STAT120C: Analysis of Variance (ANOVA)

So far we have considered only one or two samples. For one sample, we were concerned by the population mean. For two samples, we were concerned by the difference of two population means. What if there are more than two samples? In the two-sample t test, we considered only one factor. What if there is another factor that might affect the outcome variable? For example, we are interested to learn the effects of gender and smoking on body mass index. These questions can be answered using a more general framework: the analysis of variance. We will spend two weeks on this topic.

1 One-Way ANOVA

1.1 Introduction

One-way layout/design: a one-way layout / design is an experiment design in which independent measurements are made under each of several conditions. It generalizes the design of two independent samples.

Example Chlorpheniramine maleate in tablets (Kirchhoefer 1979): To study the level of chlorpheniramine maleate in tablets from seven labs, measurements of composites that had nominal dosages equal to 4mg from the seven labs. And 10 measurements were made by each lab.

<table>
<thead>
<tr>
<th>Observations(j)</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>( \bar{y}_i )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lab1</td>
<td>4.13</td>
<td>4.07</td>
<td>4.04</td>
<td>4.07</td>
<td>4.05</td>
<td>4.04</td>
<td>4.02</td>
<td>4.06</td>
<td>4.10</td>
<td>4.04</td>
<td>4.062</td>
</tr>
<tr>
<td>Lab2</td>
<td>3.86</td>
<td>3.85</td>
<td>4.08</td>
<td>4.11</td>
<td>4.08</td>
<td>4.01</td>
<td>4.02</td>
<td>4.04</td>
<td>3.97</td>
<td>3.95</td>
<td>3.997</td>
</tr>
<tr>
<td>Lab3</td>
<td>4.00</td>
<td>4.02</td>
<td>4.01</td>
<td>4.01</td>
<td>4.04</td>
<td>3.99</td>
<td>4.03</td>
<td>3.97</td>
<td>3.98</td>
<td>3.98</td>
<td>4.003</td>
</tr>
<tr>
<td>Lab4</td>
<td>3.88</td>
<td>3.88</td>
<td>3.91</td>
<td>3.95</td>
<td>3.92</td>
<td>3.97</td>
<td>3.92</td>
<td>3.90</td>
<td>3.97</td>
<td>3.90</td>
<td>3.920</td>
</tr>
<tr>
<td>Lab5</td>
<td>4.02</td>
<td>3.95</td>
<td>4.02</td>
<td>3.89</td>
<td>3.91</td>
<td>4.01</td>
<td>3.89</td>
<td>3.89</td>
<td>3.99</td>
<td>4.00</td>
<td>3.957</td>
</tr>
<tr>
<td>Lab6</td>
<td>4.02</td>
<td>3.86</td>
<td>3.96</td>
<td>3.97</td>
<td>4.00</td>
<td>3.82</td>
<td>3.98</td>
<td>3.99</td>
<td>4.02</td>
<td>3.93</td>
<td>3.955</td>
</tr>
<tr>
<td>Lab7</td>
<td>4.00</td>
<td>4.02</td>
<td>4.03</td>
<td>4.04</td>
<td>4.10</td>
<td>3.81</td>
<td>3.91</td>
<td>3.96</td>
<td>4.05</td>
<td>4.06</td>
<td>3.998</td>
</tr>
</tbody>
</table>

A boxplot of the data is shown in the figure below. There are two sources of variability in the data: variability within labs and variability between labs.

Let \( \mu_i \) be the mean level of chlorpheniramine in tablets from Lab \( i \). We might be interested in many questions. For example,

(a) Is the mean level of chlorpheniramine maleate in tablets from Lab 1 different from 4? (one-sample t test)

\[
\mu_1 = 4
\]
(b) Is the mean level of chlorpheniramine maleate in tablets from Lab 1 different from that from Lab 2? (two-sample t test)

$$\mu_1 = \mu_2$$

(c) Do mean levels differ across the seven labs? (which test?)

$$\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 = \mu_6 = \mu_7$$

Questions (a) and (b) can answered using one-sample and two-sample t-tests, respectively. To answer question (c), we need to use a more general framework.

1.2 The $F$ test under balanced designs

1.2.1 Notation

In two-sample t-test, we used notation $X_i$ and $Y_i$. When there are more than two conditions, we shall use $Y_{ij}$, which denotes the $j$th measurement under the $i$th condition.

<table>
<thead>
<tr>
<th>Treatment Group</th>
<th>Observations($j$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$y_{11}$ $y_{12}$ $\cdots$ $y_{1J}$</td>
</tr>
<tr>
<td>2</td>
<td>$y_{21}$ $y_{22}$ $\cdots$ $y_{2J}$</td>
</tr>
<tr>
<td>$\vdots$</td>
<td>$\vdots$ $\vdots$ $\cdots$ $\vdots$</td>
</tr>
<tr>
<td>$I$</td>
<td>$y_{I1}$ $y_{I2}$ $\cdots$ $y_{IJ}$</td>
</tr>
</tbody>
</table>

For two-sample t test, we assumed that

$$X_i \overset{iid}{\sim} N(\mu_X, \sigma^2),$$

$$Y_i \overset{iid}{\sim} N(\mu_Y, \sigma^2)$$

and

$$X_1, \ldots, X_m$$ and $$Y_1, \ldots, Y_n$$ are independent.

Similar assumptions will be made for the one-way layout. Here $Y_{ij}$ is the $j$th observation under the $i$th treatment/condition. The statistical model we use for one-way ANOVA is

$$y_{ij} = \mu_i + \epsilon_{ij},$$

where $\epsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2)$, for $i = 1, \ldots, I; j = 1, \cdots, J$. 

3
In the model, \( \mu_i \) is the mean of the \( i \)th treatment/condition. This is a balanced design. We say it is balanced because all groups have the same number of observations.

The model can also be written as

\[ Y_{ij} = \mu + \alpha_i + \epsilon_{ij} \]

with \( \sum \alpha_i = 0 \) and \( \epsilon_{ij} \sim N(0, \sigma^2) \), for \( i = 1, 2, \ldots, I; j = 1, \ldots, J \).

The design is balanced. In this situation, \( \mu = \overline{\mu_i} = \mu_i / I \) and \( \alpha_i = \mu_i - \mu \). Here \( \mu \) is the overall mean level, \( \alpha_i \) is the differential effect of the \( i \)th treatment, and \( \epsilon_{ij} \) is the random error in the \( j \)th observation under the \( i \)th treatment. The errors are assumed to be iid \( N(0, \sigma^2) \). Because \( \alpha_i \)'s are deviations from the overall mean:

\[ \sum_{i=1}^{I} \alpha_i = 0 \]

The expected response to the \( i \)th treatment is \( E(Y_{ij}) = \mu + \alpha_i \). Thus, if \( \alpha_i = 0 \), for \( i = 1, 2, \ldots, I \), all treatment have the same expected response. In general, \( \alpha_i - \alpha_j \) is the difference between the expected outcome values under treatments/conditions \( i \) and \( j \).

We introduce the following summary statistics:

\[
\begin{align*}
Y_b &= \sum_{j=1}^{J} Y_{ij} \\
\bar{Y}_b &= \frac{\sum_{j=1}^{J} Y_{ij}}{J} \\
\bar{Y}_s &= \frac{\sum_{i=1}^{I} \sum_{j=1}^{J} Y_{ij}}{IJ} \\
\bar{Y}_c &= \frac{\sum_{i=1}^{I} \sum_{j=1}^{J} Y_{ij} / (IJ)}{1}
\end{align*}
\]

1.2.2 Decomposition of the total sum of squares

The analysis of variance is based on the following decomposition of the Sum of Squares of TOtal (SSTO):

\[
SSTO = \sum_{i=1}^{I} \sum_{j=1}^{J} (Y_{ij} - \bar{Y}_c)^2
\]
In the following, we show that $SSTO$ can be decomposed into two sums of squares.

\[
SSTO = \sum_{i=1}^{I} \sum_{j=1}^{J} (Y_{ij} - \bar{Y}_{i\cdot})^2 \\
= \sum_{i=1}^{I} \sum_{j=1}^{J} [(Y_{ij} - \bar{Y}_{i\cdot}) + (\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})]^2 \\
= \sum_{i=1}^{I} \sum_{j=1}^{J} (Y_{ij} - \bar{Y}_{i\cdot})^2 + \sum_{i=1}^{I} \sum_{j=1}^{J} (\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2 + 2 \sum_{i=1}^{I} \sum_{j=1}^{J} (Y_{ij} - \bar{Y}_{i\cdot})(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot}) \\
= \sum_{i=1}^{I} \sum_{j=1}^{J} (Y_{ij} - \bar{Y}_{i\cdot})^2 + J \sum_{i=1}^{I} (\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2 + 2 \sum_{i=1}^{I} (\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})(\sum_{j=1}^{J} (Y_{ij} - \bar{Y}_{i\cdot}))
\]

The last term of the final expression vanished because the sum of deviations from a mean is zero.

This equation says that the sum of squares of total ($SSTO$) can be decomposed to the sum of squares within treatment groups ($SSW$) plus the sum of squares between treatment groups ($SSB$). Therefore, it is often written as

\[
SSTO = SSB + SSW,
\]

where $SSB = \sum_{i=1}^{I} \sum_{j=1}^{J} (\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2 = J \sum_{i=1}^{I} (\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2$ and $SSW = \sum_{i=1}^{I} \sum_{j=1}^{J} (Y_{ij} - \bar{Y}_{i\cdot})^2$.

Note:

- Sometimes $SSW$ is referred to as $SSE$, and $SSB$ is referred to as $SSTR$.
- Two useful terms: $MSW \equiv SSW/[I(J-1)]$, $MSB \equiv SSB/(I-1)$

**Question:** What do the values of $SSW$ and $SSB$ tell us? We will see that the idea underlying ANOVA is the comparison of the size of various sums of squares. For example, what does the following tell us?

1. Large within group variance and small between group variance
2. Small within group variance and large between group variance.

draw two boxplots here

Intuitively, large within group variance and small between group variance implies small treatment difference; small within group and large between group variance indicates large treatment effects. To see how these sums of squares are related to our hypothesis testing, we need to derive some theoretical results.