Statistical Inference

Inference for Numerical Variables

Behnam Bahrak
Spring 2020

Vocabulary Score and Class

10 question vocabulary test (scores range from 0 to 10)

	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

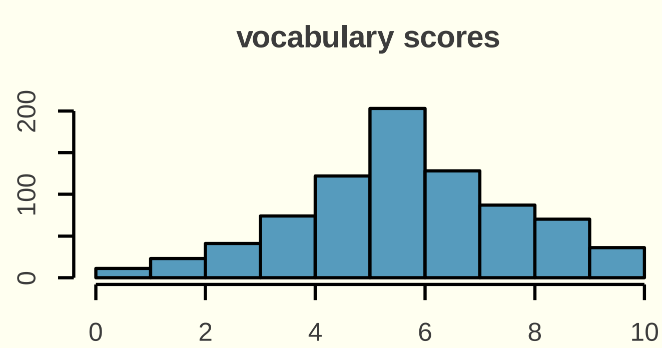
self identified social class (lower, working, middle, upper)

Vocabulary Score (`wordsum`)

➢ Choose a word from a list of provided options that comes closest to the meaning of the first word provided in capital letters.

1. SPACE (school, noon, captain, room, board, don't know)
2. BROADEN (efface, make level, elapse, embroider, widen, don't know)
3. EMANATE (populate, free, prominent, rival, come, don't know)
4. EDIBLE (auspicious, eligible, fit to eat, sagacious, able to speak, don't know)
5. ANIMOSITY (hatred, animation, disobedience, diversity, friendship, don't know)
6. PACT (puissance, remonstrance, agreement, skillet, pressure, don't know)
7. CLOISTERED (miniature, bunched, arched, malady, secluded, don't know)
8. CAPRICE (value, a star, grimace, whim, inducement, don't know)
9. ACCUSTOM (disappoint, customary, encounter, get used to, business, don't know)
10. ALLUSION (reference, dream, eulogy, illusion, aria, don't know)

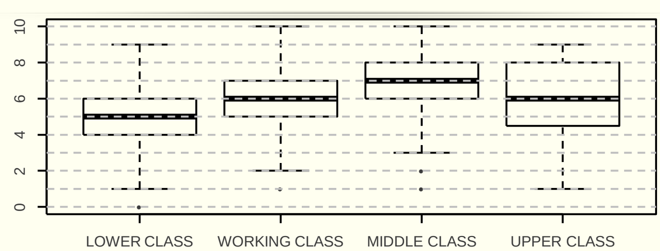
Vocabulary Score (`wordsum`)



Self Identified Social Class (`class`)

- If you were asked to use one of four names for your social class, which would you say you belong in: the lower class, the working class, the middle class, or the upper class?

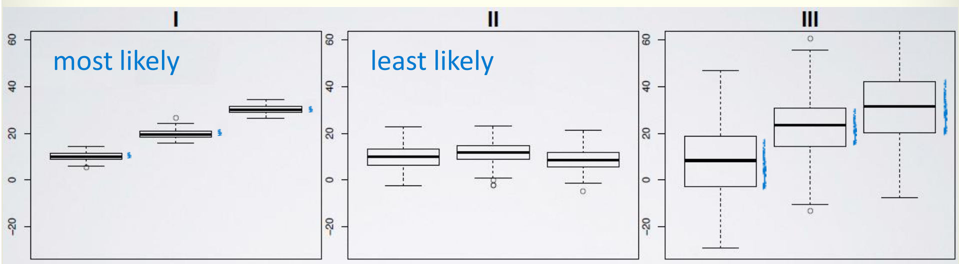
Exploratory Analysis



	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**

Question

- Which of the following plots shows groups with means that are most and least likely to be significantly different from each other?



Comparing more than two means

- Is there a difference between the average vocabulary scores of Americans from different (self reported) classes?
  - Compare means of 2 groups using a T-statistic.
  - Compare means of 3+ groups using a new test called analysis of variance (ANOVA) and a new statistic called F.

ANOVA

H_0: The mean outcome is the same across all categories

$$\mu_1 = \mu_2 = \cdots = \mu_k$$

μ_i : mean of the outcome for observations in category i

k : number of groups

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

ANOVA

T-test

Compare means from **two** groups: are so far apart that the observed difference cannot reasonably be attributed to sampling variability?

$$H_0 : \mu_1 = \mu_2$$

ANOVA

Compare means from **more than two** groups: are they so far apart that the observed differences cannot all reasonably be attributed to sampling variability?

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$$

ANOVA

اختلاف میانگین

T-test

Compute a test statistic (a ratio).

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{SE_{(\bar{x}_1 - \bar{x}_2)}}$$

ANOVA

Compute a test statistic (a ratio).

$$\boxed{F = \frac{\text{variability bet. groups}}{\text{variability w/in groups}}}$$

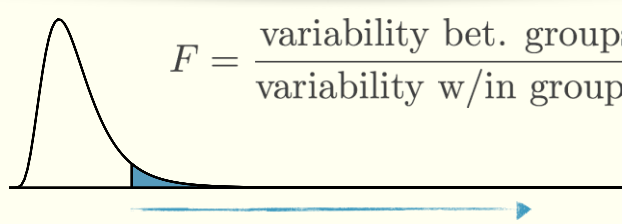
اختلاف بین گروهی

اختلاف درون گروهی

واریانس درون گروهی

- Large test statistics lead to small p-values.
- If the p-value is small enough H_0 is rejected, and we conclude that the data provide evidence of a difference in the population means.

F-statistic



$$F = \frac{\text{variability bet. groups}}{\text{variability w/in groups}} \begin{matrix} ? \\ ? \end{matrix}$$

- In order to be able to reject H_0, we need a small p-value, which requires a large F statistic.
- Obtaining a large F statistic requires that the variability between sample means is greater than the variability within the samples.

F-Distribution

$$f(x; d_1, d_2) = \frac{\sqrt{\frac{(d_1\,x)^{d_1}\,\,d_2^{d_2}}{(d_1\,x+d_2)^{d_1+d_2}}}}{x\,\mathrm{B}\!\left(\frac{d_1}{2},\frac{d_2}{2}\right)}$$

$$= \frac{1}{\mathrm{B}\!\left(\frac{d_1}{2},\frac{d_2}{2}\right)} \left(\frac{d_1}{d_2}\right)^{\frac{d_1}{2}} x^{\frac{d_1}{2}-1} \left(1+\frac{d_1}{d_2}\,x\right)^{-\frac{d_1+d_2}{2}}$$

$$f(t) = \frac{\Gamma\left(\frac{d+1}{2}\right)}{\sqrt{\pi d}\,\Gamma\left(\frac{d}{2}\right)}\left(1+\frac{t^2}{d}\right)^{-\frac{d+1}{2}}$$

Where $B(x, y)$ is the beta function:

$$\mathrm{B}(x,y) = \int_0^1 t^{x-1}(1-t)^{y-1}\,\mathrm{d}t$$

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

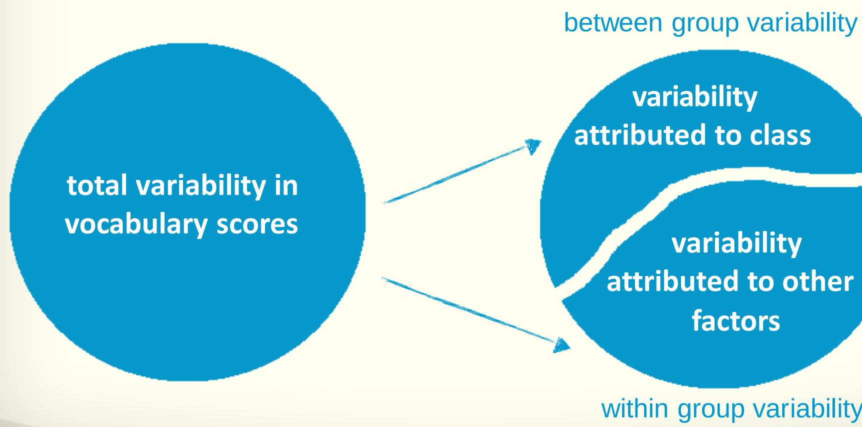
	n	mean	sd
lower class	41	5.07	2.24
working class	407	5.75	1.87
middle class	331	6.76	1.89
upper class	16	6.19	2.34
overall	**795**	**6.14**	**1.98**

H_0: The average vocabulary score is the same across all social classes.

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

H_A: The average vocabulary scores differ between at least one pair of social classes.

Variability Partitioning



ANOVA Table

		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			

- The first row is about the between group variability and the second row is the within group variability.
- We often refer to the first row as the group row and the second row as the error row.
- The third row displays the totals.

Sum of Squares Total

		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)

- SST measures the **total variability** in the response variable
- SST is 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
...	...	...
795	9	middle class

	n	mean	sd
overall	795	6.14	1.98

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

Sum of Squares Group

		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)

- SSG measures the variability **between groups**.
- **Explained variability:** squared deviation of group means 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**

$$\begin{aligned} SSG = &\left(41 \times (5.07 - 6.14)^2\right) \\ &+\left(407 \times (5.75 - 6.14)^2\right) \\ &+\left(331 \times (6.76 - 6.14)^2\right) \\ &+\left(16 \times (6.19 - 6.14)^2\right) \\ &\approx 236.56 \end{aligned}$$

Sum of Squares Error (SSE)

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

Sum of Squares Error (SSE)

$$\boxed{SSE = SST - SSG}$$

3106.36 - 236.56 = 2869.8

- SSG measures the variability **between groups**.
- **Unexplained variability:** squared deviation of group means from overall mean, weighted by sample size.

Mean Squares

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

- 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: $MSG = SSG/df_G$ → $236.56/3 \approx 78.855$
- Error: $MSE = SSE/df_E$ → $2869.8/791 \approx 3.628$

SST = SSG + SSE

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

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

$$SSE = \sum_{j=1}^{k} \sum_{i=1}^{n_j} \left(y_i - \bar{y}_j\right)^2 = \sum_{j=1}^{k} \left(n_j - 1\right) s_j^2$$

$$\Rightarrow MSE = \frac{1}{n-k} \sum_{j=1}^{k} \left(n_j - 1\right) s_j^2$$

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			

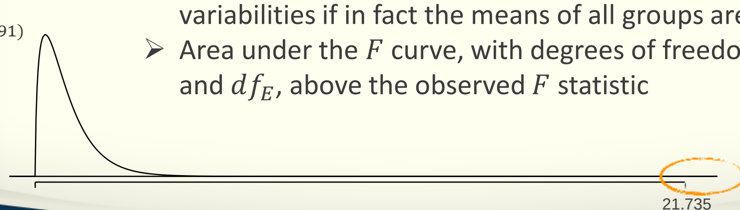
- **F statistic:** Ratio of the average between group and within group variabilities:

$$F = \frac{MSG}{MSE} \longrightarrow \frac{78.855}{3.628} \approx 21.735$$

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



Computing p-value using R

		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			

R

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

Making Decision

- 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 at least one pair of population means are different from each other, the observed differences in sample means are attributable to sampling variability (or chance).

Behnam Bahrak
bahrak@ut.ac.ir