Testing Whether the Means of Several Different Populations Are Equal: Analysis of Variance (ANOVA) and the F-Distribution

Also: the Kruskal-Wallis Test

We have seen how to test whether the means \( \mu_A \) and \( \mu_B \) of two different populations are equal. We use the 2-sample \( t \)-test based on the corresponding sample means \( \bar{X}_A \) and \( \bar{X}_B \) and their difference \( \bar{X}_A - \bar{X}_B \).

Sometimes there are more than 2 populations whose means are to be compared. Then taking a single difference does not work.

Here we treat the Analysis of Variance (ANOVA) method for comparing 2 or more population means. For the case of just 2 populations, it is equivalent to the 2-sample \( t \)-test.

Examples

(i) Comparison of two treatments. Investigate the effects of two experimental drugs, Tumostat and Inhibit, on the growth of brain tumors in humans. “Response variable” in “populations” under study.

Q. Is this a 2-sample problem?

\[ \square \text{Yes} \quad \square \text{No} \]

We need still another group, in addition to the Tumostat group and the Inhibit group. What group?

(Turn page upside down for answer: donor group)
(i), continued

So this experiment involves comparison of 3 groups, not 2.

(iii) Comparison of cleansing action of 3 detergents

Detergent A scores: 77, 81, 71, 76, 80
  B      : 72, 58, 74, 66, 70
  C      : 76, 85, 82, 80, 77

Examples (i) and (ii) illustrate the setting for **one-way** (or one-factor) ANOVA.

In a separate handout, we will consider **two-way** (or two-factor) ANOVA. That concerns the situation of comparing several different methods by carrying each of them out on each of the subjects of a sample from a single population. When the number of methods is just two, this situation reduces to the "paired-data" or "matched-data" setting considered previously.
The One-Way ANOVA Setting

- $K$ populations to be compared with respect to a variable $X$ having means $\mu_1, \mu_2, \ldots, \mu_K$ in these populations

- goal: Test the null hypothesis

$$H_0: \mu_1 = \mu_2 = \ldots = \mu_K$$

(\text{the means are equal})

versus the alternative hypothesis

$$H_A: \text{Some of the means are not equal}$$

- data: independent samples from each population

$$\begin{align*}
\text{population 1 sample: } &\ X_{11}, X_{12}, \ldots, X_{1n_1} \\
\text{population 2 sample: } &\ X_{21}, X_{22}, \ldots, X_{2n_2} \\
\vdots \\
\text{population K sample: } &\ X_{K1}, X_{K2}, \ldots, X_{Kn_K}
\end{align*}$$

The sample sizes $n_1, n_2, \ldots, n_K$ can differ.

$$n = n_1 + \cdots + n_K = \text{total sample size}$$

- sample means

$$\bar{X}_{1*} = \frac{1}{n_1} \sum_{i=1}^{n_1} X_{1i} = \text{sample mean for 1st popu}$$

$$\vdots$$

$$\bar{X}_{K*} = \frac{1}{n_K} \sum_{i=1}^{n_K} X_{Ki} = \text{sample mean for K-th poplu}$$

$$\bar{X}_{.1} = \frac{1}{K} \sum_{i=1}^{K} \frac{n_i}{n} \sum_{j=1}^{n_i} X_{ij} = \text{mean of all the observations together}$$
The One-Way ANOVA Approach

Key Idea: Write

\[
\frac{\text{Variability of all the observations about the overall sample mean } \bar{x}_.}{\text{Total Variation}} = \frac{\text{Variability of the sample means } \bar{x}_1, \bar{x}_2, \ldots, \bar{x}_k}{\text{Variation between samples}} + \frac{\text{Variability of the observations about their own sample means}}{\text{Variation within samples}}
\]

Q. How can we read this as evidence for or against \( H_0 \)?

A. We look at the relative size of the 2nd part (variation between samples). If it is relatively large, we interpret this as evidence against \( H_0 \).

Doing the above with formulas:

\[
\frac{\sum_{i=1}^{K} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_.)^2}{SS(Total)} = \frac{\sum_{i=1}^{K} n_i (x_i - \bar{x}_.)^2}{SS(Tr)} + \frac{\sum_{i=1}^{K} \sum_{j=1}^{n_i} (x_{ij} - x_i)^2}{SS(E)}
\]

"Total Sum of Squares" = "Treatment SS" + "Error SS"

- procedure: Reject \( H_0 \) if \( SS(Tr) \) is large relative to \( SS(E) \)

Q. How do we decide "relatively large" and get a \( p \)-value?
Preliminary: F-distribution

I will inflict one more continuous probability distribution on your wearied minds.

Let

$$R = \frac{V}{W}$$

where:

(i) $V = \frac{a \chi_k^2 \text{ variable}}{k}$

(ii) $W = \frac{a \chi_m^2 \text{ variable}}{m}$

(iii) $V$ and $W$ are independent chance variables.

The chance variable $R$ has a special distribution called

The F-distribution with degrees of freedom $k$ for the numerator and degrees of freedom $m$ for the denominator.

We denote it by

$$F_{k,m}$$

It looks like

$F_{k,m}$ curve ($k>2$)

(sort of like a $\chi^2$ curve)
The One-Way ANOVA Test Procedure

- For each of SS(Tr) and SSE, we form an associated "mean square":
  
  \[
  MS(Tr) = \frac{SS(Tr)}{k-1} \\
  MSE = \frac{SSE}{n-k}
  \]

- Test Statistic:
  
  \[
  f = \frac{MS(Tr)}{MSE}
  \]

- Assume that all the K populations are Normal with the same variance \( \sigma^2 \):
  
  \( N(\mu_1, \sigma^2), N(\mu_2, \sigma^2), \ldots, N(\mu_k, \sigma^2) \)

- Ho-Distribution of Test Statistic \( f \):
  
  \[
  F_{k-1, n-k}
  \]

  We reject Ho for large values of \( f \), using the \( F_{k-1, n-k} \) distribution for choosing the cutoff point for a given level of significance or for getting the \( p \)-value.

  Note. If Ho true, \( f \) tends to estimate the number 1. So "large" means considerably larger than 1.
Format for presenting the ANOVA results

Usually, a table as follows:

<table>
<thead>
<tr>
<th>Source of Variation</th>
<th>Degrees of freedom</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatments</td>
<td>K-1</td>
<td>SS(T_r)</td>
<td>MS(T_r)</td>
<td>F</td>
<td>MSE</td>
</tr>
<tr>
<td>Error</td>
<td>n-k</td>
<td>SSE</td>
<td>MSE</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>n-1</td>
<td>SS(Total)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\( n = n_1 + n_2 + \ldots + n_K \)

Illustration for the Detergent example (p.2)

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>77</td>
<td>72</td>
<td>76</td>
</tr>
<tr>
<td>81</td>
<td>58</td>
<td>85</td>
</tr>
<tr>
<td>71</td>
<td>74</td>
<td>82</td>
</tr>
<tr>
<td>76</td>
<td>66</td>
<td>80</td>
</tr>
<tr>
<td>80</td>
<td>70</td>
<td>77</td>
</tr>
</tbody>
</table>

\( K = 3 \) \( n_1 = n_2 = n_3 = 5 \)

\( \bar{X}_1 = \frac{385}{5}, \ \bar{X}_2 = \frac{340}{5}, \ \bar{X}_3 = \frac{400}{5}, \ \bar{X}_s = \frac{1125}{15} \)

SS(Total) = 666, SS(T_r) = 390, SSE = 666 - 390 = 276

MS(T_r) = \frac{390}{2} = 195, MSE = \frac{276}{12} = 23

\( f = \frac{195}{23} = 8.48 \)

relevant \( H_0 \)-distribution: \( F_{2,12} \)

\( p\)-value: 0.0051 → Reject \( H_0 \) (consistent with what dotplot suggests)

using "F calculator" on

http://www.stat.tamu.edu/~west/applets/f demo.html
The previous illustration, using statistical software (Minitab)

MTB > AOVOneWay 'detergA' 'detergB' 'detergC';
SUBC> CI Mean 99.0;
SUBC> Tukey 5.

One-way ANOVA: detergA, detergB, detergC

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Factor</td>
<td>2</td>
<td>390.0</td>
<td>195.0</td>
<td>8.48</td>
<td>0.005</td>
</tr>
<tr>
<td>Error</td>
<td>12</td>
<td>276.0</td>
<td>23.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>14</td>
<td>666.0</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\[ S = 4.796 \quad R-Sq = 58.56\% \quad R-Sq(adj) = 51.65\% \]

\[ \text{The fraction of total variation explained by the treatment effect.} \]

Individual 99% CIs For Mean Based on Pooled StDev

<table>
<thead>
<tr>
<th>Level</th>
<th>N</th>
<th>Mean</th>
<th>StDev</th>
</tr>
</thead>
<tbody>
<tr>
<td>detergA</td>
<td>5</td>
<td>77.000</td>
<td>3.937</td>
</tr>
<tr>
<td>detergB</td>
<td>5</td>
<td>68.000</td>
<td>6.325</td>
</tr>
<tr>
<td>detergC</td>
<td>5</td>
<td>80.000</td>
<td>3.674</td>
</tr>
</tbody>
</table>

Pooled StDev = 4.796

Tukey 95% Simultaneous Confidence Intervals
All Pairwise Comparisons

Individual confidence level = 97.94%

detergA subtracted from:

\[ \text{Lower} \quad \text{Center} \quad \text{Upper} \]
\[-17.086 \quad -9.000 \quad -0.914 \]
\[-5.086 \quad 3.000 \quad 11.086 \]

\[ \text{C.I. for } \text{detergA} \]

\[ \text{C.I. for } \text{detergB} \]

\[ \text{C.I. for } \text{detergC} \]

We do this kind of follow-up analysis after an Ho is rejected, to find out why.

We always take a "look" at our data. The human mind is actually a powerful tool (sometimes).

(I did not have to type this. I used a dropdown menu.)

We keep an eye on "zero". Is it in the C.I.? or not? For Mg-Mg, it is. For the others, not. We see that Ho is rejected because Mg differs from Mg and Mq, which are close.
The ANOVA procedure uses the assumption that each population is Normal: \( N(\mu, \sigma^2) \) with possibly different means but the same \( \sigma^2 \).

Can we compare the locations of several populations without this assumption?

If only 2 populations, we can use the Wilcoxon rank sum test and compare medians.

The Wilcoxon rank sum test has been extended to a K-sample approach: the Kruskal-Wallis test.

It is rank-based, too.
The Kruskal-Wallis Test

- **Goal**: test
  \[ H_0: \text{k given population distributions are identical} \]
  versus
  \[ H_A: \text{they differ in locations} \]

- **Method**:
  1. Rank all the observations in one combined collection.
  2. For each sample, compute the rank sum:
     \[ T_i = \text{sum of combined-sample ranks for the } i\text{-th sample} \]
  3. Form the test statistic
     \[ H = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{T_i^2}{n_i} - 3(n+1) \]
     \[ \text{[sample sizes } n_1, \ldots, n_k \text{ with } n = n_1 + \ldots + n_k \]}

     Larger values of \( H \) represent stronger evidence against \( H_0 \) and in favor of differences in the locations

- **Key Fact**: The \( H_0 \)-distribution of \( H \) is approximately \( \chi^2_{k-1} \)

Use this to get a p-value.
Illustration of Kruskal-Wallis Test

We revisit the detergent data.

Arrange values by increasing order within samples:

<table>
<thead>
<tr>
<th>Ranks in Combined Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>A : 4 7.5 9.5 11.5 13</td>
</tr>
<tr>
<td>71, 76, 77, 80, 81</td>
</tr>
<tr>
<td>B : 58, 66, 70, 72, 74</td>
</tr>
<tr>
<td>1, 2, 3, 5, 6</td>
</tr>
<tr>
<td>C : 75, 95, 11.5, 14, 15</td>
</tr>
<tr>
<td>76, 77, 80, 82, 85</td>
</tr>
</tbody>
</table>

\[ n_1 + n_2 + n_3 = 15 \quad \text{with} \quad k = 3, n = 15 \]

Rank sums for the 3 samples:

\[ T_A = 4 + 7.5 + 9.5 + 11.5 + 13 = 45.5 \]

\[ T_B = 1 + 2 + 3 + 5 + 6 = 17 \]

\[ T_C = 7.5 + 9.5 + 11.5 + 14 + 15 = 57.5 \]

(We see that sample B is getting most of the lowest ranks.)

Form the test statistic:

\[ H = \frac{12}{15 \times 16} \left[ \frac{(45.5)^2}{5} + \frac{(17)^2}{5} + \frac{(57.5)^2}{5} \right] - 3 \times 16 \]

\[ = \boxed{8.66} \]

Relevant \( H \)-distribution: \( \chi^2 \)

\[ \text{P-value: 0.0132} \rightarrow \text{Reject } \text{H}_0 \text{ in favor of:} \]

\( \text{H}_A : \text{differences in locations} \)

This corroborates the finding from ANOVA.

If the two findings were different, we might prefer the Kruskal-Wallis finding because the ANOVA method used a strong assumption: Normality.