Inference in One, Two, and More Than Two Populations
- A primer for Analysis of Variance -

Inference about One Population:

Consider the hypothesis test for $\mu$ for data that are normally distributed and $\sigma^2$ unknown. For this example, we will assume $n < 30$. We may have one of three possible hypotheses about $\mu$ as follows:

1. $H_0 : \mu \leq \mu_0$ Vs. $H_A : \mu > \mu_0$, a right-tailed test
2. $H_0 : \mu \geq \mu_0$ Vs. $H_A : \mu < \mu_0$, a left-tailed test
3. $H_0 : \mu = \mu_0$ Vs. $H_A : \mu \neq \mu_0$, a two-tailed test

Note:
- Recall that the null hypothesis, denoted $H_0$, is the hypothesis that we will assume to hold true unless the evidence suggests otherwise.
- We are making inference about $\mu$, a population parameter.
- The value of $\mu_0$ is the null value of $\mu$ under the null hypothesis.
- We will calculate a test statistic to make inference about $\mu$.

Test statistic when $n < 30$, the data are normally distributed (or approximately bell-shaped) and the variance is unknown:

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

Notes:
- When the data are normally distributed, the test statistic above is distributed as a $t$-distribution with $n - 1$ degrees of freedom (df).
- When the data are approximately normal, then the test statistic is approximately a $t$-distribution with $n - 1$ degrees of freedom.
- There are many different $t$-distributions. The degrees of freedom distinguish one from another.
- The $t$-distribution we will use is symmetrical about zero and has a mean of zero.
- There do exist $t$-distributions not centered at zero. These are called non-central $t$-distributions and can be used in power calculations.
• The \( t \)-distribution has a variance equal to \( \text{df}/(\text{df} - 2) \).

• Note how the variance of the \( t \)-distribution approaches that of the standard normal as \( n \) increases without bound.

Rejection Region:

• We specify (a priori) the amount of risk we are willing to accept in making a Type I error.

• Denote the Type I error as \( \alpha \).

• \( \alpha \) is the probability of rejecting \( H_0 \) when \( H_0 \) is true.

• Let \( t_\alpha \) be the critical value for the specified \( \alpha \).

• Based on \( \alpha \) and the df of the test statistic, our rejection region for the hypotheses above are, respectively,

1. Reject \( H_0 \) if \( t \geq t_\alpha \).
2. Reject \( H_0 \) if \( t \leq t_\alpha \).
3. Reject \( H_0 \) if \( |t| \geq t_{\alpha/2} \)

\( p \)-values

• The \( p \)-value can be interpreted as the probability of observing a test statistic more extreme than the one observed.

• It follows then to reject \( H_0 \) when the \( p \)-value is less than \( \alpha \).

• For the hypotheses above, the \( p \)-values are, respectively.

1. \( p \)-value = \( P(t \geq \text{computed } t) \)
2. \( p \)-value = \( P(t \leq \text{computed } t) \)
3. \( p \)-value = \( P(t \geq |\text{computed } t|) \)

Remaining steps in the hypothesis test:

• Based on the test statistic and rejection region one makes a decision to reject \( H_0 \) or conclude there is not enough evidence to refute the claim under the null.

• Follow the decision to reject \( H_0 \) (or not) by a conclusion in terms of the hypothesis.
Example (Ott and Longnecker, p. 232)

A massive multistate outbreak of food-borne illness was attributed to *Salmonella enteritidis*. Epidemiologists determined that the source of the illness was ice cream. They sampled nine production runs from the company that had produced the ice cream to determine the level of the bacteria in the ice cream. These levels (MPN/g) are as follows:

\[0.593 \quad 0.142 \quad 0.320 \quad 0.691 \quad 0.231 \quad 0.793 \quad 0.519 \quad 0.392 \quad 0.418\]

Use the data to determine whether the average level of the bacteria in the ice cream is greater than 0.3 MPN/g, a level that is considered to be very dangerous. Let \(\alpha = 0.01\).

- **Hypothesis:** \(H_0 : \mu \leq 0.3\) Vs \(H_A : \mu > 0.3\)
- The critical value for \(\alpha = 0.01\) with \(df = 9 - 1 = 8\) is 2.896
- We will reject the null when the test statistic is larger than 2.896.
- A few calculations yield \(\bar{y} = 0.456\) and \(s = 0.2128\).
- The test statistic is

\[
t = \frac{\bar{y} - \mu_0}{s/\sqrt{n}} = \frac{0.456 - 0.3}{0.2128/\sqrt{9}} = 2.21
\]

- Decision: We cannot reject \(H_0\).
- Conclusion: There is not enough evidence to conclude that the average level of *Salmonella enteritidis* in the ice cream is greater than 0.3 MPN/g.

**Inference about Two Populations:**

Example: Suppose we want to test the potency of a drug that has been in storage versus the potency of the same drug directly from the assembly line. (Management is concerned that the drug loses its potency after 1 year of storage). You conduct a hypothesis test. You randomly draw 10 bottles of the drug that have been in storage for one year and analyze the potency. You also randomly draw 10 bottles off the production line. The data regarding potency is as follows (Ott and Longnecker, p. 269).

<table>
<thead>
<tr>
<th>Fresh</th>
<th>Stored</th>
</tr>
</thead>
<tbody>
<tr>
<td>10.2</td>
<td>9.8</td>
</tr>
<tr>
<td>10.5</td>
<td>9.6</td>
</tr>
<tr>
<td>10.3</td>
<td>9.5</td>
</tr>
<tr>
<td>10.8</td>
<td>9.8</td>
</tr>
<tr>
<td>9.8</td>
<td>9.9</td>
</tr>
</tbody>
</table>

Some Notation:
• Let $n_1$ denote the number of samples from population 1. (Fresh in our case)
• Let $n_2$ denote the number of samples from population 2. (Stored)
• Let $\bar{y}_1$ and $\bar{y}_2$ denote the sample means for the data drawn from population 1 and 2, respectively.
• Let $s^2_1$ and $s^2_2$ denote the sample variances for the data drawn from populations 1 and 2, respectively.
• Let $\mu_1$ and $\mu_2$ denote the population means for population 1 and 2, respectively.

Carry-out the Hypothesis Test $H_0 : \mu_1 = \mu_2 \ Vs \ H_A : \mu_1 \neq \mu_2$.
• We will assume that the data are drawn from either normal or approximately normal distributions.
• Under $H_0$, $\mu_1 - \mu_2 = 0$.
• We will construct our test statistic based on the approximate sampling distribution of $\bar{y}_1 - \bar{y}_2$.
• From the Central Limit Theorem, $\bar{y}_1 - \bar{y}_2 \sim N(0, \sigma^2/n_1 + \sigma^2/n_2)$ under the null hypothesis.
• We can rewrite $\sigma^2/n_1 + \sigma^2/n_2$ as $\sigma^2(1/n_1 + 1/n_2)$.
• An estimate of $\sigma^2$ is given by
  \[
  s^2_{pooled} = \frac{(n_1 - 1)s^2_1 + (n_2 - 1)s^2_2}{n_1 + n_2 - 2}
  \]
• An estimate of the variance of $\bar{y}_1 - \bar{y}_2$, is given by
  \[
  \hat{\sigma}^2(1/n_1 + 1/n_2) = s^2_{pooled}(1/n_1 + 1/n_2)
  \]
• The test statistic is given by
  \[
  t.s. = \frac{\bar{y}_1 - \bar{y}_2}{\sqrt{s^2_{pooled}(1/n_1 + 1/n_2)}}
  \]
• Under the null hypothesis our test statistic follows a $t$-distribution with degrees of freedom equal to $n_1 + n_2 - 2$.
• For our example, the test statistic equates to 4.237.
• For $\alpha = .05$, our critical value (with 18 degrees of freedom) for a two-tail test is 2.101.
• Decision Reject $H_0$. 

• Conclusion: There is sufficient evidence to conclude that the potency of the drugs stored for one year is not equal to the potency of the drugs taken from production.

• Does a statistically significant difference imply that a practical difference exists? In our case, one would need to look at the sample means and ask a pharmacist about the implications behind the test.

Inference about More Than Two Populations:

Example: Our objective is to compare the mean hourly wage of three different ethnic groups (African American, Anglo-American, and Hispanic) employed by a large produce company.

• Begin by drawing independent random samples of farm laborers from each of the three ethnic populations.

• If the sample means differ, does this infer that the population means differ?

• How much need the sample means differ before we conclude that the population means do in fact differ?

• We will use Analysis of Variance (ANOVA) to answer these questions.

Consider the two tables below (Ott and Longnecker, p. 383):

\[
\begin{array}{ccc}
\text{Scenario 1} \\
\text{Af-Am} & \text{Ang-Am} & \text{His} \\
5.90 & 5.51 & 5.01 \\
5.92 & 5.50 & 5.00 \\
5.91 & 5.50 & 4.99 \\
5.89 & 5.49 & 4.98 \\
5.88 & 5.50 & 5.02 \\
\hline
\bar{y}_1 = 5.90 & \bar{y}_2 = 5.50 & \bar{y}_3 = 5.00 \\
\end{array}
\]

Note:

• Examine the variability of the sample within each ethnic group.

• The variability within the sample is very small.

• The variability among the sample means is larger.

• This situation suggests that maybe the populations are different.

• Examine the plots in Excel and compare with Scenario 2.
The sample means in Scenario 2 are the same as in Scenario 1.

Examine the variability of the sample within each ethnic group.

The variability within the sample is now high.

The variability among the sample means is smaller (relatively).

Here, it is unclear whether the population means are different.

Examine the plot in Excel and compare with Scenario 1.

To perform a statistical hypothesis, we could perform multiple $t$-tests in which case we would have the following null hypotheses:

$$H_0 : \mu_1 = \mu_2, \quad \mu_1 = \mu_3, \quad \mu_2 = \mu_3$$

Notes:

- In the case of many groups, this could become tedious and time consuming.
- The probability of falsely rejecting at least one of the hypotheses increases as the number of $t$-tests increases.
- The probability of committing a Type I error is no longer, say, .05. In fact, it is larger!
- We need a single test that “all three population means are equal” with the correct Type I error that is simple to carry-out.
- This is the starting point for Analysis of Variance (ANOVA).

ANOVA is developed under the following assumptions:

1. Each of the (three in our case) populations are normally distributed.
2. The population variances are equal. In our example, $\sigma_1^2 = \sigma_2^2 = \sigma_3^2$.
3. The sample measurements from each of the populations are drawn independently and randomly.
Estimating the common variance:

- Recall the pooled estimate of the variance for a $t$-test for two populations where we assume equal variances.
- For our example, note that we have three populations and five samples from each population.
- An extension of this to more than two populations is denoted by Mean Square Error (MSE) (and sometimes called Mean Square Within Samples).
- Denote Sum of Squared Error by $SS_{error}$.
  \[
  SS_{error} = \sum_{i=1}^{3} \sum_{j=1}^{5} (y_{ij} - \bar{y}_i)^2 = (n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + (n_3 - 1)s_3^2.
  \]
- Adding up $(n_1 - 1) + (n_2 - 1) + (n_3 - 1) = 15 - 3$.
- Let $N$ denote the total number of observations, $a$ the number of treatments
  \[
  MSE = SS_{error}/(d.f.) = SS_{error}/(N - a).
  \]

Estimating the variability among the population means.

- If the null hypothesis is true, then the means from the populations are equal (as well as the variances by assumption).
- This estimate is sometimes called Mean Square Between Samples. We will use the term Mean Square Treatments, instead.
- Let $\bar{y}_.$ denote the grand mean.
- Let $\bar{y}_i.$ denote the $i$th treatment mean.
  \[
  SS_{treat} = \sum_{i=1}^{3} n_i(\bar{y}_i. - \bar{y}_.)^2
  \]
- Let $a$ denote the number of treatments (three in our case)
  \[
  MS_{treat} = SS_{treat}/(a - 1)
  \]

The Test Statistic:

- We now have two estimates of the common variance $\sigma^2$.
- The two estimates are $MS_{treat}$ and $MS_{error}$.
- Note the following construction of our $F$ statistic:
  \[
  F = \frac{MS_{treat}}{MS_{error}} = \frac{\chi^2/(a - 1)}{\chi^2/(N - a)} \sim F(df1 = a - 1, df2 = N - a)
  \]
• This is the test statistic that we will use to test
  \( H_0 : \mu_1 = \mu_2 = \mu_3 \) Vs.
  \( H_A : \) At least one of the pop. means differs from the rest.

• These calculations are generally summarized in an ANOVA table

The ANOVA table for the Farm Wages Example - Scenario 1:

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
<th>F-value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment</td>
<td>2</td>
<td>2.0333</td>
<td>1.0167</td>
<td>5545.45</td>
<td>&lt; .0001</td>
</tr>
<tr>
<td>Error</td>
<td>12</td>
<td>.0022</td>
<td>.0002</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>14</td>
<td>2.036</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The ANOVA table for the Farm Wages Example - Scenario 2:

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
<th>F-value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment</td>
<td>2</td>
<td>2.0333</td>
<td>1.0167</td>
<td>.44</td>
<td>.6565</td>
</tr>
<tr>
<td>Error</td>
<td>12</td>
<td>27.9846</td>
<td>2.3321</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>14</td>
<td>30.0179</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Notes:

• In both scenarios, \( SS_{treat} \) is the same. Check out the formula to understand why.

• Under \( H_0 \), our \( F \) statistic should be close to one.

• Compare the \( MS_{error} \)'s in both scenarios.

• The \( F \)-test rejects \( H_0 \) for large values of the test statistic. That is, for large variation among the treatment means when compared to the variation within treatments.