ANOVA:
Analysis of Variation

The basic ANOVA situation

Two variables: 1 Categorical, 1 Quantitative

Main Question: Do the (means of) the quantitative variables depend on which group (given by categorical variable) the individual is in?

If categorical variable has only 2 values:

- 2-sample t-test

ANOVA allows for 3 or more groups

An example ANOVA situation

Subjects: 25 patients with blisters
Treatments: Treatment A, Treatment B, Placebo
Measurement: # of days until blisters heal

Data [and means]:

- A: 5,6,6,7,7,8,9,10 [7.25]
- B: 7,7,8,9,9,10,10,11 [8.875]
- P: 7,9,9,10,10,10,11,12,13 [10.11]

Are these differences significant?

Informal Investigation

Graphical investigation:

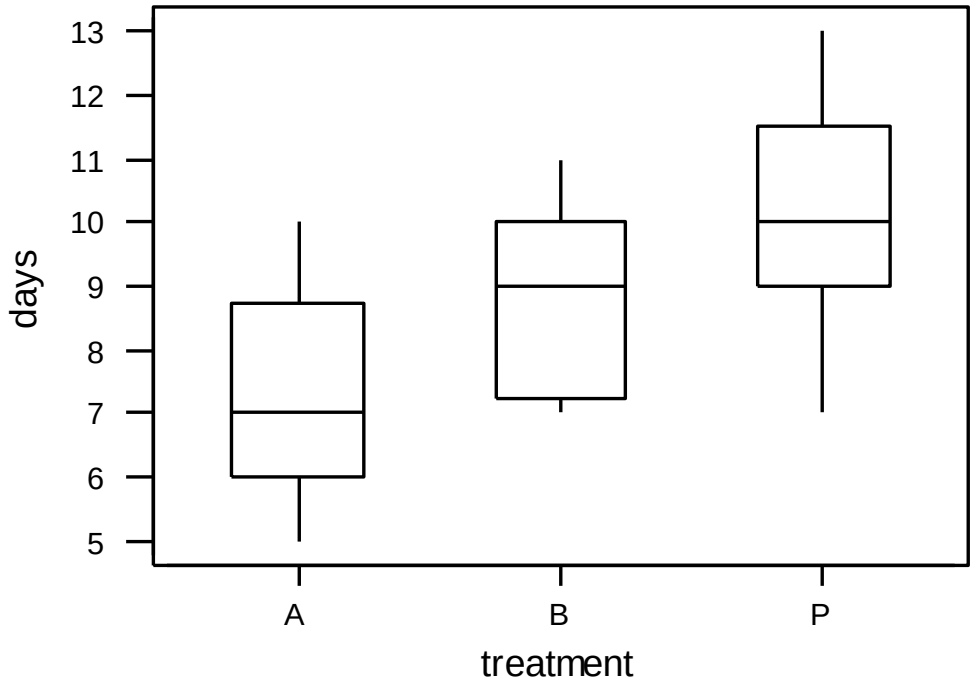
- side-by-side box plots
- multiple histograms

Whether the differences between the groups are significant depends on

- the difference in the means
- the standard deviations of each group
- the sample sizes

ANOVA determines P-value from the F statistic

Side by Side Boxplots



What does ANOVA do?

At its simplest (there are extensions) ANOVA tests the following hypotheses:

H_0: The means of all the groups are equal.

H_a: Not all the means are equal

- doesn't say how or which ones differ.
- Can follow up with "multiple comparisons"

Note: we usually refer to the sub-populations as "groups" when doing ANOVA.

Assumptions of ANOVA

- each group is approximately normal
  - check this by looking at histograms and/or normal quantile plots, or use assumptions
  - can handle some nonnormality, but not severe outliers
- standard deviations of each group are approximately equal
  - rule of thumb: ratio of largest to smallest sample st. dev. must be less than 2:1

Normality Check

We should check for normality using:

- assumptions about population
- histograms for each group
- normal quantile plot for each group

With such small data sets, there really isn't a really good way to check normality from data, but we make the common assumption that physical measurements of people tend to be normally distributed.

Standard Deviation Check

Variable	treatment	N	Mean	Median	StDev
days	A	8	7.250	7.000	1.669
	B	8	8.875	9.000	1.458
	P	9	10.111	10.000	1.764

Compare largest and smallest standard deviations:

- largest: 1.764
- smallest: 1.458
- $1.458 \times 2 = 2.916 > 1.764$

Note: variance ratio of 4:1 is equivalent.

Notation for ANOVA

- n = number of individuals all together
- I = number of groups
- $\bar{X}$ = mean for entire data set is

Group i has

- n_i = # of individuals in group i
- x_{ij} = value for individual j in group i
- $\bar{x}_i$ = mean for group i
- s_i = standard deviation for group i

How ANOVA works (outline)

ANOVA measures two sources of variation in the data and compares their relative sizes

- variation BETWEEN groups
  - for each data value look at the difference between its group mean and the overall mean

$$\left(\bar{x}_i - \bar{x}\right)^2$$

- variation WITHIN groups
  - for each data value we look at the difference between that value and the mean of its group

$$\left(x_{ij} - \bar{x}_i\right)^2$$

The ANOVA F-statistic is a ratio of the Between Group Variaton divided by the Within Group Variation:

$$F = \frac{Between}{Within} = \frac{MSG}{MSE}$$

A large F is evidence *against* H_0, since it indicates that there is more difference between groups than within groups.

Minitab ANOVA Output

Analysis of Variance for days

Source	DF	SS	MS	F	P
treatment	2	34.74	17.37	6.45	0.006
Error	22	59.26	2.69		
Total	24	94.00			

How are these computations made?

We want to measure the amount of variation due to BETWEEN group variation and WITHIN group variation

For each data value, we calculate its contribution to:

- BETWEEN group variation: $\left(\bar{x}_i - \bar{x}\right)^2$
- WITHIN group variation: $\left(x_{ij} - \bar{x}_i\right)^2$

An even smaller example

Suppose we have three groups

- Group 1: 5.3, 6.0, 6.7
- Group 2: 5.5, 6.2, 6.4, 5.7
- Group 3: 7.5, 7.2, 7.9

We get the following statistics:

SUMMARY				
Groups	*Count*	*Sum*	*Average*	*Variance*
Column 1	3	18	6	0.49
Column 2	4	23.8	5.95	0.176667
Column 3	3	22.6	7.533333	0.123333

Excel ANOVA Output

ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	5.127333	2	2.563667	10.21575	0.008394	4.737416
Within Groups	1.756667	7	0.250952			
Total	6.884	9				

1 less than number of groups

1 less than number of individuals (just like other situations)

number of data values - number of groups (equals df for each group added together)

Computing ANOVA F statistic

			WITHIN		**BETWEEN**	
			difference:		difference	
		group	data - group mean		group mean - overall mean	
data	group	mean	plain	squared	plain	squared
5.3	1	6.00	-0.70	0.490	-0.4	0.194
6.0	1	6.00	0.00	0.000	-0.4	0.194
6.7	1	6.00	0.70	0.490	-0.4	0.194
5.5	2	5.95	-0.45	0.203	-0.5	0.240
6.2	2	5.95	0.25	0.063	-0.5	0.240
6.4	2	5.95	0.45	0.203	-0.5	0.240
5.7	2	5.95	-0.25	0.063	-0.5	0.240
7.5	3	7.53	-0.03	0.001	1.1	1.188
7.2	3	7.53	-0.33	0.109	1.1	1.188
7.9	3	7.53	0.37	0.137	1.1	1.188
TOTAL				1.757		5.106
TOTAL/df				**0.25095714**		**2.55275**

overall mean: 6.44

$F = 2.5528/0.25025 = 10.21575$

Minitab ANOVA Output

Analysis of Variance for days

Source	DF	SS	MS	F	P
treatment	2	34.74	17.37	6.45	0.006
Error	22	59.26	2.69		
Total	24	94.00			

1 less than # of groups

(equals df for each group added together)

1 less than # of individuals (just like other situations)

Minitab ANOVA Output

Analysis of Variance for days

Source	DF	SS	MS	F	P
treatment	2	34.74	17.37	6.45	0.006
Error	22	59.26	2.69		
Total	24	94.00			

$$\sum_{obs}(x_{ij}-\bar{x}_i)^2$$

$$\sum_{obs}(x_{ij}-\bar{x})^2$$

$$\sum_{obs}(\bar{x}_i-\bar{x})^2$$

SS stands for sum of squares

- ANOVA splits this into 3 parts

Minitab ANOVA Output

```
Analysis of Variance for days
Source      DF         SS         MS         F         P
treatment    2      34.74      17.37      6.45     0.006
Error       22      59.26       2.69
Total       24      94.00
```

MSG = SSG / DFG
MSE = SSE / DFE

F = MSG / MSE

P-value comes from F(DFG,DFE)

(P-values for the F statistic are in Table E)

So How big is F?

Since F is
Mean Square Between / Mean Square Within

= MSG / MSE

A large value of F indicates relatively more
difference between groups than within groups
(evidence against H_0)

To get the P-value, we compare to *F(I-1,n-I)*-distribution

- *I-1* degrees of freedom in numerator (# groups -1)
- *n - I* degrees of freedom in denominator (rest of df)

Connections between SST, MST, and standard deviation

If ignore the groups for a moment and just compute the standard deviation of the entire data set, we see

$$s^2 = \frac{\sum \left(x_{ij} - \bar{x}\right)^2}{n-1} = \frac{SST}{DFT} = MST$$

So $SST = (n-1)\, s^2$, and $MST = s^2$. That is, *SST* and *MST* measure the TOTAL variation in the data set.

Connections between SSE, MSE, and standard deviation

Remember: $$s_i^2 = \frac{\sum (x_{ij} - \bar{x}_i)^2}{n_i - 1} = \frac{SS[\text{Within Group } i]}{df_i}$$

So $SS[Within\ Group\ i] = (s_i^2)\,(df_i)$

This means that we can compute SSE from the standard deviations and sizes (df) of each group:

$$SSE = SS[Within] = \sum SS[Within\ Group\ i]$$
$$= \sum s_i^2 (n_i - 1) = \sum s_i^2 (df_i)$$

Pooled estimate for st. dev

One of the ANOVA assumptions is that all groups have the same standard deviation. We can estimate this with a weighted average:

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \cdots + (n_I - 1)s_I^2}{n - I}$$

$$s_p^2 = \frac{(df_1)s_1^2 + (df_2)s_2^2 + \cdots + (df_I)s_I^2}{df_1 + df_2 + \cdots + df_I}$$

$$s_p^2 = \frac{SSE}{DFE} = MSE$$

so MSE is the pooled estimate of variance

In Summary

$$SST = \sum_{obs} (x_{ij} - \bar{x})^2 = s^2(DFT)$$

$$SSE = \sum_{obs} (x_{ij} - \bar{x}_i)^2 = \sum_{groups} s_i{}^2(df_i)$$

$$SSG = \sum_{obs} (\bar{x}_i - \bar{x})^2 = \sum_{groups} n_i(\bar{x}_i - \bar{x})^2$$

$$SSE + SSG = SST; \quad MS = \frac{SS}{DF}; \quad F = \frac{MSG}{MSE}$$

R^2 Statistic

R^2 gives the percent of variance due to between group variation

$$R^2 = \frac{SS[Between]}{SS[Total]} = \frac{SSG}{SST}$$

This is very much like the R^2 statistic that we computed back when we did regression.

Where's the Difference?

Once ANOVA indicates that the groups do not all appear to have the same means, what do we do?

```
Analysis of Variance for days
Source      DF        SS        MS        F        P
treatment    2     34.74     17.37     6.45    0.006
Error       22     59.26      2.69
Total       24     94.00

                                        Individual 95% CIs For Mean
                                        Based on Pooled StDev
Level        N      Mean     StDev  ----------+---------+---------+------
A            8     7.250     1.669  (-------*-------)
B            8     8.875     1.458            (-------*-------)
P            9    10.111     1.764                     (------*-------)
                                        ----------+---------+---------+------
Pooled StDev =    1.641               7.5       9.0      10.5
```

Clearest difference: P is worse than A (CI's don't overlap)

Multiple Comparisons

Once ANOVA indicates that the groups do not all have the same means, we can compare them two by two using the 2-sample t test

- We need to adjust our p-value threshold because we are doing multiple tests with the same data.

- There are several methods for doing this.

- If we really just want to test the difference between one pair of treatments, we should set the study up that way.

Tuckey's Pairwise Comparisons

```
Tukey's pairwise comparisons

    Family error rate = 0.0500
Individual error rate = 0.0199

Critical value = 3.55

Intervals for (column level mean) - (row level mean)

                A           B

      B      -3.685
              0.435

      P      -4.863      -3.238
             -0.859       0.766
```

95% confidence

Use alpha = 0.0199 for each test.

These give 98.01% CI's for each pairwise difference.

Only P vs A is significant (both values have same sign)

95% CI for A-P is (-0.86,-4.86)

Fisher's Pairwise Comparisons

Now we set the individual error rate (alpha) and see the overall error rate. 95% confidence on each corresponds to 88.1% confidence overall

```
Fisher's pairwise comparisons

    Family error rate = 0.119
Individual error rate = 0.0500

Critical value = 2.074

Intervals for (column level mean) - (row level mean)

                A           B

       B     -3.327
              0.077

       P     -4.515      -2.890
             -1.207       0.418
```