Math 183
Statistical Methods

Eddie Aamari
S.E.W. Assistant Professor

eaamari@ucsd.edu
math.ucsd.edu/~eaamari/
AP&M 5880A

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

Situation	1 Population	2 Populations	≥ 3 Populations
Hypotheses	$H_0\colon\ \mu = \mu_0$ $H_A\colon\ \mu \neq \mu_0$	$H_0\colon\ \mu_1 = \mu_2$ $H_A\colon\ \mu_1 \neq \mu_2$	$H_0\colon\ \mu_1 = \ldots = \mu_k$ H_A: Some μ_i different
Test Statistic	$T = \dfrac{\bar{x} - \mu_0}{SE}$	$T = \dfrac{\bar{x}_1 - \bar{x}_2 - 0}{SE}$	$F = \dfrac{MSG}{MSE}$
Sampling Distribution	t_{n-1}	$t_{\min(n_1-1,n_2-1)}$	$f_{k-1,n-k}$
Language	1 sample t-test	2 sample t-test	ANalysis Of VAriance (ANOVA)

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.

$$\begin{tikzcd}
\mu_1 \arrow[r, leftrightarrow] \arrow[d, leftrightarrow] \arrow[dr, leftrightarrow] & \mu_2 \arrow[d, leftrightarrow] \arrow[dl, leftrightarrow] \\
\mu_3 \arrow[r, leftrightarrow] & \mu_4
\end{tikzcd}$$

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).

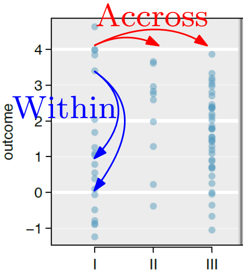
ANOVA and the F-Distribution

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 accross groups
- The data in each group are nearly normal
- The spreads in each group are roughly equal

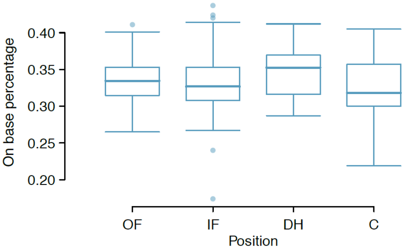


Remark: The last condition is critical and new to us. If it is not met, you must use techniques beyond the scope of this course.

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!).



	OF	IF	DH	C
Sample size (n_i)	120	154	14	39
Sample mean ($\bar{x}_i$)	0.334	0.332	0.348	0.323
Sample SD (s_i)	0.029	0.037	0.036	0.045

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

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).

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

If H_0 is true, this should be small.

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

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

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?

Year

1669	1780	1865
57.8	57.9	52.7
60.2	55.2	53.0
60.3	54.8	49.4

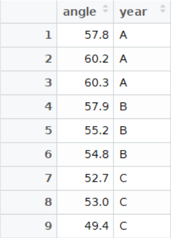
Let μ_1, μ_2 and μ_3 be the angle of Earth's magnetic field in 1669, 1780, and 1865.

Set

$$H_0\colon\ \mu_1 = \mu_2 = \mu_3 \qquad\qquad H_A\colon\ \text{Some mean is different}$$

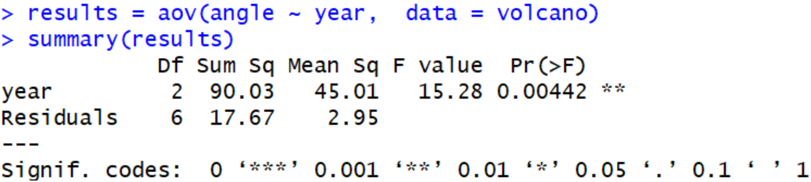
ANOVA in R

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



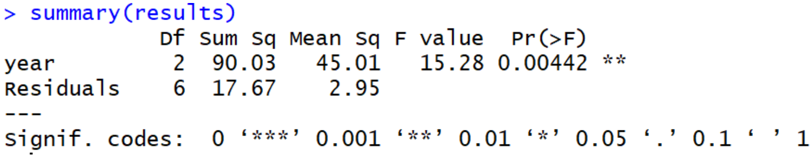
	angle	year
1	57.8	A
2	60.2	A
3	60.3	A
4	57.9	B
5	55.2	B
6	54.8	B
7	52.7	C
8	53.0	C
9	49.4	C

- 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  90.03   45.01   15.28 0.00442 **
Residuals    6  17.67    2.95                   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

ANOVA in R



```
> summary(results)
            Df Sum Sq Mean Sq F value  Pr(>F)
year         2  90.03   45.01   15.28 0.00442 **
Residuals    6  17.67    2.95
---
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 $9 - 3 = 6$)

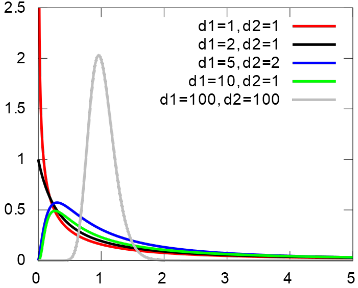
Mean Sq: $MSE = 45.01$ and $MSE = 2.95$

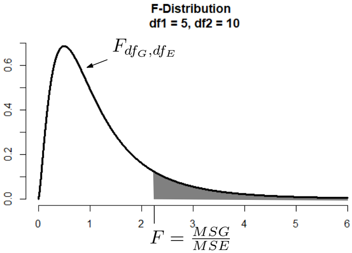
F: $F = \dfrac{MSG}{MSE} = 15.28$

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.





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 un 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.

	Caffeinated coffee consumption					
	$\leq$ 1 cup/week	2-6 cups/week	1 cup/day	2-3 cups/day	$\geq$ 4 cups/day	Total
Mean	18.7	19.6	19.3	18.9	17.5	
SD	21.1	25.5	22.5	22.0	22.0	
n	12,215	6,617	17,234	12,290	2,383	50,739

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

Set

$$H_0\text{: All the } \mu_i\text{'s are equal.} \qquad H_A\text{: Some } \mu_i \text{ is different.}$$

	Caffeinated coffee consumption					
	$\leq$ 1 cup/week	2-6 cups/week	1 cup/day	2-3 cups/day	$\geq$ 4 cups/day	Total
Mean	18.7	19.6	19.3	18.9	17.5	
SD	21.1	25.5	22.5	22.0	22.0	
n	12,215	6,617	17,234	12,290	2,383	50,739

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.

	Caffeinated coffee consumption					
	$\leq$ 1 cup/week	2-6 cups/week	1 cup/day	2-3 cups/day	$\geq$ 4 cups/day	Total
Mean	18.7	19.6	19.3	18.9	17.5	
SD	21.1	25.5	22.5	22.0	22.0	
n	12,215	6,617	17,234	12,290	2,383	50,739

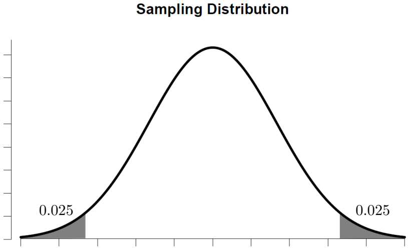
	Df	Sum Sq	Mean Sq	F value	Pr(>F)
coffee	$\boxed{k - 1 = 4}$	$\boxed{10,508}$	$\boxed{\frac{10,508}{4} = 2,627}$	$\boxed{\frac{2,627}{504} = 5.2}$	$\boxed{0.0003}$
Residuals	$\boxed{n - k = 50,734}$	$25,564,819$	$\boxed{\frac{25,564,819}{50,734} = 504}$		
Total	$\boxed{4 + 50,734}$	$25,575,327$			

```
> pf(5.2, df1=4, df2 = 50734, lower.tail = F)
 [1] 0.0003475405
```

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 $P(\text{Type I Error}) = \alpha$ 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 $P(\text{Type I Error}) = \alpha = 0.05$ 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

# of Test Runs (k)	$P(X \geq 1)$
6	0.265
10	0.401
20	0.642
40	0.871
80	0.983

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

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 $\dfrac{\alpha}{k}$ for each of the tests.

# of Test Runs (k)	$P(X \geq 1)$ Using Bonferroni
6	0.049
10	0.049
20	0.049
40	0.049
80	0.049

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

Bonferroni Correction

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

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

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!