10 One-way analysis of variance (ANOVA)

A factor is in an experiment; its values are .

A one-way analysis of variance (ANOVA) tests \( H_0: \mu_1 = \cdots = \mu_I \), where \( I \) is the for one factor, against \( H_A: \text{at least two } \mu_i\text{'s are different} \) by comparing two estimates of \( \sigma^2 \). One estimate is from variation , and the other is from variation . Their ratio is compared to an distribution.

e.g. Here are lengths (mm) for samples of three varieties of Heliconia flowers on Dominica, each fertilized by a different species of hummingbird. The flowers’ forms and birds’ beaks seem to have .

<table>
<thead>
<tr>
<th>Heliconia bihai</th>
<th>( J_i )</th>
<th>( \bar{x}_i )</th>
<th>( s_i )</th>
</tr>
</thead>
<tbody>
<tr>
<td>47.12 46.75 46.81 47.12 46.67 47.43 46.44 46.64</td>
<td>16</td>
<td>47.60</td>
<td>1.21</td>
</tr>
<tr>
<td>Heliconia caribaea red</td>
<td>41.90 42.01 41.93 43.09 41.47 41.69 39.78 40.57</td>
<td>23</td>
<td>39.71</td>
</tr>
<tr>
<td>39.63 42.18 40.66 37.87 39.16 37.40 38.20 38.07</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>38.10 37.97 38.79 38.23 38.87 37.78 38.01</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Heliconia caribaea yellow</td>
<td>36.78 37.02 36.52 36.11 36.03 35.45 38.13 37.10</td>
<td>15</td>
<td>36.18</td>
</tr>
<tr>
<td>35.17 36.82 36.66 35.68 36.03 34.57 34.63</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
| Overall: \( N = 54, \bar{x}_. = 41.07 \)

e.g. The factor here is . It has three levels: .

Notation:

- \( I = \# \)
- \( J_i = i^{th} \text{ sample} \) (\( i = 1 \) to \( I \)); \( N = \sum_{i=1}^{I} J_i = \text{sum of all sample} \)
- \( X_{ij} = i^{th} \text{ sample's } j^{th} \text{ observation} \) ( in plot)
- \( \bar{X}_i = \frac{1}{J_i} \sum_{j=1}^{J_i} X_{ij} = i^{th} \text{ sample's mean} \) ( in plot)
- \( \bar{X}_. = \frac{1}{N} \sum_{i=1}^{I} \sum_{j=1}^{J_i} X_{ij} = \frac{1}{N} \sum_{i=1}^{I} J_i \bar{X}_i = \text{grand sample mean} \) ( in plot)
- \( X_{ij} - \bar{X}_i = \text{residual of } i^{th} \text{ sample's } j^{th} \text{ observation} \) with respect to \( i^{th} \text{ sample's mean} \) (the difference from in plot)

Is the apparent difference in sample means \( \{\bar{X}_1, \bar{X}_2, \bar{X}_3\} \) statistically significant? It matters only the spread of observations. Is it due to , or significant?
Sums of Squares

Decompose an observation \( X_{ij} \) (dot in plot) as

\[
X_{ij} = \bar{X}_.. + (\bar{X}_i - \bar{X}_..) + (X_{ij} - \bar{X}_i). \]

An analysis of variance identity is \( \text{SST} = \ldots + \ldots \).

This partitions the sum of squared deviations of observations from the grand sample mean into the sum of squared deviations of sample means from \( \ldots \) plus the sum of squared deviations of observations from \( \ldots \).

- The total sum of squares is

\[
\text{SST} = \sum_{i=1}^{I} \sum_{j=1}^{J_i} (X_{ij} - \bar{X}_..)^2
\]

- Variation \( \ldots \) samples is measured by the \( \ldots \) sum of squares (SSE)

\[
\text{SSE} = \sum_{i=1}^{I} \sum_{j=1}^{J_i} (X_{ij} - \bar{X}_i)^2 = \sum_{i=1}^{I} (J_i - 1)s_i^2
\]

Its average is the \( \ldots \) square,

\[
\text{MSE} = \frac{\text{SSE}}{N - I}
\]

where \( N - I \) is the degrees of freedom for SSE. (It’s “\( \ldots \)”, not “\( \ldots \”).)

\( s^2 = \text{MSE} \) is an estimate of \( \ldots \).

e.g. MSE is the variance of the residuals (\( \ldots \) minus respective \( \ldots \));

\[
\text{SSE} = \text{MSE} = \ldots
\]

- Variation \( \ldots \) samples is measured by the \( \ldots \) sum of squares,

\[
\text{SSTR} = \sum_{i=1}^{I} J_i(\bar{X}_i - \bar{X}_.)^2
\]

Its average is the \( \ldots \) square,

\[
\text{MSTR} = \frac{\text{SSTR}}{I - 1}
\]

where \( I - 1 \) is the degrees of freedom for SSTR.

A \( \ldots \) MSTR is evidence \( \ldots \) \( H_0 \), while a \( \ldots \) MSTR leaves \( H_0 \) \( \ldots \).


e.g. MSTR measures variation of the \( \ldots \) about the \( \ldots \):

\[
\text{SSTR} = \text{MSTR} = \ldots
\]
Assumptions

- All treatment populations are ________
- All populations all have the same ________
- All observations are ____________

Checks include:

- A normal QQ plot of residuals \{X_{ij} - \bar{X}_i\} should be \approx ________
- The ________ sample standard deviation shouldn’t be more than about ________ as the ________ sample standard deviation.
- A residual plot of residuals \{X_{ij} - \bar{X}_i\} against their respective sample means \bar{X}_i should show reasonably ________ spread across samples and no ________

The F-test of \(H_0: \mu_1 = \cdots = \mu_I\) for One-Way ANOVA

Under \(H_0\), MSTr and MSE are both estimates of \(\sigma^2\), the common variance of the populations.

- MSTr ________ on the truth of \(H_0\):
  - \(H_0\) true \(\implies\) \(\mu_{\text{MSTr}} = \sigma^2\)
    Here is why. Suppose \(H_0\) is true, so \(\mu_1 = \cdots = \mu_I\), and let \(\mu\) denote this common mean. Suppose also, to treat only the easiest case, all the sample sizes \(J_1, \ldots, J_I\), are the same, and let \(J\) denote this common sample size. Then \(\bar{X}_i \sim N(\mu, \sigma^2/J)\), so \(\frac{1}{J} \sum_{i=1}^{I} (\bar{X}_i - \bar{X})^2\) is really \(s_{\bar{X}_i}^2\), an estimate of \(\sigma^2_{\bar{X}_i} = \sigma^2/J\). Multiply both sides by \(J\) and we have that, under \(H_0\), \(\text{MSTr} = Js_{\bar{X}_i}^2\) is an estimate of \(\sigma^2\).
  - \(H_0\) false \(\implies\) \(\mu_{\text{MSTr}} \lessgtr \sigma^2\)

- MSE ________ on the truth of \(H_0\). Either way, \(\mu_{\text{MSE}} = \sigma^2\).

- Use the test statistic \(F = \frac{\text{MSTr}}{\text{MSE}} = \frac{\text{average variation among sample means}}{\text{average variation among individuals in the same sample}}\)
  - \(H_0\) true \(\implies\) \(F\) should be near ________
  - \(H_0\) false \(\implies\) \(F\) should be ________

Under \(H_0\), \(F \sim F_{I-1,N-I}\). (It’s “________”, not “\(N-1\)”.)

e.g. Here is R code that simulates an \(F_{4-1,7-1}\) distribution:

```r
F = replicate(n=100000, expr={x=rnorm(4); y=rnorm(7); var(x)/var(y)})
plot(density(F), xlim=c(0, 6)) # Show simulated F values.
curve(df(x, 4-1, 7-1), add=TRUE, col="red") # Show \(F_{4-1,7-1}\) in red.
```
e.g. For the flower data, find $F$ and its $P$-value, and draw a conclusion.

$$F = \quad \Rightarrow \quad P\text{-value} = \quad \Rightarrow$$

The preceding work can be summarized in this ANOVA table:

<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>Treatment</td>
<td>2</td>
<td>1084.00</td>
<td>541.50</td>
<td>259.09</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>Error</td>
<td>51</td>
<td>106.58</td>
<td>2.09</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>53</td>
<td>1190.58</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Rejecting $H_0: \mu_1 = \cdots = \mu_I$, doesn’t tell

(A balanced experiment has equal sample sizes and is more robust against ______________ than an unbalanced one. $I = 2 \quad \Rightarrow \quad F\text{-test is __________ to } t\text{-test of } H_0: \mu_X - \mu_Y = 0$, with $F = ____$.)

**ANOVA review, by hand**

e.g. Here is a quick one-way ANOVA to test $H_0: \mu_A = \mu_B = \mu_C$.

<table>
<thead>
<tr>
<th>Sample</th>
<th>Data</th>
<th>$\bar{x}_{i.}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0 1 2</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>1 2 3</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>5 6 7</td>
<td></td>
</tr>
</tbody>
</table>

| Overall $\bar{x}_{..} =$ |

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
<th>$F$</th>
<th>$P$</th>
<th>$P(F_{I-1,N-I} &gt; ____)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment</td>
<td>$I-1$</td>
<td>SSt r $= \sum_{i=1}^{I} J_i (\bar{X}<em>{i.} - \bar{X}</em>{..})^2$</td>
<td>MST r $= \frac{SSt r}{I-1}$</td>
<td>$F = \frac{MSt r}{MSE}$</td>
<td>$P(F_{I-1,N-I} &gt; ____)$</td>
<td></td>
</tr>
<tr>
<td>Error</td>
<td>$N-I$</td>
<td>SSE $= \sum_{i=1}^{I} \sum_{j=1}^{J_i} (X_{ij} - \bar{X}_{i.})^2$</td>
<td>MSE $= \frac{SSE}{N-I}$ ($= s^2$)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>$N-1$</td>
<td>SST $= \sum_{i=1}^{I} \sum_{j=1}^{J_i} (X_{ij} - \bar{X}_{..})^2$</td>
<td></td>
<td>$= SSt r + SSE$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Conclusion: ____________________________
The problem of multiple comparisons

- mind-reading
- butterfly effect
- jelly beans
- salmon
- “Why Published Research Findings Are ________”
- Publication ________

Pairwise comparisons after one-way ANOVA rejects $H_0 : \mu_1 = \cdots = \mu_I$

A problem:

- One-way ANOVA tests $H_0 : \mu_1 = \cdots = \mu_I$ (for $I$ samples), but rejecting $H_0$ doesn’t tell ________ are different.
- If we calculate a level-$(1-\alpha)$ confidence interval for the difference of each pair of means, $\mu_i - \mu_j$, the probability of ________ intervals containing their respective differences ________ is usually ________.

The Tukey-Kramer Method of Multiple Comparisons

Experts say the Tukey-Kramer level-$(1-\alpha)$ ________ confidence intervals for $\mu_i - \mu_j$ are

$$\bar{X}_i - \bar{X}_j \pm q_{I,N-I,\alpha} \sqrt{\frac{\text{MSE}}{2} \left( \frac{1}{J_i} + \frac{1}{J_j} \right)}$$

where $q_{I,N-I,\alpha}$ cuts off a right tail area $\alpha$ from the ________ distribution with $I$ and $N-I$ degrees of freedom.

The Studentized Range $q$ table gives this number in column ________, row $\nu_2 = ________$, and subrow ________. R gives $q_{I,N-I,\alpha}$ via, e.g.,

```r
alpha=.05; I=3; N=9; qtukey(p=1-alpha, nmeans=I, df=N-I)
```

which returns $\approx 4.34$.

We have $(1-\alpha)$ confidence that these intervals contain $\mu_i - \mu_j$ ________.

Reject $H_0 : \mu_i - \mu_j = 0 \iff$ the interval ________ $\iff |\bar{X}_i - \bar{X}_j| > q_{I,N-I,\alpha} \sqrt{\frac{\text{MSE}}{2} \left( \frac{1}{J_i} + \frac{1}{J_j} \right)}$.

A convenient way to indicate which pairs of means differ is to sort treatment means and then add letter codes (A, B, C, ...) so that two treatments ________ if they do not differ significantly.
e.g. Which pairs of population means from the “quick one-way ANOVA” example, above, differ at the 5% level?

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Sample mean</th>
<th>Letter</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

e.g. Which pairs of population means from the flowers example, above, differ at the 5% level? (Hint: The R code

\[
\text{alpha}=.05; \ I=3; \ N=54; \ \text{qtukey}(p=1-\text{alpha}, \ nmeans=I, \ df=N-I)
\]
returns \(\approx 3.41\).)

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Sample mean</th>
<th>Letter</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The Kruskal-Wallis test, an ANOVA alternative not requiring normal populations

The Kruskal-Wallis test generalizes the rank-based Wilcoxon Rank Sum test to the case of \_______________ groups. As a test of equality of \_______________, it requires independent SRSs from the treatment populations, which must have the same \_______________. It has low power for tiny sample size but power almost as good as ANOVA’s power for large samples. e.g. See the R code, below.
R code for ANOVA

# Here is one way to do the calculations and make the graphs for "10 ANOVA.

bihai = c(47.12,46.75,46.81,47.12,46.67,47.43,46.44,46.64,48.07,48.34,48.15,
          50.26,50.12,46.34,46.94,48.36)
red = c(41.90,42.01,41.93,43.09,41.47,41.69,39.78,40.57,39.63,42.18,40.66,
       37.87,39.16,37.40,38.20,38.07,38.10,37.97,38.79,38.23,38.87,37.78,38)
yellow = c(36.78,37.02,36.52,36.11,36.03,35.45,38.13,37.10,35.17,36.82,36.66,
          35.68,36.03,34.57,34.6)
(flower.lengths = c(bihai, red, yellow))
(N=length(flower.lengths))

(bihai.labels = rep(x="bihai", times=length(bihai)))
red.labels = rep(x="red", times=length(red))
yellow.labels = rep(x="yellow", times=length(yellow))
(flower.labels = c(bihai.labels, red.labels, yellow.labels)) # vector of words

# R wants a categorical variable stored as a "factor" vector.
(flower.species = factor(flower.labels))

# Here the formula "flower.lengths ~ flower.species" means
# "make three stripcharts of flower.lengths, one for each
# category in flower.species."
stripchart(flower.lengths ~ flower.species, vertical=TRUE)

# Here are sample means. This tapply() function call gives the sample
# mean flower.lengths for each species in the flower.species variable.
tapply(X=flower.lengths, INDEX=flower.species, FUN=mean)

# Here are sample standard deviations. This tapply() function call
# gives the sample sd() flower.lengths for each species in the
# flower.species variable.
tapply(X=flower.lengths, INDEX=flower.species, FUN=sd)

# Make ANOVA model and table.
flower.model = aov(flower.lengths ~ flower.species)
anova(flower.model)

# Plot residuals vs. fitted values.
plot(residuals(flower.model) ~ fitted(flower.model), ylab = "Residuals",
     xlab = "Fitted Values", main = "Residuals vs Fitted Values")

# Make QQ plot of residuals.
qqnorm(residuals(flower.model), main = "QQ Plot of Residuals")
# Run pairwise tests.
TukeyHSD(flower.model, conf.level = 0.95)

# find Tukey multiplier #
alpha=.05; I=3; N=54; qtukey(p=1-alpha, nmeans=I, df=N-I)

# Kruskal-Wallis, a rank-based alternative to ANOVA that does not
# require normal populations
kruskal.test(flower.lengths ~ flower.species)

# Here is a check on our "ANOVA review, by hand" example.
data = c(c(0, 1, 2), c(1, 2, 3), c(5, 6, 7))
sample = factor(c(rep(x="1", times=3), rep(x="2", times=3), rep(x="3", times=3)))
review.model = aov(data ~ sample)
anova(review.model)
TukeyHSD(review.model, conf.level = 0.95)