Secondary Analysis and Public Use Data Sets

Paul E. Johnson[1], B. Wade Garrison[2]

[1]Center for Research Methods and Data Analysis
[2]Center for Faculty Initiatives and Engagement
University of Kansas

April 30, 2014

Using Other Peoples' Data

This happens A LOT to SOME PEOPLE in SOME FIELDS

- Advisor says "go get Roper Poll from 1996 on death penalty"
- Student goes looking, finds a "fixed field" text file with numbers like:

```
12134362614451116561 1
12145256236223523422 5
12153342577373326783 9
12167457432533211457 1
12177467321567876868 6
```

- Student needs a magical decoder ring
  - Definition of information in each column
  - Codebook that explains what the columns mean

No Commonly Accepted Format Exists

Even after "all this time"

- There is still no "mandatory" data storage format that will always be accepted
- Efforts to create generic, commonly accepted formats have been resisted by commercial software developers
- Consumers are generally "oblivious" until it is too late to escape
- Hence, anybody can use anything and expect other people to learn to use it
- Expect lots of jargon about "metadata," "compressed storage," and so forth

Some Examples We Encounter Regularly

- American National Election Study (ANES)
- General Social Survey (GSS)
- American Social Capital Community Benchmark Survey
- Survey of Income Program Participation (SIPP)
- National Longitudinal Survey

Sections are "open", "public", unrestricted
Some material closely held.

- DOE projects: NLTS (National Longitudinal Transition Study)
- DOD projects: Surveys in Afghanistan, Iraq, etc

"Secret" Confidential data that is very closely guarded

Web Access For "Taste Testing"

- http://sda.berkeley.edu "SDA: Survey Documentation and Analysis"

SDA: Survey Documentation and Analysis

Home | Archive | Community | Documentation | Projects | News

SDA is a set of programs for the documentation and Web-based analysis of survey data. SDA is developed and maintained by the Computer-assisted Survey Methods Program (CSM) at the University of California, Berkeley. CSM also develops the CASES software package.

To see how it all works, test-drive SDA at our demonstration SDA Archive. Browse the documentation for a survey and get *fast* data analysis results. The SDA Archive includes several datasets, including the **General Social Survey (GSS)** and the **American National Election Study (ANES)**. You can also look at some other archives that use SDA software.

SDA Features

Documentation:

- **Codebooks**: SDA can produce both HTML and print-format codebooks. The documentation for each study contains a full description of each variable, indexes to the variables, and links to study-level information.
- **DDI (Data Documentation Initiative) compatibility**: SDA programs can produce DDI-format metadata from SDA datasets and from other metadata formats. SDA also provides an online utility that converts DDI metadata to SDA's own metadata format (DDL).

Analysis:

- **Various analysis types are available**: frequencies and crosstabulation, comparison of means, correlation matrix, comparison of correlations, multiple regression, logit/probit regression.
- **Fast results**: SDA was designed to produce analysis results *very quickly* -- within seconds -- even for large datasets with millions of cases and thousands of variables. Although many of our users assume we are using some sort of super computer to achieve these speeds, the secret lies solely in the method of storing the data and the design of the programs. The SDA Archive on our site runs on a low-cost (Intel) Linux server -- although versions of SDA are also available for Windows and (Sparc-based) Solaris.

SDA (cont.)

- Several large general purpose datasets have been put into an SDA format for interactive exploration.
- Probably not sufficient to do a whole project, but sufficient to explore and decide if you really want to use this data.

American National Election Study

Paul E. Johnson[1], B. Wade Garrison[2]

[1]Center for Research Methods and Data Analysis
[2]Center for Faculty Initiatives and Engagement
University of Kansas

April 30, 2014

American National Election Study

- University of Michigan Election Studies 1950s $\rightarrow$
  - Pre/Post surveys sandwich presidential elections
  - "Off year" Congressional election surveys
- Keystone of "behavioral movement" in political science
- Seminal publication: Angus Campbell, Philip Converse, Warren Miller, Donald Stokes, *The American Voter* (1960)
- Current Website: `http://www.electionstudies.org`

Capsule History

- Pre 1977, U. Michigan funded & 'controlled' this project
- 1977, NSF funded:
  - Michigan Election Study →American National Election Study
  - Community participation in design
  - Constant concern about costs and affordability of face-to-face interviews

The ANES is like a Box of Chocolates

You Never Know What You'll Get (Gump, 1994)

- The ANES has some "permanent" questions
- Many variations in design & administration from year-to-year
- Occasional special features
  - "Validated" voter turnout data
  - "Snowball" samples
  - Contextual data
  - Randomized ordering of questions (check for order and framing effects)

How to get the ANES

- Website: `http://www.electionstudies.org/studypages/download/datacenter_all.htm`
- Cumulative Data File (CDF)
- "Time Series" datasets are the "snapshots" that go into the CDF
- Other "panel" datasets
- I chose the 2008 "Time Series" dataset: "anes2008prepost.zip"

The 2008 Time Series study

"The sample of the ANES 2008 Time Series Study consisted of a new cross-section of respondents that yielded 2,323 face-to-face interviews in the pre-election study; 2,102 of which later provided a face-to-face interview in the post-election study.

STUDY CONTENT HIGHLIGHTS: In addition to content on electoral participation, voting behavior, and public opinion, the 2008 ANES Time Series Study contains questions in other areas such as media exposure, cognitive style, and values and predispositions. Special-interest and topical content provided significant coverage of foreign policy, including the war on terrorism and the war in Iraq. In addition, the study carried expanded instrumentation on organizational membership, unemployment, the federal budget, modern sexism, and race and gender politics. The Post-Election interview also included Module 3 from the Comparative Study of Electoral Systems (CSES). "(2008 Study Website)

What's in "anes2008prepost.zip"

2008prepost UsersGuide.pdf	Info on Sampling and Special Features

ANES2008TS.por	SPSS "portable" file
anes2008TS_dat.txt	comma separated variable text (CSV)
sas.zip	zip archive: SAS command files
spss.zip	zip archive: 5 SPSS command files
stata.zip	archive: 5 Stata 'do' files and a Stata data dictionary
documentation.zip	archive: pre/post Codebooks: Surveys that were administered with coding information

This one only offers one "ready to use" file, the SPSS portable file. They expect users of other programs to adapt code & do work to "read in" the CSV file.

The Data is 2323 Lines with 1963 Columns

The pre-election survey had 2323 respondents, some of whom were also interviewed in the post election wave.

They always use generic names like V081212 for their variables

	Resp. ID	Pre-election Questions		⋮	Post-election
version	V080001	V080101	V080101a	⋮	V085409a
2008TS	1	3	4		
2008TS	2	6	2		
2008TS	3	8	4		
2008TS	4	1	2		
2008TS	⋮	⋮	⋮		
2008TS	2321	2	4		NAP
2008TS	2322	1	9		NAP
2008TS	2323	3	9		NAP

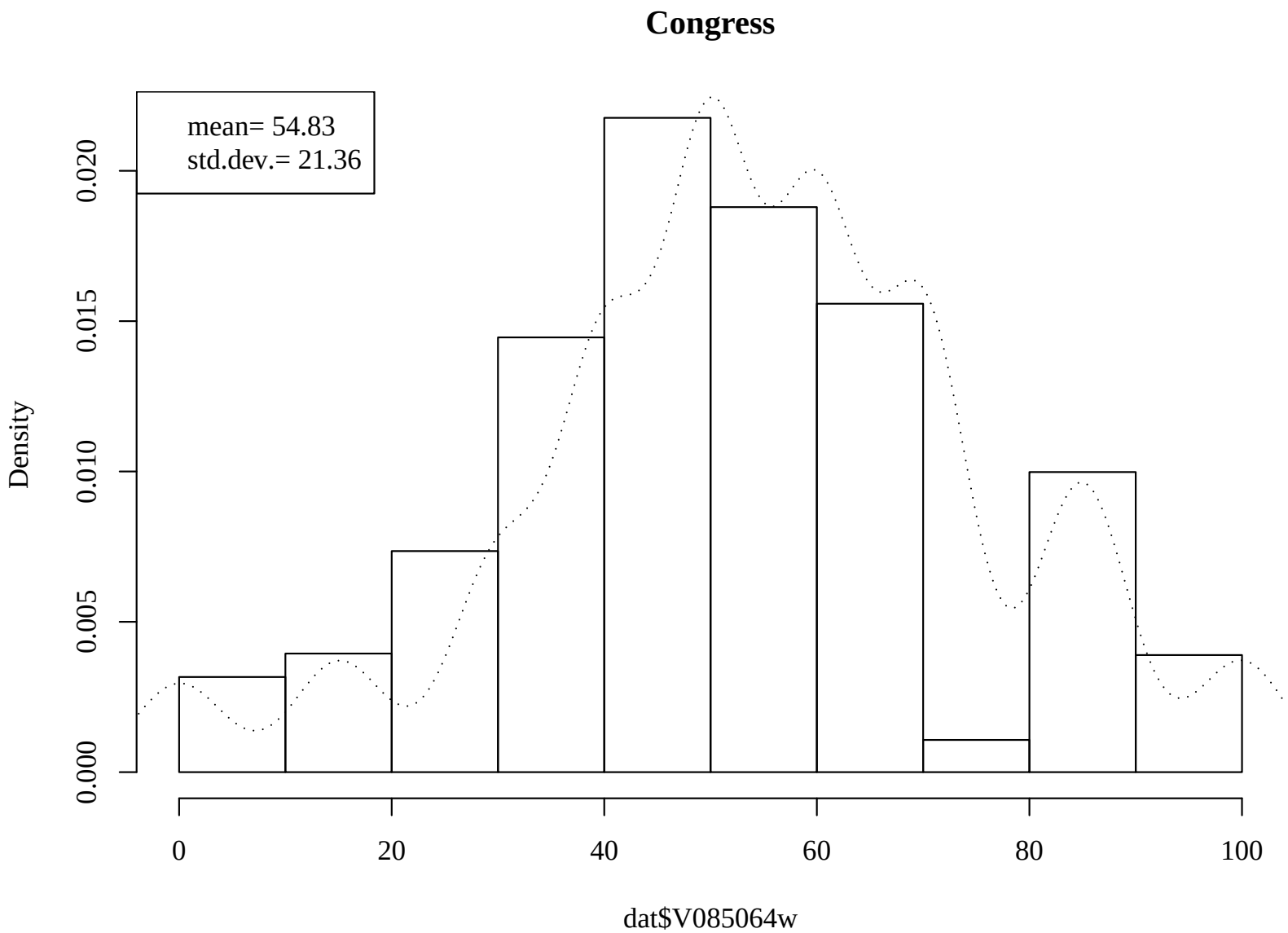
You Need to Read the Codebook/Questionnaire

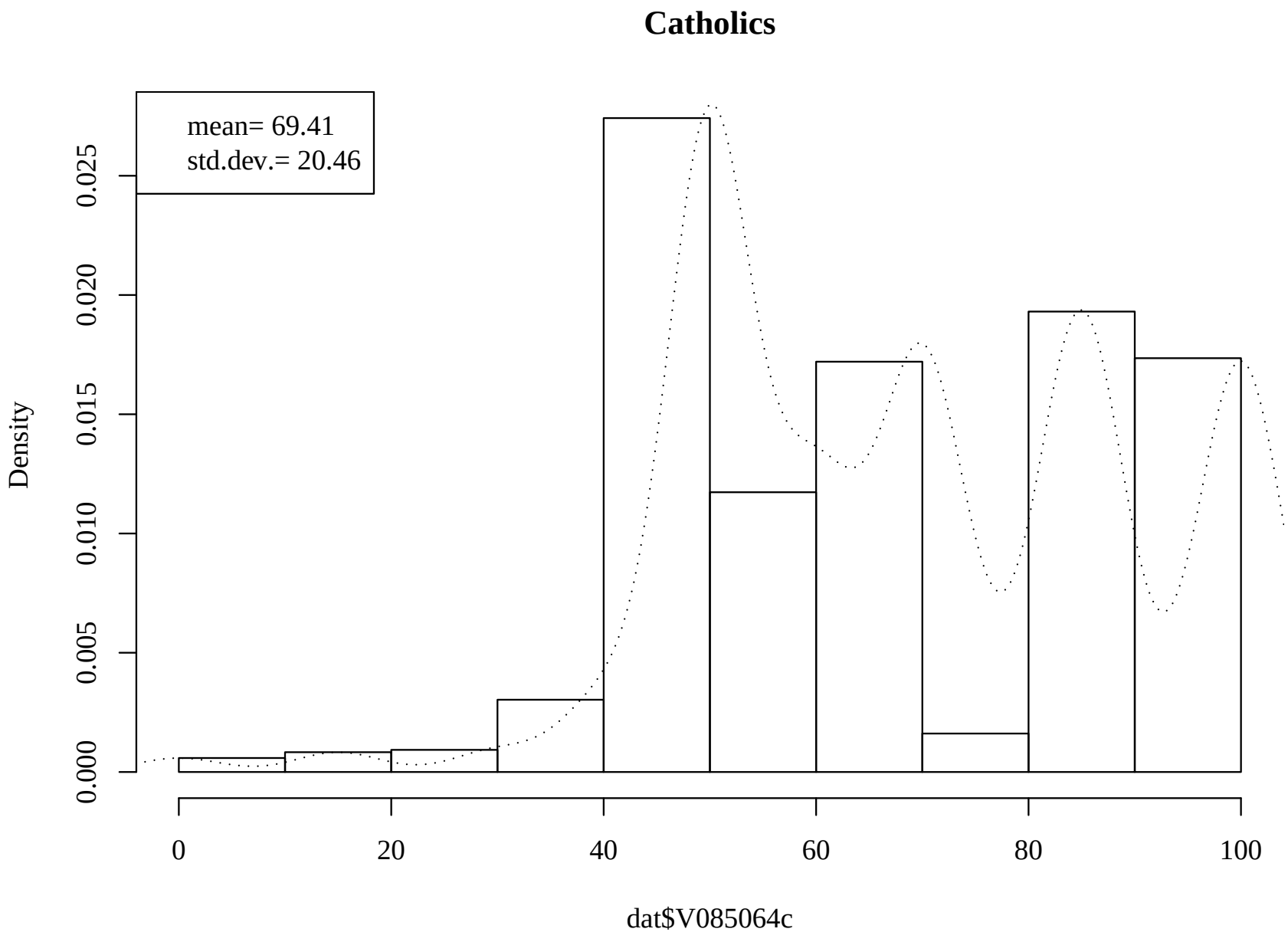
- Codebook provides variable names
- Numerical values assigned for the answers
- Exact question wording

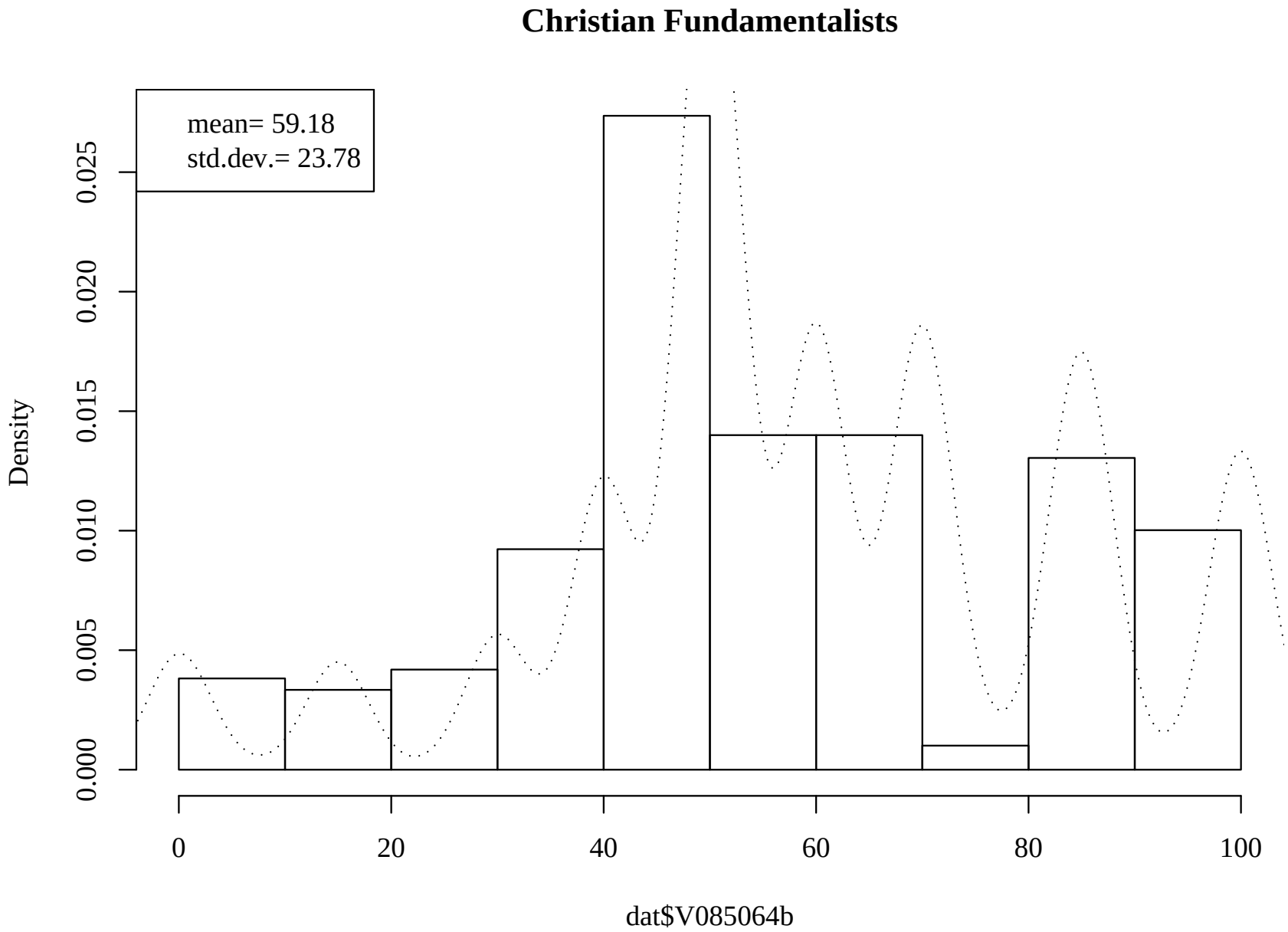
Feeling Thermometers
are a special feature of the ANES

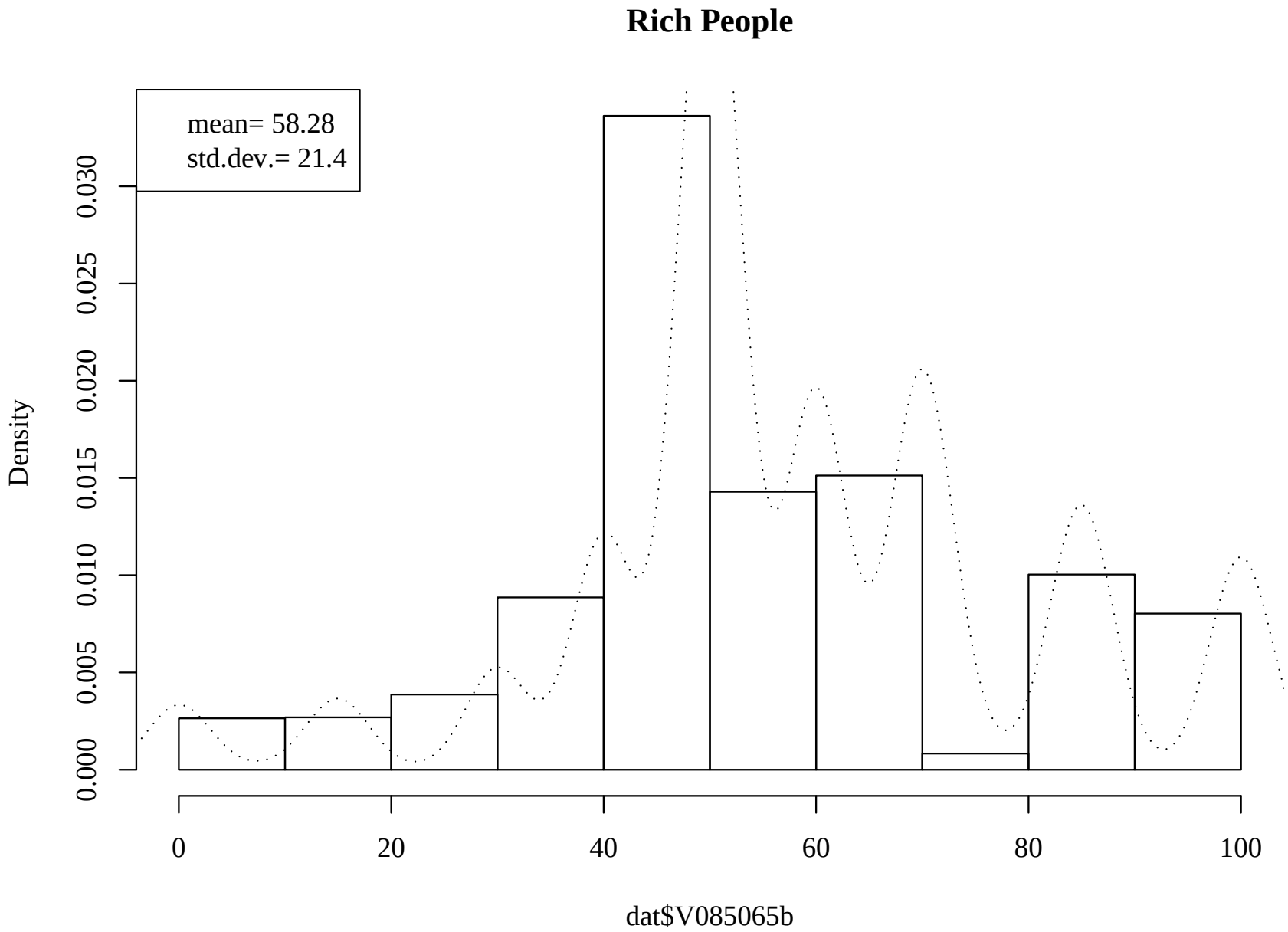
- Many demographic variables
- All of the usual "political" stuff, whether respondents
  - paid attention to the campaign
  - consider selves member of a political party
  - voted
  - preferred a candidate
- Also they have a raft of "Thermometer" variables
  - 0-100 scores indicating favorability toward a group or entity
  - 2008 survey "randomized" the presentation of the groups

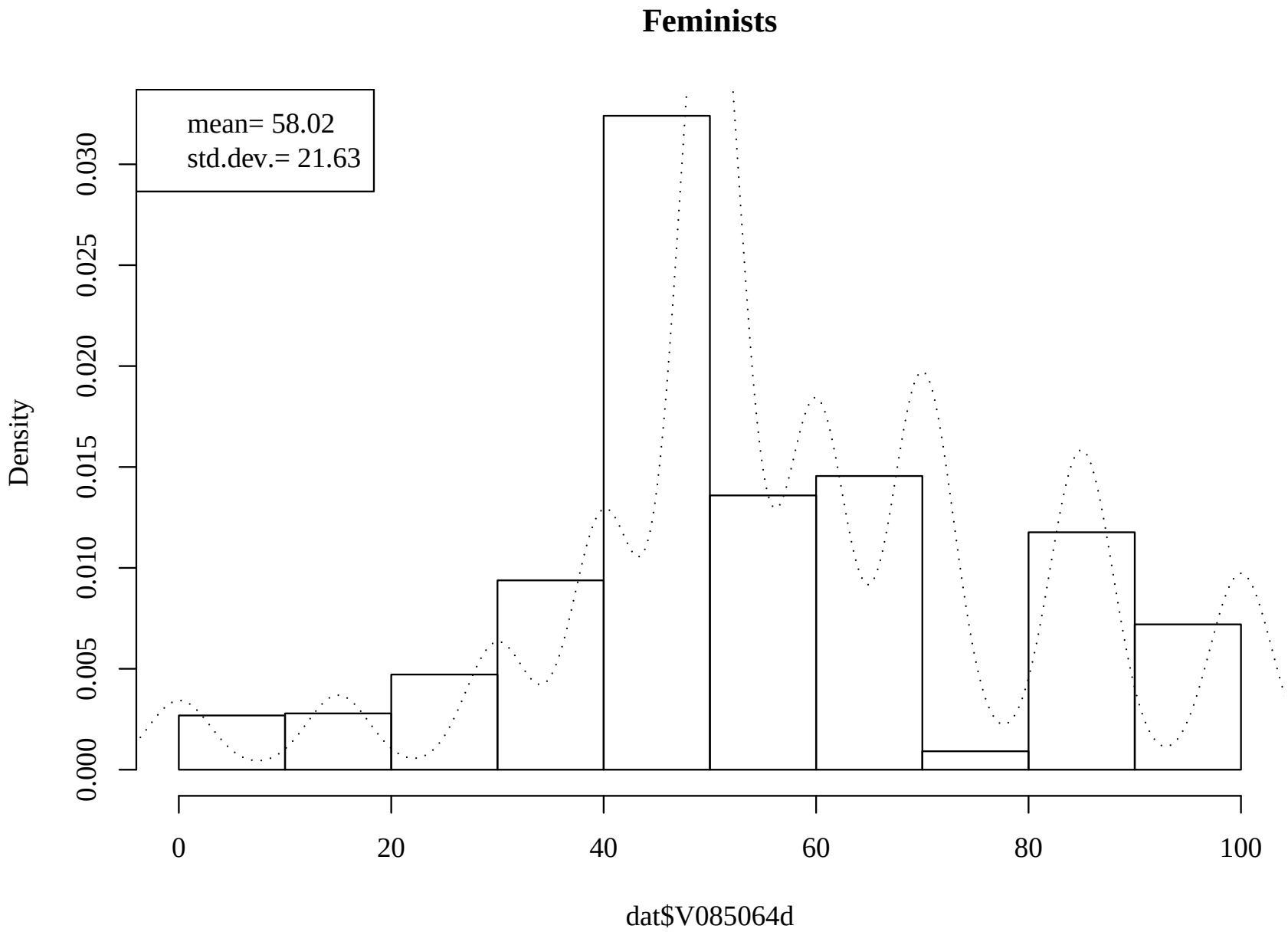
Brief Walkabout in Thermometer Land

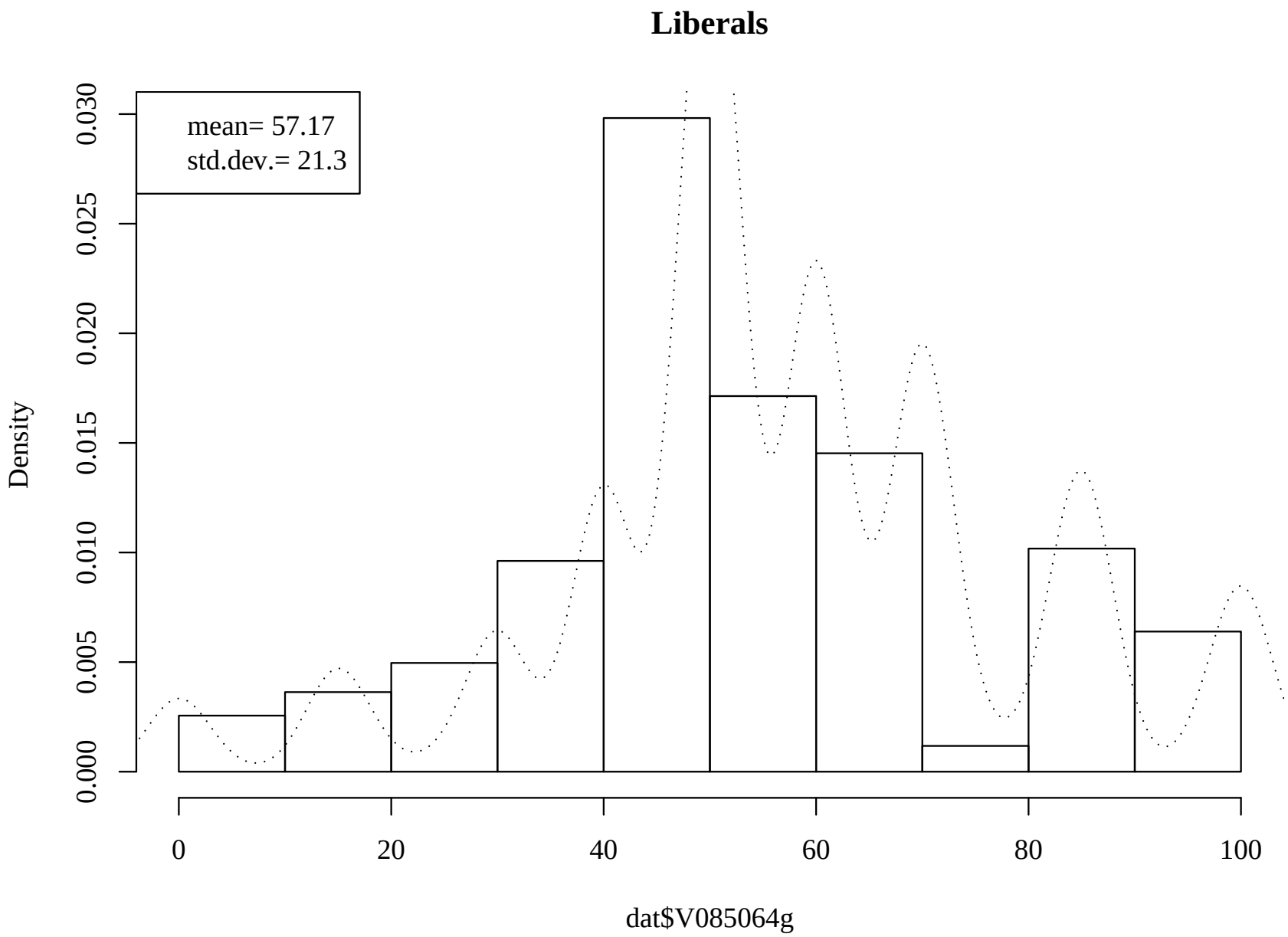



Catholics
mean= 69.41
std.dev.= 20.46
Density
0.000
0.005
0.010
0.015
0.020
0.025
0
20
40
60
80
100
dat$V085064c

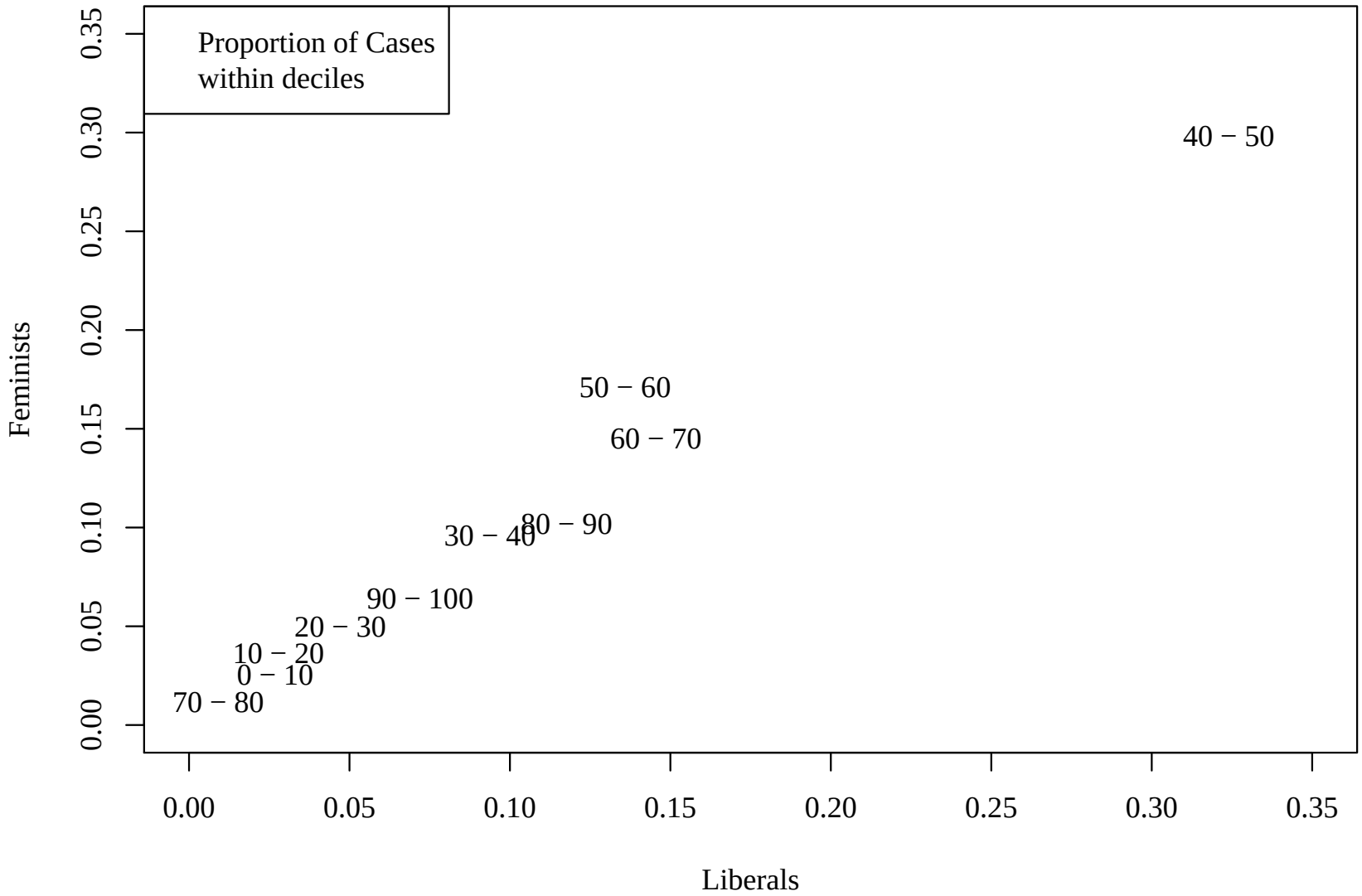

Christian Fundamentalists
mean= 59.18
std.dev.= 23.78
Density
0.000
0.005
0.010
0.015
0.020
0.025
0
20
40
60
80
100
dat$V085064b


Rich People
mean= 58.28
std.dev.= 21.4
Density
0.000
0.005
0.010
0.015
0.020
0.025
0.030
0
20
40
60
80
100
dat$V085065b

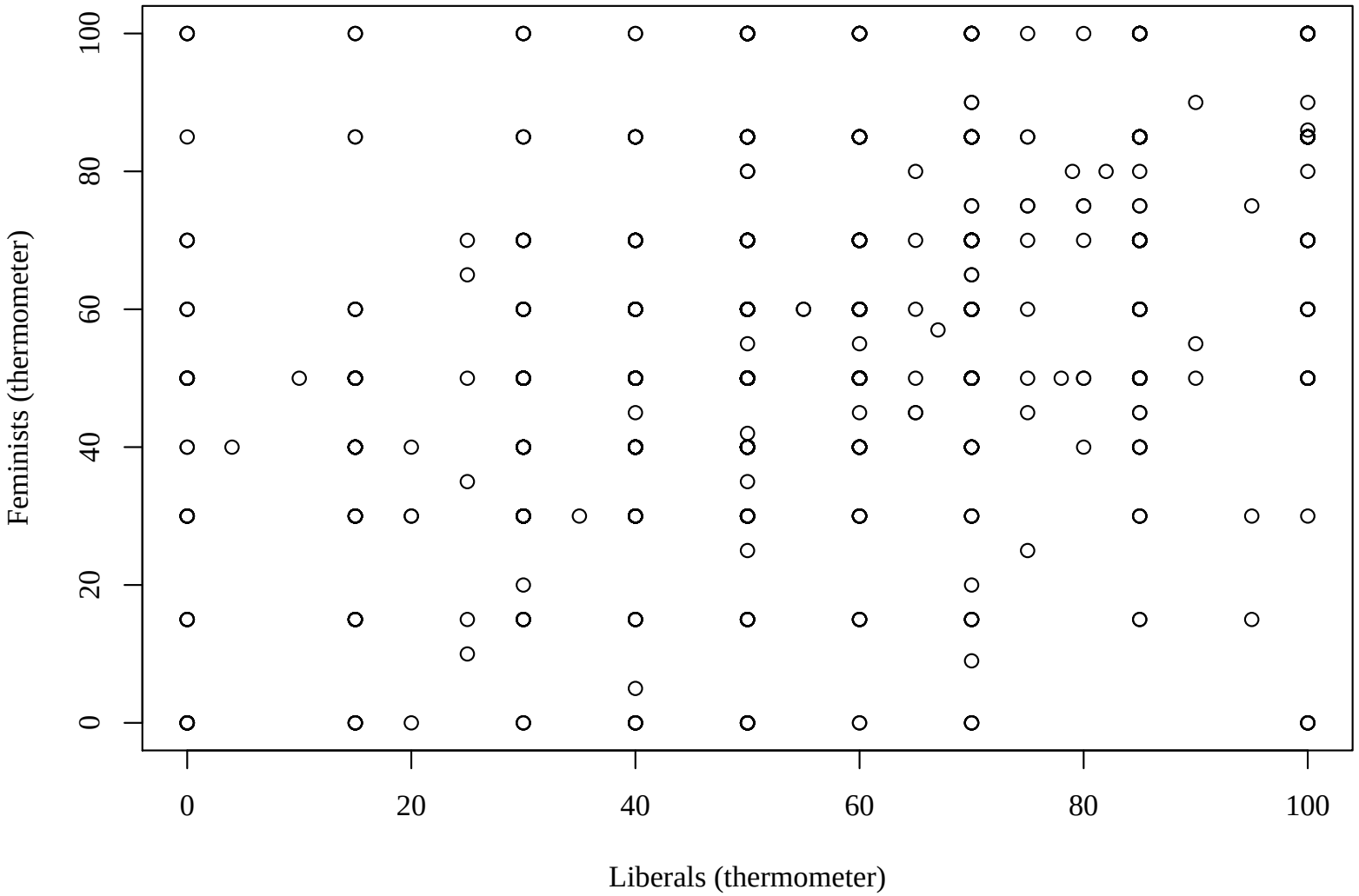

Feminists
mean= 58.02
std.dev.= 21.63
Density
0.000
0.005
0.010
0.015
0.020
0.025
0.030
0
20
40
60
80
100
dat$V085064d


Liberals
mean= 57.17
std.dev.= 21.3
Density
0.000
0.005
0.010
0.015
0.020
0.025
0.030
0
20
40
60
80
100
dat$V085064g

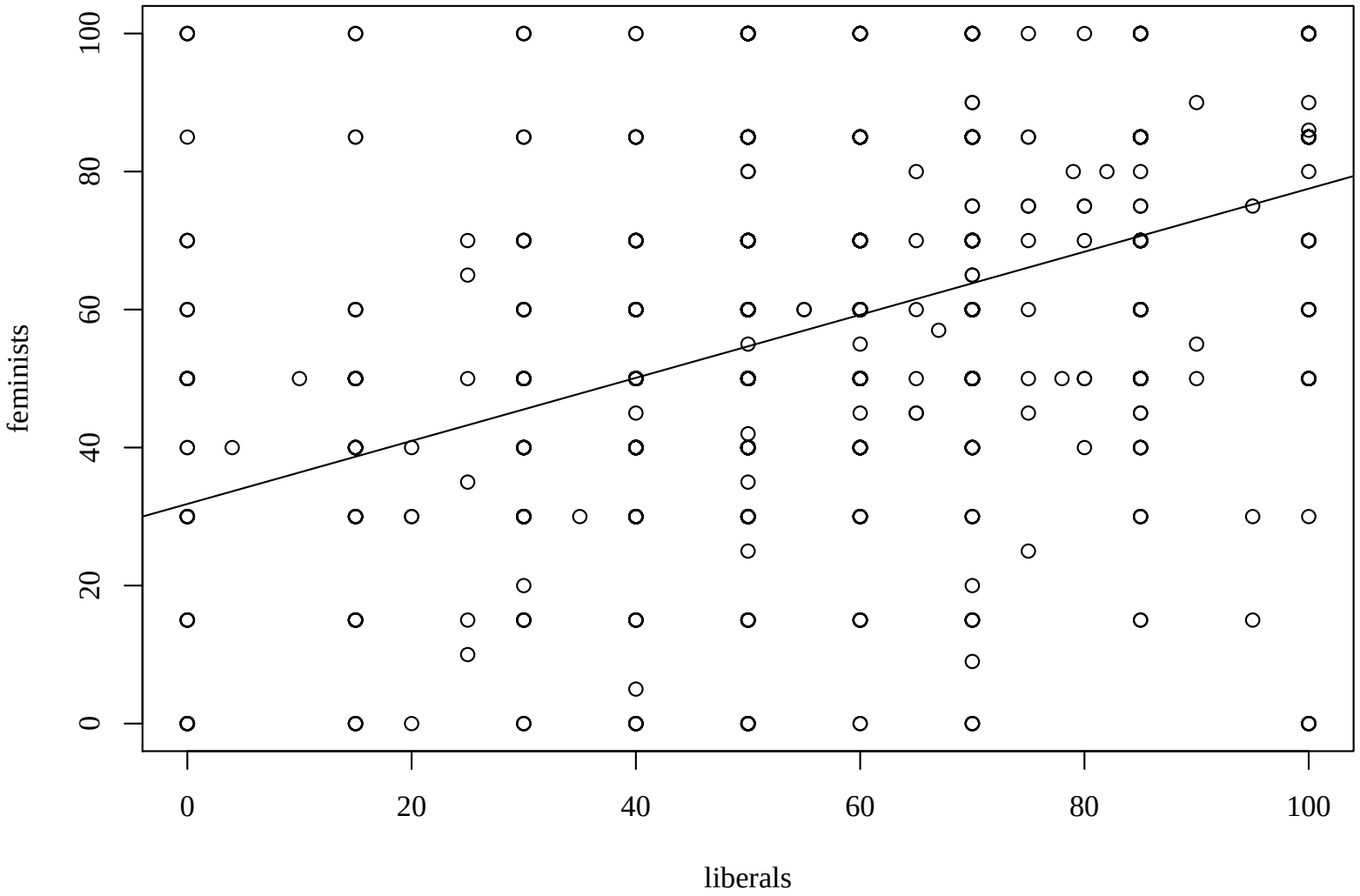
The Feminist and Liberals Plots are Very Similar



But the Scatterplot is a Bit Disappointing



But Statistical Tools Can Discern Pattern



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

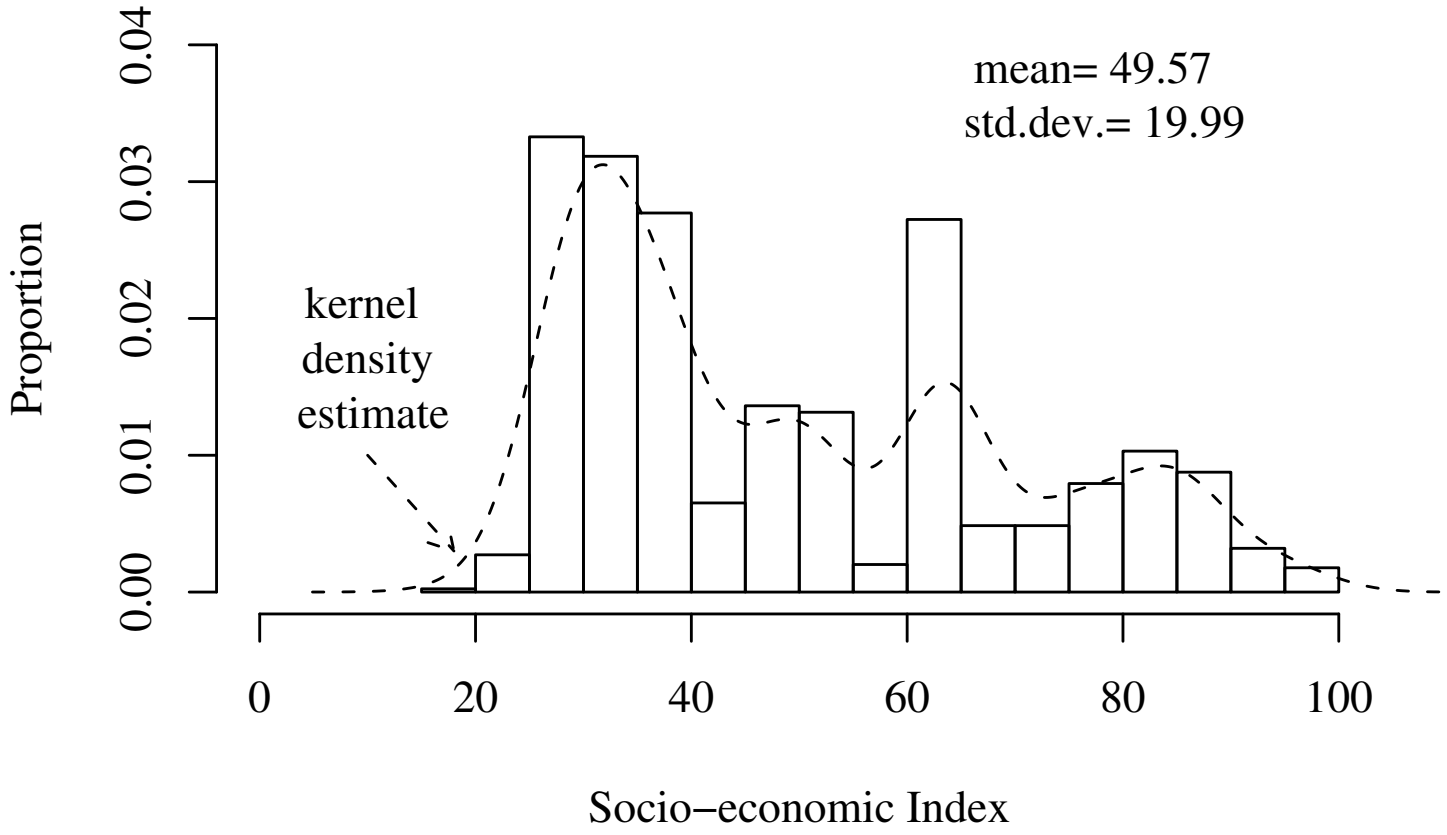
	VOTED	DID NOT VOTE	INELIGIBLE
N	1826	715	389
N / Table Total	0.623	0.244	0.133

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

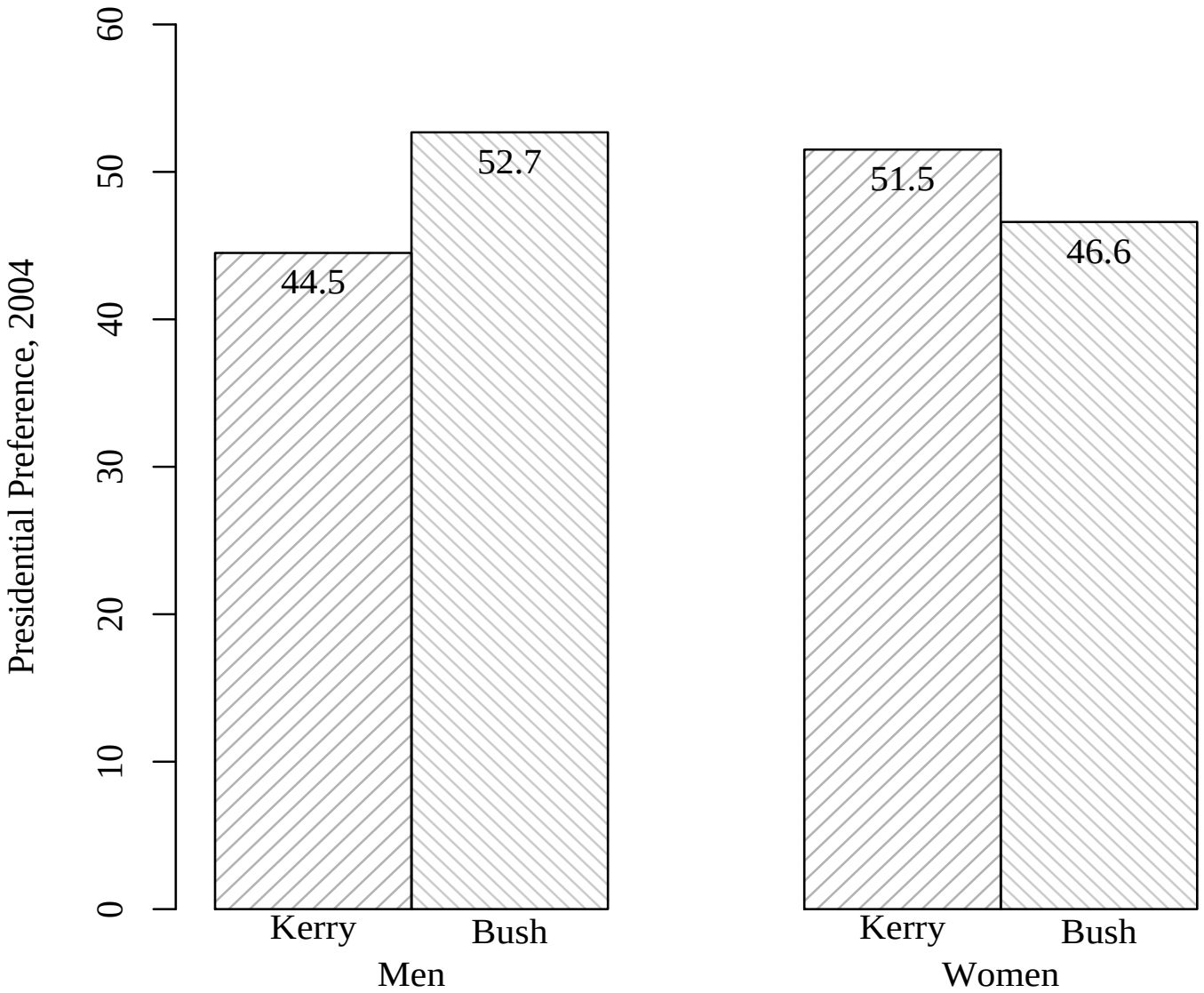
Voter Participation by Sex in 2006

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

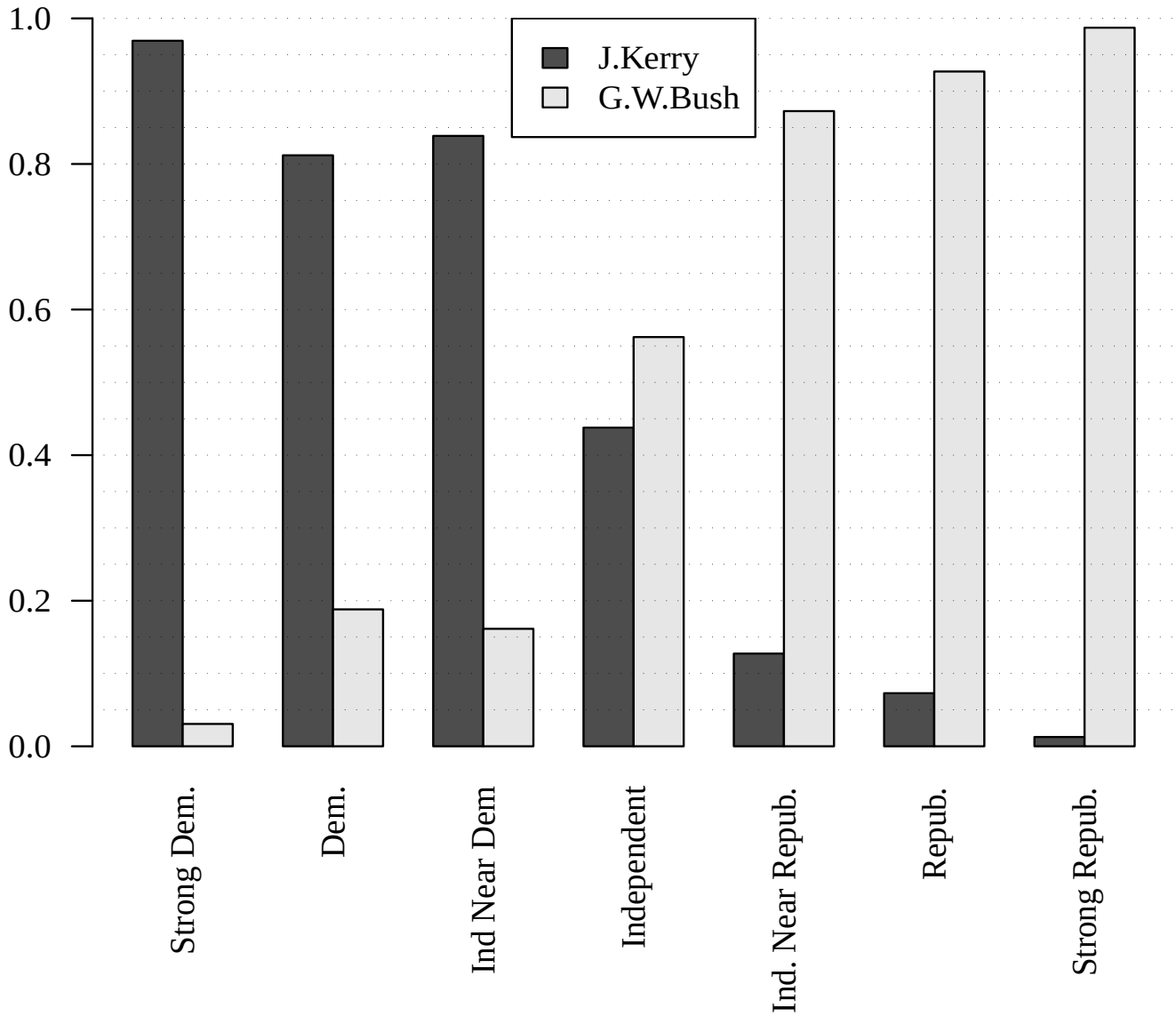
Socio-Economic Status of Men



The Gender Gap in Voting



The BarPlot of the Century (so far)



Probing Promiscuity (Box & Whisker Plot)