Unit 6: Basic Concepts of Inference

Statistics 571: Statistical Methods
Ramón V. León
Statistical Inference

• Deals with methods for making statements about a population based on a sample drawn from the population
  – **Estimation:** Estimating an unknown population parameter
  – **Hypothesis testing:** Testing a *hypothesis* about an unknown population parameter

• Examples
  – **Estimation:** Estimating the mean package weight of a cereal box filled during a production shift
  – **Hypothesis testing:** Do the cereal boxes meet the minimum mean weight specification of 16 oz?

• Inference
  – Informal using summary statistics (Chapter 4)
  – Formal which uses methods of probability and sampling distributions to develop measures of statistical accuracy (Chapter 6 and beyond)
Estimation Problems

- **Point estimation**: Estimation of an unknown population parameter by a single statistic calculated from the sample data.

- **Confidence interval estimation**: Calculation of an interval from sample data that includes the unknown population parameter with a preassigned probability.
Point Estimation

Let $X_1, X_2, \ldots, X_n$ be a random sample from a population with an unknown parameter $\theta$.

A point estimator $\hat{\theta}$ of $\theta$ is a statistic computed from sample data

$$
\hat{\theta} = \bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}
$$

is an estimator of $\theta = \mu$

$$
\hat{\theta} = S^2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n-1}
$$

is an estimator of $\theta = \sigma^2$

$\hat{\theta}$ is a r.v. because it is a function of the $X_i$s which are r.v.’s
Point Estimation Terminology

**Estimator** = the random variable $\hat{\theta}$

**Estimate** = the numerical value of $\hat{\theta}$ calculated from the observed sample data $X_1 = x_1, \ldots, X_n = x_n$

Example:

An alternative estimator of the population mean $\theta$ is the **midrange**, defined as $\hat{\theta} = \frac{1}{2}(X_{\text{min}} + X_{\text{max}})$. Suppose a sample of size five is observed with $x_1 = 2.3$, $x_2 = 3.3$, $x_3 = 1.9$, $x_4 = 2.8$, and $x_5 = 2.4$. Then $x_{\text{min}} = 1.9$, $x_{\text{max}} = 3.3$, and the midrange estimate equals $\hat{\theta} = \frac{1}{2}(1.9 + 3.3) = 2.6$. The sample mean estimate equals $\hat{\theta} = \bar{x} = \frac{1}{5}(2.3 + 3.3 + 1.9 + 2.8 + 2.4) = 2.54$. ◆
Estimators are Random Variables

A random sample of 100 transistors from a large shipment is inspected to estimate the proportion of defectives in the shipment. Let $\theta$ be the fraction of defective transistors in the shipment and $\hat{\theta}$ be the proportion of defectives in the sample. Suppose that four independent samples are drawn with replacement from the shipment, yielding the following results.

<table>
<thead>
<tr>
<th>Sample</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proportion defectives $\hat{\theta}$</td>
<td>0.02</td>
<td>0.00</td>
<td>0.01</td>
<td>0.04</td>
</tr>
</tbody>
</table>

Although each value of $\hat{\theta}$ in this table is an estimate of the proportion of defectives $\theta$ in the shipment, the estimate varies from sample to sample. Thus the population parameter $\theta$ is a constant, but the point estimator $\hat{\theta}$ is an r.v.
Bias and Variance

\[ \text{Bias}(\hat{\theta}) = E(\hat{\theta}) - \theta \]

- The bias measures the average accuracy of an estimator
- An estimator whose bias is zero is called unbiased
- An unbiased estimator may fluctuate greatly from sample to sample

\[ \text{Var}(\hat{\theta}) = E[\hat{\theta} - E(\hat{\theta})]^2 \]

- The lower the variance the more precise (reliable) the estimator
- A “good” estimator should have low bias and low variance
- Among unbiased estimator the one with the lowest variance should be chosen
Bull’s Eye Analogy

Figure 6.1  Illustration of Bias and Variance of an Estimator: (a) Low bias, low variance; (b) Low bias, high variance; (c) High bias, low variance; (d) High bias, high variance.
Mean Square Error

\[ \text{MSE}(\hat{\theta}) = E(\hat{\theta} - \theta)^2 \]

This is commonly referred to as the standard error of the mean (SEM). For example, the standard errors are estimated from the sample

Thus the MSE is the sum of the variance and the bias-squared terms.
A Bias Estimator Can Have a Smaller MSE Than a Bias One

\[ S_1^2 = \left( \sum_{i=1}^{n} (X_i - \bar{X})^2 \right) / (n - 1) \] is an unbiased estimator of \( \sigma^2 \)

\[
\begin{align*}
\text{unbiased} & \quad S_1^2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n - 1} \\
\text{and} & \quad S_2^2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n} \quad \text{biased}
\end{align*}
\]

\[ MSE(S_1^2) = Var(S_1^2) = \frac{2\sigma^2}{n - 1}, \quad MSE(S_2^2) = \frac{2n - 1}{n^2} \sigma^2 \]

Simple algebra shows that \( MSE(S_1^2) > MSE(S_2^2) \) when \( n > 1 \). Although \( S_2^2 \) is biased, it has a smaller MSE! However, the difference is very slight and approaches zero as \( n \) becomes large. In practice, \( S_1^2 \) is almost universally used.

\[
0.5 = \frac{2}{5 - 1} > \frac{2(5) - 1}{5^2} = 0.36
\]
Standard Error (SE)

• The standard deviation of an estimator is called the **standard error** of the estimator.

• The estimated standard error is also often called standard error (SE).

• The precision of an estimator (or estimate?) is measured by the SE.

**Example 1:** \( \bar{X} \) is an unbiased estimator of \( \mu \)

\[
SE(\bar{X}) = \frac{\sigma}{\sqrt{n}}
\]

(estimated) \( SE(\bar{X}) = \frac{s}{\sqrt{n}} \)

**Example 2:**
\( \hat{p} \) is an unbiased estimate of \( p \) and \( SE(\hat{p}) = \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} \)
Precision and Standard Error

• A precise estimate has a small standard error, but exactly how are the precision and standard error related?

• If the sampling distribution of an estimator is normal with mean equal to the true parameter value (i.e., unbiased). Then we know that about 95% of the time the estimator will be within two SE’s from the true parameter value
Confidence Interval Estimation

• We want an interval \([L, U]\), where \(L\) and \(U\) are two statistics calculated from \(X_1, X_2, \ldots, X_n\), such that

\[
P[L \leq \theta \leq U] \geq 1 - \alpha
\]

regardless of the true value of \(\theta\).

• \([L, U]\) is called a \textbf{100}(1-\alpha)\% confidence interval (CI)

• 1-\(\alpha\) is called the confidence level of the interval

• After the data is observed \(X_1 = x_1, \ldots, X_n = x_n\)

the confidence limits \(L = l, U = u\) can be calculated
Illustration of the Meaning of CI

Notice that for any particular confidence interval the probability of $\mu$ being in the interval is either 0 or 1. However, the process that generates the confidence intervals will produce an interval that contains $\mu$ 95% of the time.
When you click the sample button, 100 samples of the specified sample size (10, 15, or 20) will be taken from a population with a mean of 50 and a standard deviation of 10. The confidence interval on the mean will be computed for each. If the 95% confidence interval contains the population mean of 50 then a line will show the 95% confidence interval in orange and the 99% confidence interval in blue. If the 95% confidence interval does not contain the population mean then it will be shown in red. If the 99% interval does not contain the population mean it will be shown in white.

Cumulative Results

<table>
<thead>
<tr>
<th></th>
<th>99% Conf. Int</th>
<th>95% Conf. Int</th>
</tr>
</thead>
<tbody>
<tr>
<td>Contained 50</td>
<td>99</td>
<td>96</td>
</tr>
<tr>
<td>Did Not Contain 50</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>Proportion Contained</td>
<td>0.990</td>
<td>0.960</td>
</tr>
</tbody>
</table>
5 out of 50 do not cover 0 with alpha = 0.05.
95% Confidence Interval – Normal Case – $\sigma^2$ Known

Consider a random sample $X_1, X_2, \ldots, X_n$ from an $N(\mu, \sigma^2)$ where $\sigma^2$ is assumed to be known and the mean, $\mu$, is an unknown parameter to be estimated. Then

$$P \left[ -1.96 \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq 1.96 \right] = 0.95$$

By the CLT even if the observation are not normal this result is approximately correct for large $n$.

$$P \left[ L = \bar{X} - 1.96 \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + 1.96 \frac{\sigma}{\sqrt{n}} = U \right] = 0.95$$

$$l = \bar{x} - 1.96 \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + 1.96 \frac{\sigma}{\sqrt{n}} = u$$

is a 95% CI for $\mu$
Example of Confidence Interval

Example 6.7 (Airline Revenues: 95% Confidence Interval)

Airlines use sampling to estimate their mean share of the revenue for passengers traveling on two or more airlines. Suppose that the revenues for a certain airline are normally distributed with $\sigma = 50$ dollars. To estimate its mean share per ticket, the airline uses a sample of 400 tickets resulting in sample mean $\bar{x} = 175.60$ dollars. Calculate a 95% CI for the true mean share $\mu$.

A 95% CI for $\mu$ is

$$\left[ 175.60 - 1.96 \times \frac{50}{\sqrt{400}}, 175.60 + 1.96 \times \frac{50}{\sqrt{400}} \right] = [170.70, 180.50].$$

This can be considered as a plausible set of values for the true mean share of revenue that is consistent with the observed sample mean $\bar{x} = 175.60$. 🌟
Frequentist Interpretation of CI’s

Note that before this CI was calculated, one could make the statement that the probability is 0.95 that the random interval $[\bar{X} \pm 1.96 \times \frac{\sigma}{\sqrt{n}}]$ will include $\mu$, which is fixed but unknown. However, after the limits of the interval are calculated, the interval is no longer random because its limits are now fixed numbers. Either $\mu$ does or does not lie in the calculated interval. Thus $P(170.70 \leq \mu \leq 180.50)$ is either 1 or 0, but not 0.95. Therefore it is incorrect to say that the probability is 0.95 that the true $\mu$ is in $[170.70,180.50]$.

What then exactly is meant by the statement that the confidence level is 95%? The answer lies in the frequentist interpretation of a CI: In an infinitely long series of trials in which repeated samples of size $n$ are drawn from the same population and 95% CI’s for $\mu$ are calculated using the same method, the proportion of intervals that actually include $\mu$ will be 95%. However, for any particular CI, it is not known whether or not that CI includes $\mu$ (since $\mu$ is unknown).

Bayesian Approach to Statistics:
A 95% credibility interval $170.70 \leq \mu \leq 180.50$ is correctly interpreted as "there is a .95 probability of $\mu$ being between 170.70 and 180.50"
Arbitrary Confidence Level for CI – $\sigma^2$ Known

100(1-$\alpha$)% CI for $\mu$ based on the observed sample mean $\bar{x}$

$$\bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

$z_{0.005} = 2.576$ is used for 99% CI

See the JMP tutorial “Tabled Values of Common Distributions”
1954 Salk vaccine trial

- Sample of grade school children randomly divided into two groups, each with about 200,000 children
- Treatment group received the vaccine
- Control group received a placebo
- To test the claim that the population of vaccinated children would have a lower rate of new polio cases than the population of unvaccinated children
- 2.84 new cases per 10,000 for vaccinated, 7.06 new cases per 10,000 for unvaccinated
- Is there sufficiently strong evidence to support the efficacy claim for the Salk vaccine? Could rate difference be due to chance? The probability that change would produce a difference at least this large in less than one in a million. Therefore there is strong evidence of efficacy
Null and Alternative Hypothesis

- Objective of hypothesis testing is to assess the validity of a claim against a counterclaim using sample data.
- The claim to be proved is the alternative hypothesis ($H_1$) (or research hypothesis)
- The competing claim is called the null hypothesis ($H_0$)

Polio vaccine:  $H_0 : p_1 = p_2$ vs. $H_1 : p_1 > p_2$
Null and Alternative Hypothesis

• One begins by assuming that $H_0$ is true. If the data fails to contradict $H_0$ beyond a reasonable doubt, then $H_0$ is not rejected. However, failing to reject $H_0$ does not mean that we accept it as true. It simply means that $H_0$ cannot be ruled out as a possible explanation for the observed data. A proof by insufficient data is not a proof at all.

• Only when the data strongly contradict $H_0$ is this hypothesis rejected and $H_1$ is accepted. Thus the proof of $H_1$ is by contradiction of $H_0$.

• U.S. justice system analogy:
  – Was O.J Simpson innocent?
Which is the Null Hypothesis?

Example 6.10  (Acceptance Sampling: Formulation of Hypotheses)

An electrical parts manufacturer receives a large lot of fuses from a vendor. The lot is regarded as “satisfactory” if the fraction defective $p$ is no more than 1%; otherwise it is regarded as “unsatisfactory.” Since it is not feasible to inspect all the fuses in the lot, $p$ is an unknown parameter. Therefore sampling inspection is done to decide the fate (accept/reject) of the lot. There are two hypotheses: (i) the lot is “satisfactory” ($p \leq 1\%$) and (ii) the lot is “unsatisfactory” ($p > 1\%$). If the vendor has an established quality record, then he is given the benefit of doubt, and the hypotheses are set up as $H_0: \ p \leq 1\%$ vs. $H_1: \ p > 1\%$. On the other hand, if the vendor is new, then the burden of proof is on him to show that the quality standard is met. Therefore the hypotheses are set up as $H_0: \ p \geq 1\%$ vs. $H_1: \ p < 1\%$.

The null hypothesis is what we choose to believe unless there is overwhelming evidence against it, in which case we accept the alternative hypothesis. Notice, that a null hypothesis that is not rejected is not accepted. We simply do not have sufficient evidence to reject it. Lack of evidence is not proof.
Which is the Null Hypothesis?

Example 6.11 (SAT Coaching: Formulation of Hypotheses)

There is considerable debate about the effectiveness of various coaching programs in improving the SAT scores. Powers\(^4\) reviewed the literature on this subject. Here we consider only the verbal part (SAT-V) of the test. Based on the 1990–91 test data, it is known that the changes in the test scores for the general population of high school juniors and seniors retaking the test without any special coaching have a mean \(\mu\) of about 15 points and a standard deviation (SD) \(\sigma\) of about 40 points.\(^5\) The changes in the test scores are approximately normally distributed. Thus 15 points represent the average improvement due to the learning effect and the natural increase in the verbal ability that occurs over time. If a coaching program does not improve the test scores by more than 15 points on average, then we say that the coaching program has no effect.

Suppose that an SAT coaching company claims that its coaching program will improve the test scores, on the average, by more than 15 points, i.e., the coaching program is effective. (The advertised claim may be much stronger, e.g., improvement by at least 40 points.) Let \(\mu\) denote the mean change in the test scores for the population of students who are coached. Since the company’s claim requires proof, the hypotheses are set up as \(H_0: \mu = 15\) (the coaching program has no effect) vs. \(H_1: \mu > 15\).

At this point one may ask, why not set up \(H_0\) as \(\mu \leq 15\) rather than as \(\mu = 15\), because it is possible that the coaching program may actually reduce the mean score? We will see later in this chapter that we can replace \(H_0: \mu = 15\) by \(H_0: \mu \leq 15\) without any loss. Until then we will continue to use the formulation \(H_0: \mu = 15\).
Hypothesis Tests (Neyman Pearson Theory)

- A **hypothesis test** is a data-based rule to decide between $H_0$ and $H_1$
- A **test statistic** calculated from the data is used to make this decision
- The values of the test statistics for which the test rejects $H_0$ comprises the **rejection region** of the test
- The complement of the rejection region is called the **acceptance region**. (Does this name bother you?)
- The boundary of the rejection region are defined by one or more **critical constants**

Refer to Example 6.10. Suppose that the hypotheses are set up as $H_0: p \leq 1\%$ (the lot is satisfactory) vs. $H_1: p > 1\%$ (the lot is not satisfactory), where $p$ is the unknown fraction defective in the lot. A sample of 100 items is inspected. The decision rule is to accept the lot if the number of defectives in the sample is 0 or 1, otherwise reject the lot. The number of defectives in the sample, denoted by $x$, is the test statistic. The rejection region of the test is $x > 1$. ♦
Another Example and Foundational Issue

**Example 6.14 (SAT Coaching: A Hypothesis Test)**

Refer to Example 6.11. Consider conducting the following experiment: A sample of 20 students take the SAT before and after participating in the coaching program.\(^6\) Let \(\bar{x}\) be the sample mean of changes in the test scores of the 20 students and let \(\mu\) be the corresponding unknown population mean. The hypothesis test is: Reject \(H_0: \mu = 15\) in favor of \(H_1: \mu > 15\) if \(\bar{x} > 25\). The test statistic is \(\bar{x}\) and the rejection region is \(\bar{x} > 25\).

Historically, the formulation of a hypothesis test as a **two-decision problem** is due to two great statisticians, Jerzy Neyman (1894–1981) and Egon Pearson\(^7\) (1895–1980). However, many statisticians prefer to think of a hypothesis test as a method to weigh evidence in the data against \(H_0\) rather than as a decision procedure. Typically, \(H_0\) is a hypothesis of no difference (hence the name “null” hypothesis) between a new method and an existing method or no effect of an intervention under study (e.g., \(H_0: \mu = 15\) in Example 6.11 represents no improvement due to coaching, or \(H_0: p_1 = p_2\) in Example 6.12 represents no effect of vaccine). The truth of \(H_0\) is never seriously entertained; it is set up simply as a “straw man.” Therefore \(H_0\) is never formally accepted. The purpose of the test is to determine if the observed data can be reasonably explained by chance alone, assuming that \(H_0\) is true. If not, then we have a **statistically significant** proof of the research claim (\(H_1\)). When a hypothesis test is used in this fashion, it is called a **significance test**.
Type I and Type II Error Probabilities

When a hypothesis test is viewed as a decision procedure, two types of error are possible:

<table>
<thead>
<tr>
<th>Decision</th>
<th>Do not reject $H_0$</th>
<th>Reject $H_0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H_0$ True</td>
<td>Correct Decision</td>
<td>Type I Error</td>
</tr>
<tr>
<td>$H_0$ False</td>
<td>Type II Error</td>
<td>Correct Decision</td>
</tr>
</tbody>
</table>

Since we do not know in fact whether $H_0$ is true or $H_1$ is true, we cannot tell whether a test has made a correct decision or not. If a test rejects $H_0$, we do not know whether the decision is correct or a type I error is made. If a test fails to reject $H_0$, we do not know whether the decision is correct or a type II error is made. We can only assess the long term accuracy of decisions made by a test by adopting the same frequentist approach that we used to interpret the confidence intervals.
Probabilities of Type I and II Errors

Consider a long series of trials in which the same decision rule is applied to repeated samples drawn under identical conditions (either under $H_0$ or under $H_1$). The limiting proportion of trials in which $H_0$ is rejected when $H_0$ is true is the probability of a type I error, also called the $\alpha$-risk. Similarly, the limiting proportion of trials in which $H_0$ is not rejected when $H_1$ is true is the probability of a type II error, also called the $\beta$-risk. This frequentist interpretation of error probabilities leads to the following definitions:

$$\alpha = P\{\text{Type I error}\} = P\{\text{Reject } H_0 \text{ when } H_0 \text{ is true}\}$$

$$= P\{\text{Reject } H_0 \mid H_0\}$$

(6.8)

and

$$\beta = P\{\text{Type II error}\} = P\{\text{Fail to reject } H_0 \text{ when } H_1 \text{ is true}\}$$

$$= P\{\text{Fail to reject } H_0 \mid H_1\}.$$ \hspace{1cm} (6.9)

$$= P\{\text{“accept } H_0\text{” when } H_1 \text{ is true}\}$$

Instead of $\beta$ it is common to work with the power of the test:

$$\pi = 1 - \beta = P\{\text{Reject } H_0 \mid H_1\}.$$ \hspace{1cm} (6.10)

Measure of the ability of the test to “prove” $H_1$ when $H_1$ is true.
Acceptance Sampling Example

Suppose that the long run average defective rate for the lots supplied by a vendor is 1% and we are testing

\[ H_0: p = .01 \text{ vs. } H_1: p > .01 \]

where \( p \) is the unknown fraction defective of the current lot.

The decision rule is: Do not reject \( H_0 \) (accept the lot) if the number defective \( x \) in a random sample of 100 is 0 or 1

Note that the number defective has a \( Bin(100, .01) \) distribution.
\[ \alpha = P(\text{Type I Error}) = P(\text{Reject } H_0 \mid H_0: p = .01) \]

\[ = 1 - P(X = 0 \text{ or } 1 \mid H_0: p = .01) \]

\[ = 1 - \sum_{i=0}^{1} \binom{100}{i} (.01)^i (.99)^{100-i} = 0.264. \]

In lot acceptance sampling \( \alpha \) is called the \textbf{producer's risk}
\[ \beta = P(\text{Type II Error}) \]

**Acceptance Sampling Example**

Next suppose that if the fraction defective in the lot is as high as .03, then the lot is regarded as clearly unsatisfactory. So it is of interest to calculate the \( \beta \)-risk for \( p = .03 \):

\[
\beta = P\{\text{Type II error}\} = P\{\text{Fail to reject } H_0 \mid H_1: p = .03\}
\]

\[
= P\{X = 0\text{ or } 1 \mid H_1: p = .03\}
\]

\[
= \sum_{i=0}^{1} \binom{100}{i} (.03)^i (.97)^{100-i} = 0.195.
\]

The power is

\[
\pi = 1 - \beta = 1 - 0.195 = 0.805.
\]

In lot acceptance sampling \( \beta \) is called the **consumer’s risk**

Note that there are different values of \( \beta \) and \( \pi \) for different values of \( p > .1 \).
\[ \alpha = P(\text{Type I Error}): \]
SAT Coaching Example

**Example 6.17 (SAT Coaching: \( \alpha \)- and \( \beta \)-Risks)**

Refer to Example 6.14. The hypotheses \( H_0: \mu = 15 \) (coaching program is ineffective) vs. \( H_1: \mu > 15 \) (coaching program improves SAT-V scores) are tested by rejecting \( H_0 \) if \( \bar{x} > 25 \), where \( \bar{x} \) is the mean change in SAT-V scores for a sample of 20 students. Calculate the error probabilities of this decision rule.

Note that the r.v. \( \bar{X} \) is normally distributed with mean = \( \mu \) and SD = \( \sigma / \sqrt{n} = 40 / \sqrt{20} \). Therefore, \( \frac{\bar{x} - \mu}{40 / \sqrt{20}} = Z \sim N(0, 1) \). Using this fact, we can write

\[
\alpha = P\{\text{Type I error}\} = P\{\bar{X} > 25 \mid \mu = 15\}
\]

\[
= P\left\{ Z = \frac{\bar{X} - 15}{40 / \sqrt{20}} > \frac{25 - 15}{40 / \sqrt{20}} \mid \mu = 15 \right\}
\]

\[
= P\{Z > 1.118\} = 1 - \Phi(1.118) = 0.132.
\]

Thus there is a 13\% chance that this test will conclude that the coaching program is effective when in fact it is not.
\[ \beta = P(\text{Type II Error}): \]
SAT Coaching Example

Next suppose that the coaching company claims that the actual mean improvement is at least 40 points. The \( \beta \)-risk for \( \mu = 40 \) is the probability that the test will fail to prove this claim even if it is in fact true. It is calculated as follows:

\[
\beta = P\{\text{Type II error}\} = P\{\bar{X} \leq 25 \mid \mu = 40\} \\
= P\left\{ Z = \frac{\bar{X} - 40}{40/\sqrt{20}} \leq \frac{25 - 40}{40/\sqrt{20}} \mid \mu = 40 \right\} (\sigma = 40) \\
= P\{Z \leq -1.677\} = 0.047.
\]

The power, which is the probability that the claim will be proved when \( \mu = 40 \), equals \( \pi = 1 - \beta = 1 - 0.047 = 0.953 \).
Operating Characteristic and Power Functions

6.3.4 *Operating Characteristic and Power Functions

When $H_0$ and $H_1$ are composite, the two error probabilities are not just single numbers but are functions of the test parameter. For instance, in Example 6.16 they are functions of $p$ and in Example 6.17 they are functions of $\mu$. Instead of dealing with these functions separately, we can combine them into a single function called the operating characteristic (OC) function of the test. In general, if $\theta$ denotes the test parameter, then the OC function is the probability of failing to reject $H_0$ as a function of $\theta$:

$$OC(\theta) = P\{\text{Test fails to reject } H_0 \mid \theta\}. \quad (6.11)$$

Note that for $\theta$ values under $H_1$ the OC function is simply the $\beta$-risk. The power function is

$$\pi(\theta) = P\{\text{Test rejects } H_0 \mid \theta\} = 1 - OC(\theta). \quad (6.12)$$

The following examples show how the OC function can be used to compute the error probabilities.
Acceptance Sampling:
OC Function

Probability of type I error is smaller for $H_0 : p = p_0$
with $p_0 < .01$. So one can write $H_0 : p \leq .01$ with maximum
type I error probability of .264. This maximum is called
the $\alpha$-level of the test.

Figure 6.5  OC Curve for the Test in the Acceptance Sampling Example
Type I error probabilities

$\alpha$-level of the test for $H_0 : \mu \leq 15$ is 0.132

SAT Coaching: OC Function

$\alpha = 0.132$

$\beta = 0.047$

$H_0$ true

$\mu = \text{mean change in SAT-V score}$

Figure 6.6 OC Curve for the Test in the SAT Coaching Example

Probability of failing to reject $H_0$
Simple and Composite Hypothesis
(What we explained in the previous two pictures)

The above two examples show how to calculate the error probabilities of a test when exactly one parameter value (i.e., one probability distribution) is specified under each hypothesis. Such a hypothesis is called a simple hypothesis, e.g., $H_0: p = .01$ and $H_1: p = .03$ or $H_0: \mu = 15$ and $H_1: \mu = 40$. If a hypothesis specifies more than one probability distribution, then it is called a composite hypothesis, e.g., $H_1: p > .01$ or $H_1: \mu > 15$.

We next show how to calculate the error probabilities as functions of the test parameters for composite hypotheses. These functions are monotone (i.e., increasing or decreasing) in the test parameters for many standard testing problems. This enables us to specify $H_0$ more naturally as a composite hypothesis rather than as a simple hypothesis (as we have done thus far for mathematical convenience), e.g., $H_0: p \leq .01$ instead of $H_0: p = .01$ or $H_0: \mu \leq 15$ instead of $H_0: \mu = 15$. The reason is that the maximum type I error probability of the tests we use occurs at the boundary parameter value of $H_0$; see the discussion in the following two sections. Since we are concerned with controlling this maximum type I error probability, we may as well assume from now on, that both $H_0$ and $H_1$ are specified as composite hypotheses.
Level of Significance

- Practice of test of hypothesis is to put an upper bound on the $P$(Type I error) and subject to that constraint find a test with the lowest possible $P$(Type II error)
- Motivated by the fact that the Type I error is usually the more serious
- The least upper bound on $P$(Type I error) is called the level of significance of the test and is denoted by $\alpha$
  (usually some small number such as 0.01, 0.05, or 0.10)

The test is required to satisfy:

$$\Pr\{\text{Type I error}\} = \Pr\{\text{Test Rejects } H_0 \mid H_0\} \leq \alpha$$

- Note that $\alpha$ is now used to denote an upper bound on $P$(Type I error)
- A hypothesis test with a significance level $\alpha$ is called an $\alpha$-level test
Acceptance Sampling Example

Example 6.20 (Acceptance Sampling: 0.10-Level Test)

Suppose that we want a 0.10-level test of $H_0: p \leq .01$ vs. $H_1: p > .01$. We saw in Example 6.16 that if $x$ denotes the number of defectives in a sample of 100 fuses, then the test that rejects $H_0$ when $x > 1$ has $P\{\text{Type I error}\} = 0.264$, which is too high. If we change the rejection region to $x > 2$, we get

$$P\{\text{Type I error}\} = P\{\text{Test rejects } H_0 \mid H_0\} = P\{X > 2 \mid p = .01\}$$

$$= 1 - \sum_{i=0}^{2} \binom{100}{i}(.01)^i(.99)^{100-i} = 0.079,$$

which is less than the upper bound $\alpha = 0.10$. In fact, this test has $\alpha = 0.079$. The $P\{\text{Type II error}\}$ for this test can be computed to be 0.420, which is greater than $P\{\text{Type II error}\} = 0.195$ that we obtained in Example 6.16 for the test with $\alpha = 0.264$. when $p = 0.03$
Example: $\alpha = 0.05$ Test

**Example 6.21 (SAT Coaching: 0.05-Level Test)**

Suppose we want a 0.05-level test of $H_0: \mu \leq 15$ vs. $H_1: \mu > 15$. Consider the test that rejects $H_0$ if $\bar{x} > c$, where the critical value $c$ is to be determined so that the test will have level $\alpha = 0.05$.

Thus we need to solve the equation

$$P\{\text{Type I error}\} = P\{\bar{X} > c \mid \mu = 15\} = 0.05.$$

The solution $c$ can be obtained by standardizing $\bar{X}$. This results in the equation

$$P \left\{ Z = \frac{\bar{X} - 15}{40/\sqrt{20}} > \frac{c - 15}{40/\sqrt{20}} \right\} = 0.05 \quad (\sigma = 40)$$

which is satisfied when

$$\frac{c - 15}{40/\sqrt{20}} = z_{0.05} = 1.645 \quad \text{or} \quad c = 15 + 1.645 \times \frac{40}{\sqrt{20}} = 29.71.$$

Thus a 0.05-level test of $H_0: \mu \leq 15$ vs. $H_1: \mu > 15$ rejects $H_0$ when $\bar{x} > 29.71$.

If an actual experiment resulted in a mean improvement of $\bar{x} = 35$, then that result would be significant at $\alpha = 0.05$ and the coaching program would be declared to be effective. •
Alternative Version of the Test

The above test can be alternatively expressed in terms of the standardized test statistic

\[ z = \frac{\bar{x} - 15}{\sigma / \sqrt{n}} = \frac{\bar{x} - 15}{40 / \sqrt{20}} \]

as follows:

Reject \( H_0 \) if \( z > z_{.05} = 1.645 \).

In the above example

\[ z = \frac{35 - 15}{40 / \sqrt{20}} = 2.236 > 1.645, \]

so \( H_0 \) is rejected at \( \alpha = .05 \).
General Version of the Test

More generally, an $\alpha$-level test of

$$H_0: \mu \leq \mu_0 \text{ vs. } H_1: \mu > \mu_0$$  \hspace{1cm} (6.14)

where $\mu_0$ is a specified number ($\mu_0 = 15$ in the above example) is:

Reject $H_0$ if $z > z_\alpha$ or equivalently if $\bar{x} > \mu_0 + z_\alpha \frac{\sigma}{\sqrt{n}}$  \hspace{1cm} (6.15)

where $z$ is the standardized test statistic,

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}.$$  \hspace{1cm} (6.16)

Here we are able to obtain an exact test for any level $\alpha$, because the test statistic $\bar{X}$ (or equivalently $Z$) has a continuous distribution.
What $\alpha$ Level Should One Use?

• Fisher said: $\alpha = 0.05$

• Recall that as $P(\text{Type I error})$ decreases $P(\text{Type II error})$ increases. A proper choice of $\alpha$ should take into account the relative cost of type I and type II errors. However, these cost are difficult to determine in practice.

• Today $\alpha = .10, .05, .01, .001$ are commonly used depending on how much proof against the null hypothesis we want to have before rejecting it.

<table>
<thead>
<tr>
<th>Rejection $\alpha$</th>
<th>Common Terminology</th>
</tr>
</thead>
<tbody>
<tr>
<td>.10</td>
<td>There is some statistical evidence against the null hypothesis</td>
</tr>
<tr>
<td>.05</td>
<td>There is statistical evidence against the null hypnosis</td>
</tr>
<tr>
<td>.01</td>
<td>There is strong statistical evidence against the null hypothesis</td>
</tr>
<tr>
<td>.001</td>
<td>There is extremely strong statistical evidence against the null hypothesis</td>
</tr>
</tbody>
</table>
Observed Level of Significance or P-value

Simply rejecting or not rejecting $H_0$ at a specified $\alpha$ level does not fully convey the information in the data.

Example: $H_0 : \mu = 15$ vs. $H_1 : \mu > 15$ is rejected at the $\alpha = 0.05$ when

$$\bar{x} > 15 + 1.645 \times \frac{40}{\sqrt{20}} = 29.71$$

Is a sample with a mean of 30 equivalent to a sample with a mean of 50? (Note that both lead to rejection at the $\alpha$-level of 0.05)

More useful to report the smallest $\alpha$-level for which the data would reject (called the observed level of significance or P-value).

- Reject $H_0$ at the $\alpha$-level if P-value $< \alpha$
Example 6.22 (Acceptance Sampling: P-Value)

\( H_0 : p \leq 0.01 \) vs. \( H_1 : p > 0.01 \)

Test based on random sample of 100 fuses which has 1 defective

\[
P\text{-value} = P\{ X \geq 1 \mid p = .01 \} \]
\[
= 1 - P\{ X = 0 \mid p = .01 \} 
\]
\[
= 1 - (0.99)^{100} 
\]
\[
= 0.634 
\]

Since the P-value > 0.10, \( H_0 \) is not reject at \( \alpha = 0.10 \) or at any other \( \alpha \)-level smaller than 0.634

An outcome of \( x = 1 \) or more is probable under \( H_0 \)
Example 6.23 (SAT Coaching: P-value)

\( H_0 : \mu \leq 15 \) vs. \( H_1 : \mu > 15 \)

Random sample of \( n = 20 \) students has \( \bar{x} = 35 \)

\[ P\text{-value} = P\{ \bar{X} \geq 35 \mid \mu = 15 \} \]

\[ = P \left\{ Z = \frac{\bar{X} - 15}{40/\sqrt{20}} \geq z = \frac{35 - 15}{40/\sqrt{20}} \mid \mu = 15 \right\} \]

\[ = P\{Z \geq 2.236\} \]

\[ = 1 - \Phi(2.236) \]

\[ = 0.013. \]

\( \bar{x} = 35 \) is possible but not probable under \( H_0 \)

Reject at \( \alpha\)-level = 0.05 but not at \( \alpha\)-level = 0.01
One-Sided and Two-Sided Tests

$H_0 : \mu = 15$ can have three possible alternative hypotheses:

$H_1 : \mu > 15$, $H_1 : \mu < 15$, or $H_1 : \mu \neq 15$

(upper one-sided) (lower one-sided) (two-sided test)

Example 6.27 (SAT Coaching: $\alpha = 0.05$ Two-Sided Test)

$H_0 : \mu = 15 \text{ vs. } H_1 : \mu \neq 15$

Reject if

$\left| \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} \right| = \left| \frac{\bar{x} - 15}{40 / \sqrt{20}} \right| > z_{0.025} = 1.96$

or if $\bar{x} < -2.53$ or $\bar{x} > 32.53$
Example 6.27 Continued
(SAT Coaching: Two-sided Test P-value)

Suppose $\bar{x} = 35$

$$P\text{-value} = P\left\{|Z| > \frac{35-15}{40/\sqrt{20}} = 2.236\right\} = 0.026$$

Recall that in Example 6.23 one-sided P-value = 0.013

The two-sided P-value is twice the one-sided P-value

- In general two-sided alternatives should be used when the deviation in either direction is worth detecting.
- A research hypothesis $H_1$ has more support when a two-sided test rejects than when a one-sided test rejects.
JMP Example

![JMP Table]

<table>
<thead>
<tr>
<th>Problem4</th>
<th>Rainfall</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1468</td>
</tr>
<tr>
<td>2</td>
<td>909</td>
</tr>
<tr>
<td>3</td>
<td>841</td>
</tr>
<tr>
<td>4</td>
<td>475</td>
</tr>
<tr>
<td>5</td>
<td>846</td>
</tr>
<tr>
<td>6</td>
<td>452</td>
</tr>
<tr>
<td>7</td>
<td>3830</td>
</tr>
<tr>
<td>8</td>
<td>1397</td>
</tr>
<tr>
<td>9</td>
<td>556</td>
</tr>
<tr>
<td>10</td>
<td>978</td>
</tr>
<tr>
<td>11</td>
<td>1715</td>
</tr>
<tr>
<td>12</td>
<td>747</td>
</tr>
<tr>
<td>13</td>
<td>909</td>
</tr>
<tr>
<td>14</td>
<td>2002</td>
</tr>
<tr>
<td>15</td>
<td>1331</td>
</tr>
<tr>
<td>16</td>
<td>1227</td>
</tr>
</tbody>
</table>
JMP Example: Test of Hypothesis for the Mean

Go to **Distribution of Y** in the Analyze menu
JMP Example: Test of Hypothesis for the Mean

Hypothesized Value: 1000
Actual Estimate: 1369.11
df: 46
Std Dev: 693.67
Sigma given: 700

Test Statistic: 3.6150
Prob > |z|: 0.0003
Prob > z: 0.0002
Prob < z: 0.9998

H₁: μ ≠ 1000
H₁: μ ≥ 1000
H₁: μ ≤ 1000
JMP Example: Confidence Interval for the Mean

Unlike the test on the previous page this confidence interval does not assume a known standard deviation of 700 but estimates this from the data. As far as I know, JMP does not allow one to use a known standard deviation in confidence interval calculations.
Relationship Between Confidence Interval and Hypothesis Test

A $\alpha$-level two-sided test reject a null hypothesis $H_0: \mu = \mu_0$ if and only if the $(1 - \alpha)100\%$ confidence interval does not contain $\mu_0$

Example 6.7 (Airline Revenues)
Suppose a 95% confidence interval for $\mu$ is $170.70 \leq \mu \leq 180.50$
Would we reject the hypothesis $H_0 : \mu = 200$ vs. $H_0 : \mu \neq 200$ at $\alpha = .05$ level?

Would we reject the hypothesis $H_0 : \mu = 175$ vs. $H_0 : \mu \neq 175$ at $\alpha = .05$ level?
Use an Misuse of Hypothesis Test in Practice

- Difficulties of Interpreting Test on Nonrandom Samples and Observational Data
  - Valid hypothesis tests in comparative studies require that the experimental units be randomly assigned to the experimental groups
  - In observational studies that do not use randomization, calculated P-values and confidence levels are tenuous and may at best be taken as rough indicators of statistical confidence or significance
  - In observational studies one can not draw cause-effect conclusions even when we reject a null hypothesis. Remember confounding factors.
  - In pretest/posttest studies one needs a parallel control group
Use an Misuse of Hypothesis Test in Practice

• Statistical Significance vs. Practical Significance
  – **Statistical significance is a function of the sample size.** With a very large sample size, even a small, practically unimportant difference can be shown to be statistically significant
  – On the other hand with a small sample size, a test may lack the power for even a large, practically important, difference to be statistically significant
  – In practice, an estimate of the difference between two alternatives is usually much more useful than a test for its statistical significance. **Confidence intervals are preferred to hypothesis tests.**
Use an Misuse of Hypothesis Test in Practice

• Perils of searching for Significance
  – Significance at the 5% level has become almost mandatory for publication of research findings in many applied fields
  – If enough tests are done then just by chance some will be significant (with $\alpha = .05$, 5% of then will be significant)
  – Many variables are measured in many studies and test are done in all of them. Only significant differences are reported. This is a poor practice. Problem of multiple comparisons.
  – Bonferroni method: To have a simultaneous probability of a Type I error of $\alpha$ assign to each of $k$ independent test an alpha level of $\alpha/k$. 
Use an Misuse of Hypothesis Test in Practice

• Ignoring Lack of Significance
  – Nonsignificant results go mostly unreported, although they may be equally important
  – If an experiment is designed to have sufficient power to detect a specified difference, but does not detect it, that is an important finding and should be reported
  – For instance, in an epidemiological study it is important to know that a suspected risk factor is not associated with a disease
Use an Misuse of Hypothesis Test in Practice

Scientific Research and Hypothesis Testing. In scientific research, a single experiment generally does not make or break a theory. Therefore the purpose of a hypothesis test in such experiments is not to make a decision, but to quantify the evidence in support of a theory. Usually, what makes or breaks a theory is the reproducibility of results from independently replicated experiments by several investigators. There are situations, however, where a single scientific experiment is used to make a critical decision. Some examples are: (i) the experiment is prohibitively costly to repeat, (ii) the results of the experiment overwhelmingly support (or refute) a theory, and (iii) the problem that the experiment addresses requires immediate solution. The Salk Polio vaccine trial had all three elements present, and resulted in an enormously important health policy decision of vaccinating all young children. The drug AZT for AIDS was approved by the Food and Drug Administration of U.S. in 1987 based on a rather small trial before the trial was even completed.
Meta Analysis

A seminal paper on meta analysis by Lau et al, published in the New England Journal of Medicine (1992) dealt with the use of Thrombolytics for treatment of myocardial infarction. Thrombolytics are the so-called clot busters that are administered immediately subsequent to a heart attack, in the hope that by dissolving the clot, the drug will limit the damage to the heart. In the years between 1959 and 1988 there were a total of 31 studies that looked at the impact of these drugs, all using essentially the same method: Patients were assigned at random to receive either the drug or a placebo, and the clinician recorded whether or not the patient died. Lau et al conducted a meta analysis using the data from the 31 studies and reported the drug substantially increases the chances of survival. The odds ratio was .77 (.72, .82) with a p-value < .00001. As a result, the use of these drugs today is considered standard treatment for a heart attack patient throughout the developed world.
Meta Analysis Forrest Plot

6/20/2003