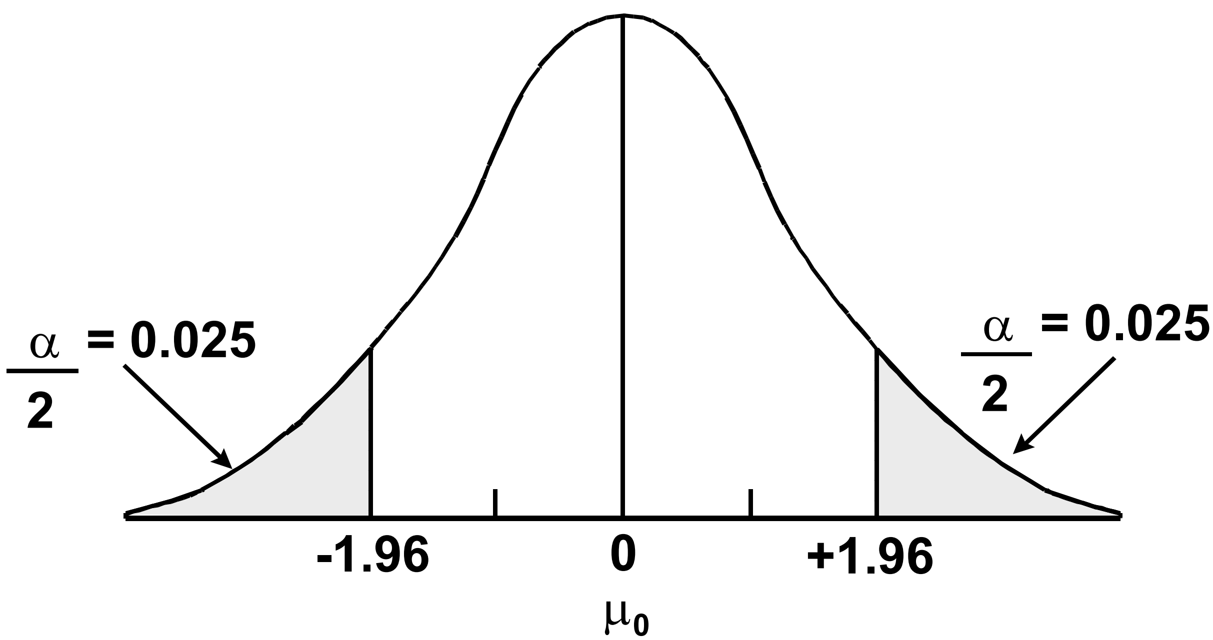


Determine if the true mean is within the α critical region.

Two-Tail Test

If a null hypothesis is established to test whether a population shift has occurred, in either direction, then a two-tail test is required. The allowable α error is generally divided into two equal parts.

H_0: levels are = **H_1: levels are ≠**



Determine if the true mean is within either the upper or lower α critical regions.