anova

- variability partitioning
- anova output

Dr. Mine Çetinkaya-Rundel
Duke University

vocabulary score and class

	wordsum	class
1	6	middle class
2	9	working class
3	6	working class
4	5	working class
5	6	working class
6	6	working class
…	…	…
795	9	middle class

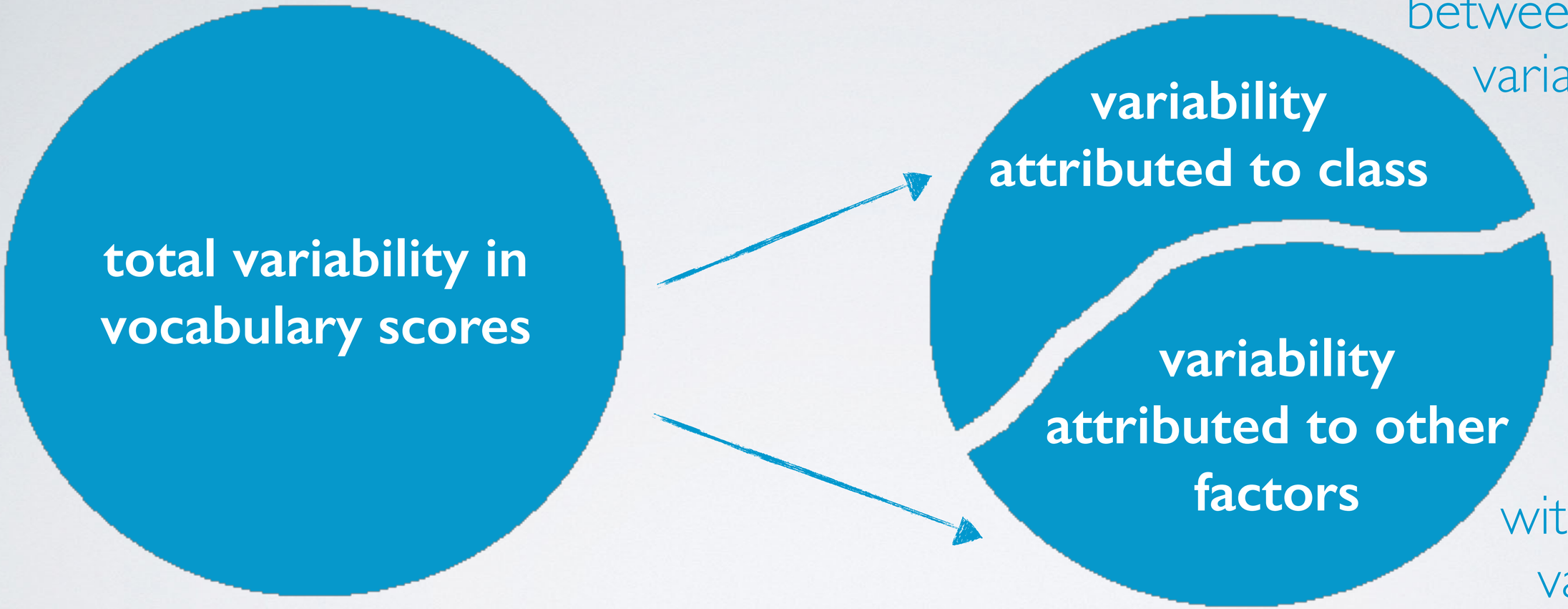
	n	mean	sd
lower class	41	5.07	2.24
working class	407	5.75	1.87
middle class	331	6.76	1.89
upper class	16	6.19	2.34
overall	795	6.14	1.98

H_0: The mean outcome is the same across all categories

$$\mu_1 = \mu_2 = \mu_3 = \mu_4$$

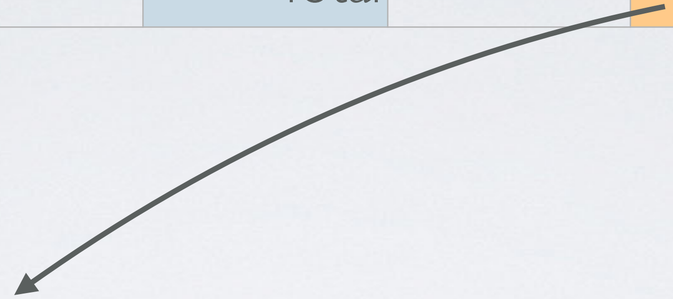
H_A: At least one pair of means are different from each other

variability partitioning



		Df	Sum Sq	Mean Sq	F value	Pr(> F)
Group	class	3	236.56	78.855	21.735	<0.0001
Error	Residuals	791	2869.80	3.628		
	Total	794	3106.36			

		Df	Sum Sq	Mean Sq	F value	Pr(> F)
Group	class		236.56			
Error	Residuals		2869.80			
	Total		**3106.36**			



sum of squares total (SST)

- measures the **total variability** in the response variable
- calculated very similarly to variance (except not scaled by the sample size)

Sum of squares total (SST):

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

y_i : value of the response variable for each observation

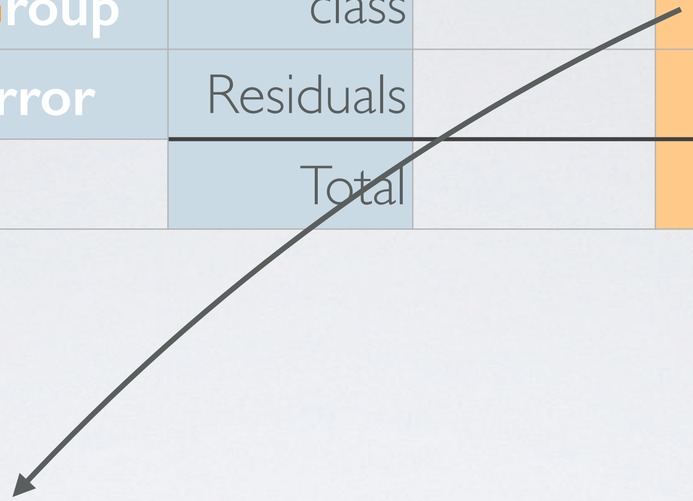
$\bar{y}$: grand mean of the response variable

	wordsum	class
1	6	middle class
2	9	working class
3	6	working class
…	…	…
795	9	middle class

	n	mean	sd
overall	795	6.14	1.98

$$SST = (6-6.14)^2 + (9-6.14)^2 + (6-6.14)^2 + \cdots + (9-6.14)^2 = 3106.36$$

		Df	Sum Sq	Mean Sq	F value	Pr(> F)
Group	class		236.56			
Error	Residuals		2869.80			
	Total		3106.36			



sum of squares groups (SSG)

- measures the variability **between groups**
- **explained variability:** deviation of group mean from overall mean, weighted by sample size

Sum of squares group (SSG):

$$SSG = \sum_{j=1}^{k} n_j(\bar{y}_j - \bar{y})^2$$

n_j : number of observations in group *j*

$\bar{y}_j$: mean of the response variable for group *j*

$\bar{y}$: grand mean of the response variable

	n	mean	sd
lower class	41	5.07	2.24
working class	407	5.75	1.87
middle class	331	6.76	1.89
upper class	16	6.19	2.34
overall	795	6.14	1.98

$$SSG = (41 \times (5.07 - 6.14)^2) + (407 \times (5.75 - 6.14)^2) + (331 \times (6.76 - 6.14)^2) + (16 \times (6.19 - 6.14)^2)$$

$$\approx 236.56$$

		Df	Sum Sq	Mean Sq	F value	Pr(> F)
Group	class		236.56			
Error	Residuals		**2869.8**			
	Total		3106.36			

Sum of squares error (SSE):

$$SSE = SST - SSG$$

3106.36 - 236.56 = 2869.8

sum of squares error (SSE)

- measures the variability **within groups**
- **unexplained variability:** unexplained by the group variable, due to other reasons

		Df	Sum Sq	Mean Sq	F value	Pr(> F)
Group	class		236.56	?		
Error	Residuals		2869.8	?		
	Total		3106.36	?		

- now we need a way to get from these measures of total variability to average variability
- scaling by a measure that incorporates sample sizes and number of groups → degrees of freedom

degrees of freedom

		Df	Sum Sq	Mean Sq	F value	Pr(> F)
Group	class	3	236.56			
Error	Residuals	791	2869.80			
	Total	794	3106.36			

Degrees of freedom associated with ANOVA:

- total: $df_T = n - 1$ → 795 − 1 = 794
- group: $df_G = k - 1$ → 4 − 1 = 3
- error: $df_E = df_T - df_G$ → 794 − 3 = 791

mean square error

		Df	Sum Sq	Mean Sq	F value	Pr(> F)
Group	class	3	236.56	78.855		
Error	Residuals	791	2869.80	3.628		
	Total	794	3106.36			

Mean squares: Average variability between and within groups, calculated as the total variability (sum of squares) scaled by the associated degrees of freedom.

- group: $MSG = SSG/df_G$ → 236.56 / 3 ≈ 78.855
- error: $MSE = SSE/df_E$ → 2869.8 / 791 ≈ 3.628

F statistic

		Df	Sum Sq	Mean Sq	F value	Pr(> F)
Group	class	3	236.56	78.855	21.735	
Error	Residuals	791	2869.80	3.628		
	Total	794	3106.36			

F statistic: Ratio of the between group and within group variability:

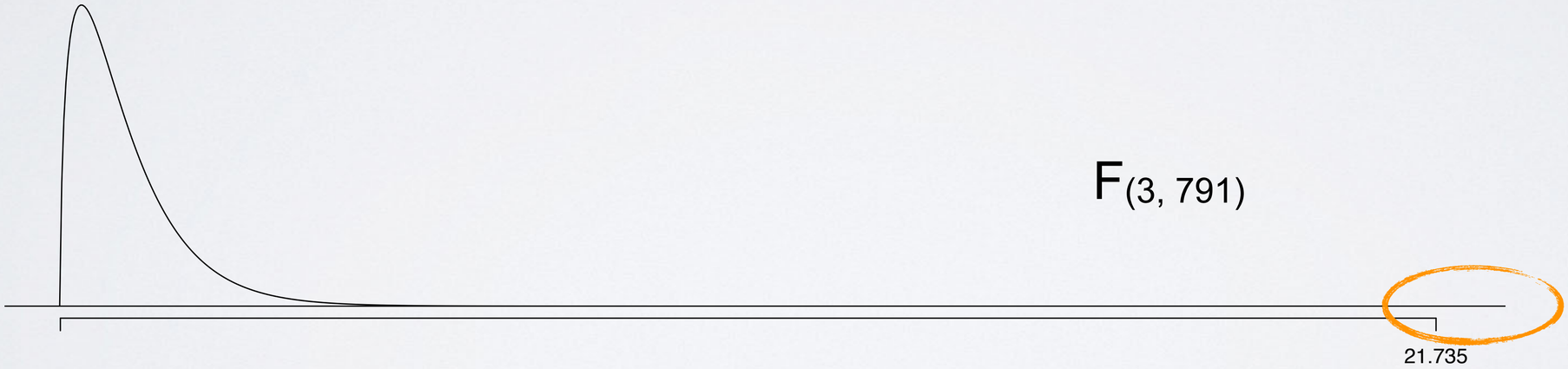
$$F = \frac{MSG}{MSE}$$

$$\frac{78.855}{3.628} \approx 21.735$$

p-value

		Df	Sum Sq	Mean Sq	F value	Pr(> F)
Group	class	3	236.56	78.855	21.735	<0.0001
Error	Residuals	791	2869.80	3.628		
	Total	794	3106.36			

- p-value is the probability of at least as large a ratio between the "between" and "within" group variabilities if in fact the means of all groups are equal
- area under the F curve, with degrees of freedom df_G and df_E, above the observed F statistic.



		Df	Sum Sq	Mean Sq	F value	Pr(> F)
Group	class	3	236.56	78.855	21.735	<0.0001
Error	Residuals	791	2869.80	3.628		
	Total	794	3106.36			

using R

R

```
> pf(21.735, 3, 791, lower.tail = FALSE)
[1] 1.559855e-13
```

using the applet

http://bitly.com/dist_calc

conclusion

		Df	Sum Sq	Mean Sq	F value	Pr(> F)
Group	class	3	236.56	78.855	21.735	<0.0001
Error	Residuals	791	2869.80	3.628		
	Total	794	3106.36			

- If p-value is small (less than α), reject H_0.
  - The data provide convincing evidence that at least one pair of population means are different from each other (but we can't tell which one).
- If p-value is large, fail to reject H_0.
  - The data do not provide convincing evidence that one pair of population means are different from each other, the observed differences in sample means are attributable to sampling variability (or chance).