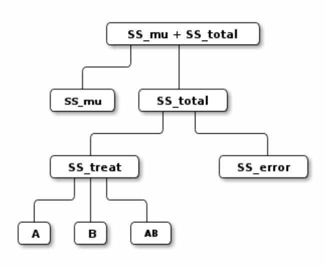
When is ANOVA applicable?

- When you wish to assess the independent/joint effects of one or more *categorical* factors on a single *continuous* dependent variable
- Strictly speaking, ANOVA is not applicable to count or categorical DVs (but that stops few researchers from using it anyway!)
- ANOVA is a special case of linear regression, ultimately a more flexible approach

ANOVA as variance partitioning



$$SS_{total} = SS_{treat} + SS_{error}$$
$$SS_{treat} = SS_A + SS_B + SS_{AB}$$

Source	SS	df	MS	F
A		$k_A - 1$		
B		$k_B - 1$		
AB		$df_A \times df_B$		
Error		$N_{subj} - N_{groups}$		
Total				

How the GLM represents relationships

Component of GLM	Notation
DV	Y
Grand Average	μ "mu"
Main Effects	$A, B, C, \ldots$
Interactions	$AB, AC, BC, ABC, \ldots$
Random Error	$S(Group)$

Score	=	Grand Avg.	+	Main Effects	+	Interactions	+	Error
Y	=	μ	+	$A + B + C + \ldots$	+	$AB + AC + BC + ABC + \ldots$	+	$S(Group)$

- Components of the model are estimated from the observed data
- Tests are performed (F) to see whether its variability is too large to be introduced by chance

Making comparisons across groups

Example (Spelling)

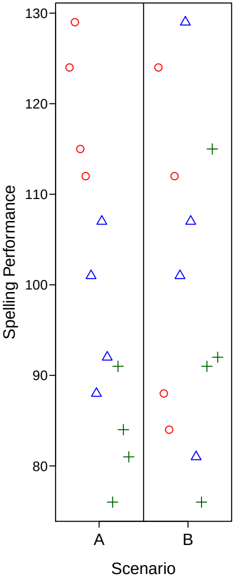
You wish to compare the benefits of three different spelling programs. Do these programs yield differences in spelling performance?

$H_0 : \mu_1 = \mu_2 = \mu_3$

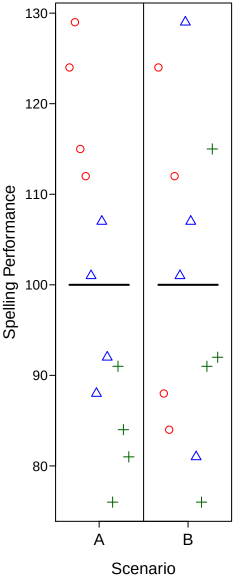
Factors and Levels

Factor: a categorical variable that is used to divide subjects into groups, usually to draw some comparison. Factors are composed of different *levels*. Do not confuse factors with levels!

Means, Variability, and Deviation Scores

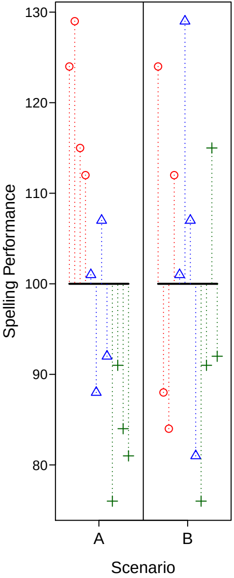


Means, Variability, and Deviation Scores



$$Y_{..} = \frac{\sum_{ij} Y_{ij}}{N}$$

Means, Variability, and Deviation Scores

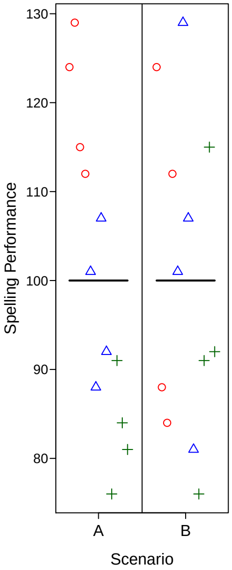


grand mean $Y_{..} = \frac{\sum_{ij} Y_{ij}}{N}$

$$SD_Y = \sqrt{\frac{\sum_{ij}\left(Y_{ij} - Y_{..}\right)^2}{N}}$$

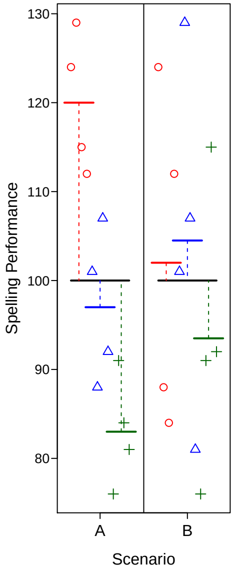
deviation score: $Y_{ij} - Y_{..}$

GLM for One-Factor ANOVA



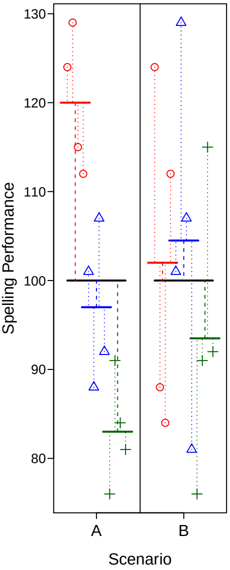
$$Y_{ij} = \mu$$

GLM for One-Factor ANOVA



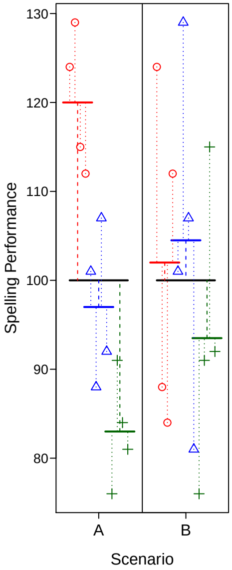
$$Y_{ij} = \mu + A_i$$

GLM for One-Factor ANOVA



$$Y_{ij} = \mu + A_i + S(A)_{ij}$$

GLM for One-Factor ANOVA



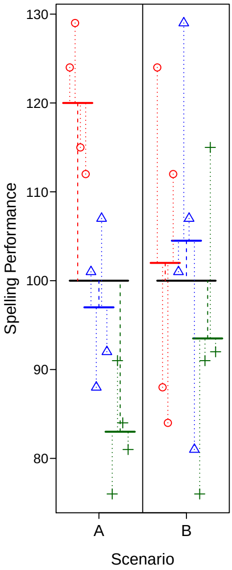
$$Y_{ij} = \mu + A_i + S(A)_{ij}$$

Estimation Equations

$$\begin{aligned} \hat{\mu} &= Y_{..} \\ \hat{A}_i &= Y_{i.} - \hat{\mu} \\ \widehat{S(A)}_{ij} &= Y_{ij} - \hat{\mu} - \hat{A}_i \end{aligned}$$

Note that $\sum_i \hat{A}_i = 0$ and $\sum_{ij} \widehat{S(A)}_{ij} = 0$

Sources of Variance



$$Y_{ij} = \mu + A_i + S(A)_{ij}$$

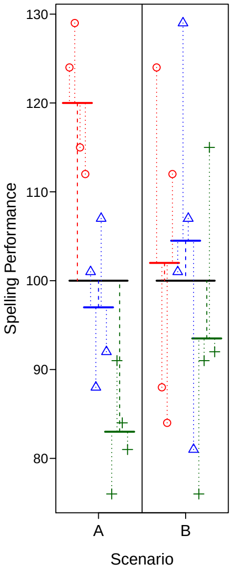
$$\begin{aligned} Y_{ij} - \mu &= A_i + S(A)_{ij} \\ \textit{individual} &= \textit{group} + \textit{random} \end{aligned}$$

Sum of Squares (SS)

A measure of variability consisting of the sum of squared *deviation* scores, where a deviation score is a score minus a mean.

$$SS_A = \sum (Y_{i.} - \mu)^2$$

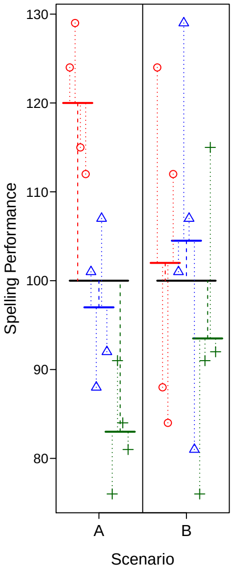
Decomposition Matrix



$$\hat{\mu} = 100$$
$$\hat{A}_1 = 120 - 100 = 20$$
$$\hat{A}_2 = 97 - 100 = -3$$
$$\hat{A}_3 = 83 - 100 = -17$$

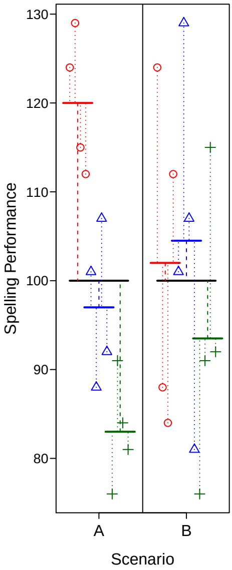
	Y_{ij}	=	$\hat{\mu}$	+	$\hat{A}_i$	+	$\widehat{S(A)}_{ij}$
	124	=	100	+	20	+	4
	129	=	100	+	20	+	9
	115	=	100	+	20	+	-5
	112	=	100	+	20	+	-8
	101	=	100	+	-3	+	4
	88	=	100	+	-3	+	-9
	107	=	100	+	-3	+	10
	92	=	100	+	-3	+	-5
	76	=	100	+	-17	+	-7
	91	=	100	+	-17	+	8
	84	=	100	+	-17	+	1
	81	=	100	+	-17	+	-2
$SS =$	123318	=	120000	+	2792	+	526

Logic of ANOVA



- Compare two estimates of the variability, the *between-group* estimate ($SS_{between}$) and the *within-group* estimate (SS_{within})
- If $H_0 : \mu_1 = \mu_2 = \mu_3$ is true, then these two measures estimate the same quantity.
- The extent to which the between-group variability exceeds the within-group variability gives evidence against $H_0 : \mu_1 = \mu_2 = \mu_3$.

Calculating $SS_{between}$ and SS_{within}

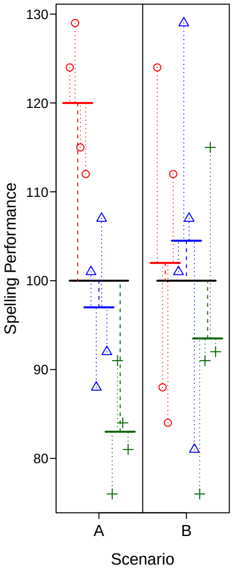


	Y_{ij}	=	$\hat{\mu}$	+	$\hat{A}_i$	+	$\widehat{S(A)}_{ij}$
	124	=	100	+	20	+	4
	129	=	100	+	20	+	9
	115	=	100	+	20	+	-5
	112	=	100	+	20	+	-8
	101	=	100	+	-3	+	4
	88	=	100	+	-3	+	-9
	107	=	100	+	-3	+	10
	92	=	100	+	-3	+	-5
	76	=	100	+	-17	+	-7
	91	=	100	+	-17	+	8
	84	=	100	+	-17	+	1
	81	=	100	+	-17	+	-2
$SS =$	123318	=	120000	+	2792	+	526

check your math

$$SS_Y = SS_\mu + SS_A + SS_{S(A)}$$

H_0 and Sums of Squares



$$Y_{ij} - \mu = A_i + S(A)_{ij}$$

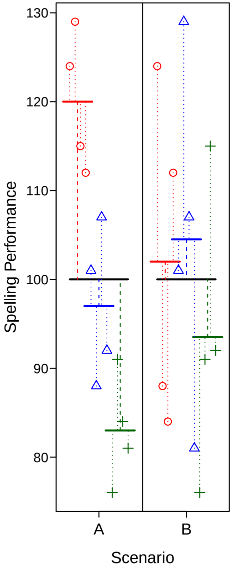
Scenario A

$SS_A = 2792$
$SS_{S(A)} = 526$
$SS_A + SS_{S(A)} = 3318$

Scenario B

$SS_A = 266$
$SS_{S(A)} = 3052$
$SS_A + SS_{S(A)} = 3318$

Mean Square and Degrees of Freedom



Degrees of Freedom (df)

The number of observations that are “free to vary”.

$df_A = K - 1$

$df_{S(A)} = N - K$

where N is the number of subjects and K is the number of groups.

Mean Square (MS)

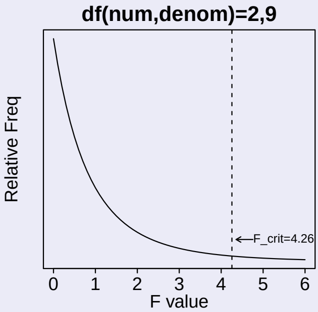
A sum of squares divided by its degrees of freedom.

$$MS_A = \frac{SS_A}{df_A} = \frac{2792}{2} = 1396$$

$$MS_{S(A)} = \frac{SS_{S(A)}}{df_{S(A)}} = \frac{526}{9} = 58.4$$

The F-ratio

F density function



If $F_{obs} > F_{crit}$, then reject H_0

F ratio

A ratio of mean squares, with $df_{numerator}$ and $df_{denominator}$ degrees of freedom.

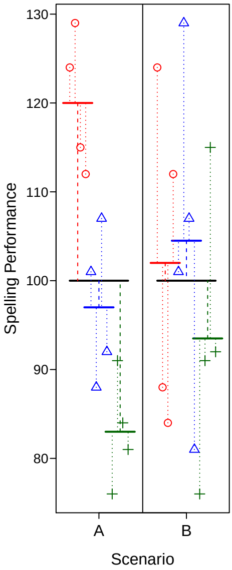
$$F_A = \frac{MS_A}{MS_{S(A)}} = \frac{1396}{58.4} = 23.886$$

df in denominator	df in numerator 1	2	3	4	5	6	7	8
1	161.40	199.50	215.70	224.60	230.20	234.00	236.80	238.90
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23

Density/Quantile functions for *F*-distribution

name	function
`pf(x, df1, df2, lower.tail = FALSE)`	density (get p given F_{obs})
`qf(p, df1, df2, lower.tail = FALSE)`	quantile (get F_{crit} given p)

Summary Table



Scenario A

Source	*df*	*SS*	*MS*	*F*	*p*	Error
μ	1	120000	120000.0	2053.232	<.001	$S(A)$
A	2	2792	1396.0	23.886	<.001	$S(A)$
$S(A)$	9	526	58.4			
Total	12	123318				

Scenario B

Source	*df*	*SS*	*MS*	*F*	*p*	Error
μ	1	120000	120000.0	353.878	<.001	$S(A)$
A	2	266	133.0	.392	.687	$S(A)$
$S(A)$	9	3052	339.1			
Total	12	123318				

Overview of One-Way ANOVA

1. Write the GLM: $Y_{ij} = \mu + A_i + S(A)_{ij}$
2. Write down the estimating equations:
   - $\hat{\mu} = Y_{..}$
   - $\hat{A}_i = Y_{i.} - \hat{\mu}$
   - $\widehat{S(A)}_{ij} = Y_{ij} - \hat{\mu} - \hat{A}_i$
3. Compute estimates for all terms in model.
4. Create *decomposition matrix.*
5. Compute *SS*, *MS*, *df*.
   - $df_{\mu} = 1$
   - $df_A = K - 1$
   - $df_{S(A)} = N - K$
   - $MS = SS/df$
6. Construct a summary ANOVA table.
7. Compare F_{obs} with F_{crit}.

R

use the `aov()` function, e.g.:

```
spelling$A <- factor(spelling$A)
mod <- aov(Y ~ A, data = spelling)
summary(mod)
```

http://talklab.psy.gla.ac.uk/stats/onefactoranova.html#sec-3-2

ANOVA assumptions

- Normality
- Conditional independence
- Homoskedasticity
- Sphericity (RM-designs only, where $k > 2$)