Chapter 14: Analysis of Variance

Yu Yang

School of Statistics
University of Minnesota

November 7, 2022

Example 14.1

What is the most effective method for treating anorexia? 72 anorexic teenage girls were randomly assigned to one of three treatments.

The first treatment group received cognitive behavorial therapy in which girls are taught to identify the thinking that triggers their eating disorder and to replace it with other thoughts means to prevent this behavior.

The second treatment group attended family therapy and the third group served as a control group and did not receive any therapy. Each girl was weighed before her treatment began and weighed again upon completion of the treatment.

These data can be found at
`http://stat.umn.edu/~wuxxx725/data/anorexia.txt`

Example 14.1

Let:

μ_1 = mean weight gain after completing cognitive behavior therapy
μ_2 = mean weight gain after completing family therapy
μ_3 = mean weight gain with no treatment

Are any of the treatments better of worse than the others? That is, are there any significant differences among μ_1, μ_2 and μ_3?

Goal

Use Analysis of Variance (ANOVA) to compare the means of **more than two groups** .

Why not just do a two sample t-test for each pair of means?

1. If g was the number of groups, we would have to do $\binom{g}{2}$ pairwise tests. BUT: Type I error rate increases as the number of tests increases.

2. This would only let us to compare 2 groups as a time, but we want know if means differ among all groups and we want to answer this with one test.

One-Way ANOVA

Notation:
g = numbers of groups we are comparing
n_i = sample size of the ith group
N = overall sample size $= n_1 + n_2 + \ldots n_g$
$\bar{y}_i$ = sample mean of the ith group
$\bar{y}$ = overall sample mean (sample mean of all the observations)
s_i = sample standard deviation of the ith group

One-Way ANOVA compares population means among g different groups.

One-Way ANOVA

Hypotheses:
$H_0 : \mu_1 = \mu_2 = \cdots = \mu_g$
H_a : at least 2 of the population means are unequal

To test H_0 we do an analysis of variance. That is, we compare

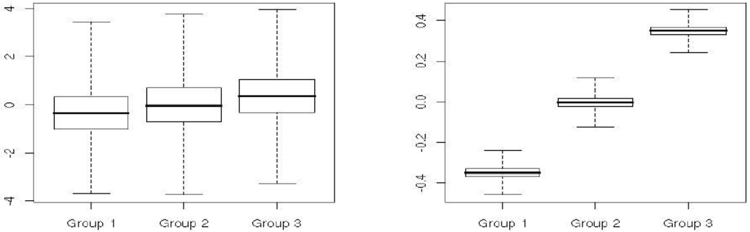
1. variation *between* groups, i.e. how much do $\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_g$ differ?
2. variation *within* groups, i.e. how much variability is there within each separate group?

One-Way ANOVA

If our goal is to compare the **means** of several populations, why are we doing analysis of **variance**? In the following box plots, the means of the groups are same between the two graphs.

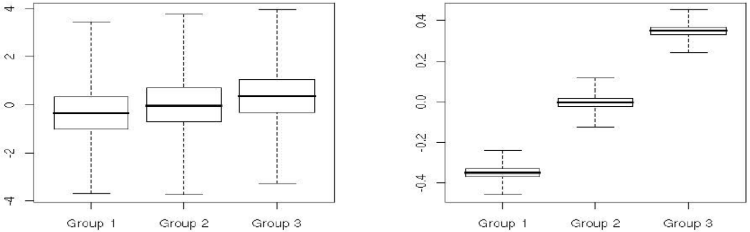
One-Way ANOVA

If our goal is to compare the **means** of several populations, why are we doing analysis of **variance**? In the following box plots, the means of the groups are same between the two graphs.



One-Way ANOVA

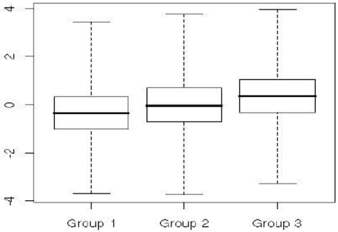
If our goal is to compare the **means** of several populations, why are we doing analysis of **variance**? In the following box plots, the means of the groups are same between the two graphs.

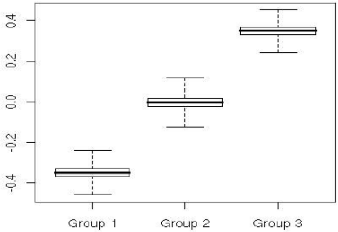


1. Variation between groups is *small* compared to variability within groups
2. Variation between groups is *large* compared to variability within groups

One-Way ANOVA

If our goal is to compare the **means** of several populations, why are we doing analysis of **variance**? In the following box plots, the means of the groups are same between the two graphs.





1. Variation between groups is *small* compared to variability within groups
2. Variation between groups is *large* compared to variability within groups

Conclusions: Differences in plot 1 means are probably due to chance whereas those in plot 2 are probably due to true difference among groups and not to chance. Two means are significantly different only if their difference is large **relative** to the variability within each group.

Measuring Variability

Between Groups

SSG = Sum of squared deviations for groups = $\sum_{i=1}^{g} n_i(\bar{y}_i - \bar{y})^2$

$MSG = \frac{SSG}{g-1}$ = a measure of how much means vary from group to group

Measuring Variability

Between Groups

SSG = Sum of squared deviations for groups = $\sum_{i=1}^{g} n_i(\bar{y}_i - \bar{y})^2$

$MSG = \frac{SSG}{g-1}$ = a measure of how much means vary from group to group

Within Groups

SSE = Sum of Squared Deviations for Error = $\sum_{i=1}^{g}(n_i - 1)s_i^2$

$MSE = \frac{SSE}{N-g}$ = a measure of how much observations vary within each group.

MSE is also an estimate of σ^2, the population variance for each group.

Measuring Variability

Between Groups

SSG = Sum of squared deviations for groups $= \sum_{i=1}^{g} n_i(\bar{y}_i - \bar{y})^2$

$MSG = \frac{SSG}{g-1}$ = a measure of how much means vary from group to group

Within Groups

SSE = Sum of Squared Deviations for Error $= \sum_{i=1}^{g}(n_i - 1)s_i^2$

$MSE = \frac{SSE}{N-g}$ = a measure of how much observations vary within each group.
MSE is also an estimate of σ^2, the population variance for each group.

Overall Variability

$SST = SSG + SSE = \sum(y - \bar{y})^2$ = a measure of how much observations vary from the overall mean.

Nov 7 Lecture Stopped Here

The ANOVA F Test Statistic

The F statistic measures how compatible the data is with the null $H_0 : \mu_1 = \mu_2 = \cdots = \mu_g$ by comparing the variability between groups to the variability within groups.

$$\boxed{F = \frac{\text{Variability Between Groups}}{\text{Variability Within Groups}} = \frac{MSG}{MSE}}$$

Interpretation: When F is large, variability between groups is large compared to variability within groups. Large F suggests that at least 2 of the means are different.

ANOVA Table

Source of Variation	df	SS	MS	F
Group				
Error				
Total				

ANOVA Table

Source of Variation	df	SS	MS	F
Group	$g-1$	SSG	$MSG = \dfrac{SSG}{g-1}$	$F = \dfrac{MSG}{MSE}$
Error	$N-g$	SSE	$MSE = \dfrac{SSE}{N-g}$	
Total	$N-1$	SST		

The One-Way ANOVA F Test

Assumptions:

1. g independent random sample from g populations
2. Normality: Each population has a normal distribution with unknown mean
3. Equal Variance: Each population has (unknown) equal standard deviation σ.

Note:

- F test is robust to departures from normality and equal variance assumptions
- Graphical methods can be used to check these assumptions

Hypothesis:

$H_0 : \mu_1 = \mu_2 = \cdots = \mu_g$
H_a : at least 2 of the population means are unequal

The One-Way ANOVA F test

Test Statistic

$$\boxed{F = \frac{MSG}{MSE} \sim F_{g-1,N-g} \text{ when } H_0 \text{ is true}}$$

p-value

Recall that the larger F is, the more the evidence *against* H_0

$$\boxed{\text{p}-\text{value} = P(F_{g-1,N-g} \geq F)}$$

In R: `pf(F, df1 = g-1, df2 = N-g, lower.tail = FALSE)`

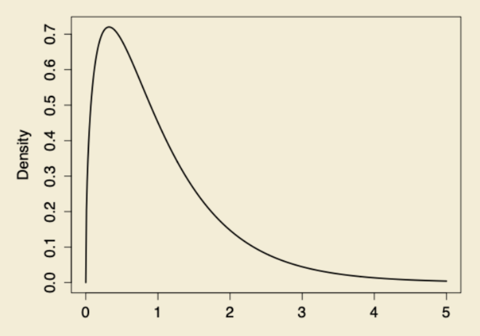
Conclusion:

If p-value $< \alpha$, reject H_0

If p-value $\geq \alpha$ fail to reject H_0

The F distribution

The $F_{3,60}$ density curve



1. Non-negative
2. Right Skewed
3. Mean is approximately 1 (exact value: $\frac{df_2}{df_2-2}$ for $df_2 > 2$)
4. Mode is also approximately 1 (exact value: $\frac{df_1-2}{df_1}\frac{df_2}{df_2+2}$ for $df_1 > 2$)

Calculating F values

We will only be using R to calculate p-values

$P(F_{2,20} > 3.493)$

```
> pf(3.493, df1 = 2, df2 = 20, lower.tail = FALSE)
[1] 0.04999364
```

$P(F_{5,15} > 7.42)$

```
> pf(7.42, df1 = 5, df2 = 15, lower.tail = FALSE)
[1] 0.001103249
```

Example 14.1 Continued

Recall that we want to compare anorexia treatment methods. Let

change = weight change
therapy = treatment groups (3 groups)

Assumptions

Example 14.1 Continued

Recall that we want to compare anorexia treatment methods. Let

change = weight change
therapy = treatment groups (3 groups)

Assumptions
Independent random samples, and box plot indicates assumption of normality and equal variance are OK.

Hypotheses

Example 14.1 Continued

Recall that we want to compare anorexia treatment methods. Let

change = weight change
therapy = treatment groups (3 groups)

Assumptions
Independent random samples, and box plot indicates assumption of normality and equal variance are OK.

Hypotheses
$H_0 : \mu_1 = \mu_2 = \mu_3$
H_a : at least 2 of the means are unequal

Example 14.1 Continued

We are given $SSG = 614.6$ and $SSE = 3910.7$ and $N = 72$. Fill in the ANOVA Table:

Example 14.1 Continued

We are given $SSG = 614.6$ and $SSE = 3910.7$ and $N = 72$. Fill in the ANOVA Table:

Source of variation	df	SS	MS	F
groups	**2**	614.6	**307.3**	**5.42**
error	**69**	3910.7	**56.7**	
Total	**71**	**4525.3**		

Example 14.1 Continued

Test Statistic:
From the table, we calculated

Example 14.1 Continued

Test Statistic:
From the table, we calculated $F = 5.42$

p-**value**
p-value $= P(F_{2,69} > 5.42)$
$=$ `pf(5.42, df1 = 2, df2 = 69, lower.tail = FALSE)`
$= 0.006511565$

Conclusion:
p-value $< .05$, thus we reject H_0 and conclude that the mean weight gain is different for at least two of the treatment groups.

But, which ones?

Example 14.1 in R

```
> dat <- read.table("http://www.stat.umn.edu/~wuxxx725/data/anorexia.txt",
+                   header = TRUE)
> # ANOVA format: aov(numerical var ~ group var)
> mod <- aov(change ~ therapy, data = dat)
> summary(mod)
            Df Sum Sq Mean Sq F value Pr(>F)
therapy      2    615  307.32   5.422 0.0065 **
Residuals   69   3911   56.68
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Follow-up to the ANOVA F-Test

When we reject we do not know which groups, or even how much they differ! What do we do?

We rejected $H_0 : \mu_1 = \mu_2 = \mu_3$. To determine which means are different, we could use Chapter 10 methods, making 1 - α confidence intervals for each pair:

$$\mu_1 - \mu_2 \qquad \mu_1 - \mu_3 \qquad \mu_2 - \mu_3$$

What is the problem with this?

Follow-up to the ANOVA F-Test

When we reject we do not know which groups, or even how much they differ! What do we do?

We rejected $H_0 : \mu_1 = \mu_2 = \mu_3$. To determine which means are different, we could use Chapter 10 methods, making 1 - α confidence intervals for each pair:

$$\mu_1 - \mu_2 \qquad \mu_1 - \mu_3 \qquad \mu_2 - \mu_3$$

What is the problem with this?

Each interval has an error probability of α. Thus, the **overall** probability that at least one of the interval misses the true difference is $> \alpha$.

Multiple Comparisons

Multiple comparison methods perform several separate statistical analyses with a confidence level that applies simultaneously to the entire analysis.

For confidence intervals: We are $100(1-\alpha)\%$ confident that ALL the intervals contain the truth.

A common method is the **Tukey Honest Significant Difference** (Tukey HSD) method.

Example 14.1 Continued

Use R to construct Tukey HSD multiple comparison confidence intervals for the mean weight gain for the 3 anorexia treatments. Overall confidence level is 0.95.

```
> mod <- aov(change ~ therapy, data = dat)
> TukeyHSD(mod, "therapy", conf.level = 0.95)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = change ~ therapy, data = dat)

$therapy
                    diff       lwr       upr     p adj
control-cog    -3.456897 -8.327276  1.413483 0.2124428
family-cog      4.257809 -1.250554  9.766173 0.1607461
family-control  7.714706  2.090124 13.339288 0.0045127
```

Example 14.1

```
                  diff       lwr       upr     p adj
control-cog    -3.456897 -8.327276  1.413483 0.2124428
family-cog      4.257809 -1.250554  9.766173 0.1607461
family-control  7.714706  2.090124 13.339288 0.0045127
```

`diff` = $\bar{y}_{control} - \bar{y}_{cog}$ etc = difference in group sample means
`lwr` = lower bound of confidence interval
`upr` = upper bound of confidence interval
`p adj` = p-value after adjustment for the multiple comparisons

If 0 is in the CI, there is no difference between the two groups.
If 0 is not in the CI, the two groups are significantly different.

Which groups are different here?

Example 14.1

```
                   diff       lwr       upr     p adj
control-cog    -3.456897 -8.327276  1.413483 0.2124428
family-cog      4.257809 -1.250554  9.766173 0.1607461
family-control  7.714706  2.090124 13.339288 0.0045127
```

`diff` $= \bar{y}_{control} - \bar{y}_{cog}$ etc = difference in group sample means
`lwr` = lower bound of confidence interval
`upr` = upper bound of confidence interval
`p adj` = p-value after adjustment for the multiple comparisons

If 0 is in the CI, there is no difference between the two groups.
If 0 is not in the CI, the two groups are significantly different.

Which groups are different here?
The mean weight gain for the **family** treatment is significantly **higher** than the mean weight gain for the **control** group.

Example 14.1

	diff	lwr	upr	p adj
control-cog	-3.456897	-8.327276	1.413483	0.2124428
family-cog	4.257809	-1.250554	9.766173	0.1607461
family-control	7.714706	2.090124	13.339288	0.0045127

diff = $\bar{y}_{control} - \bar{y}_{cog}$ etc = difference in group sample means
lwr = lower bound of confidence interval
upr = upper bound of confidence interval
p adj = p-value after adjustment for the multiple comparisons

If 0 is in the CI, there is no difference between the two groups.
If 0 is not in the CI, the two groups are significantly different.

Which groups are different here?
The mean weight gain for the **family** treatment is significantly **higher** than the mean weight gain for the **control** group.

Note: Hypothesis testing is not transitive!, i.e. if family is different from control, and cognitive is same as family, **it does not mean** cognitive is different from control!