GSS Data Importation & Exploration

Paul E. Johnson, CRMDA
B. Wade Garrison, CFIE

University of Kansas

What In The World is the General Social Survey?

- NORC (Nat'l Opinion Research Center)
- Administered Annually or Bi-annually
  - Permanent Questions
  - Question Modules (addressed to subsets of respondents)
  - Some Questions only asked once or twice
- Davis, James A., Tom W. Smith, and Peter V. Marsden. *General Social Surveys, 1972-2006 [Cumulative File]* Storrs, CT: Roper Center for Public Opinion Research, [Computer file]. ICPSR04697-v4 University of Connecticut/Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributors], 2009-12-04.

Where In The World is the General Social Survey?

- It turns out it is available in many places, in various subsets
- The ICPSR (U. of Michigan) is a canonical source
  - `http://www.icpsr.umich.edu/icpsrweb/ICPSR/`
  - There are Individual GSS sets for some individual years, but I want 2006
  - `GeneralSocialSurveys,1972-2006[CumulativeFile]`
    - ★ "Download All" button grabs a file called "10805932.zip"

What's in that Zip File?

- Unzip that to reveal the contents:
  - TermsOfUse.html
  - ICPSR_04697: a folder
  - Inside ICPSR_04697

04697-descriptioncitation.pdf	how to cite
04697-manifest.txt	list of files
04697-related_literature.txt	citations
series-28-related_literature.txt	citations
DS0001	Another Folder

Its Like Peeling an Onion: Inside DS0001

04697-0001-Codebook.pdf	List of Variables & Summary Info
04697-0001-Data.dta	Stata Format
04697-0001-Data.sav	SPSS save Format
04697-0001-Data.stc	Terrasoft SAS data file
04697-0001-Data.tsv	tab-separated text
04697-0001-Data.txt	"fixed field" text
04697-0001-Setup.dct	Stata "data dictionary"
04697-0001-Setup.do	Example Stata code uses txt*
04697-0001-Setup.sas	Example SAS code uses txt*
04697-0001-Setup.sps	Example SPSS code uses tsv*
04697-0001-Supplemental_syntax.do	Cleans up missing values
04697-0001-Supplemental_syntax.sas	Cleans up missing values

Wrestle that Data into some Stats Program

- Consider "R", `http://www.r-project.org`: "free", "open source"
- Some POLS profs prefer Stata
- Some in Quant Psych seem to prefer SAS
- The SPSS sav and the Stata dta files are "predigested" (for good or ill)
- The SAS control file (or Stata do file) will require editing to import data

Got a Bull By the Horns? Or Does It Have You?

This data set includes responses from several thousand people in each of 26 surveys
In each survey, there will be 1500-3000 respondents.
Laptop can't handle this file (even with 2GB memory)

Imagine a spreadsheet with 51020 rows and 5137 columns

year	id	Q1	Q2	Q3	Q4	Q5	$\cdots$	Q5135
1972	1	3	1	NAP	NAP	NAP		3
1972	2	2	2	NAP	NA	NAP		2
$\vdots$								
1974	3455	3	3	5	NAP	NAP		NA
1974	3456	1	2	6	NAP	NAP		NA
$\vdots$								
2006	44555	NAP	4	NAP	1	4		3
2006	44556	NAP	5	NAP	2	3		1

NAP: question not included in survey for that year or for that respondent

NAs: ordinary missings are also observed

Need to Extract a Subset

- Use your software to extract the questions & years you want
- Annual Extracts on the NORC website
  - `http://www.norc.org/GSS+Website/Download`
    - ★ Problem: don't always include all questions/modules
- Can build "customized" sets of variables and years with the SDA at Berkeley
  - `http://sda.berkeley.edu/cgi-bin/hsda?harcsda+gss08`
- Either way, you have to "dig around" to see which variables exist, for which respondents

"SEXFREQ" Sounds More Interesting Than It Really Is

- SEXFREQ : About how often did you have sex during the last 12 months?

Val	Label	1989	1990	⋮	1994	⋮	2006
0	not at all	298	110		563		595
1	once or twice	99	39		188		205
2	once a month	114	57		290		265
3	2-3 times a month	221	91		416		361
4	weekly	258	110		483		343
5	2-3 per week	307	108		538		430
6	4+ per week	64	37		155		134
8	don't know	0	0		3		6
-1/9	NAP/NA	136/40	199/621		201/155		2096/75
	Valid N	1361	552		2533		2333

- Never included in surveys before 1989
- Asked of *some* respondents in other years

Voter Participation in 2006

Cell Contents

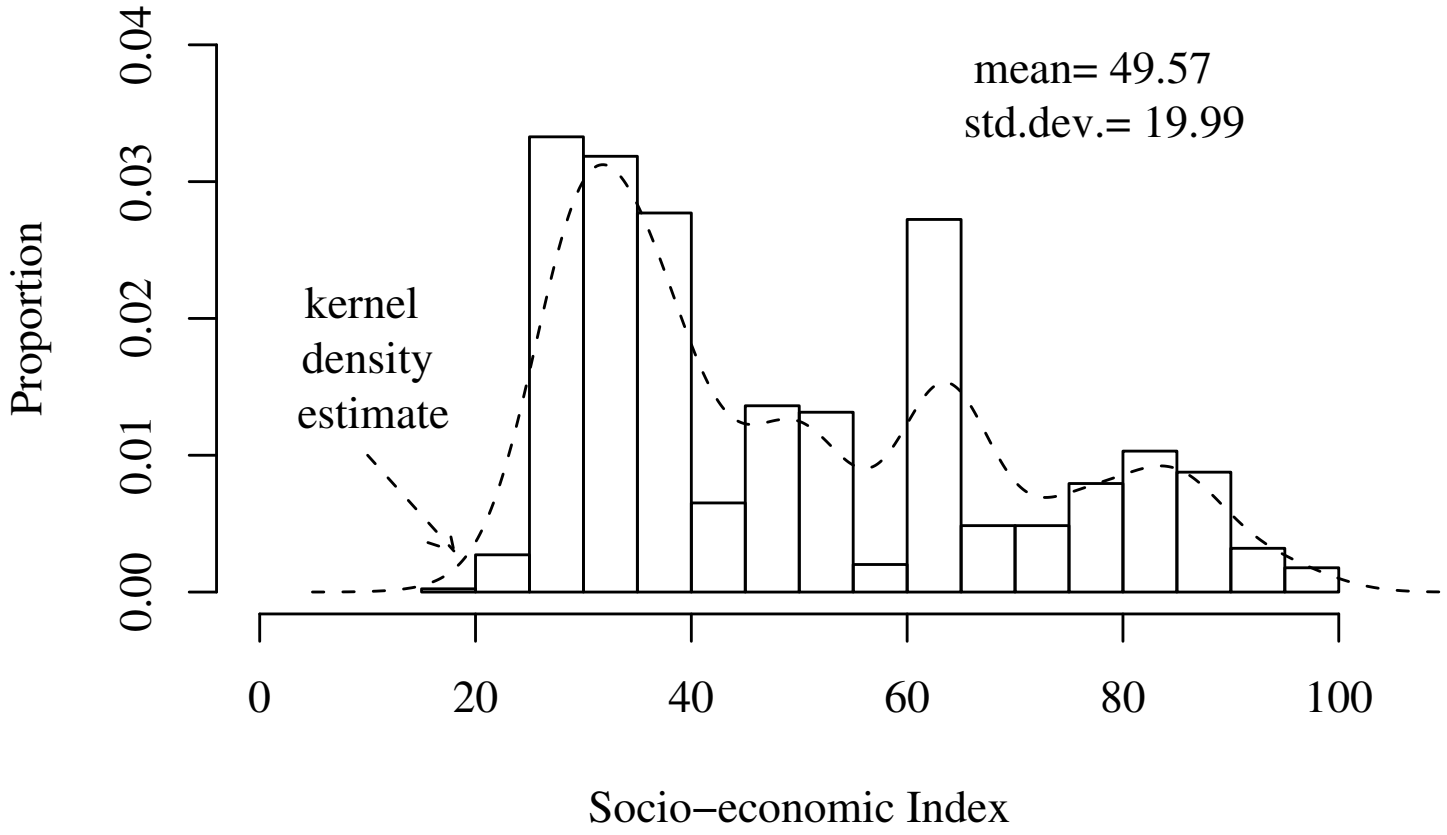
	VOTED	DID NOT VOTE	INELIGIBLE
N	1826	715	389
N / Table Total	0.623	0.244	0.133

I Terrorize the Students with my "Iron Law of Crosstabs"

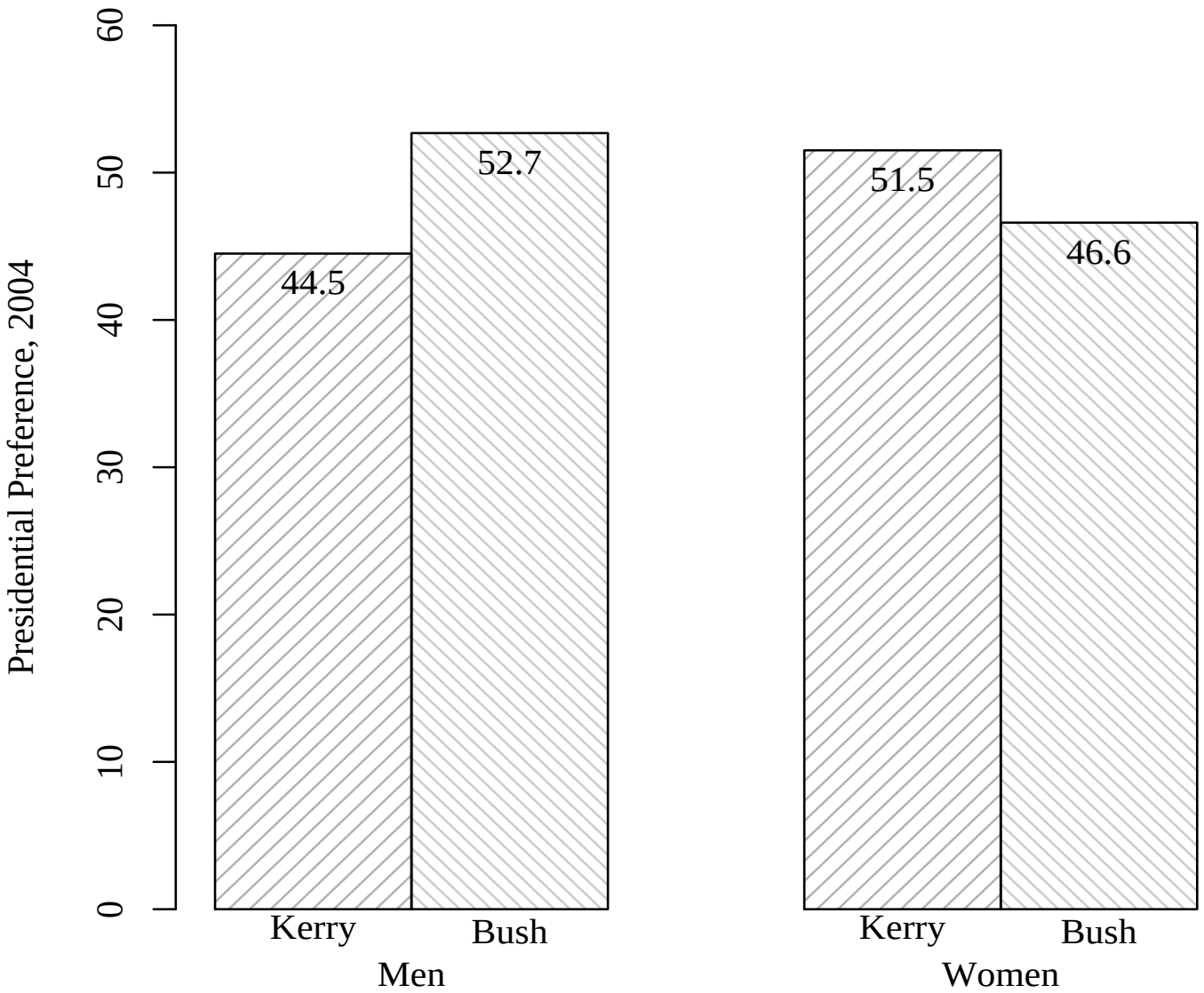
Voter Participation by Sex in 2006

	MALE	FEMALE
VOTED	61%	63%
DID NOT VOTE	25	24
INELIGIBLE	14	13
REFUSED TO ANSWER	0	0
N	1273	1657

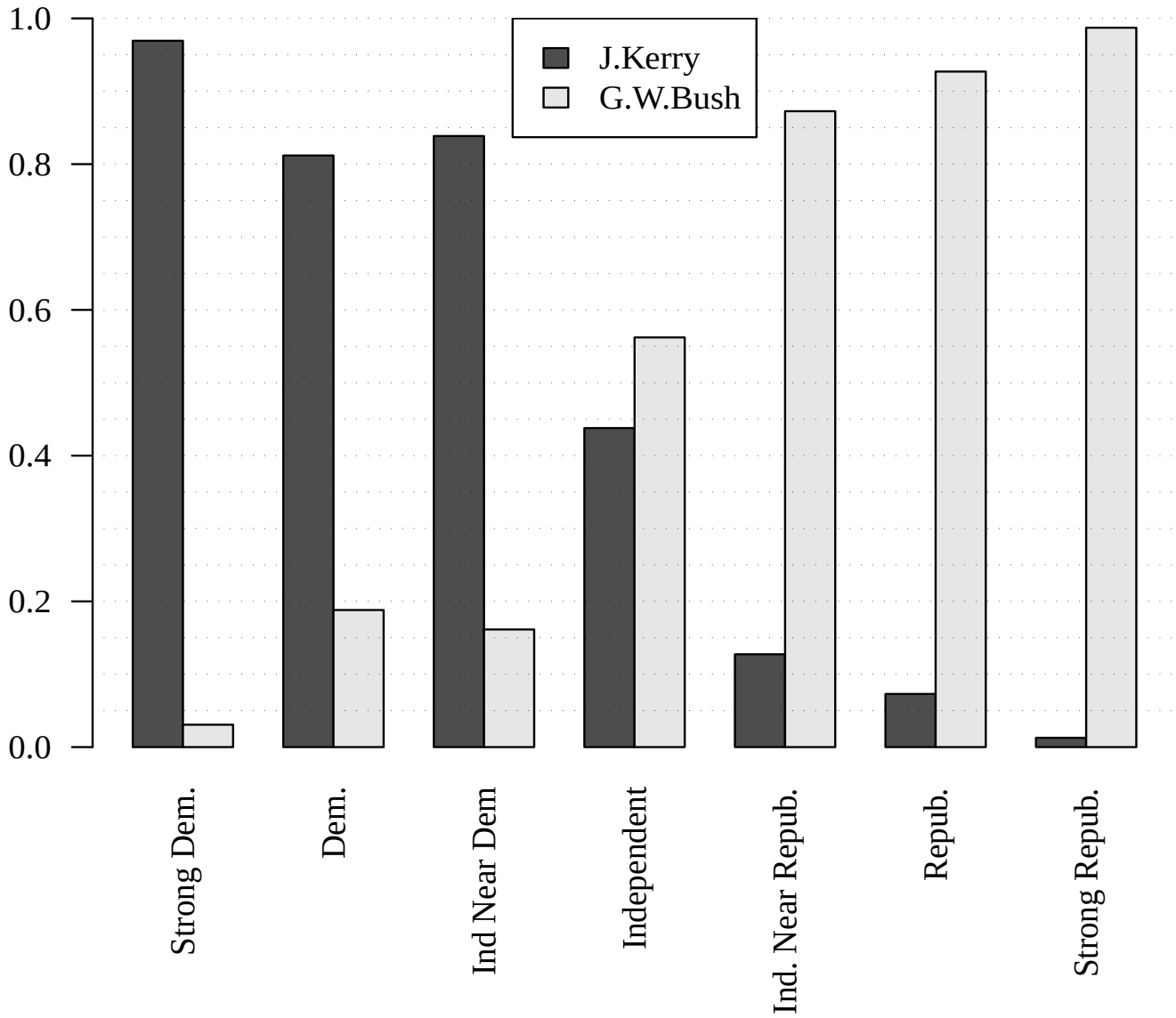
Socio-Economic Status of Men



The Gender Gap in Voting



The BarPlot of the Century (so far)



Probing Promiscuity (Box & Whisker Plot)