Scatter Box Bar Plots

And Cross Tabulations!

Paul Johnson[1]

[1]University of Kansas

2020

Outline

Outline

Levels of Measurement

- Measurement Defined: Observations are associated with "letters" or "numbers" with which we remember them.
- Numeric variables
  - real numbers: continuous, (respondent's weight, in kilograms)
  - discrete valued: 0,1,2,3 (how many arms does respondent have?)
- Hopefully, proportional re-scaling $(k_1 + k_2 \times X_i)$ does not alter the meaning of measurements.
- Natural science abounds with variables that are truly numeric, such as velocity, mass, volume, thickness, etc.
- Social science has fewer "real valued" numeric variables, but we often pretend that discrete variables or ordinal scales are numeric (feeling thermometer variables, left/right ideological placements, etc).

KU

Categorical Measurements

- Categorical variables: Observations are drawn from a discrete list of possible observations
  - Nominal. Ordering of levels is completely unimportant (e.g., candidate names)
  - Ordinal. There is substantive significance to the ordering of the levels (conservative, moderate, liberal)
- The measurement process will often keep a categorical variable in 2 formats

religion	label	numeric score
	Catholic	1
	Protestant	2
	Jewish	3

- In some software (SPSS), the numeric score is the primary referent, while the label is incorporated in presentations.
- In R, the label is the primary referent, the user is discouraged from acting as though the numeric score is meaningful (except for differentiating observations).

Unordered: Category Order does not matter

- Many equivalent numeric scores can be used

religion	label	numeric score	numeric score	numeric score
	Catholic	1	2	-1
	Protestant	2	1	0
	Jewish	3	3	1

- All of the above have the same information
- Any analysis which purports to gain "power" or "statistical significance" by choosing one over the other is wrong.

Ordinal

- A truly ordinal variable has the same meaning as long as the numbers representing it retain the order.

Party Ideology	X	Y	Z
Democrat	1	-10,000	399
Republican	2	0	653
Libertarian	3	1	1,000,000

- X, Y, Z are equivalent numeric scores.
- Reminder: if conclusions differ from analysis of X, Y, or Z, the procedure is probably wrong.

Factor = R Term For Categorical Variable

- In R, categorical variables are called factors (see functions factor(), ordered(), levels())
- Data values as levels, terms like "male", "female"
- Most R statistical procedures try to automatically handle the "behind the scenes" conversion into numeric variables.
- Examples, $sex_i \in \{Male, Female\}$. When R analyzes that variable, many procedures will report a result for a numeric variable that R constructs automatically, $sexFemale$. (see next slide)

KU

How to make a Categorical variable into Numeric Variables

- A "dummy variable" is usually coded 0 or 1, to mean that a quality is present (or not).
- A variable "sex" may be $Male$ or $Female$, but we often focus on a $0, 1$ numeric representation, $sexFemale_i \in \{0, 1\}$(remembering 0 is for $Male$ and 1 is for $Female$).
- $sexFemale_i$ often called an "indicator variable," or "binary variable", or "dichotomous variable", or "dummy variable".
- Note the variable sex_i can beget 2 indicators, $sexMale_i = \{0, 1\}$ or $sexFemale_i \in \{0, 1\}$.

Foreshadow multi-Category treatment

- A multi-category variable like religion may be used to create several separate indicators

religion	numeric score	Cath	Prot	Jewish
Catholic	1	1	0	0
Protestant	2	0	1	0
Jewish	3	0	0	1

- Expect the "too many dummy variables" problem in regression.

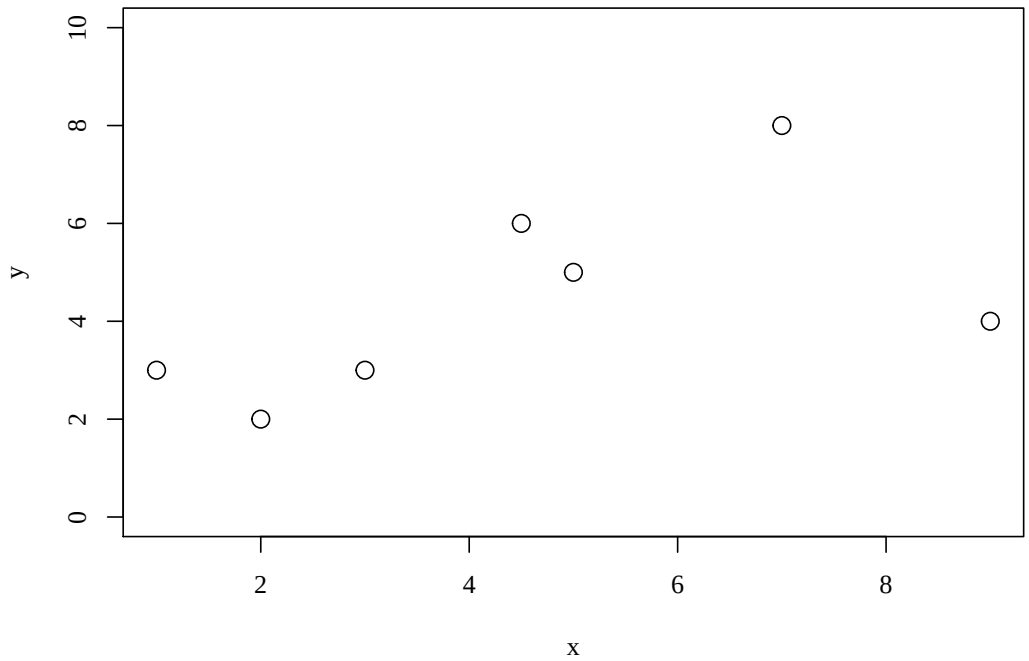
Summarizing "categorical" variables ("factors") is an art form.

- Only the "mode" appears to be universally accepted as one indicator of central tendency.
- No universally accepted index of "diversity"
- A table of frequencies (either raw counts or proportions) may be best.
- rockchalk::summarize() tries to handle that.

Outline

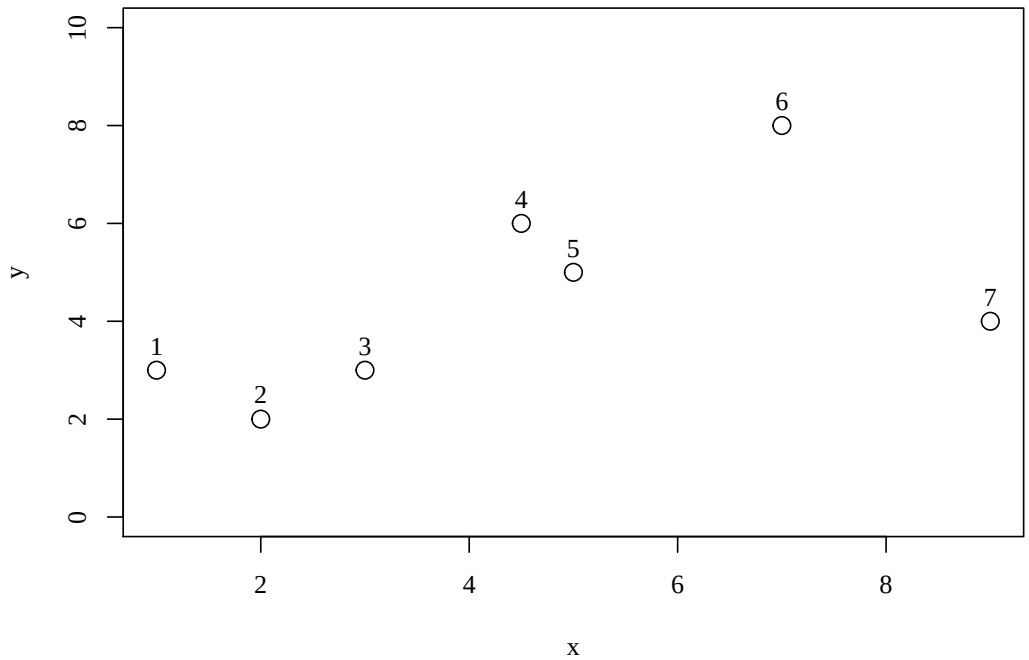
Scatterplot 2 Numeric Variables

- Each observation is one "point"
- x and y appear to positively related
- they "go together", but not perfectly

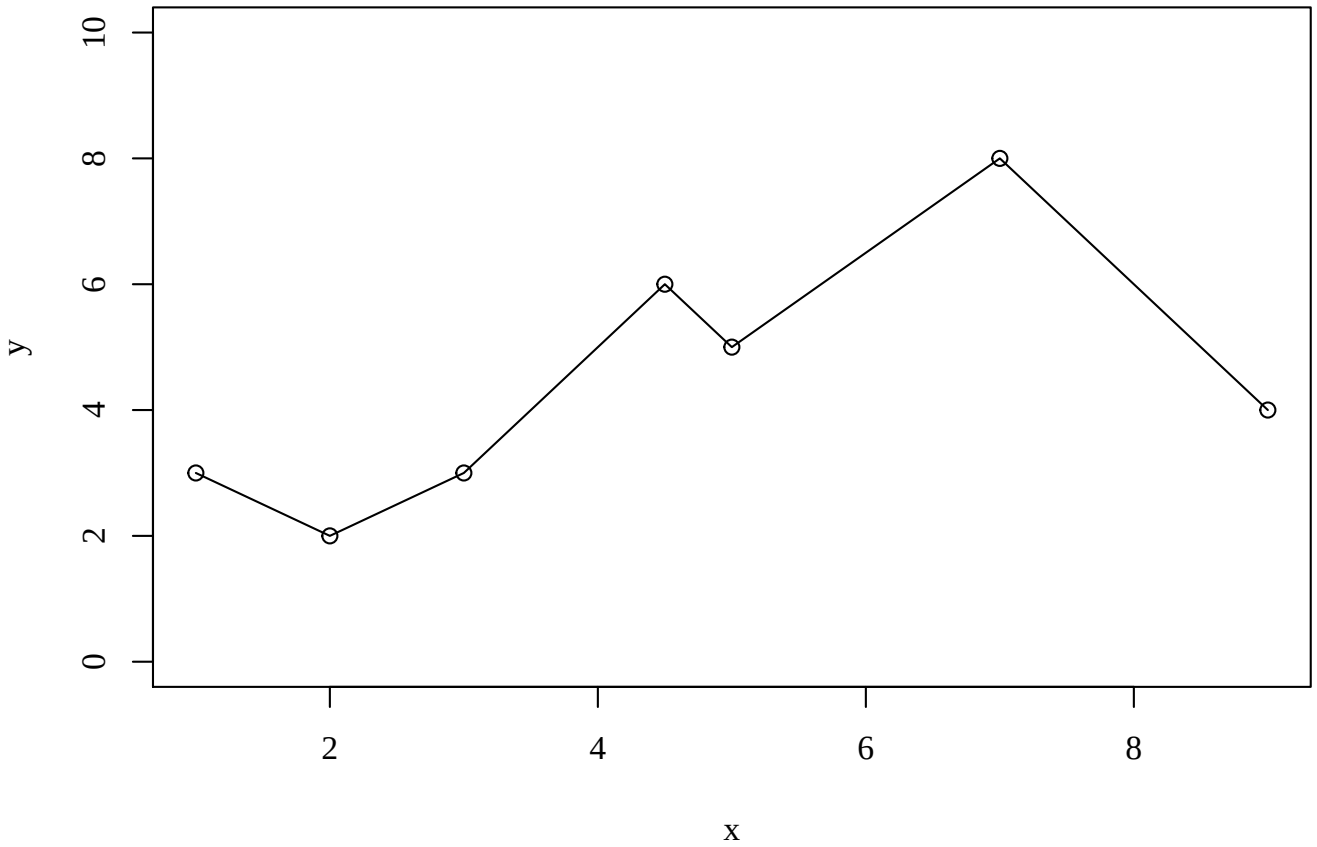


Perhaps you want to number the points

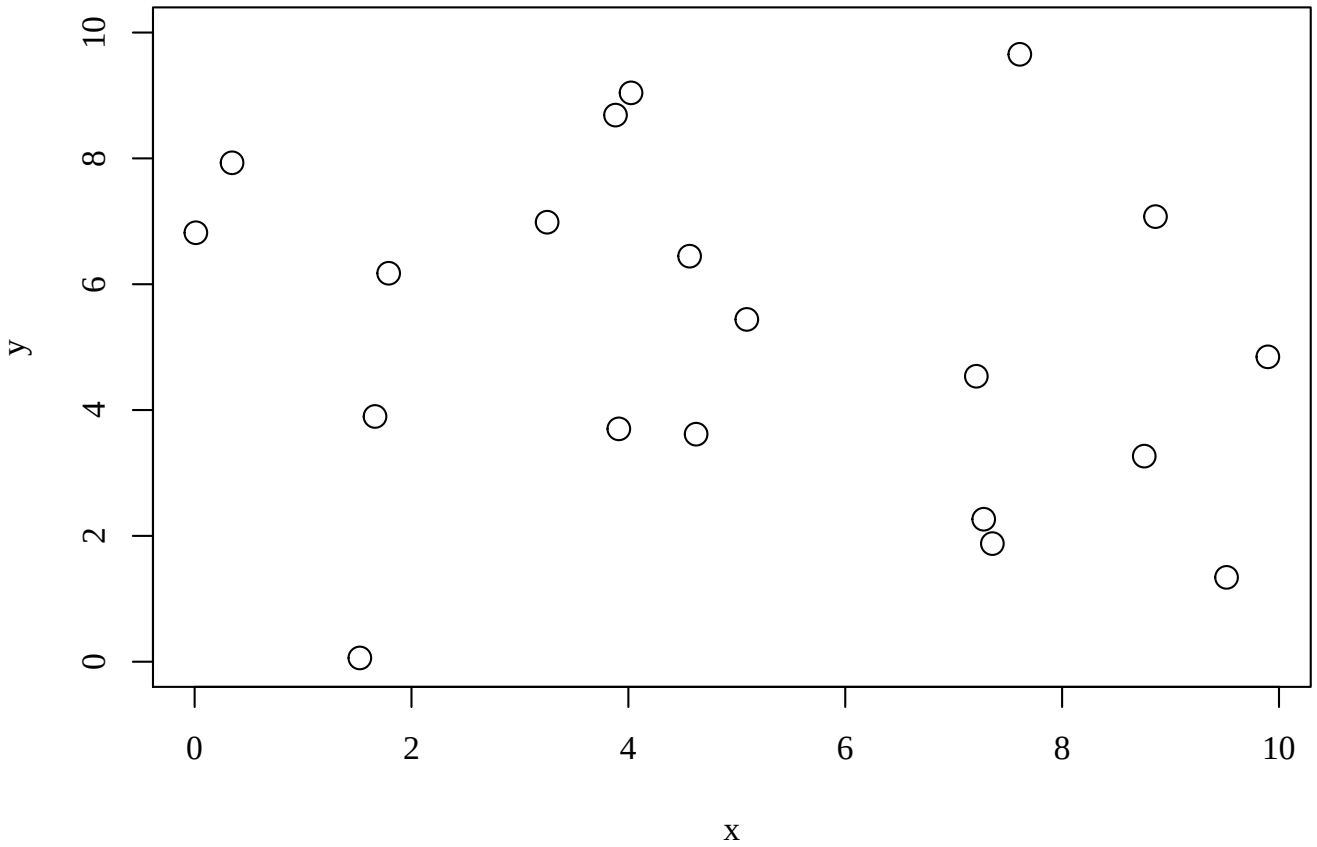
- I used row numbers for the points



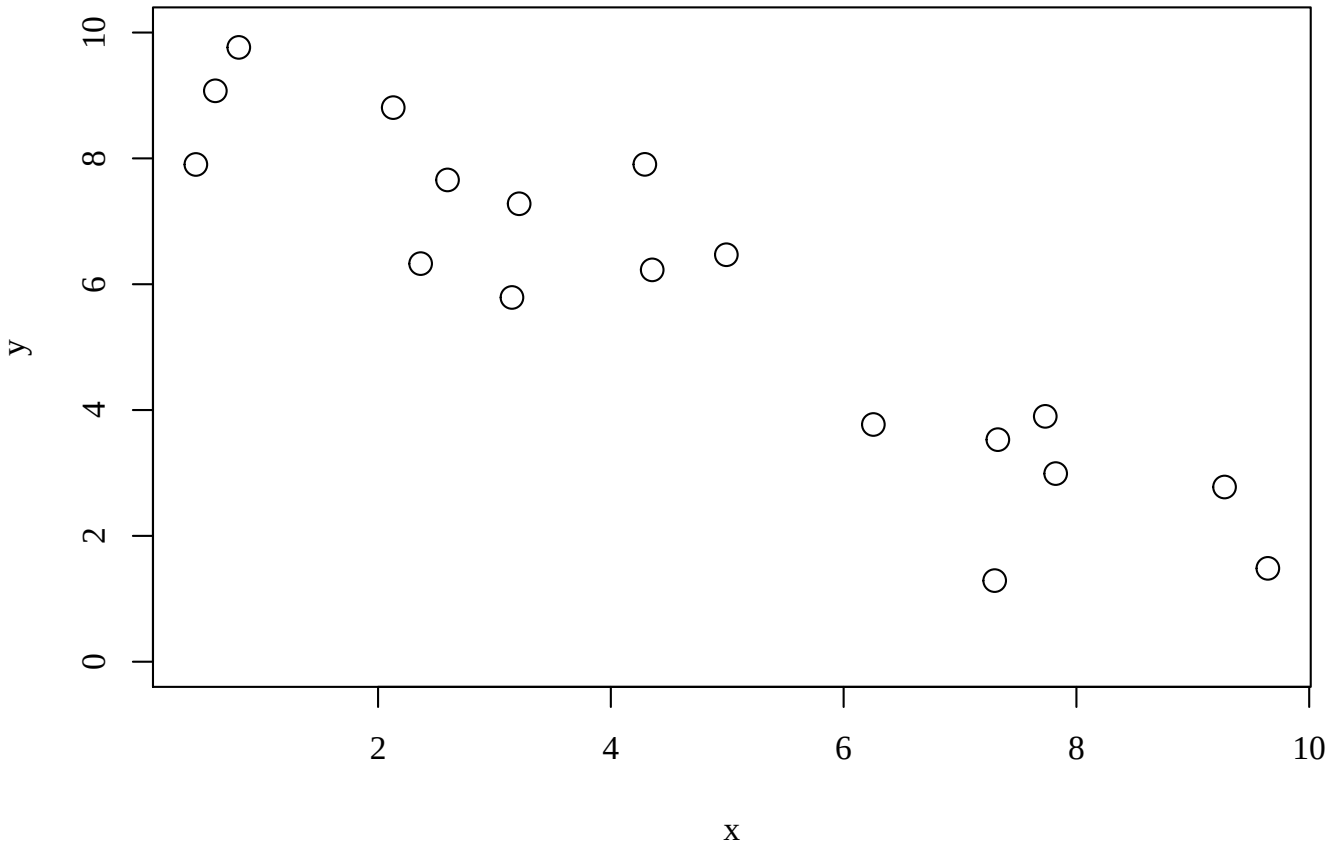
Line graph: Connect the dots



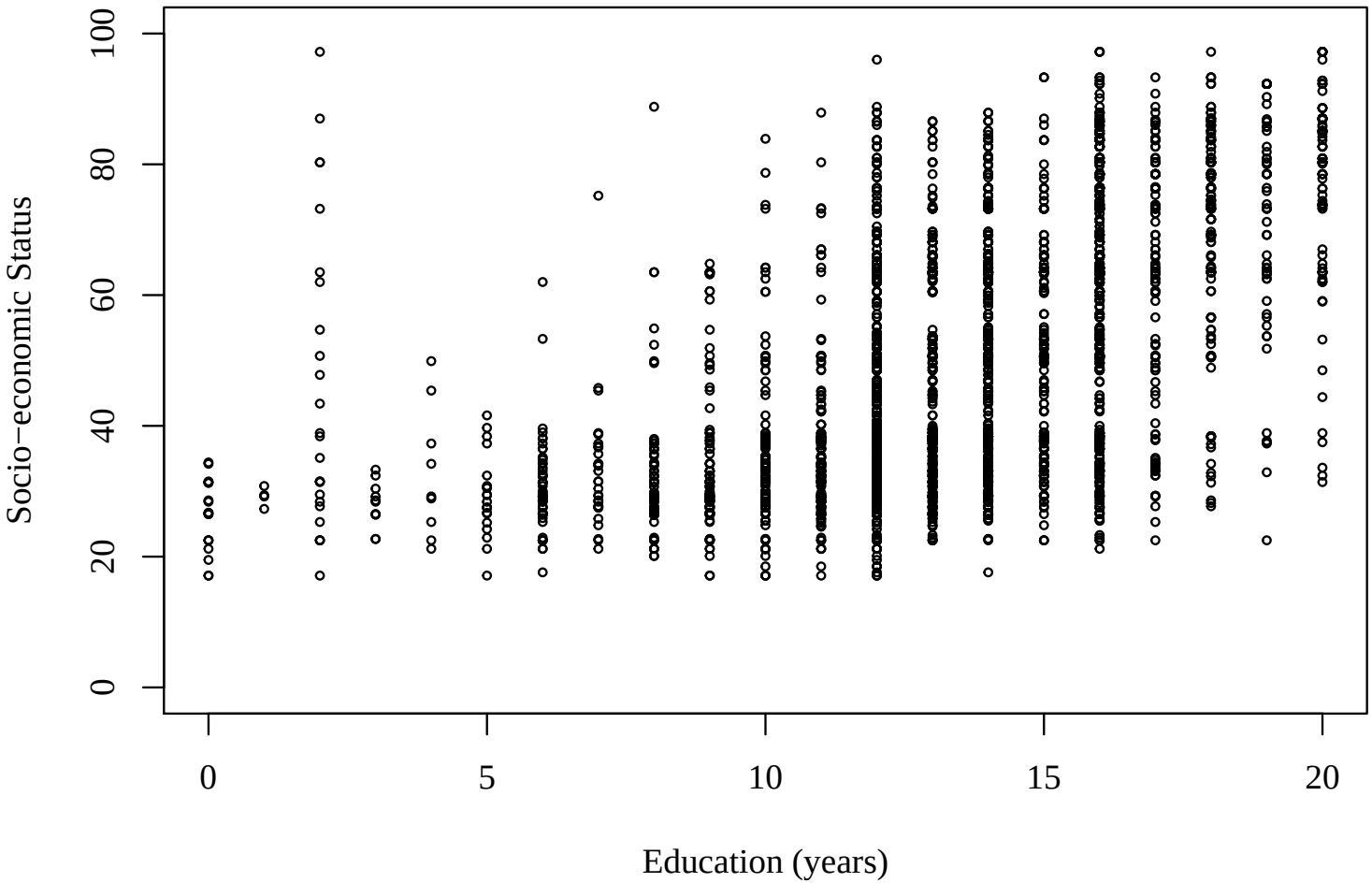
Another Scatterplot: No relationship



Another Scatterplot: Negative relationship



2 Variables from GSS

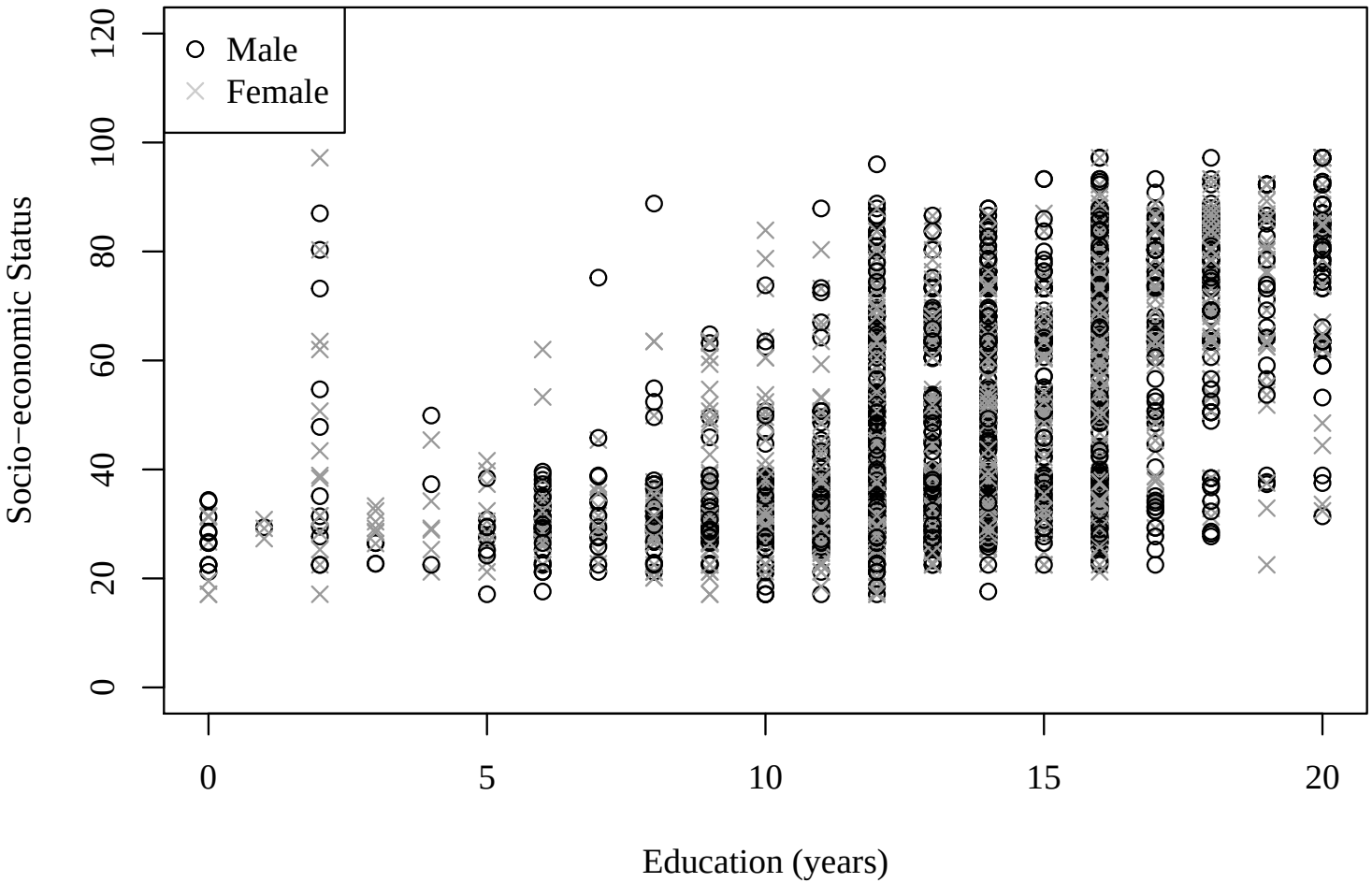


R plot tools

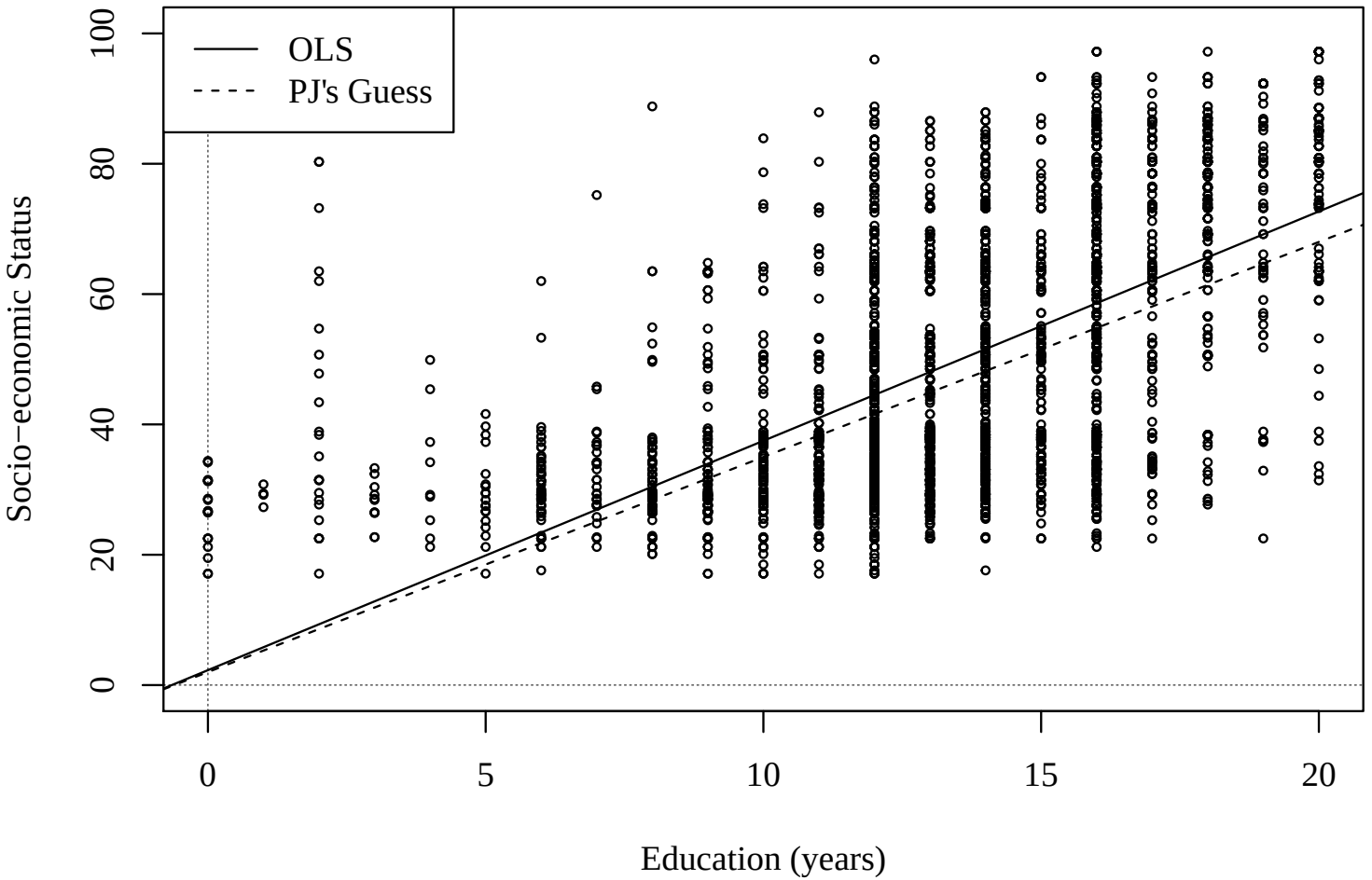
- R's plot() function is a rather interesting case: it doesn't actually draw plots, it re-organizes the input and sends it to other plot functions (for scatterplots, barplots, and so forth)
- Numeric variable plots are created by the function plot.default(), which we could use directly instead if we wanted to.
- plot functions share a common syntax, we specify (optionally) xlim, ylim, col, type, and so forth.
- Once plot is created, add details with "points", "text", "lines", "polygon", "legend" and so forth.

KU

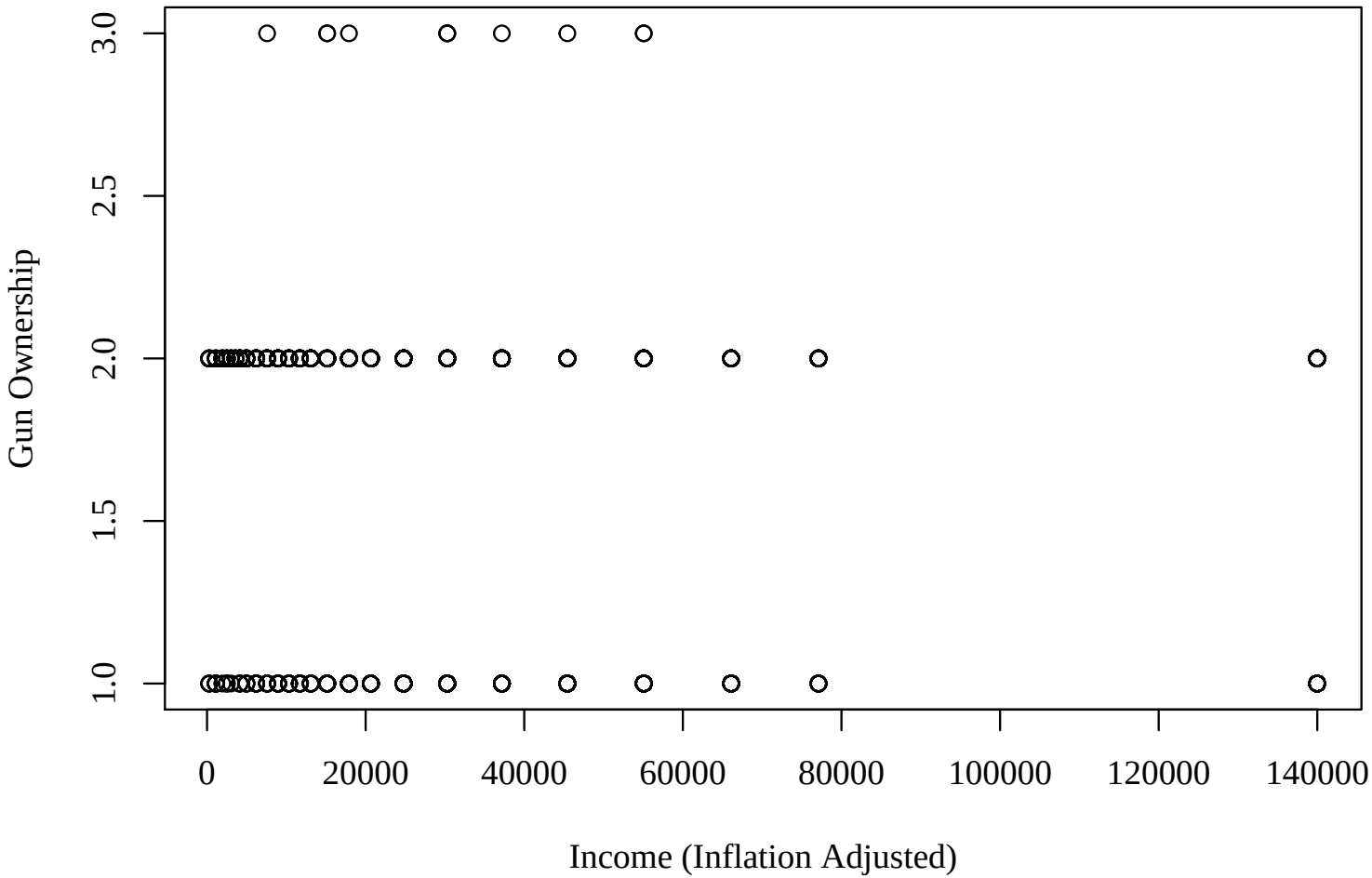
Color Coded Men and Women



Can Add Predictive Lines

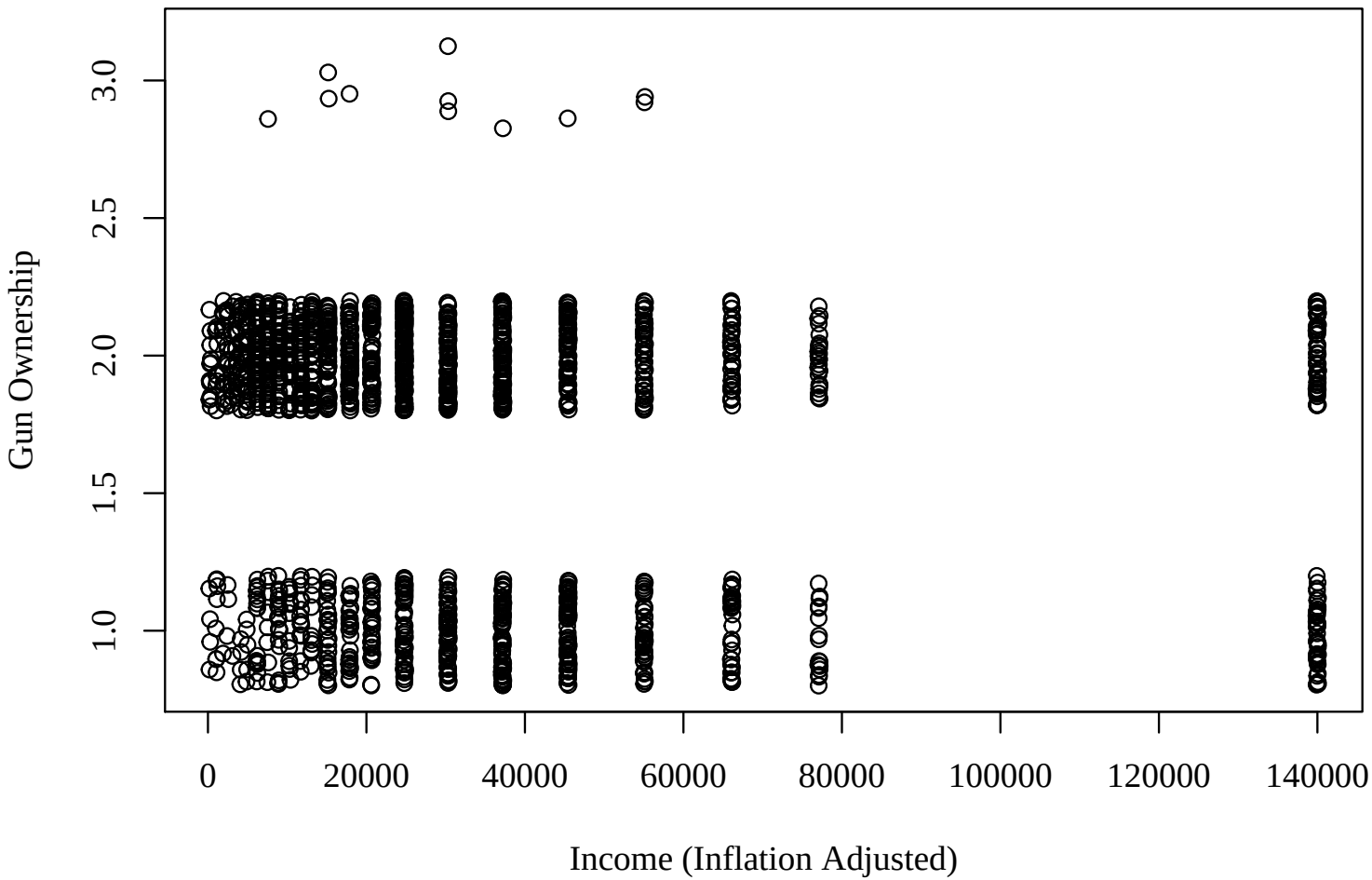


Discrete Variable Problem: Piled Up Observations



Jitter Observations

The jitter function adds random noise to scores, so they don't overlap anymore



Is Jittering Good? Or Dubious!

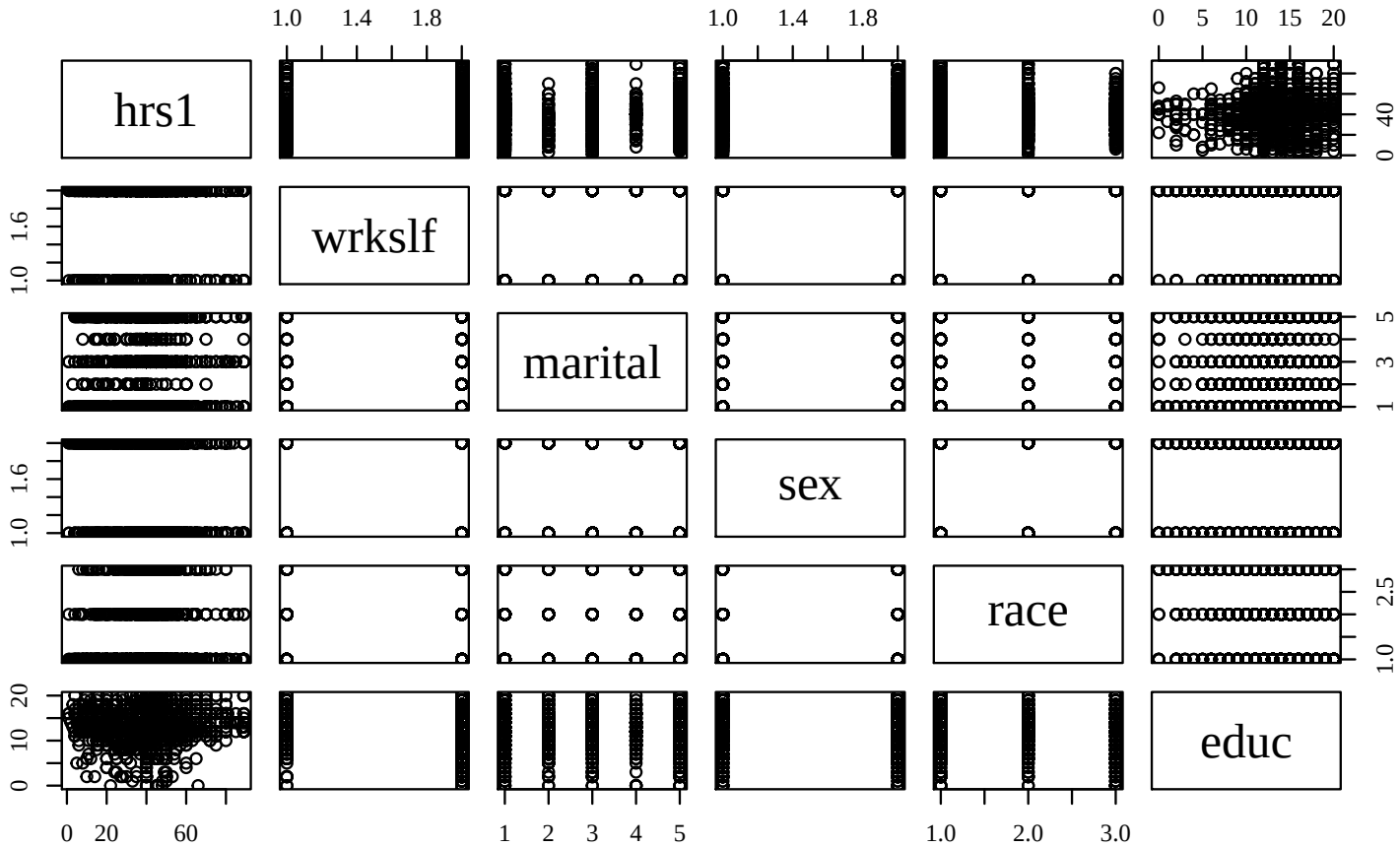
- If variables are truly numeric (discrete), jittering may help.
- But if variables are categorical, it may be deceptive.
- Giving numbers 1, 2, 3 to gun ownership does not mean the numbers are meaningful!
  - 1 : Yes
  - 2 : No
  - 3 : Refused to Answer
- Better to use methods intended for truly categorical outcomes
- Nevertheless, common "solution" is to add numerical random noise to 0, 1 in order to make a better looking scatterplot

KU

Scatterplot Matrix

- Some programs offer a quick way to see a lot of scatterplots in a single picture.
- Usually doesn't help me too much.

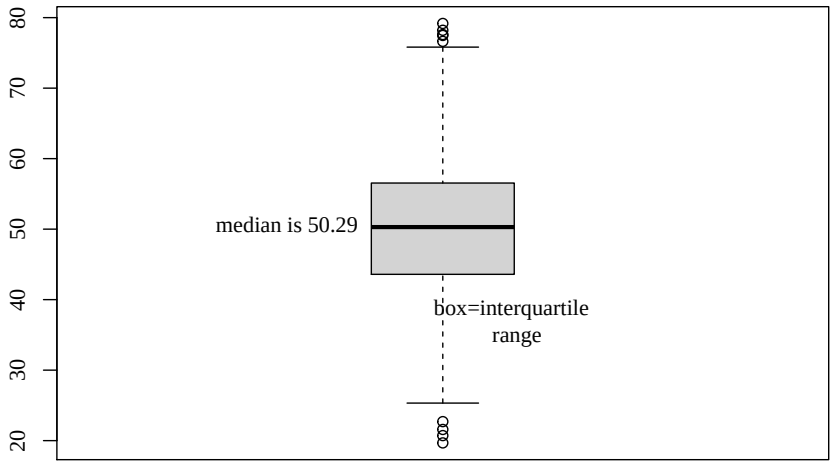
S.M. for 6 variables



Outline

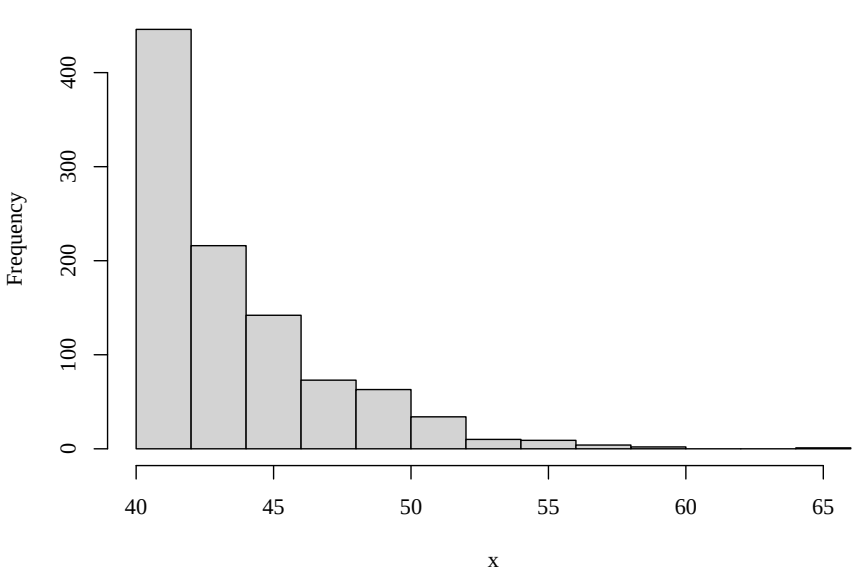
Boxplot: Like a Histogram Turned on its Side

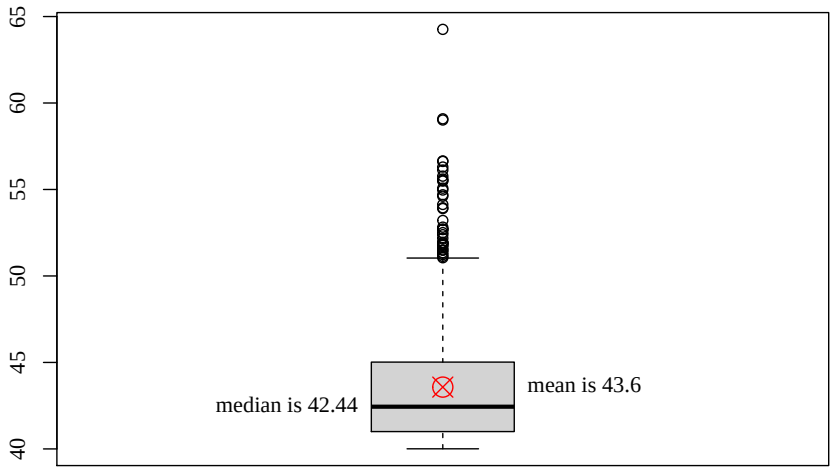
- A boxplot is
- with weird shaped bars and funny markings.
- Dark line at Median
- Box has 25% of cases above and below
- "Whiskers" default to reach out 1.5*interquartile range
- Dots represent extreme cases.



This variable is symmetric, with mean near median of 50.

Boxplot: For a Nonsymmetric Variable

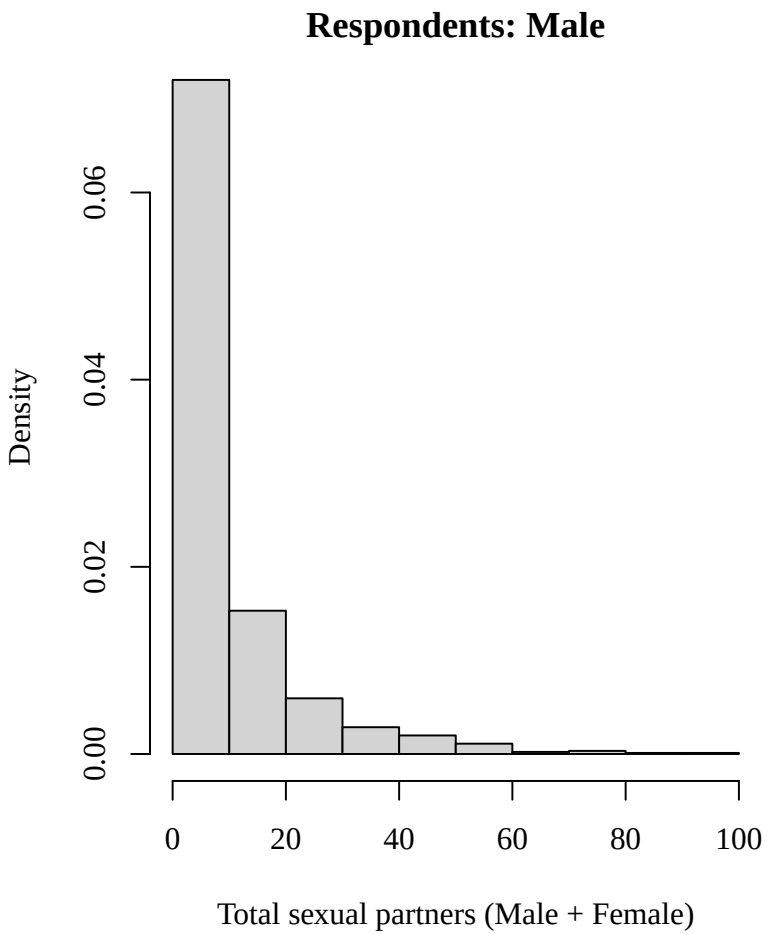


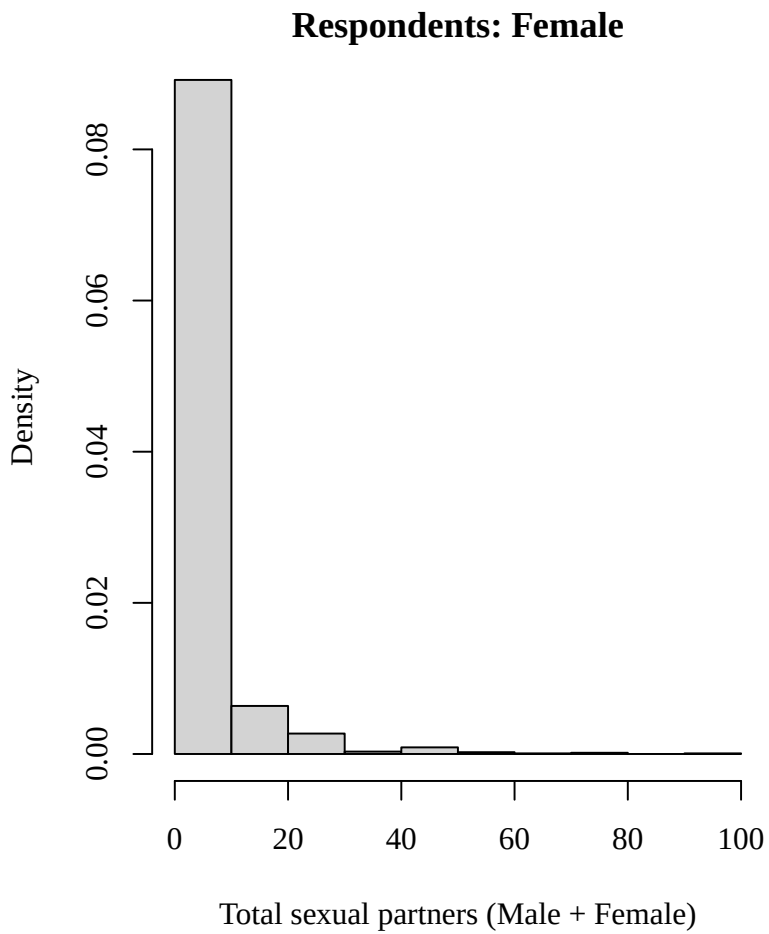


Boxplot: View Several Subsets

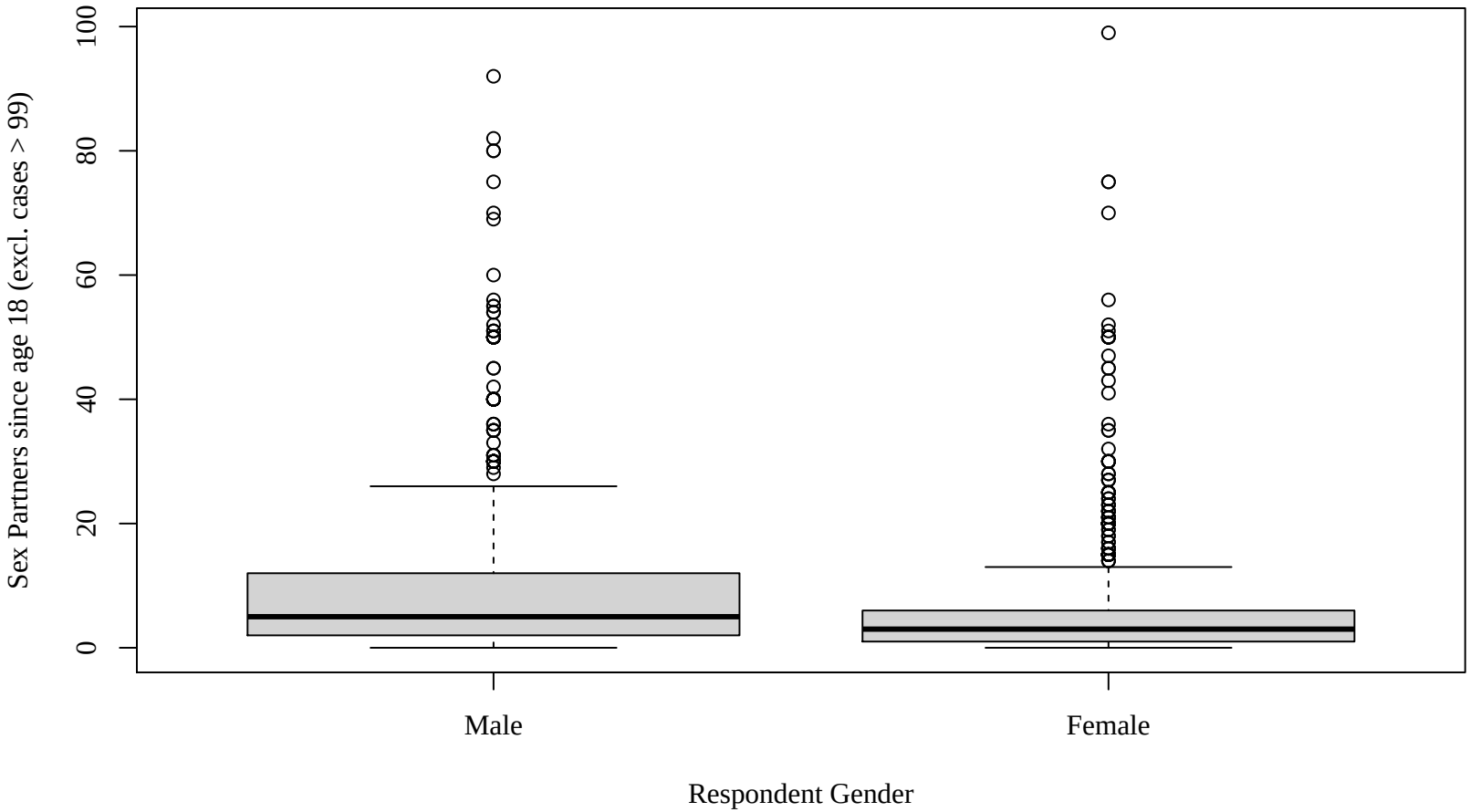
- A histogram can display only one group of respondents
- If you get used to boxplots, you gain the benefit that more groups can be fit into a single display.

Histograms for Number of Sexual Partners(GSS 2006)

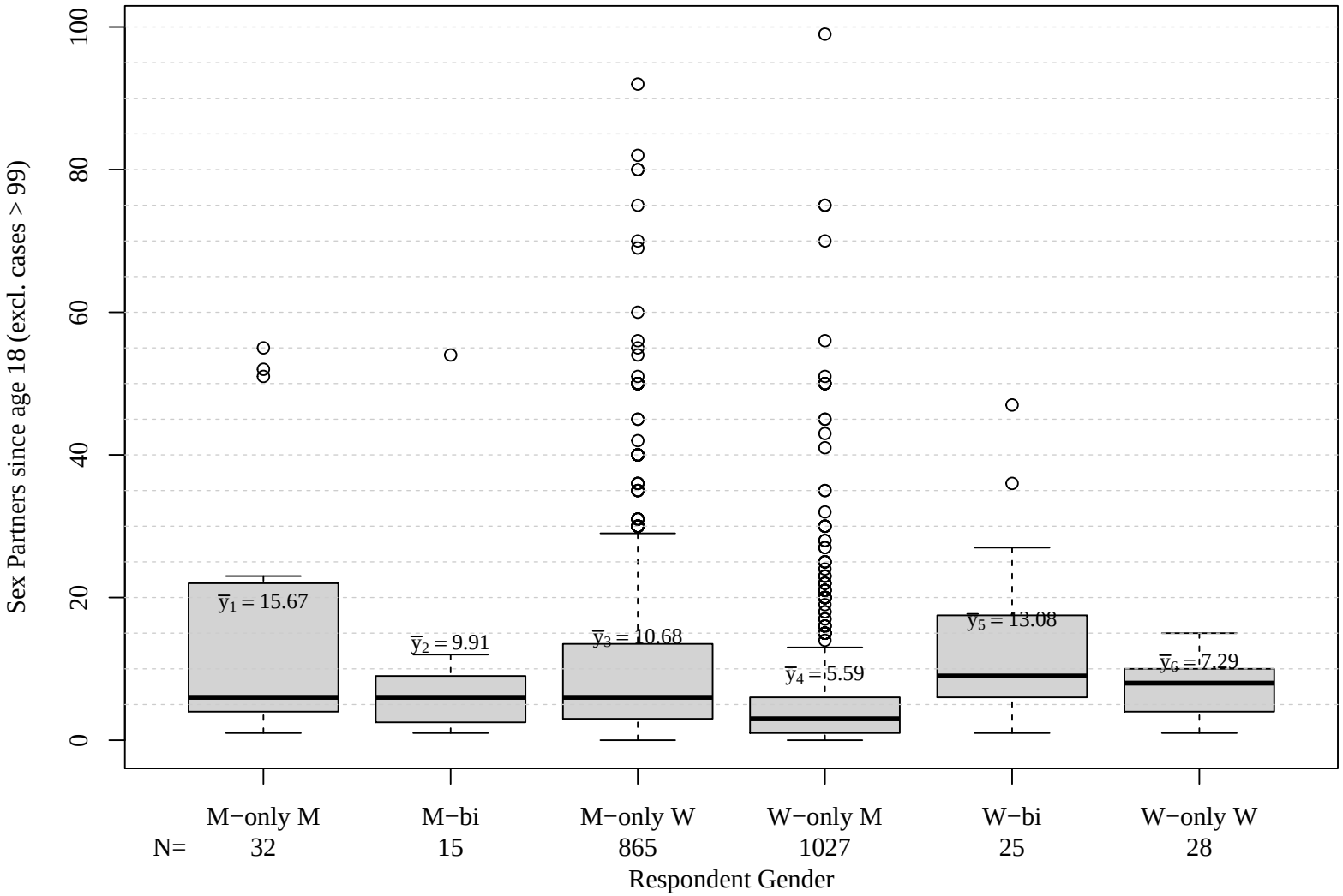




Use a Box Plot Instead



My Most Informative Box Plot



Frequencies (Counts)

- A Frequency Count is a simple table that lists the number of observations within each category

Hair Color Among POLS110 Students				
Brown	Black	Blond	Red	Unknown
155	67	48	10	63

- The MODE is the most frequently occurring value. (Here, Brown)

Cross-Tabulation Table

- A Cross-Tabulation shows cross-classified counts.

Hair Color Among POLS110 Students					
	Brown	Black	Blond	Red	Unknown
From Kansas	98	40	18	5	10
Not	57	27	30	5	53
	155	67	48	10	63

- This simple descriptive table does not necessarily have a dependent or independent variable.
- There have been many efforts to summarize the observed relationship numerically (Google for correlation coefficients like phi, Gamma or Kendall's tau)

The Iron Laws of Crosstabs

For analytical purposes, Gerry Lowenberg taught me to always follow 3 rules. You should too.

1. IV on top, DV on left
2. Convert to percentages (or proportions) on the columns
3. Compare the across, find if columns are distributed differently

The FX Network is	Column Percentages	
	Respondent	Sex
	male	female
really horrible	25%	80%
OK	50%	18%
really great	25%	2%
N	343	288

The 1980 American presidential election (ANES)

		Party Identification	
	Republican	Independent	Democrat
Percentages voting for			
Reagan	86%	55	26
Carter	9	30	67
Anderson	4	12	6
N	544	355	721

Example: Patriotism and Voting

American National Election Study of 1988.

How does seeing the American Flag make you feel? *

Percentages voting for	Extremely good	Very good	Somewhat good	Not very good
Bush	60.7%	50%	30%	8%
Dukakis	39.3%	50%	70%	92%
N	403	666	299	28

*Excludes voters who did not select either Bush or Dukakis

There Are Standards for Tables

- There is a literature on the correct format for tables in publications
- See Simon Fear, "Publication quality tables in LaTeX" (in the LaTeX booktabs package)
- Advice: Avoid vertical lines

Example: GSS 2006 Gun Data

Table: Opinions on Gun Registration Laws

Stance on Gun Registration	Does Respondent Own a Gun? Yes	No	Refused To Say
Favor	71%	85	57
Oppose	29	15	43
Number of Cases	600	1128	23

- Reminder how I did that. Using gss data "dat" (from above)

```
library(memisc)
guntab <- with(dat, genTable(percent(gunlaw) ~
    owngun))
toLatex(guntab)
```

Making tables in R. PITA

- The rockchalk function "pctable" was introduced to help with this problem!

```
library(rockchalk)
p1 <- pctable(gunlaw ~ owngun, data=dat)
```

```
Count (column %)
         owngun
gunlaw    YES          NO           REFUSED  Sum
  FAVOR   464(70.7%)   1085(84.9%)  17(63%)  1566
  OPPOSE  192(29.3%)   193(15.1%)   10(37%)  395
  Sum     656          1278         27       1961
```

- Convert that into something that gracefully goes into a table.

```
p1sum <- summary(p1)
```

- xtable works! example output:

Making tables in R. PITA ...

```
library(xtable)
p1xt <- xtable(p1sum)
outfn <- file.path(tdir, "pctable1.tex")
print(p1xt, type="latex", file = outfn)
```

	YES	NO	REFUSED	Sum
FAVOR	464(70.7%)	1085(84.9%)	17(63%)	1566
OPPOSE	192(29.3%)	193(15.1%)	10(37%)	395
Sum	656	1278	27	1961

- Can also write as CSV

```
outfn2 <- file.path(tdir, "pctable1.csv")
write.csv(p1xt, file=outfn2, row.names=FALSE)
```

If you want to manufacture your own percentage tables

- table(rowvar, colvar) gives the raw counts, completely unbeautified

```
load("../../DataSets/GSS/gss-subset2.Rda")
with(dat, table(gunlaw, owngun) )
```

```
        owngun
gunlaw     YES    NO REFUSED
  FAVOR    464  1085       17
  OPPOSE   192   193       10
```

Get column percentages

- prop.table(table(rowvar, colvar), margin=2) converts the table into column percentages. Proportions are reported in numbers with 6 decimal digits, so I convert them to percentages and round to 1 digit (my taste, not a hard rule)

```
t1 <- with(dat, table(gunlaw, owngun,
    exclude=NULL) )
t1.prop <- 100 * prop.table(t1, margin=2)
t1.prop <- round(t1.prop, 1)
t1.prop
```

```
        owngun
gunlaw     YES    NO REFUSED <NA>
  FAVOR   69.9 83.0    56.7  0.1
  OPPOSE  28.9 14.8    33.3  0.0
  <NA>     1.2  2.2    10.0 99.9
```

Oops. we need column totals, so go get them

Add Column totals

```
t1.marg <- margin.table(t1, margin=2)
t1.result <- rbind(t1.prop, t1.marg)
t1.result
```

```
          YES      NO REFUSED    <NA>
FAVOR    69.9    83.0    56.7     0.1
OPPOSE   28.9    14.8    33.3     0.0
<NA>      1.2     2.2    10.0    99.9
t1.marg 664.0 1307.0    30.0 2509.0
```

After all of this, the table still is not great. Hence, I use `rockchalk::pctable`

Other packages

- gmodels: the original CrossTable function
- memisc: My favorite for making tables (can make LaTeX)
- "vcd" & "vcdExtra" (VCD="visualize categorical data"
- "descr"

example with gmodels::CrossTable

```
library(gmodels)
with(dat, CrossTable(gunlaw, owngun))
```

```
   Cell Contents
|-------------------------|
|                       N |
| Chi-square contribution |
|           N / Row Total |
|           N / Col Total |
|         N / Table Total |
|-------------------------|


Total Observations in Table:  1961


             | owngun
      gunlaw |       YES |        NO |   REFUSED | Row Total |
-------------|-----------|-----------|-----------|-----------|
       FAVOR |       464 |      1085 |        17 |      1566 |
             |     6.841 |     4.067 |     0.965 |           |
             |     0.296 |     0.693 |     0.011 |     0.799 |
             |     0.707 |     0.849 |     0.630 |           |
             |     0.237 |     0.553 |     0.009 |           |
```

example with gmodels::CrossTable ...

```
-------------|-----------|-----------|-----------|-----------|
      OPPOSE |       192 |       193 |        10 |       395 |
             |    27.121 |    16.123 |     3.826 |           |
             |     0.486 |     0.489 |     0.025 |     0.201 |
             |     0.293 |     0.151 |     0.370 |           |
             |     0.098 |     0.098 |     0.005 |           |
-------------|-----------|-----------|-----------|-----------|
Column Total |       656 |      1278 |        27 |      1961 |
             |     0.335 |     0.652 |     0.014 |           |
-------------|-----------|-----------|-----------|-----------|
```

descr CrossTable similar

```
library(descr);
descrCT <- with(dat , descr::CrossTable(gunlaw ,
    owngun))
descrCT
```

```
   Cell Contents
|-------------------------|
|                       N |
| Chi-square contribution |
|           N / Row Total |
|           N / Col Total |
|         N / Table Total |
|-------------------------|

=========================================
          owngun
gunlaw       YES        NO   REFUSED    Total
-----------------------------------------
FAVOR        464      1085        17     1566
            6.841    4.067     0.965
            0.296    0.693     0.011    0.799
            0.707    0.849     0.630
            0.237    0.553     0.009
```

descr CrossTable similar ...

```
-----------------------------------------
OPPOSE        192       193        10      395
           27.121    16.123     3.826
            0.486     0.489     0.025    0.201
            0.293     0.151     0.370
            0.098     0.098     0.005
-----------------------------------------
Total         656      1278        27     1961
            0.335     0.652     0.014
=========================================
```

Outline

Barplot

- Barplot: Make a graphic presentation of a cross tabulation table
- Use
  - Any discrete variable that can classify respondents
  - Any summary score (mean, proportion, count) calculated from the subgroups of respondents
- The width of the bar has no "substantive" meaning
- Unlike a histogram, where the width $\times$ height represents the area

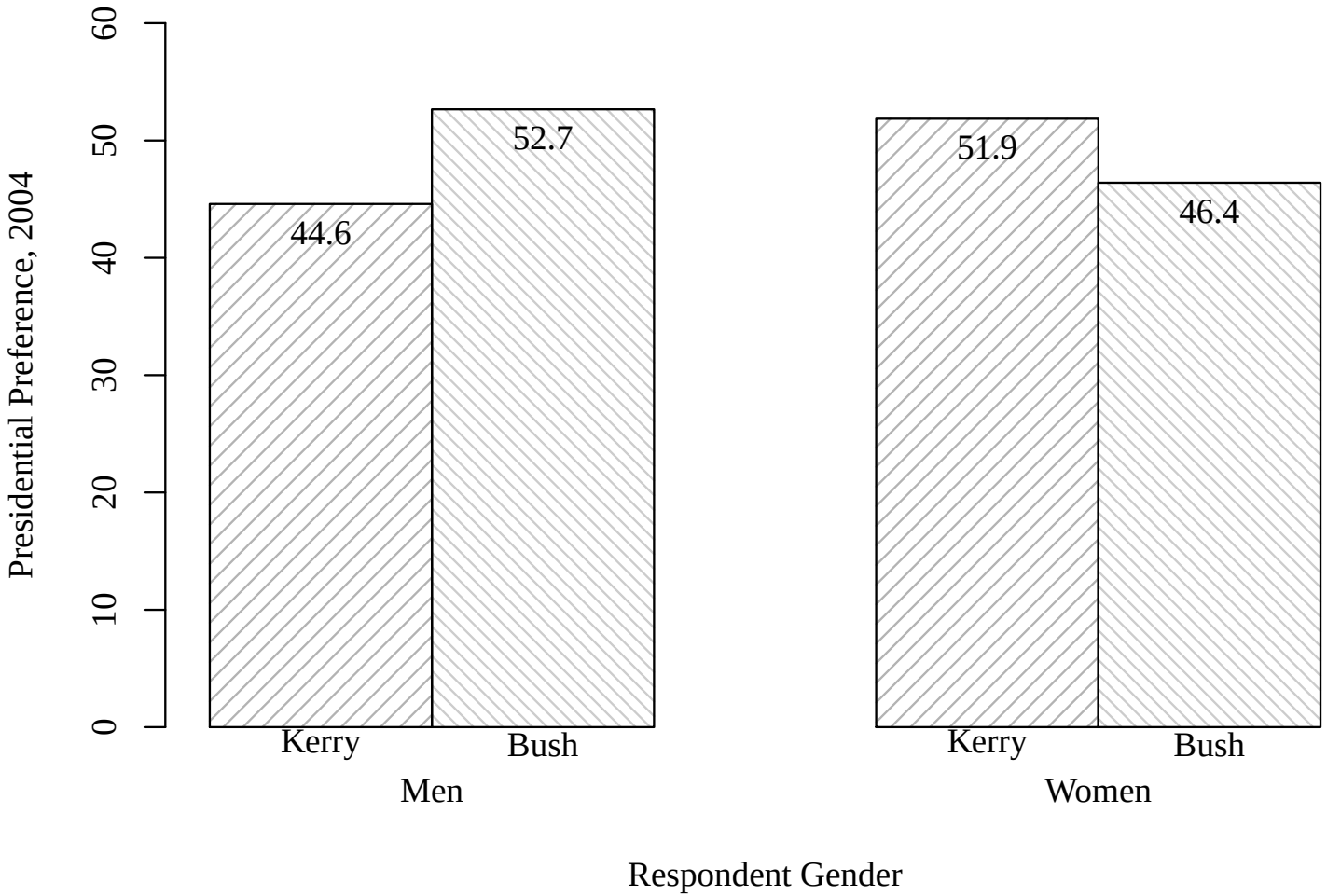
KU

Table Demonstrating the Gender Gap in 2004

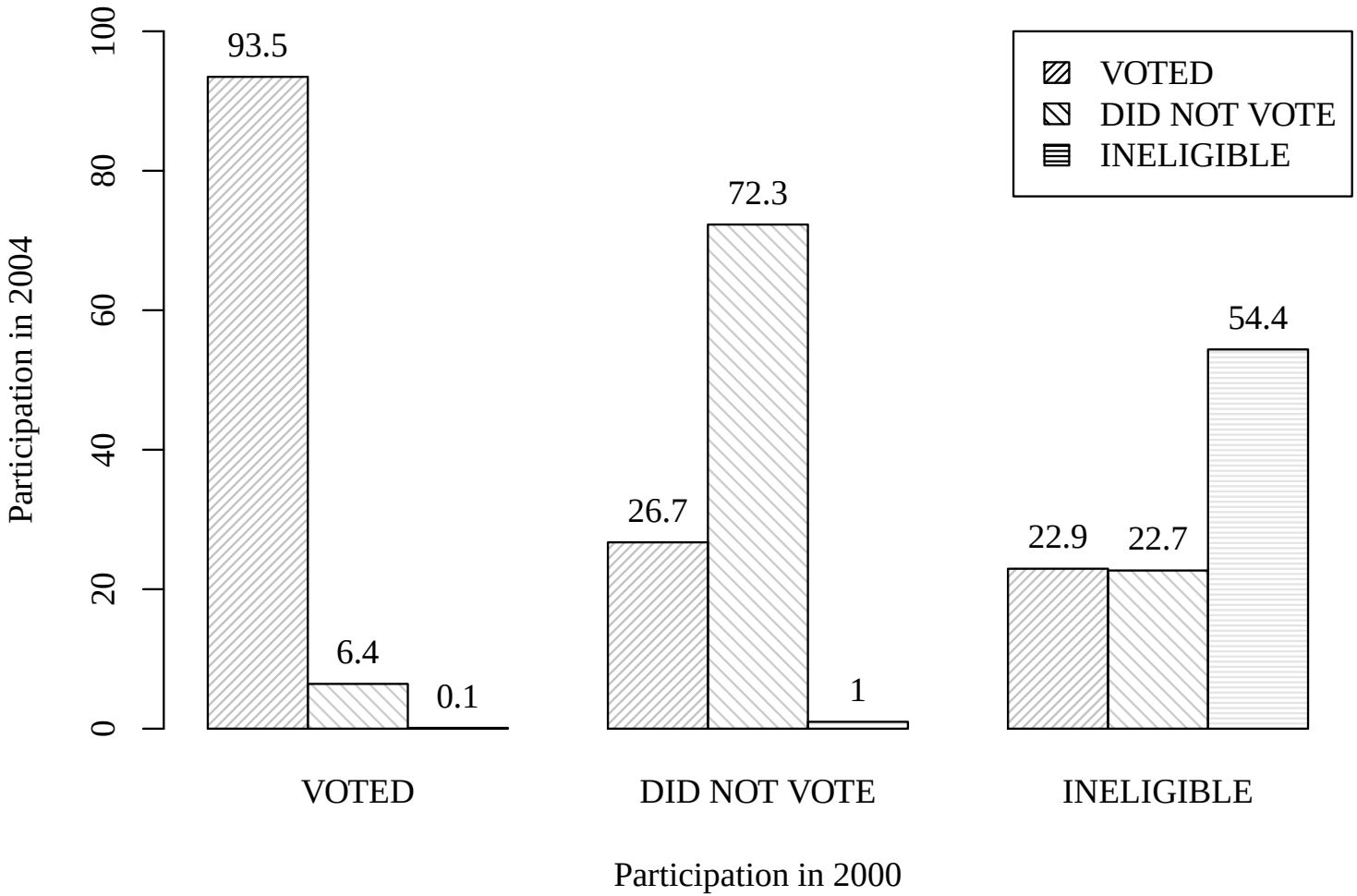
Presidential Choice	Respondent Gender Male	Female
Kerry	45%	52
Bush	53	47
Nader	2	1
Didn't Vote*	1	1
Number of Cases	1137	1487

* Respondent voted, but did not cast vote in Presidential contest

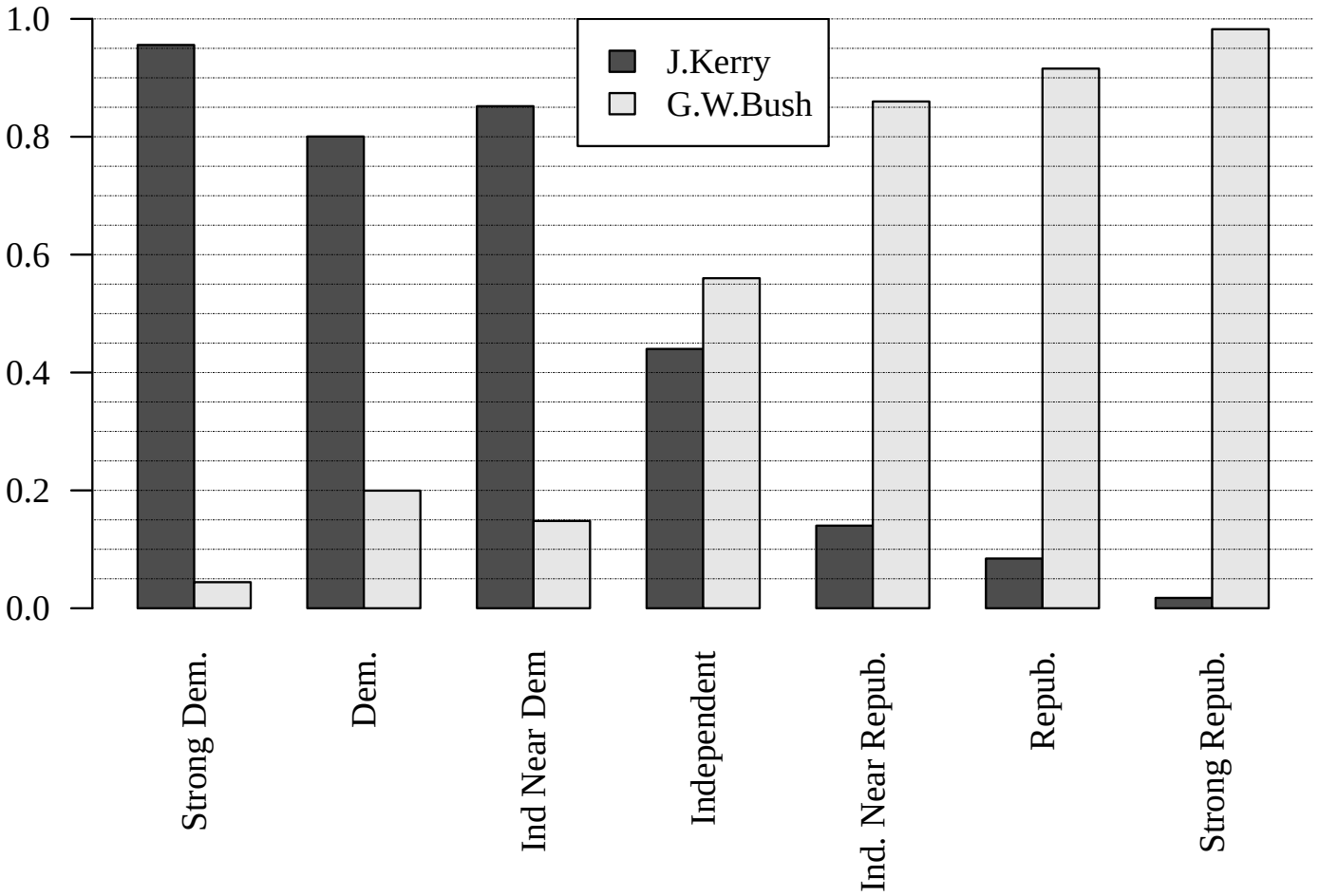
Barplot Representing Gender Gap in 2004



Barplot: Voter Participation Dynamics



Barplot: Partisanship in 2004



To Find Out More

- Check Rcourse for lectures plot-1, plot-2, and plot-3
- In http://pj.freefaculty.org/R/WorkingExamples, see plot-barplot*.R examples. They have very detailed step-by-step instructions.

NB: Many Other Types of Plots

- "spinogram" is a barplot that scales the widths of the bars according to the numbers of observations
- dot plot replaces the "big boxy bars" with smaller dots to mark the tops of the bars.
- pie charts are awful, every reasonable person would agree they should never be used for anything. (my definition of reasonable is based on your answer: "do you hate pie charts?").

KU

Session

```
sessionInfo()
```

```
R version 4.0.2 (2020-06-22)
Platform: x86_64-pc-linux-gnu (64-bit)
Running under: Ubuntu 20.04.1 LTS

Matrix products: default
BLAS:   /usr/lib/x86_64-linux-gnu/blas/libblas.so.3.9.0
LAPACK: /usr/lib/x86_64-linux-gnu/lapack/liblapack.so.3.9.0

locale:
 [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
 [3] LC_TIME=en_US.UTF-8        LC_COLLATE=en_US.UTF-8
 [5] LC_MONETARY=en_US.UTF-8    LC_MESSAGES=en_US.UTF-8
 [7] LC_PAPER=en_US.UTF-8       LC_NAME=C
 [9] LC_ADDRESS=C               LC_TELEPHONE=C
[11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods
[7] base

other attached packages:
[1] memisc_0.99.25.6   MASS_7.3-51.6      lattice_0.20-41
```

Session ...

```
[4] descr_1.1.4          gmodels_2.18.1      xtable_1.8-4
[7] rockchalk_1.8.144    stationery_0.98.30

loaded via a namespace (and not attached):
 [1] zip_2.0.4           Rcpp_1.0.4.6        cellranger_1.1.0
 [4] pillar_1.4.6        compiler_4.0.2      nloptr_1.2.2.2
 [7] plyr_1.8.6          forcats_0.5.0       base64enc_0.1-3
[10] tools_4.0.2         boot_1.3-25         digest_0.6.25
[13] lme4_1.1-23         statmod_1.4.34      tibble_3.0.3
[16] lifecycle_0.2.0     jsonlite_1.7.0      evaluate_0.14
[19] nlme_3.1-147        pkgconfig_2.0.3     rlang_0.4.7
[22] openxlsx_4.1.5      Matrix_1.2-18       curl_4.3
[25] haven_2.3.1         xfun_0.15           rio_0.5.16
[28] repr_1.1.0          knitr_1.29          hms_0.5.3
[31] vctrs_0.3.2         gtools_3.8.2        grid_4.0.2
[34] data.table_1.13.0   readxl_1.3.1        foreign_0.8-79
[37] rmarkdown_2.3       gdata_2.18.0        carData_3.0-4
[40] minqa_1.2.4         magrittr_1.5        car_3.0-9
[43] ellipsis_0.3.1      htmltools_0.5.0     splines_4.0.2
[46] kutils_1.70         abind_1.4-5         stringi_1.4.6
[49] crayon_1.3.4
```