Introduction to Inference

Tests of Significance

Objectives (IPS Chapter 6.2)

Tests of significance

- The reasoning of significance tests
- Stating hypotheses
- The $P$-value
- Statistical significance
- Tests for a population mean
- Confidence intervals to test hypotheses
How Statistical Inference Works

Statistical inference deals with two types of problems.
- Estimate the value of a population parameter.
- Test hypotheses about given values of population parameters.

Population under study
$\mu = ?, \ p = ?$

1. Take a SRS of size $n$

2. Compute the value of a sample statistic $\bar{x}, \hat{p}$

3. Use Statistical Inference to draw conclusions

Hypothesis Testing

Example
- Cobra Cheese Company buys milk from several suppliers as the essential raw material for its cheese. Cobra suspects that some producers are adding water to their milk to increase their profit.

Excess water can be detected by determining the freezing point of the milk. The freezing temperature of natural milk varies normally, with a mean $\mu = -0.545$ degrees Celsius, and a standard deviation of $\sigma = 0.008$ degrees Celsius. Added water raises the freezing temperature toward 0, the freezing point of water.

Cobra’s laboratory manager measures the freezing temperature of five consecutive lots of milk from one producer. The mean measurement is $-0.538$.

Is this good evidence that the producer is adding water to the milk?
Two Hypotheses to Choose From…

Null Hypothesis (Ho):
- The producer is not adding water to the milk

Alternative Hypothesis (Ha):
- The producer is adding water to the milk

The Benefit of the Doubt…
Assumed Innocent Till Proven Guilty!

Null Hypothesis (Ho):
- The producer is not adding water to the milk
Burden of Proof Falls on Us…

**Null Hypothesis (Ho):**
- The producer *is not adding water* to the milk

### The Evidence Against Ho

<table>
<thead>
<tr>
<th>x1</th>
<th>x2</th>
<th>x3</th>
<th>x4</th>
<th>x5</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.537</td>
<td>-0.538</td>
<td>-0.543</td>
<td>-0.537</td>
<td>-0.536</td>
<td>-0.538</td>
</tr>
</tbody>
</table>

How Significant is a 0.007° Increase in the Freezing Temperature of Milk?

**Null Hypothesis (Ho):**
- The producer *is not adding water* to the milk

\[-0.538 - (-0.545) = 0.007° C\]

<table>
<thead>
<tr>
<th>x1</th>
<th>x2</th>
<th>x3</th>
<th>x4</th>
<th>x5</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.537</td>
<td>-0.538</td>
<td>-0.543</td>
<td>-0.537</td>
<td>-0.536</td>
<td>-0.538</td>
</tr>
</tbody>
</table>
What’s chance got to do with it…

Null Hypothesis (Ho):

- The producer is not adding water to the milk

How rare is it to observe a sample mean with a value of -0.538 or higher if the 5 specimens of milk come from natural milk?

Based on this simulation, 4 out of 100 samples (or 4%) have a sample mean equal to or greater than the one we observed with a difference from the average freezing temperature of natural milk equal to or greater than 0.007°.
Doing the Math – The zTest Statistic…

- Sampling distribution of the sample mean in samples of size 5

\[ z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} = \frac{-0.538 + 0.545}{0.003578} = 1.96 \]

\[ \bar{x} = -0.538 \degree C \]

Doing the Math – The P-value…

- \[ P[\bar{x} \geq -0.538] = P[Z \geq 1.96] = ? \]

\[ P[Z > 1.96] = 0.0250 \text{ Or 2.5\%} \]

This probability is called the P-Value of the test.
P-Value & Evidence Against $H_0$

<table>
<thead>
<tr>
<th>P-Value</th>
<th>Evidence against $H_0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Greater than 10%</td>
<td>Very weak to none</td>
</tr>
<tr>
<td>Between 10% - 5%</td>
<td>Weak</td>
</tr>
<tr>
<td>Between 5% - 1%</td>
<td>Strong</td>
</tr>
<tr>
<td>Less than 1%</td>
<td>Very strong</td>
</tr>
</tbody>
</table>

- A P-Value of 0.05 or less is typically considered statistically significant.

Tests of Significance – Introduction

- To decide whether the nature of a population has changed or to compare two populations, in a matter that is valid to a researcher, we need to use statistical evidence which prove that any suspected changes in a population, or any differences that seem logical, are not just the results of chance variation or due to anecdotal evidence.

- The statistical procedure used to decide whether a change in the population has indeed occurred or an observed effect is the result of a treatment, is known as hypothesis testing or tests of significance.
The Null Hypothesis ($H_0$)

- In every test of significance, the **null hypothesis** is the hypothesis we must disprove in order to prove our claim.

**Example**
- The average age of a college student in the 1970's was 23 years old. A researcher believes that in the 1980's and 1990's the average age of a student attending college increased due to the fact that more older adults decided to go back to college.

- The hypothesis the researcher must disprove in this case to prove his claim is: $H_0$
  - "The average age of a college student in the 1980’s and 1990’s **is the same** as the average age of a college student in the 1970’s".
  - Stated in statistical terms, $\mu_{80’s & 90’s} = 23$ yrs

The Alternative Hypothesis ($H_a$)

- The **alternative hypothesis** is the claim itself that a change in the population has occurred or that an observed effect is the result of a treatment.

**Example**
- The average age of a college student in the 1970's was 23 years old. A researcher believes that in the 1980's and 1990's the average age of a student attending college increased due to the fact that more older adults decided to go back to college.

- The researcher's claim is: $H_a$
  - "The average age of a college student in the 1980’s and 1990’s **is greater than** the average age of a college student in the 1970’s".
  - Stated in statistical terms, $\mu_{80’s & 90’s} > 23$ yrs
Choosing the Null Hypothesis ($H_0$)

- The null hypothesis is a statement about the status quo. i.e., that no change has occurred in the population, or that a treatment has no effect on a subject.

- In other words, the process of testing a null hypothesis is similar to a jury trial. The assumption is that the “defendant” is innocent (that’s the null hypothesis). To disprove it, the prosecution must establish, beyond any reasonable doubt, that the null hypothesis is false.

- Reasonable doubt means that events (what could be considered evidence against the null hypothesis) can occur by chance alone.

- Eliminating reasonable doubt means to rule out chance. That is, show that observed differences are very unlikely to occur under the assumption that the null hypothesis true.

One-sided and two-sided tests

- A **two-tail or two-sided** test of the population mean has these null and alternative hypotheses:
  
  $H_0 : \mu = [a \text{ specific number}]$  \hspace{1cm}  $H_a : \mu \neq [a \text{ specific number}]$

- A **one-tail or one-sided** test of a population mean has these null and alternative hypotheses:
  
  $H_0 : \mu = [a \text{ specific number}]$  \hspace{1cm}  $H_a : \mu < [a \text{ specific number}]$  \hspace{1cm}  OR  \hspace{1cm}  $H_0 : \mu = [a \text{ specific number}]$  \hspace{1cm}  $H_a : \mu > [a \text{ specific number}]$

- Example

  The FDA tests whether a generic drug has an absorption extent similar to the known absorption extent of the brand-name drug it is copying. Higher or lower absorption would both be problematic, thus we test:

  $H_0 : \mu_{\text{generic}} = \mu_{\text{brand}}$  \hspace{1cm}  $H_a : \mu_{\text{generic}} \neq \mu_{\text{brand}}$  \hspace{1cm}  two-sided
How to choose?

What determines the choice of a one-sided versus a two-sided test is what we know about the problem before we perform a test of statistical significance.

A health advocacy group tests whether the mean nicotine content of a brand of cigarettes is greater than the advertised value of 1.4 mg.

Here, the health advocacy group suspects that cigarette manufacturers sell cigarettes with a nicotine content higher than what they advertise in order to better addict consumers to their products and maintain revenues.

Thus, this is a one-sided test: \( H_0: \mu = 1.4 \text{ mg} \quad H_a: \mu > 1.4 \text{ mg} \)

Reasoning of Hypotheses Tests

Example
- Cobra Cheese Company buys milk from several suppliers as the essential raw material for its cheese. Cobra suspects that some producers are adding water to their milk to increase their profit.
- Excess water can be detected by determining the freezing point of the milk. The freezing temperature of natural milk varies normally, with a mean \( \mu = -0.545 \) degrees Celsius, and a standard deviation of \( \sigma = 0.008 \) degrees Celsius. Added water raises the freezing temperature toward 0, the freezing point of water.
- Cobra’s laboratory manager measures the freezing temperature of five consecutive lots of milk from one producer. The mean measurement is -0.538.

Is this good evidence that the producer is adding water to the milk?
Setting up the test

Formulating the null & alternative hypothesis

- Null hypothesis
  - Milk from this producer has the same freezing temperature as natural milk.
  
  \[ H_0: \mu = -0.545°C \]

- Alternative hypothesis
  - Milk from this producer has a higher freezing temperature than natural milk.
  
  \[ H_a: \mu > -0.545°C \]

Measuring Statistical Significance

- Our decision whether to accept or reject the null hypothesis is based on the statistical significance of the observed value of the sample statistic (sample mean in this case) under the assumption that the null hypothesis is true.

- In this example the sample data is the freezing points of 5 consecutive lots of milk from this producer with a sample mean measurement of -0.538°C.
Which population does the sample come from?

Population Distribution of freezing temperature of milk
- Natural milk
- Water-down milk

Sampling distribution of X-bar for samples of size 5
- Sample mean $\bar{x} = 0.538^\circ$ C
- Standard error $\sigma_{\bar{x}} = 0.004$

Measuring Statistical Significance (P-Value)

- What is the statistical significance of the occurrence of the event \( \{X \text{-bar} \geq -0.538\} \)?

Sampling distribution of X-bar for samples of size 5 (assuming $H_o$ is true)
- Test statistic $Z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \frac{0.538 - (-0.545)}{0.004} = 1.96$
- P-value $P[Z > 1.96] = 0.0250$ or 2.5%
Interpreting a P-value

Could random variation alone account for the difference between the null hypothesis and observations from a random sample?

- A small P-value implies that random variation due to the sampling process alone is not likely to account for the observed difference.

- With a small p-value we reject $H_0$. The true property of the population is significantly different from what was stated in $H_0$.

Thus, small P-values are strong evidence AGAINST $H_0$.

But how small is small…?

P- Value & Evidence Against $H_0$

<table>
<thead>
<tr>
<th>P-Value</th>
<th>Evidence against $H_0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Greater than 10%</td>
<td>Very weak to none</td>
</tr>
<tr>
<td>Between 10% - 5%</td>
<td>Weak</td>
</tr>
<tr>
<td>Between 5% - 1%</td>
<td>Strong</td>
</tr>
<tr>
<td>Less than 1%</td>
<td>Very strong</td>
</tr>
</tbody>
</table>

- A P-Value of 0.05 or less is typically considered statistically significant.
Tests for a population mean – Summary

To test the hypothesis $H_0: \mu = \mu_0$ based on an SRS of size $n$ from a Normal population with unknown mean $\mu$ and known standard deviation $\sigma$, we rely on the properties of the sampling distribution $N(\mu, \sigma/\sqrt{n})$.

The P-value is the area under the sampling distribution for values at least as extreme, in the direction of $H_a$, as that of our random sample.

Again, we first calculate a $z$-value and then use Table A.

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

P-value in one-sided and two-sided tests

One-sided (one-tailed) test

$H_a: \mu > \mu_0$ is $P(Z \geq z)$

$H_a: \mu < \mu_0$ is $P(Z \leq z)$

Two-sided (two-tailed) test

$H_a: \mu \neq \mu_0$ is $2P(Z \geq |z|)$

To calculate the P-value for a two-sided test, use the symmetry of the normal curve. Find the P-value for a one-sided test and double it.
Example:
You are in charge of quality control in your food company. You sample randomly four packs of cherry tomatoes, each labeled 1/2 lb. (227 g).

The average weight (x-bar) from the four boxes that you examine is 222 g. Obviously, you don’t expect boxes filled with whole tomatoes to all weigh exactly half a pound.

- Is the somewhat smaller average weight simply due to chance variation?
- Is it evidence that the calibrating machine that sorts cherry tomatoes into packs needs revision?

Does the packaging machine need revision?
- $H_0: \mu = 227g$ versus $H_a: \mu \neq 227g$
- What is the probability of drawing a random sample such as yours if $H_0$ is true?

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{222 - 227}{5/\sqrt{4}} = -2$$

From table A, the area under the standard normal curve to the left of $z$ is 0.0228.
Thus, P-value = 2*0.0228 = 4.56%.

The probability of getting a random sample average so different from $\mu$ is so low that we reject $H_0$.

⇒ The machine does need recalibration.
The significance level $\alpha$

The significance level, $\alpha$, is the largest $P$-value tolerated for rejecting a true null hypothesis (how much evidence against $H_0$ we require). This value is decided arbitrarily before conducting the test.

- If the $P$-value is equal to or less than $\alpha$ ($P \leq \alpha$), then we **reject** $H_0$.
- If the $P$-value is greater than $\alpha$ ($P > \alpha$), then we **fail to reject** $H_0$.

**Does the packaging machine need revision?**

Two-sided test. The $P$-value is 4.56%.

* If $\alpha$ had been set to 5%, then the $P$-value would be significant.
* If $\alpha$ had been set to 1%, then the $P$-value would **not** be significant.

### Z-Tests based on a given level of significance $\alpha$

When the $z$ score falls within the rejection region (shaded area on the tail-side), the $p$-value is smaller than $\alpha$ and you have shown statistical significance.
Rejection region for a two-tail test of $\mu$ with $\alpha = 0.05$ (5%)

- A two-sided test means that $\alpha$ is spread between both tails of the curve.
  - Middle area $C = 1 - \alpha = 95\%$
  - Upper tail area $= \alpha / 2 = 0.025$

- Critical $z$-values
  - $-z_{\alpha/2} = -1.96$
  - $z_{\alpha/2} = 1.96$

Rejection Rule
- Reject $H_0$ if:
  \[ z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} < -1.96 \]
  OR
  \[ z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} > 1.96 \]

Confidence intervals to test hypotheses

Because a two-sided test is symmetrical, you can also use a confidence interval to test a two-sided hypothesis.

In a two-sided test, $C = 1 - \alpha$.

- **C confidence level**
- **$\alpha/2$** significance level

Decision Rule:
Reject the null hypothesis if the sample mean $\bar{x}$-bar falls outside the $C = 1 - \alpha$ confidence interval.
Example

- Packs of cherry tomatoes ($\sigma = 5 \text{ g}$)
  
  $H_0 : \mu = 227 \text{g} \text{ Vs } H_a : \mu \neq 227 \text{g}$

- Sample of size 4 has average 222g
  
  95% CI for $\mu = 222 \pm 1.96 \times (5/\sqrt{4}) = 222 \pm 4.9 \text{g} = (217.1 \text{g}, 226.9 \text{g})$

- Decision:
  
  Since $x$-bar = 222g does not fall inside the 95% CI we reject $H_0$.

A confidence interval gives a black and white answer: Reject or don’t reject $H_0$.
But it also estimates a range of likely values for the true population mean $\mu$.

A P-value quantifies how strong the evidence is against the $H_0$. But if you reject $H_0$, it doesn’t provide any information about the true population mean $\mu$.

Significance Tests Using a TI-83

6.2 Tests of Significance

We now show how to perform one-sided and two-sided hypothesis tests about the mean $\mu$ of a normally distributed population for which the standard deviation $\sigma$ is known. To do so, we will use the Z-Test feature (item 1) from the STAT TESTS menu. This menu is found using [2nd][F1][F6] in the Stats/List Editor on a TI-89. We can use this feature to work with either summary statistics or data sets.

1. Press STAT.
2. Select TESTS $\rightarrow$ (1) Z-Test ...
   - Select Inpt: Data; enter the value for $\mu_0$, $\sigma$, the list (Li) where the sample data is stored, and select the form of the alternative.
   - Select Inpt: Stats; enter the value for $\mu_0$, $\sigma$, $x$-bar, $n$, and select the form of the alternative.
3. Select Calculate or Draw & press Enter
Example 6.4 Executives’ Blood Pressure. The mean systolic blood pressure for males 35 to 44 years of age is 128 and the standard deviation is 15. But for a sample of 72 company executives in this age group, the mean systolic blood pressure is $\bar{x} = 126.07$. Is this evidence that the company’s executives in this age group have a different mean systolic blood pressure from the general population?

Solution. To test if the mean is different from 128, we use the null hypothesis $H_0: \mu = 128$ with a two-sided alternative $H_A: \mu \neq 128$. Bring up the Z-Test screen and adjust the Inpt to STATS, which allows us to enter the statistics. Enter the values $\mu_0 = 128$, $\sigma = 15$, $\bar{x} = 126.07$, and $n = 72$. Set the alternative to $\neq \mu_0$ then press [ENTER] on either Calculate or Draw.

Example 6.5 California SATs. An SRS of 500 California high school seniors gave an average SAT mathematics score of $\bar{x} = 461$. Is this good evidence against the claim that the mean for all California seniors is no more than 450? Assuming that $\sigma = 100$ for all such scores, perform the test $H_0: \mu = 450$, $H_A: \mu > 450$. Give the z test statistic and the $p$-value.

Solution. Bring up the Z-Test screen from the STAT TESTS menu and check that Inpt is STATS. Enter the value of $\mu_0 = 450$ and the summary statistics, set the alternative to $> \mu_0$, then scroll down to Calculate and press [ENTER].

We obtain a z test statistic of 2.46 and a $p$-value of 0.00695. Because the $p$-value is so small, we have significant evidence to reject $H_0$. If the true mean were 450, then there would be only a 0.00695 probability of obtaining a sample mean as high as $\bar{x} = 461$ with an SRS of 500 students.
Example 6.6 DRP Scores. The following table gives the DRP scores for a sample of 44 third-grade students in a certain district. It is known that \( \sigma = 11 \) for all such scores in the district. A researcher believes that the mean score of all third-graders in this district is higher than the national mean of 32. State the appropriate \( H_0 \) and \( H_a \), then conduct the test and give the \( p \)-value.

<p>| | | | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>40</td>
<td>26</td>
<td>39</td>
<td>14</td>
<td>42</td>
<td>18</td>
<td>25</td>
<td>43</td>
<td>46</td>
<td>27</td>
</tr>
<tr>
<td>47</td>
<td>19</td>
<td>26</td>
<td>35</td>
<td>34</td>
<td>15</td>
<td>44</td>
<td>40</td>
<td>38</td>
<td>31</td>
</tr>
<tr>
<td>52</td>
<td>25</td>
<td>35</td>
<td>35</td>
<td>33</td>
<td>29</td>
<td>34</td>
<td>41</td>
<td>49</td>
<td>28</td>
</tr>
<tr>
<td>47</td>
<td>35</td>
<td>48</td>
<td>22</td>
<td>33</td>
<td>41</td>
<td>51</td>
<td>27</td>
<td>14</td>
<td>54</td>
</tr>
</tbody>
</table>

Solution. Here we test \( H_0: \mu = 32 \) with a one-sided alternative \( H_a: \mu > 32 \). Enter the data into a list, say list L1, then call up the Z-Test screen and change Inpt to Data. Enter the values \( \mu_0 = 32 \) and \( \sigma = 11 \), set the list to L1 with frequencies 1, and set the alternative to \( \mu > \mu_0 \). Press ENTER on Calculate or Draw.

We obtain a \( p \)-value of 0.0312. If the average of the district were equal to 32, then there would be only a 3.117% chance of a sample group of 44 averaging as high as \( \bar{x} = 35.09 \). There is evidence to reject \( H_0 \) and conclude that the district’s average is higher than 32.
**Example 6.7 SAT Coaching.** Suppose that SATM scores vary normally with $\sigma = 100$. Calculate the $p$-value for the test of $H_0$: $\mu = 480$, $H_A$: $\mu > 480$ in each of the following situations:

(a) A sample of 100 coached students yielded an average of $\bar{x} = 483$.

(b) A sample of 1000 coached students yielded an average of $\bar{x} = 483$.

(c) A sample of 10,000 coached students yielded an average of $\bar{x} = 483$.

---

**Solution.** We adjust the settings in the Z-Test screen from the STAT TESTS menu and calculate, changing $n$ each time. Below are the results using the three different sample sizes. Notice that with increasing sample size, the $p$-value changes dramatically. We see that the rise in the average score to $\bar{x} = 483$ is significant ($p$ very small) only when the results stem from the very large sample of 10,000 coached students. With the sample of only 100 students, there is 38.2% chance of obtaining a sample mean as high as $\bar{x} = 483$, even if the true mean were still 480.
Example 6.8 More SAT Coaching. For the same hypothesis test as in Example 6.7 above, consider the sample mean of 100 coached students. (a) Is $\bar{x} = 496.4$ significant at the 5% level? (b) Is $\bar{x} = 496.5$ significant at the 5% level?

Solution: we perform the Z-Test for both values of $\bar{x}$:

<table>
<thead>
<tr>
<th>Z-Test</th>
<th>Z-Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mu &gt; 480$</td>
<td>$\mu &gt; 480$</td>
</tr>
<tr>
<td>$z = 1.64$</td>
<td>$z = 1.65$</td>
</tr>
<tr>
<td>$p = .0505025692$</td>
<td>$p = .0494714509$</td>
</tr>
<tr>
<td>$\bar{x} = 496.4$</td>
<td>$\bar{x} = 496.5$</td>
</tr>
<tr>
<td>$n = 100$</td>
<td>$n = 100$</td>
</tr>
</tbody>
</table>

In the first case, $p = 0.0505 > 0.05$; so the value of $\bar{x} = 496.4$ is not significant at the 5% level. However, in the second case, $p = 0.04947 < 0.05$; so the value of $\bar{x} = 496.5$ is significant at the 5% level. However, for SATM scores, there is no real “significant” difference between means of 496.4 and 496.5.