Today: Chapter 5 (end)

• Analysis of variance (ANOVA) as the generalization of inference for 3 or more means
• Conditions for running ANOVA
• ANOVA in R
• Danger of running multiple tests, and how to correct this using Bonferroni correction
## Testing for Means

<table>
<thead>
<tr>
<th>Situation</th>
<th>1 Population</th>
<th>2 Populations</th>
<th>≥ 3 Populations</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Hypotheses</strong></td>
<td>$H_0: \mu = \mu_0$</td>
<td>$H_0: \mu_1 = \mu_2$</td>
<td>$H_0: \mu_1 = \ldots = \mu_k$</td>
</tr>
<tr>
<td></td>
<td>$H_A: \mu \neq \mu_0$</td>
<td>$H_A: \mu_1 \neq \mu_2$</td>
<td>$H_A$: Some $\mu_i$ different</td>
</tr>
<tr>
<td><strong>Test Statistic</strong></td>
<td>$T = \frac{\bar{x} - \mu_0}{SE}$</td>
<td>$T = \frac{\bar{x}_1 - \bar{x}_2 - 0}{SE}$</td>
<td>$F = \frac{MSG}{MSE}$</td>
</tr>
<tr>
<td><strong>Sampling Distribution</strong></td>
<td>$t_{n-1}$</td>
<td>$t_{\min(n_1-1,n_2-1)}$</td>
<td>$f_{k-1,n-k}$</td>
</tr>
<tr>
<td><strong>Language</strong></td>
<td>1 sample $t$-test</td>
<td>2 sample $t$-test</td>
<td>ANalysis Of VAriance (ANOVA)</td>
</tr>
</tbody>
</table>
Why Do We Need Some Complicated New Test?

It is natural to think that you should compare, say, 4 means by comparing each population with each other, doing many 2-sample $t$-tests.

Comparing 4 means in sets of 2 requires $\binom{4}{2} = 6$ different tests. This number of tests quickly escalates, since $\binom{k}{2} = \frac{k(k-1)}{2} \asymp k^2$.

When you have many tests, one is likely to show a significant difference just because of random variation that occurs (and not because of a true effect).
ANOVA and the F-Distribution

An ANOVA runs a single test to see if the means are all equal \((H_0)\), with the alternative that (at least) one mean differs \((H_A)\).

Drawback: If you move to \(H_A\), you don’t know which mean differs or how many differ. (At this point, we often move to pairwise comparison with a twist!)

To use ANOVA, you must meet these conditions:

- Data are independent within and across groups
- The data in each group are nearly normal
- The spreads in each group are roughly equal

Remark: The last condition is critical and new to us. If it is not met, you must use techniques beyond the scope of this course.
How Do I Check All These Conditions?

**Independence:** We usually check for Randomization and < 10% Condition in each population.

**Normality and Equal Spreads:** Side-by-side boxplots, data tables (or quantile-quantile plots!).

On-base percentage (which is roughly equal to the fraction of times a player gets on base or hits a home run) of players who are:

- **OF:** outfield
- **IF:** infield
- **DH:** designated hitter
- **C:** catcher

<table>
<thead>
<tr>
<th>Position</th>
<th>OF</th>
<th>IF</th>
<th>DH</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample size ($n_i$)</td>
<td>120</td>
<td>154</td>
<td>14</td>
<td>39</td>
</tr>
<tr>
<td>Sample mean ($\bar{x}_i$)</td>
<td>0.334</td>
<td>0.332</td>
<td>0.348</td>
<td>0.323</td>
</tr>
<tr>
<td>Sample SD ($s_i$)</td>
<td>0.029</td>
<td>0.037</td>
<td>0.036</td>
<td>0.045</td>
</tr>
</tbody>
</table>
From Data to the $F$-Value

The $F$-statistic is:

$$F = \frac{MSG}{MSE}.$$  

– The **Mean Square across Groups** is an estimate of the variability we see in the $k$ different means (for each group).

$$MSG = \frac{1}{k-1} \sum_{i=1}^{k} n_i (\bar{x}_i - \bar{x})^2,$$

If $H_0$ is true, this should be small.

– The **Mean Square Error** is a (combined) measure of the variability within each group.

$$MSE = \frac{1}{n-k} \sum_{i=1}^{k} (n_i - 1) s_i^2.$$
An Explosive Example

When lava hardens, it reveals the direction of Earth’s magnetic field. Taking three lava samples from Mt Etna eruptions from the years 1669, 1780, and 1865, volcanologists recorded the below data on the direction of Earth’s magnetic field. Do these data support the idea that the direction has changed over time?

<table>
<thead>
<tr>
<th>Year</th>
<th>1669</th>
<th>1780</th>
<th>1865</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>57.8</td>
<td>57.9</td>
<td>52.7</td>
</tr>
<tr>
<td></td>
<td>60.2</td>
<td>55.2</td>
<td>53.0</td>
</tr>
<tr>
<td></td>
<td>60.3</td>
<td>54.8</td>
<td>49.4</td>
</tr>
</tbody>
</table>

Let $\mu_1$, $\mu_2$ and $\mu_3$ be the angle of Earth’s magnetic field in 1669, 1780, and 1865.

Set

$H_0: \mu_1 = \mu_2 = \mu_3 \quad H_A: \text{Some mean is different}$
ANOVA in R

Given how R does the ANOVA, you must have your data in this format:

- 1 column for the values being studied (quantitative variable = angle of magnetic field)
- 1 column specifying what group each observation belongs to (categorical variable = year, where A = 1669, B = 1780, C = 1865).

```
> results = aov(angle ~ year, data = volcano)
> summary(results)

            Df Sum Sq Mean Sq F value   Pr(>F)
year         2 90.032  45.016  15.278 0.00442 **
Residuals    6 17.669  2.945
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```
ANOVA in R

> summary(results)

<table>
<thead>
<tr>
<th></th>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>year</td>
<td>2</td>
<td>90.03</td>
<td>45.01</td>
<td>15.28</td>
<td>0.00442</td>
</tr>
<tr>
<td>Residuals</td>
<td>6</td>
<td>17.67</td>
<td>2.95</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

**Df:** There are two degrees of freedom related to this problem:
- $df_G = k - 1$ (here $3 - 1 = 2$)
- $df_E = n - k$ (here $9 - 3 = 6$)

**Mean Sq:** $MSG = 45.01$ and $MSE = 2.95$

**F:** $F = \frac{MSG}{MSE} = 15.28$

**Pr(>F):** This is the $p$-value for the hypothesis test. Here, we would reject the null if $\alpha = 5\%$.
The $p$-value is the area under the $F$-distribution with the degrees of freedom $df_G$ and $df_E$. 
The F-Distribution Family

The $F$-distribution family is critically important when doing ANOVA and indexed by two degree of freedom parameters.

$$F = \frac{\text{MSG}}{\text{MSE}}$$

With an ANOVA ($F$-test), you always shade to the right, and there is no idea of one-sided or two-sided alternative hypotheses.
Your Turn!

A researcher is doing an $F$-test on the $F$-distribution with $df_G = 7$ and $df_E = 112$. How many groups is the researcher’s data set divided into?
Answer: $df_G = k - 1 = 7$, so $k = 8$.

How many total data observations did the researcher make?
Answer: $df_E = n - k = n - 8 = 112$, so $n = 120$. 
Your Turn!

You collect salary info on 1200 total people split across 5 different states in the US. When you run your ANOVA ($F$-test), what curve is the $p$-value found on?

Answer: On the plot of $F_{k-1,n-k} = F_{5-1,1200-5} = F_{4,1995}$. (the order is $F_{df_G,df_E}$)
Your Turn!

If an ANOVA ($F$-test) suggests a move to the alternative hypothesis, then we will have identified the mean that is different than the rest.

- True
- False

Answer: False. $H_A$ only says that some difference exists, not what it is.

If an ANOVA ($F$-test) suggests that some mean is different, then we will be able to find it by using pairwise-comparison of means.

- True
- False

Answer: False. Amazingly, it is possible for the ANOVA to suggest a difference is present, but for all pairwise tests to be unable to detect the difference!

This is like sensing corruption is present in the government, but being unable to identify the particular source of it.
Your Turn!

Do people who drink different amounts of coffee tend to exercise different amounts?
Researchers looked at cups of coffee consumed and average exercise per week (measured in metabolic equivalent tasks, METs) in 50,739 women.

<table>
<thead>
<tr>
<th>Caffeinated coffee consumption</th>
<th>≤ 1 cup/week</th>
<th>2-6 cups/week</th>
<th>1 cup/day</th>
<th>2-3 cups/day</th>
<th>≥ 4 cups/day</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>18.7</td>
<td>19.6</td>
<td>19.3</td>
<td>18.9</td>
<td>17.5</td>
<td>17.5</td>
</tr>
<tr>
<td>SD</td>
<td>21.1</td>
<td>25.5</td>
<td>22.5</td>
<td>22.0</td>
<td>22.0</td>
<td>22.0</td>
</tr>
<tr>
<td>n</td>
<td>12,215</td>
<td>6,617</td>
<td>17,234</td>
<td>12,290</td>
<td>2,383</td>
<td>50,739</td>
</tr>
</tbody>
</table>

1) Set up parameters and hypotheses.
Let $\mu_i$ ($i = 1, \ldots, 5$) be the mean weekly MET levels in women in the five different coffee categories.
( Remember, $\mu_i$ should measure a numeric variable. And here, for the groups, coffee consumption has been made into an ordinal variable)

Set

$H_0$: All the $\mu_i$’s are equal.  $H_A$: Some $\mu_i$ is different.
2) Discuss the conditions for inference and whether they are met.

- Independence within and across the groups: we don’t know.
- Normality: we don’t know, but with samples this large, we’re ok.
- Roughly constant spread in groups: The SD’s look somewhat similar, be we are worried.

<table>
<thead>
<tr>
<th></th>
<th>≤ 1 cup/week</th>
<th>2-6 cups/week</th>
<th>1 cup/day</th>
<th>2-3 cups/day</th>
<th>≥ 4 cups/day</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>18.7</td>
<td>19.6</td>
<td>19.3</td>
<td>18.9</td>
<td>17.5</td>
<td></td>
</tr>
<tr>
<td>SD</td>
<td>21.1</td>
<td>25.5</td>
<td>22.5</td>
<td>22.0</td>
<td>22.0</td>
<td></td>
</tr>
<tr>
<td>n</td>
<td>12,215</td>
<td>6,617</td>
<td>17,234</td>
<td>12,290</td>
<td>2,383</td>
<td>50,739</td>
</tr>
</tbody>
</table>
Since 0.0003 < 0.05, we reject $H_0$ in favor of $H_A$. It appears that MET levels are not the same across different levels of coffee consumption.
Why Use a Single Test Instead Multiple Pairwise Tests?

Suppose $H_0$ is actually true and that $\alpha = 0.05$.

If $\bar{x}$ falls in the shaded area, it means that $p < 0.05$. Thus, we (incorrectly) reject $H_0$ (Type I Error).

Otherwise, we will (correctly) retain $H_0$ (no error made).

So $P$(Type I Error) $= \alpha$ by construction.
Running Multiple Tests

Suppose we run 6 tests on the same situation. If there really is nothing going on ($H_0$ is true), how likely are we to wrongly believe something interesting is going on?

We know $P$(Type I Error) = $\alpha = 0.05$ on each single test. Let $X$ be the number of errors we make in our 6 tests.

$$P(X \geq 1) = 1 - P(X = 0) = 1 - (0.95)^6 = 0.265.$$ 

If we run 6 tests, then we think we have an interesting finding (that is, we move to $H_A$) 26.5% of the time when no real effect is present!

If we run $k$ tests, then

$$P(X \geq 1) = 1 - P(X = 0) = 1 - (1 - \alpha)^k.$$
The danger of Multiple Tests

<table>
<thead>
<tr>
<th># of Test Runs (k)</th>
<th>$P(X \geq 1)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>0.265</td>
</tr>
<tr>
<td>10</td>
<td>0.401</td>
</tr>
<tr>
<td>20</td>
<td>0.642</td>
</tr>
<tr>
<td>40</td>
<td>0.871</td>
</tr>
<tr>
<td>80</td>
<td>0.983</td>
</tr>
</tbody>
</table>

$P(X \geq 1) = 1 - (1 - 0.05)^k$.

As an example, doing pairwise tests on 8 groups yields $\binom{8}{2} = 28$ tests and will have 76.2% chance of getting a false finding (Type I Error).

Bonferroni Correction: If you want to perform $k$ total tests on some situation, use an altered significance level of $\frac{\alpha}{k}$ for each of the tests.

<table>
<thead>
<tr>
<th># of Test Runs (k)</th>
<th>$P(X \geq 1)$ Using Bonferroni</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>0.049</td>
</tr>
<tr>
<td>10</td>
<td>0.049</td>
</tr>
<tr>
<td>20</td>
<td>0.049</td>
</tr>
<tr>
<td>40</td>
<td>0.049</td>
</tr>
<tr>
<td>80</td>
<td>0.049</td>
</tr>
</tbody>
</table>

$P(X \geq 1) = 1 - (1 - 0.05/k)^k$. 
Bonferroni Correction

Bonferroni Correction: If you want to perform $k$ total tests on some situation, use an altered significance level of $\alpha/k$ for each of the tests.

By construction, using the Bonferroni correction $\alpha/k$ always yields a Type I Error smaller than $\alpha$.

**But nothing is free:** Doing so, the power of each test gets worse (since the level of confidence is better). Hence, one has less chance to actually discover something interesting (conclude $H_A$) if $H_0$ is actually not true!