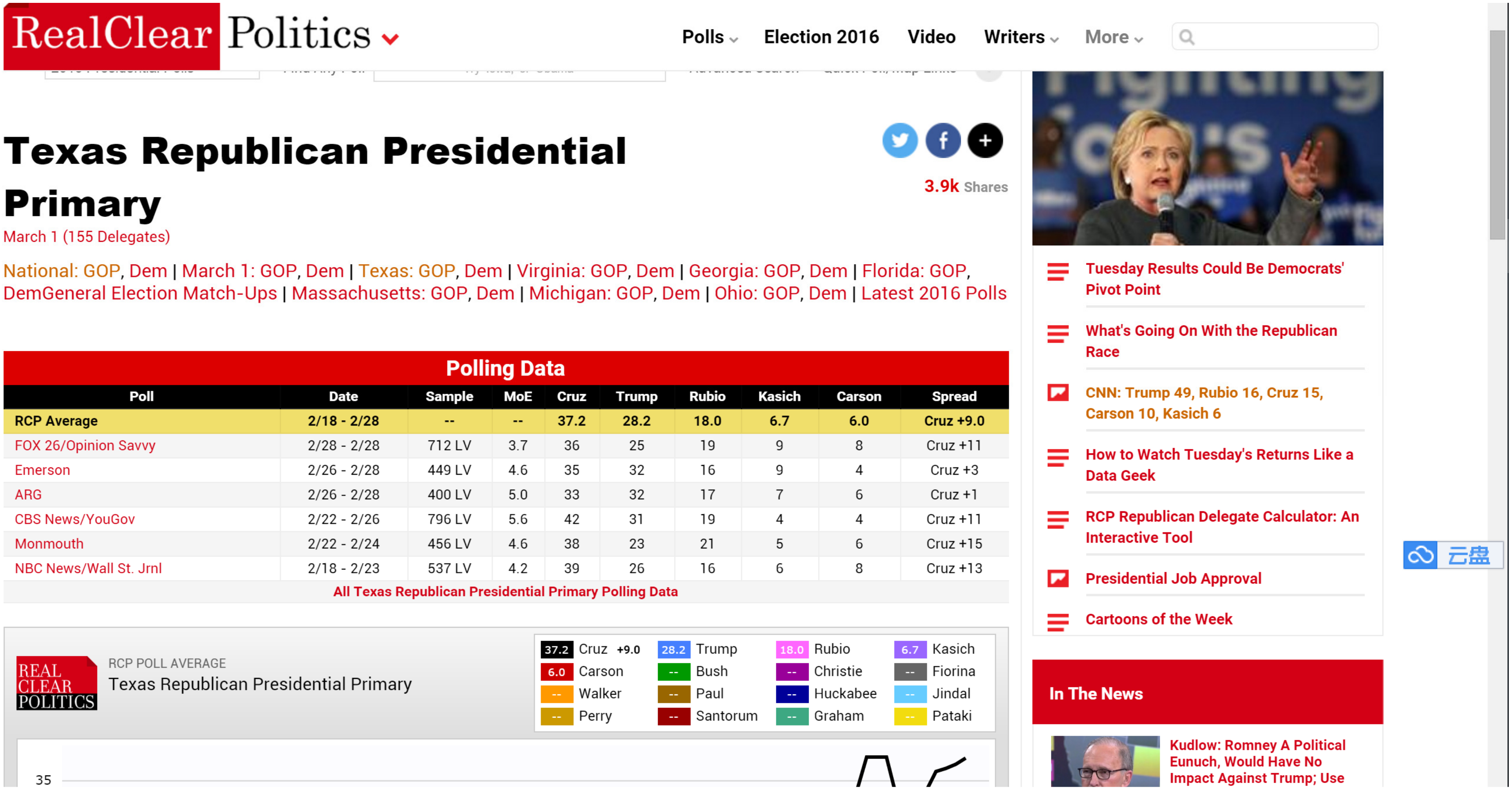
Inference of Numerical Data V

Dajiang Liu

@PHS 525

Mar-1st-2016

Something Fun



RealClear Politics

Polls ⌄ Election 2016 Video Writers ⌄ More ⌄

Texas Republican Presidential Primary

3.9k Shares

March 1 (155 Delegates)

National: GOP, Dem | March 1: GOP, Dem | Texas: GOP, Dem | Virginia: GOP, Dem | Georgia: GOP, Dem | Florida: GOP, DemGeneral Election Match-Ups | Massachusetts: GOP, Dem | Michigan: GOP, Dem | Ohio: GOP, Dem | Latest 2016 Polls

Polling Data

Poll	Date	Sample	MoE	Cruz	Trump	Rubio	Kasich	Carson	Spread
RCP Average	**2/18 - 2/28**	**--**	**--**	**37.2**	**28.2**	**18.0**	**6.7**	**6.0**	**Cruz +9.0**
FOX 26/Opinion Savvy	2/28 - 2/28	712 LV	3.7	36	25	19	9	8	Cruz +11
Emerson	2/26 - 2/28	449 LV	4.6	35	32	16	9	4	Cruz +3
ARG	2/26 - 2/28	400 LV	5.0	33	32	17	7	6	Cruz +1
CBS News/YouGov	2/22 - 2/26	796 LV	5.6	42	31	19	4	4	Cruz +11
Monmouth	2/22 - 2/24	456 LV	4.6	38	23	21	5	6	Cruz +15
NBC News/Wall St. Jrnl	2/18 - 2/23	537 LV	4.2	39	26	16	6	8	Cruz +13

All Texas Republican Presidential Primary Polling Data

REAL CLEAR POLITICS

RCP POLL AVERAGE

Texas Republican Presidential Primary

37.2 Cruz +9.0	28.2 Trump	18.0 Rubio	6.7 Kasich
6.0 Carson	-- Bush	-- Christie	-- Fiorina
-- Walker	-- Paul	-- Huckabee	-- Jindal
-- Perry	-- Santorum	-- Graham	-- Pataki

Tuesday Results Could Be Democrats' Pivot Point

What's Going On With the Republican Race

CNN: Trump 49, Rubio 16, Cruz 15, Carson 10, Kasich 6

How to Watch Tuesday's Returns Like a Data Geek

RCP Republican Delegate Calculator: An Interactive Tool

Presidential Job Approval

Cartoons of the Week

In The News

Kudlow: Romney A Political Eunuch, Would Have No Impact Against Trump; Use

Motivational Problem

- We have thoroughly discussed how to perform two-sample inference
  - How to compare if the sample mean in two different groups differ
    - We have learnt how to perform
      - T-test
        - When sample sizes are small, but sample distribution is near normal
      - Normal test:
        - When sample sizes are large, but sample distribution does not have to be normal

- But how to compare the sample mean differences between multiple groups
  - What is the ideas:
    - Compare pairwise differences
    - Compare if at least one pair have different sample mean value

ANOVA

- ANOVA stands for analysis of variance
- ANOVA compares if the sample means differ across multiple groups
- ANOVA uses a different statistic
  - F-statistic

- Hypotheses tested:
  - H_0: The mean outcome is the same across different groups, i.e. $\mu_1 = \mu_2 = \cdots = \mu_N$
  - H_A: At least one pairs of mean values are different

Three Conditions to be Verified Before ANOVA

- Samples are independent within and between groups

- Samples within each group are nearly normal

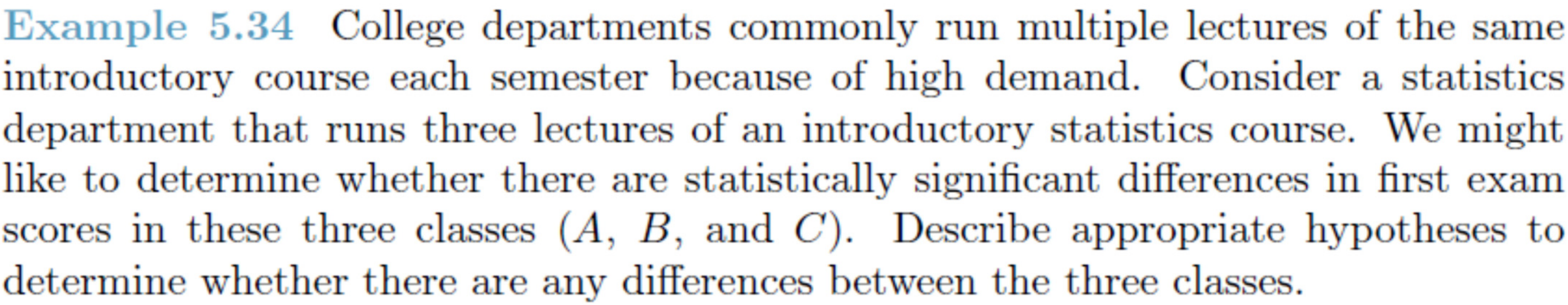
- Variability across group are about equal

How to Check for These Conditions

- Sample independence:
  - Samples are chosen from <10% of the population

- Sample normality:
  - qqnorm command
- Variability across groups
  - boxplot

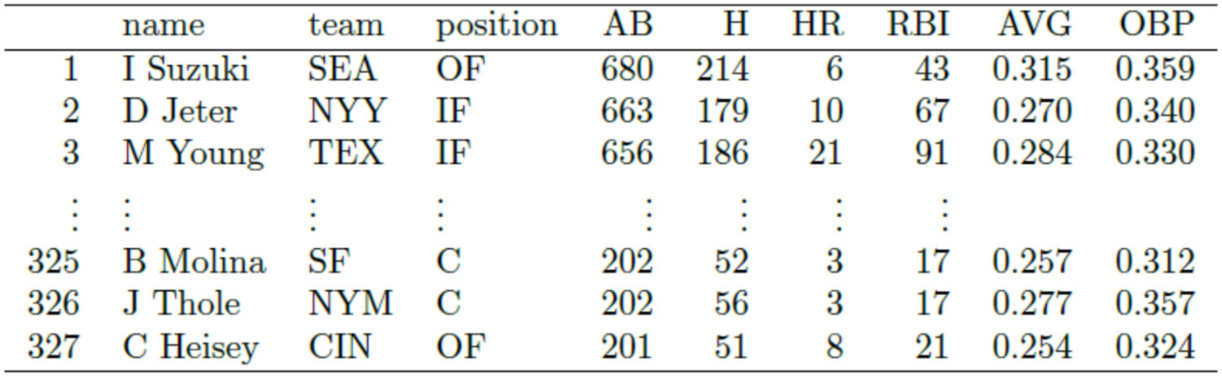
Example



- Example 5.34 College departments commonly run multiple lectures of the same introductory course each semester because of high demand. Consider a statistics department that runs three lectures of an introductory statistics course. We might like to determine whether there are statistically significant differences in first exam scores in these three classes (A, B, and C). Describe appropriate hypotheses to determine whether there are any differences between the three classes.

Example – Examine if Batting Performance Differ between Positions

- Dataset: bat10
- Batting performance is evaluated by the statistic OBP (on-base percentage)



	name	team	position	AB	H	HR	RBI	AVG	OBP
1	I Suzuki	SEA	OF	680	214	6	43	0.315	0.359
2	D Jeter	NYY	IF	663	179	10	67	0.270	0.340
3	M Young	TEX	IF	656	186	21	91	0.284	0.330
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮		
325	B Molina	SF	C	202	52	3	17	0.257	0.312
326	J Thole	NYM	C	202	56	3	17	0.277	0.357
327	C Heisey	CIN	OF	201	51	8	21	0.254	0.324

Table 5.25: Six cases from the bat10 data matrix.

Guiding Questions

- What is the hypothesis to be tested in order to examine if the OBP differs between groups?

- What is the appropriate point estimate for mean value of OBP within each group?

- How to estimate it in R?

ANOVA and F-test

- Questions answered: Is the sample means between group so far that it cannot be due to chance alone?
- Notations are a bit different from the textbook
- Subjects in group $i = 1, \dots, I$
  - $X_{ij}, j = 1, \dots, n_i$
  - $\sum_i n_i = N$

ANOVA and F-test

- Sum of squares between groups: (SSG)

$$SSG = \sum_i n_i(\bar{X}_i - \bar{X})^2$$

- Total sum of squares (SST)

$$SST = \sum_{i,j} \left(X_{ij} - \bar{X}\right)^2$$

- Residual sum of squares (SSE)

$$SSE = SST - SSG$$

F-statistic

- MSG: Mean squares between groups:
  - $df_G = I - 1$

$$MSG = \frac{SSG}{df_G} = \frac{SSG}{I - 1}$$

- MSE: Mean squared error
  - $df_E = N - I$

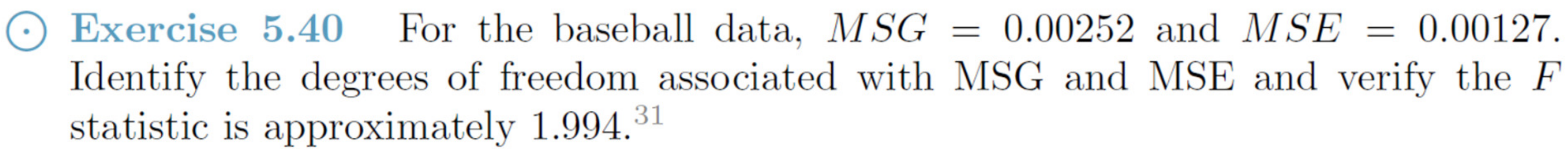
$$MSE = \frac{SSE}{df_E} = \frac{SSE}{N - I}$$

- F statistic is equal to

$$F = MSG/MSE$$

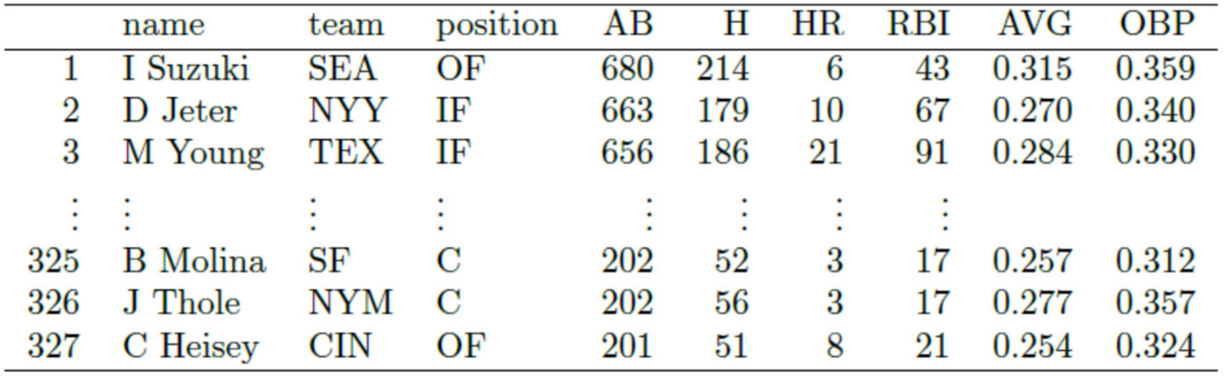
- F statistic follows a F-distribution with $df_1 = df_G, df_2 = df_E$

Exercise:



⊙ **Exercise 5.40** For the baseball data, $MSG = 0.00252$ and $MSE = 0.00127$. Identify the degrees of freedom associated with MSG and MSE and verify the F statistic is approximately 1.994.[31]

What is the p-value for the F-statistic?



	name	team	position	AB	H	HR	RBI	AVG	OBP
1	I Suzuki	SEA	OF	680	214	6	43	0.315	0.359
2	D Jeter	NYY	IF	663	179	10	67	0.270	0.340
3	M Young	TEX	IF	656	186	21	91	0.284	0.330
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮		
325	B Molina	SF	C	202	52	3	17	0.257	0.312
326	J Thole	NYM	C	202	56	3	17	0.277	0.357
327	C Heisey	CIN	OF	201	51	8	21	0.254	0.324

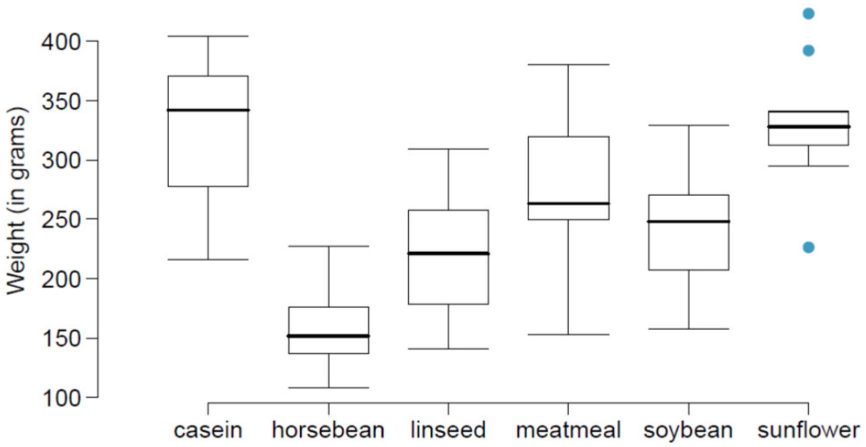
Table 5.25: Six cases from the `bat10` data matrix.

Exercise

5.37 Chicken diet and weight, Part III. In Exercises 5.29 and 5.31 we compared the effects of two types of feed at a time. A better analysis would first consider all feed types at once: casein, horsebean, linseed, meat meal, soybean, and sunflower. The ANOVA output below can be used to test for differences between the average weights of chicks on different diets.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
feed	5	231,129.16	46,225.83	15.36	0.0000
Residuals	65	195,556.02	3,008.55		

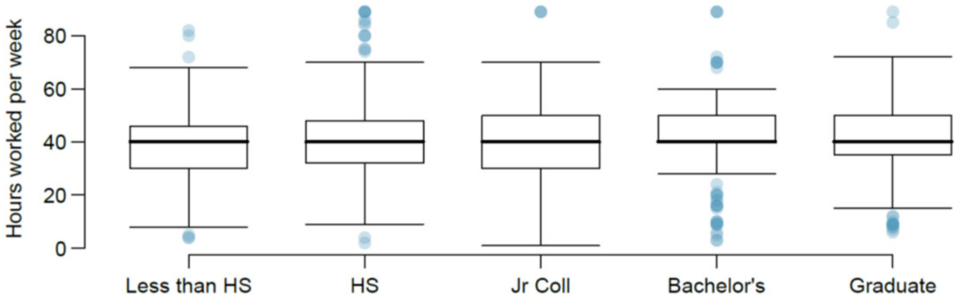
Conduct a hypothesis test to determine if these data provide convincing evidence that the average weight of chicks varies across some (or all) groups. Make sure to check relevant conditions. Figures and summary statistics are shown below.



	Mean	SD	n
casein	323.58	64.43	12
horsebean	160.20	38.63	10
linseed	218.75	52.24	12
meatmeal	276.91	64.90	11
soybean	246.43	54.13	14
sunflower	328.92	48.84	12

5.40 Work hours and education, Part III. In Exercises 5.8 and 5.10 you worked with data from the General Social Survey in order to compare the average number of hours worked per week by US residents with and without a college degree. However, this analysis didn't take advantage of the original data which contained more accurate information on educational attainment (less than high school, high school, junior college, Bachelor's, and graduate school). Using ANOVA, we can consider educational attainment levels for all 1,172 respondents at once instead of re-categorizing them into two groups. Below are the distributions of hours worked by educational attainment and relevant summary statistics that will be helpful in carrying out this analysis.

	Educational attainment					
	Less than HS	HS	Jr Coll	Bachelor's	Graduate	Total
Mean	38.67	39.6	41.39	42.55	40.85	40.45
SD	15.81	14.97	18.1	13.62	15.51	15.17
n	121	546	97	253	155	1,172



(a) Write hypotheses for evaluating whether the average number of hours worked varies across the five groups.

(b) Check conditions and describe any assumptions you must make to proceed with the test.

(c) Below is part of the output associated with this test. Fill in the empty cells.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
degree			501.54		0.0682
Residuals		267,382			
Total					

(d) What is the conclusion of the test?