Chapter 15
Inference in Practice

z Procedures

- If we know the standard deviation $\sigma$ of the population, a confidence interval for the mean $\mu$ is:

$$\bar{x} \pm z \frac{\sigma}{\sqrt{n}}$$

- To test a hypothesis $H_0: \mu = \mu_0$ we use the one-sample $z$ statistic:

$$z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$$

- These are called $z$ procedures because they both involve a one-sample $z$ statistic and use the standard Normal distribution.

Where Did the Data Come From?

- When you use statistical inference, you are acting as if your data are a probability sample or come from a randomized experiment.

- Statistical confidence intervals and tests cannot remedy basic flaws in producing data, such as voluntary response samples or uncontrolled experiments.

- If the data do not come from a probability sample or a randomized experiment, the conclusions may be open to challenge. To answer the challenge, ask whether the data can be trusted as a basis for the conclusions of the study.
Case Study
Mammary Artery Ligation


Surgeons tested a procedure to alleviate pain caused by inadequate blood supply to the heart, and the patients reported a statistically significant reduction in angina pain.

Statistical significance indicates that something other than chance is at work, but it does not say what that something is. Since this experiment was not controlled, the reduction in pain could be due to the placebo effect. A controlled experiment showed that this was the case, and surgeons immediately stopped performing the operation.

Cautions About the $z$ Procedures

- The data must be an SRS from the population.
  - Different methods are needed for different designs.
  - The $z$ procedures are not correct for samples other than SRS.
- Outliers can distort the result.
  - The sample mean is strongly influenced by outliers.
  - Always explore your data before performing an analysis.
- The shape of the population distribution matters.
  - Skewness and outliers make the $z$ procedures untrustworthy unless the sample is large.
  - In practice, the $z$ procedures are reasonably accurate for any sample of at least moderate size from a fairly symmetric distribution.
- The population standard deviation $\sigma$ must be known.
  - Unfortunately $\sigma$ is rarely known, so $z$ procedures are rarely useful.
  - Chapter 16 will introduce procedures for when $\sigma$ is unknown.
Cautions About Confidence Intervals

**The margin of error does not cover all errors.**

- The margin of error in a confidence interval covers only random sampling errors. No other source of variation or bias in the sample data influence the sampling distribution.
- Practical difficulties such as undercoverage and nonresponse are often more serious than random sampling error. The margin of error does not take such difficulties into account.

*Be aware of these points when reading any study results.*

Cautions About Significance Tests

**How small a P-value is convincing?**

- If \( H_0 \) represents an assumption that people have believed in for years, strong evidence (small \( P \)-value) will be needed to persuade them otherwise.
- If the consequences of rejecting \( H_0 \) are great (such as making an expensive or difficult change from one procedure or type of product to another), then strong evidence as to the benefits of the change will be required.

*Although \( \alpha = 0.05 \) is a common cut-off for the P-value, there is no set border between "significant" and "insignificant," only increasingly strong evidence against \( H_0 \) (in favor of \( H_a \)) as the P-value gets smaller.*

Cautions About Significance Tests

**Statistical Significance & Practical Significance**

- When the sample size is very large, tiny deviations from the null hypothesis (with little practical consequence) will be statistically significant.
- When the sample size is very small, large deviations from the null hypothesis (of great practical importance) might go undetected (statistically insignificant).

*Statistical significance is not the same thing as practical significance.*
Case Study: Drug Use in American High Schools

Alcohol Use


---

Case Study Alcohol Use

- **Alternative Hypothesis:** The percentage of high school students who used alcohol in 1993 is less than the percentage who used alcohol in 1992.
- **Null Hypothesis:** There is no difference in the percentage of high school students who used in 1993 and in 1992.

---

Case Study Alcohol Use

1993 survey was based on 17,000 seniors, 15,500 10th graders, and 18,500 8th graders.

<table>
<thead>
<tr>
<th>Grade</th>
<th>1992</th>
<th>1993</th>
<th>Diff</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>8th</td>
<td>53.7</td>
<td>51.6</td>
<td>-2.1</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>10th</td>
<td>70.2</td>
<td>69.3</td>
<td>-0.9</td>
<td>.04</td>
</tr>
<tr>
<td>12th</td>
<td>76.8</td>
<td>76.0</td>
<td>-0.8</td>
<td>.04</td>
</tr>
</tbody>
</table>
Case Study: Memory Loss in American Hearing, American Deaf, and Chinese Adults

Memory Loss


Case Study

Average Memory Test Scores (higher is better)

30 subjects were sampled from each population

<table>
<thead>
<tr>
<th></th>
<th>Hearing</th>
<th>Deaf</th>
<th>Chinese</th>
</tr>
</thead>
<tbody>
<tr>
<td>Young</td>
<td>1.69</td>
<td>0.98</td>
<td>1.34</td>
</tr>
<tr>
<td>Old</td>
<td>-2.97</td>
<td>-1.55</td>
<td>0.50</td>
</tr>
</tbody>
</table>
Case Study
Memory Loss

◆ Young Americans (hearing and deaf) have significantly higher mean scores.
◆ Science News (July 2, 1994, p. 13): “Surprisingly, ...memory scores for older and younger Chinese did not statistically differ.”

Since the sample sizes are very small, there is an increased chance that the test will result in no statistically significance difference being detected even if indeed there is a difference between young and old subjects’ mean memory scores.

The “surprising” result could just be because the sample size was too small to statistically detect a difference. A larger sample may yield different results.

Cautions About Significance Tests

Beware of Multiple Analyses

◆ Suppose that 20 null hypotheses are true.
◆ Each test has a 5% chance of being significant at the 5% level. That’s what \( \alpha = 0.05 \) means: results this extreme occur only 5% of the time just by chance when the null hypothesis is true.
◆ Thus, we expect about 1 in 20 tests (which is 5%) to give a significant result just by chance.
◆ Running one test and reaching the \( \alpha = 0.05 \) level is reasonably good evidence against \( H_0 \); running 20 tests and reaching that level only once is not.

Similarly, the probability that all of twenty 95% confidence intervals will capture their true mean is much less than 95%.
The Power of a Test

- The probability that a fixed level $\alpha$ significance test will reject $H_0$ when a particular alternative value of the parameter is true is called the **power** of the test against that specific alternative value.
- While $\alpha$ gives the probability of wrongly rejecting $H_0$ when in fact $H_0$ is true, power gives the probability of correctly rejecting $H_0$ when in fact $H_0$ should be rejected (because the value of the parameter is some specific value satisfying the alternative hypothesis).
- When $\mu$ is close to $\mu_0$, the test will find it hard to distinguish between the two (low power); however, when $\mu$ is far from $\mu_0$, the test will find it easier to find a difference (high power).

Case Study

**Sweetening Colas (Ch. 14)**

- The cola maker determines that a sweetness loss is too large to be acceptable if the mean response for all tasters is $\mu = 1.1$ (or larger).
- Will a 5% significance test of the hypotheses
  
  $H_0: \mu = 0$  
  $H_a: \mu > 0$

  based on a sample of 10 tasters usually detect a change this great (rejecting $H_0$)?

Case Study

**Sweetening Colas**

1. Write the rule for rejecting $H_0$ in terms of $X$.
   
   We know that $\sigma = 1$, so the $z$ test rejects $H_0$ at the $\alpha = 0.05$ level when
   
   $$z = \frac{X - 0}{1/\sqrt{10}} \geq 1.645$$

   This is the same as:

   Reject $H_0$ when $X \geq 0 + 1.645 \cdot \frac{1}{\sqrt{10}} = 0.520$

   This step just restates the rule for the test. It pays no attention to the specific alternative we have in mind.
Case Study
Sweetening Colas

2. The power is the probability of rejecting $H_0$ under the condition that the alternative $\mu = 1.1$ is true.

To calculate this probability, standardize $\bar{x}$ using $\mu = 1.1$:

$$P(\bar{x} \geq 0.520 \text{ when } \mu = 1.1) = P \left( \frac{\bar{x} - 1.1}{\sqrt{1/10}} \geq \frac{0.520 - 1.1}{\sqrt{1/10}} \right)$$

$$= P(Z \geq -1.83) = 1 - 0.0336 = 0.9664$$

96.64% of tests will declare that the cola loses sweetness when the true mean sweetness loss is 1.1 (power = 0.9664).

Decision Errors: Type I

◆ If we reject $H_0$ when in fact $H_0$ is true, this is a **Type I error**.

◆ If we decide there is a significant relationship in the population (reject the null hypothesis):
  – This is an incorrect decision only if $H_0$ is true.
  – The probability of this incorrect decision is equal to $\alpha$.

◆ If the null hypothesis is true and $\alpha = 0.05$:
  – There really is no relationship and the extremity of the test statistic is due to chance.
  – About 5% of all samples from this population will lead us to wrongly reject chance and conclude significance.
Decision Errors: Type II

- If we fail to reject $H_0$ when in fact $H_a$ is true, this is a **Type II error**.
- If we decide not to reject chance and thus allow for the plausibility of the null hypothesis
  - This is an incorrect decision only if $H_a$ is true.
  - The probability of this incorrect decision is computed as 1 minus the power of the test.

Decision Errors: Type I & Type II

<table>
<thead>
<tr>
<th>Truth about the population</th>
<th>$H_0$ true</th>
<th>$H_a$ true</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reject $H_0$</td>
<td>Type I error</td>
<td>Correct decision</td>
</tr>
<tr>
<td>Accept $H_0$</td>
<td>Correct decision</td>
<td>Type II error</td>
</tr>
</tbody>
</table>