Analysis of Varience (ANOVA)

ANOVA

- Sometimes we want to know whether the mean level on one continuous variable (such as income) is different for each group relative to the others in a nominal variable (such as degree received).
- We could use descriptive statistics (the mean income) to compare the groups (Ex. sociology BA vs. MA vs. PhD).
- However, as sociologists, we usually want to use a sample to determine whether groups are different in the population.

ANOVA

- ANOVA is an inferential statistics technique that allows you to compare the mean level on one interval-ratio variable (such as income) for each group relative to the others in a nominal variable (such as degree).
- If you had only two groups to compare, ANOVA would give the same answer as an independent samples t-test.

ANOVA

- One typically uses ANOVA in experiments because these typically involve comparing persons in experimental conditions with those in control conditions to see if the experimental conditions affect people.

  Independent → Dependent
  Nominal Variable → Interval-ratio Variable
  Experimental Grouping → Outcome Variable

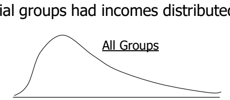
- For example: Is "Diff'rent Strokes" funnier than "Charles in Charge?"
  Experiment:
  Do kids exposed to "Diff'rent Strokes" laugh more than those who watch "Charles in Charge?"

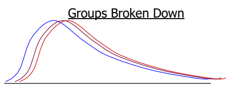
  Expose Groups to a Show → Record Amount of Laughter

- We then use the sample to make inferences about the population.

ANOVA

What if three racial groups had incomes distributed like this in your sample?

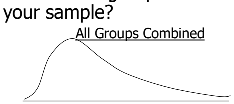


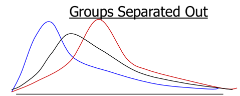


Isn't it conceivable that the differences are due to natural random variability between samples? Would you want to claim they are different in the population?

ANOVA

Now…What if three racial groups had incomes distributed like this in your sample?

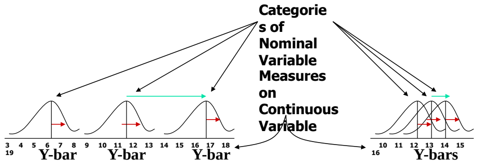




Doesn't it now appear that the groups may be different regardless of sampling variability? Would you feel comfortable claiming the groups are different in the population?

ANOVA

- Conceptually, ANOVA compares the variance within groups to the overall variance between all the groups to determine whether the groups appear distinct from each other or if they look quite the same.

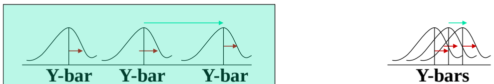


Different groups, different means.

Similar groups, similar means.

ANOVA

- When the groups have little variation within themselves, but large variation between them, it would appear that they are distinct and that their means are different.

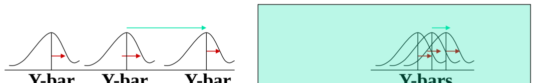


Different groups, different means.

Similar groups, similar means.

ANOVA

- When the groups have a lot of variation within themselves, but little variation between them, it would appear that they are similar and that their means are not really different (perhaps they differ only because of peculiarities of the particular sample).



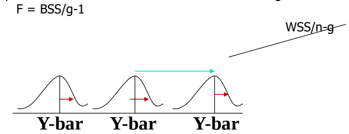
Different groups, different means.

Similar groups, similar means.

ANOVA

- Let's call the the between groups variation: Between Variance: Between Sum of Squares, BSS/df
- Let's call the within groups variation: Within Variance: Within Sum of Squares, WSS/df
- ANOVA compares Between Variance to Within Variance through a ratio we will call F. F = BSS/g-1

WSS/n-g



ANOVA for comparing means between more than 2 groups

Hypotheses of One-Way ANOVA

- $H_0 : \mu_1 = \mu_2 = \mu_3 = \cdots = \mu_c$
  - All population means are equal
  - i.e., no treatment effect (no variation in means among groups)
- H_1 : Not all of the population means are the same
  - At least one population mean is different
  - i.e., there is a treatment effect
  - Does not mean that all population means are different (some pairs may be the same)

The F-distribution

- A ratio of variances follows an F-distribution:

$$\boxed{\frac{\sigma^2_{between}}{\sigma^2_{within}} \sim F_{n,m}}$$

- The F-test tests the hypothesis that two variances are equal.
- F will be close to 1 if sample variances are equal.

$$H_0: \sigma^2_{between} = \sigma^2_{within}$$
$$H_a: \sigma^2_{between} \neq \sigma^2_{within}$$

How to calculate ANOVA's by hand…

Treatment 1	Treatment 2	Treatment 3	Treatment 4
y_{11}	y_{21}	y_{31}	y_{41}
y_{12}	y_{22}	y_{32}	y_{42}
y_{13}	y_{23}	y_{33}	y_{43}
y_{14}	y_{24}	y_{34}	y_{44}
y_{15}	y_{25}	y_{35}	y_{45}
y_{16}	y_{26}	y_{36}	y_{46}
y_{17}	y_{27}	y_{37}	y_{47}
y_{18}	y_{28}	y_{38}	y_{48}
y_{19}	y_{29}	y_{39}	y_{49}
y_{110}	y_{210}	y_{310}	y_{410}

n=10 obs./group

k=4 groups

$$\bar{y}_{1\bullet} = \frac{\sum_{j=1}^{10} y_{1j}}{10} \quad \bar{y}_{2\bullet} = \frac{\sum_{j=1}^{10} y_{2j}}{10} \quad \bar{y}_{3\bullet} = \frac{\sum_{j=1}^{10} y_{3j}}{10} \quad \bar{y}_{4\bullet} = \frac{\sum_{j=1}^{10} y_{4j}}{10}$$

The group means

$$\frac{\sum_{j=1}^{10}(y_{1j}-\bar{y}_{1\bullet})^2}{10-1} \quad \frac{\sum_{j=1}^{10}(y_{2j}-\bar{y}_{2\bullet})^2}{10-1} \quad \frac{\sum_{j=1}^{10}(y_{3j}-\bar{y}_{3\bullet})^2}{10-1} \quad \frac{\sum_{j=1}^{10}(y_{4j}-\bar{y}_{4\bullet})^2}{10-1}$$

The (within) group variances

Sum of Squares Within (SSW), or Sum of Squares Error (SSE)

$$\frac{\sum_{j=1}^{10}(y_{1j}-\bar{y}_{1\bullet})^2}{10-1} \quad \frac{\sum_{j=1}^{10}(y_{2j}-\bar{y}_{2\bullet})^2}{10-1} \quad \frac{\sum_{j=1}^{10}(y_{3j}-\bar{y}_{3\bullet})^2}{10-1} \quad \frac{\sum_{j=1}^{10}(y_{4j}-\bar{y}_{4\bullet})^2}{10-1}$$

The (within) group variances

$$\sum_{j=1}^{10}(y_{1j}-\bar{y}_{1\bullet})^2 + \sum_{j=1}^{10}(y_{2j}-\bar{y}_{2\bullet})^2 + \sum_{j=3}^{10}(y_{3j}-\bar{y}_{3\bullet})^2 + \sum_{j=1}^{10}(y_{4j}-\bar{y}_{4\bullet})^2$$

$$= \sum_{i=1}^{4}\sum_{j=1}^{10}(y_{ij}-\bar{y}_{i\bullet})^2$$

Sum of Squares Within (SSW) (or SSE, for chance error)

Sum of Squares Between (SSB), or Sum of Squares Regression (SSR)

Overall mean of all 40 observations ("grand mean")

$$\bar{\bar{y}}_{\bullet\bullet} = \frac{\sum_{i=1}^{4}\sum_{j=1}^{10} y_{ij}}{40}$$

$$10x\sum_{i=1}^{4}(\bar{y}_{i\bullet} - \bar{\bar{y}}_{\bullet\bullet})^2 \longleftarrow$$

Sum of Squares Between (SSB). Variability of the group means compared to the grand mean (the variability due to the treatment).

Total Sum of Squares (SST)

$$\sum_{i=1}^{4}\sum_{j=1}^{10}(y_{ij} - \bar{\bar{y}}_{\bullet\bullet})^2$$

Total sum of squares(TSS). Squared difference of every observation from the overall mean. (numerator of variance of Y!)

Partitioning of Variance

$$\sum_{i=1}^{4}\sum_{j=1}^{10}(y_{ij}-\bar{y}_{i\bullet})^2 + 10x\sum_{i=1}^{4}(\bar{y}_{i\bullet}-\bar{\bar{y}}_{\bullet\bullet})^2 = \sum_{i=1}^{4}\sum_{j=1}^{10}(y_{ij}-\bar{\bar{y}}_{\bullet\bullet})^2$$

SSW + SSB = TSS

ANOVA Table

Source of variation	d.f.	Sum of squares	Mean Sum of Squares	F-statistic	p-value
Between (k groups)	**k-1**	**SSB (sum of squared deviations of group means from grand mean)**	SSB/k-1	$\frac{SSB/k-1}{SSW/nk-k}$	**Go to $F_{k-1,nk-k}$ chart**
Within (n individuals per group)	**nk-k**	**SSW (sum of squared deviations of observations from their group mean)**	**s^2=SSW/nk-k**		
Total variation	**nk-1**	**TSS (sum of squared deviations of observations from grand mean)**		TSS=SSB + SSW	

Example

Treatment 1	Treatment 2	Treatment 3	Treatment 4
60 inches	50	48	47
67	52	49	67
42	43	50	54
67	67	55	67
56	67	56	68
62	59	61	65
64	67	61	65
59	64	60	56
72	63	59	60
71	65	64	65

Example

Step 1) calculate the sum of squares between groups:

Treatment 1	Treatment 2	Treatment 3	Treatment 4
60 inches	50	48	47
67	52	49	67
42	43	50	54
67	67	55	67
56	67	56	68
62	59	61	65
64	67	61	65
59	64	60	56
72	63	59	60
71	65	64	65

Mean for group 1 = 62.0

Mean for group 2 = 59.7

Mean for group 3 = 56.3

Mean for group 4 = 61.4

Grand mean= 59.85

SSB = $[(62-59.85)^2 + (59.7-59.85)^2 + (56.3-59.85)^2 + (61.4-59.85)^2]$ *xn per group*= 19.65*x10* = **196.5**

Example

Step 2) calculate the sum of squares within groups:

Treatment 1	Treatment 2	Treatment 3	Treatment 4
60 inches	50	48	47
67	52	49	67
42	43	50	54
67	67	55	67
56	67	56	68
62	59	61	65
64	67	61	65
59	64	60	56
72	63	59	60
71	65	64	65

$(60-62)^2+(67-62)^2+(42-62)^2+(67-62)^2+(56-62)^2+(62-62)^2+(64-62)^2+(59-62)^2+(72-62)^2+(71-62)^2+(50-59.7)^2+(52-59.7)^2+(43-59.7)^2+67-59.7)^2+(67-59.7)^2+(69-59.7)^2$...+....(sum of 40 squared deviations) = **2060.6**

Step 3) Fill in the ANOVA table

Source of variation	d.f.	Sum of squares	Mean Sum of Squares	F-statistic	p-value
Between	3	**196.5**	65.5	1.14	**.344**
Within	36	**2060.6**	57.2	-	-
Total	39	2257.1	-	-	-

Step 3) Fill in the ANOVA table

Source of variation	d.f.	Sum of squares	Mean Sum of Squares	F-statistic	p-value
Between	3	**196.5**	65.5	1.14	**.344**
Within	36	**2060.6**	57.2	-	-
Total	39	2257.1	-	-	-

INTERPRETATION of ANOVA:

How much of the variance in height is explained by treatment group?

R^2="Coefficient of Determination" = SSB/TSS = 196.5/2275.1=9%

Coefficient of Determination

$$R^2 = \frac{SSB}{SSB + SSE} = \frac{SSB}{SST}$$

The amount of variation in the outcome (response) variable (dependent variable) that is explained by the predictor (factor) (independent variable).

Terminology

- Experimental design in general, and analysis of variance in particular, has its own language. We'll quickly review some important terms.
- We'll use a series of increasingly complex study designs to introduce the most significant concepts.
- We are interested in studying the treatment of anxiety. Two popular therapies for anxiety are cognitive behavior therapy (CBT) and eye movement desensitization and reprocessing (EMDR).
- We recruit 10 anxious individuals and randomly assign half of them to receive five weeks of CBT and half to receive five weeks of EMDR.
- At the conclusion of therapy, each patient is asked to complete the State-Trait Anxiety Inventory (STAI), a self report measure of anxiety.
- The design is outlined in table 9.1.

Terminology

- In this design, Treatment is a between-groups factor with two levels (CBT, EMDR).
- It's called a between-groups factor because patients are assigned to one and only one group.
- No patient receives both CBT and EMDR. These characters represent the subjects (patients).
- STAI is the dependent variable, and Treatment is the independent variable.
- Because there is an equal number of observations in each treatment condition, we have a balanced design.
- When the sample sizes are unequal across the cells of a design, you have an unbalanced design.

Table 9.1 One-way between-groups ANOVA

Treatment	
CBT	EMDR
s1	s6
s2	s7
s3	s8
s4	s9
s5	s10

Terminology

- The statistical design in table 9.1 is called a *one-way ANOVA* because there's a single classification variable.
- Specifically, it's a one-way between-groups ANOVA.
- Effects in ANOVA designs are primarily evaluated through F tests.
- If the F test for Treatment is significant, you can conclude that the mean STAI scores for two therapies differed after five weeks of treatment.

Table 9.1 One-way between-groups ANOVA

Treatment	
CBT	EMDR
s1	s6
s2	s7
s3	s8
s4	s9
s5	s10

Terminology

- If you were interested in the effect of CBT on anxiety over time, you could place all 10 patients in the CBT group and assess them at the conclusion of therapy and again six months later.
- This design is displayed in table 9.2.
- Time is a *within-groups* factor with two levels (five weeks, six months).
- It's called a within-groups factor because each patient is measured under both levels.
- The statistical design is a *one-way within-groups ANOVA*.
- Because each subject is measured more than once, the design is also called a *repeated measures ANOVA*.
- If the F test for Time is significant, you can conclude that patients' mean STAI scores changed between five weeks and six months.

Table 9.2 One-way within-groups ANOVA

Patient	Time	
	5 weeks	6 months
s1		
s2		
s3		
s4		
s5		
s6		
s7		
s8		
s9		
s10		

Terminology

- If you were interested in both treatment differences and change over time, you could combine the first two study designs and randomly assign five patients to CBT and five patients to EMDR, and assess their STAI results at the end of therapy (five weeks) and at six months (see table 9.3).
- By including both Therapy and Time as factors, you're able to examine the impact of Therapy (averaged across time), Time (averaged across therapy type), and the interaction of Therapy and Time.
- The first two are called the main effects, whereas the interaction is (not surprisingly) called an interaction effect.

Table 9.3 Two-way factorial ANOVA with one between-groups and one within-groups factor

		Patient	Time	
			5 weeks	6 months
Therapy	CBT	s1		
		s2		
		s3		
		s4		
		s5		
	EMDR	s6		
		s7		
		s8		
		s9		
		s10		

Terminology

- When we cross two or more factors, as is done here, you have a factorial ANOVA design. Crossing two factors produces a two-way ANOVA, crossing three factors produces a three-way ANOVA, and so forth.
- When a factorial design includes both between-groups and within-groups factors, it's also called a mixed-model ANOVA. The current design is a two-way mixed-model factorial ANOVA.
- In this case, you'll have three F tests: one for Therapy, one for Time, and one for the Therapy × Time interaction.
- A significant result for Therapy indicates that CBT and EMDR differ in their impact on anxiety.
- A significant result for Time indicates that anxiety changed from week five to the six-month follow-up.
- A significant Therapy × Time interaction indicates that the two treatments for anxiety had a differential impact over time (that is, the change in anxiety from five weeks to six months was different for the two treatments).

Fitting ANOVA models

- The syntax of the aov() function is aov(formula, data=dataframe).
- The following Table describes special symbols that can be used in the formulas. In this table, y is the dependent variable and the letters A, B, and C represent factors.

Table 9.4 Special symbols used in R formulas

Symbol	Usage
~	Separates response variables on the left from the explanatory variables on the right. For example, a prediction of y from A, B, and C would be coded `y ~ A + B + C`
:	Denotes an interaction between variables. A prediction of y from A, B, and the interaction between A and B would be coded `y ~ A + B + A:B`
*	Denotes the complete crossing variables. The code `y ~ A*B*C` expands to `y ~ A + B + C + A:B + A:C + B:C + A:B:C`
^	Denotes crossing to a specified degree. The code `y ~ (A+B+C)^2` expands to `y ~ A + B + C + A:B + A:C + A:B`
.	Denotes all remaining variables. The code `y ~ .` expands to `y ~ A + B + C`

Fitting ANOVA models

- The following Table provides formulas for several common research designs.
- In this table, lowercase letters are quantitative variables, uppercase letters are grouping factors, and Subject is a unique identifier variable for subjects.

Table 9.5 Formulas for common research designs

Design	Formula
One-way ANOVA	`y ~ A`
One-way ANCOVA with 1 covariate	`y ~ x + A`
Two-way factorial ANOVA	`y ~ A * B`
Two-way factorial ANCOVA with 2 covariates	`y ~ x1 + x2 + A * B`
Randomized block	`y ~ B + A` (where B is a blocking factor)
One-way within-groups ANOVA	`y ~ A + Error(Subject/A)`
Repeated measures ANOVA with 1 within-groups factor (W) and 1 between-groups factor (B)	`y ~ B * W + Error(Subject/W)`