Data Visualization I

Principles of Visualization & Univariate Graphics

Visualization: Three Hows

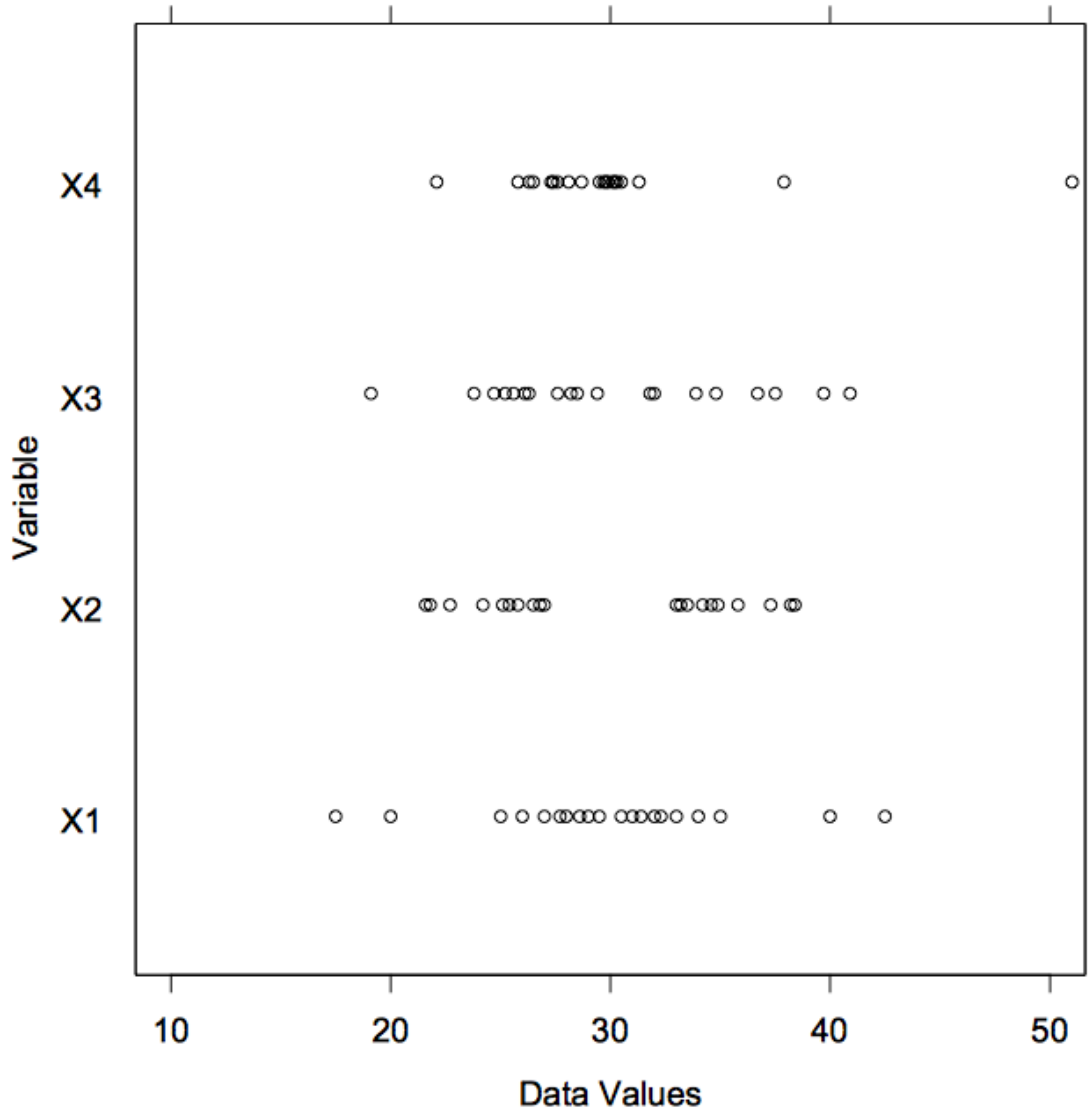
1. ***How*** humans ***perceive*** graphical information
2. ***How*** human perception means you should ***design*** your graphics
3. ***How***, practically, you ***draw*** your graphics with software

Visualization

- **Today: Graphics I: Univariate Graphics**
  - Conceptual: Theory of Graphics and Psychology of Perception
  - Practical: What kinds of graphics to use, when, and ideas

Why are graphics useful?

Observation:	X_1	X_2	X_3	X_4
1	32.3	33.2	24.7	29.7
2	28.0	34.2	29.4	30.2
3	31.4	27.0	28.5	28.7
4	29.5	33.0	25.6	27.3
5	40.0	35.8	27.6	31.3
6	20.0	34.6	32.0	29.5
7	26.0	24.2	28.2	26.3
8	28.6	34.9	40.9	29.9
9	27.7	25.1	37.5	29.8
10	27.0	37.3	26.3	30.1
11	17.5	22.7	33.9	37.9
12	31.0	25.4	36.7	27.6
13	32.0	25.8	25.2	30.3
14	30.5	38.2	23.8	22.1
15	34.0	26.5	26.1	28.1
16	42.5	38.4	28.2	26.5
17	35.0	26.8	31.8	30.5
18	29.0	21.6	39.7	27.4
19	25.0	33.5	19.1	51.0
20	33.0	21.8	34.8	25.8
Mean:	30.0	30.0	30.0	30.0
Standard Deviation:	5.8	5.8	5.8	5.8


X4
X3
X2
X1
Variable
10
20
30
40
50
Data Values

The role of graphics

Tufte (2001)

- Show the data/results
- Induce the viewer to think about the *substance* of the data, rather than the technical process behind the data (or graphics) production
- Avoid distortion
- Present many numbers in a small space
- Make large data sets coherent

The role of graphics

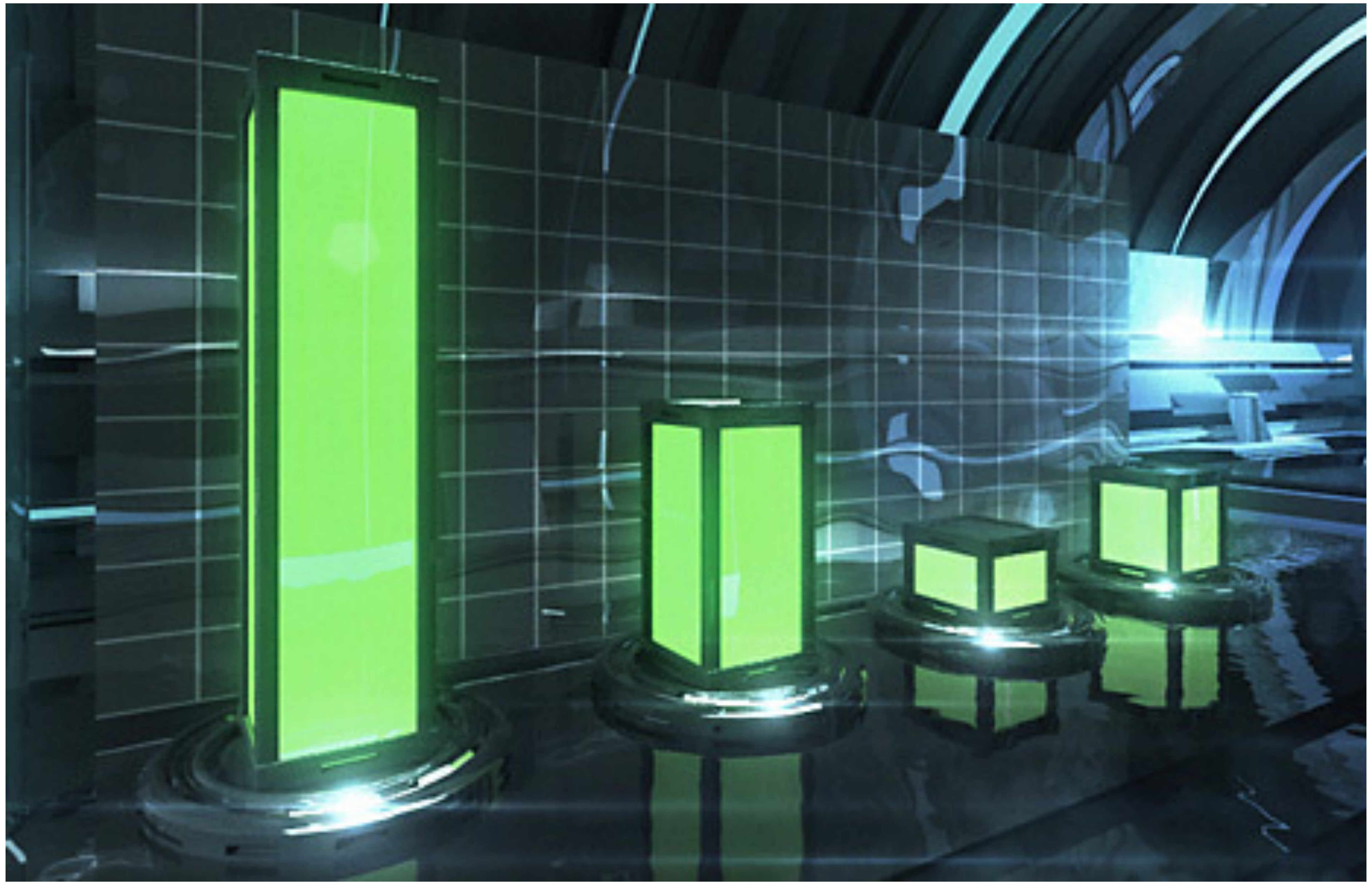
Tufte (2001)

- Encourage visual comparison of many different pieces of data
- Reveal data at several levels of detail, from a broad overview to the fine structure
- Serve a clear purpose: description, exploration, tabulation, or decoration
- Be closely integrated with the statistical and verbal descriptions of a data set

Core Principle: The Data-Ink Ratio

- *Maximize* your data-ink ratio
    - Data-ink ratio = data ink/total ink use to print the graphic
- Erase non-data ink
- Erase redundant data-ink
- Revise and edit, revise and edit, revise and edit.

An egregious example

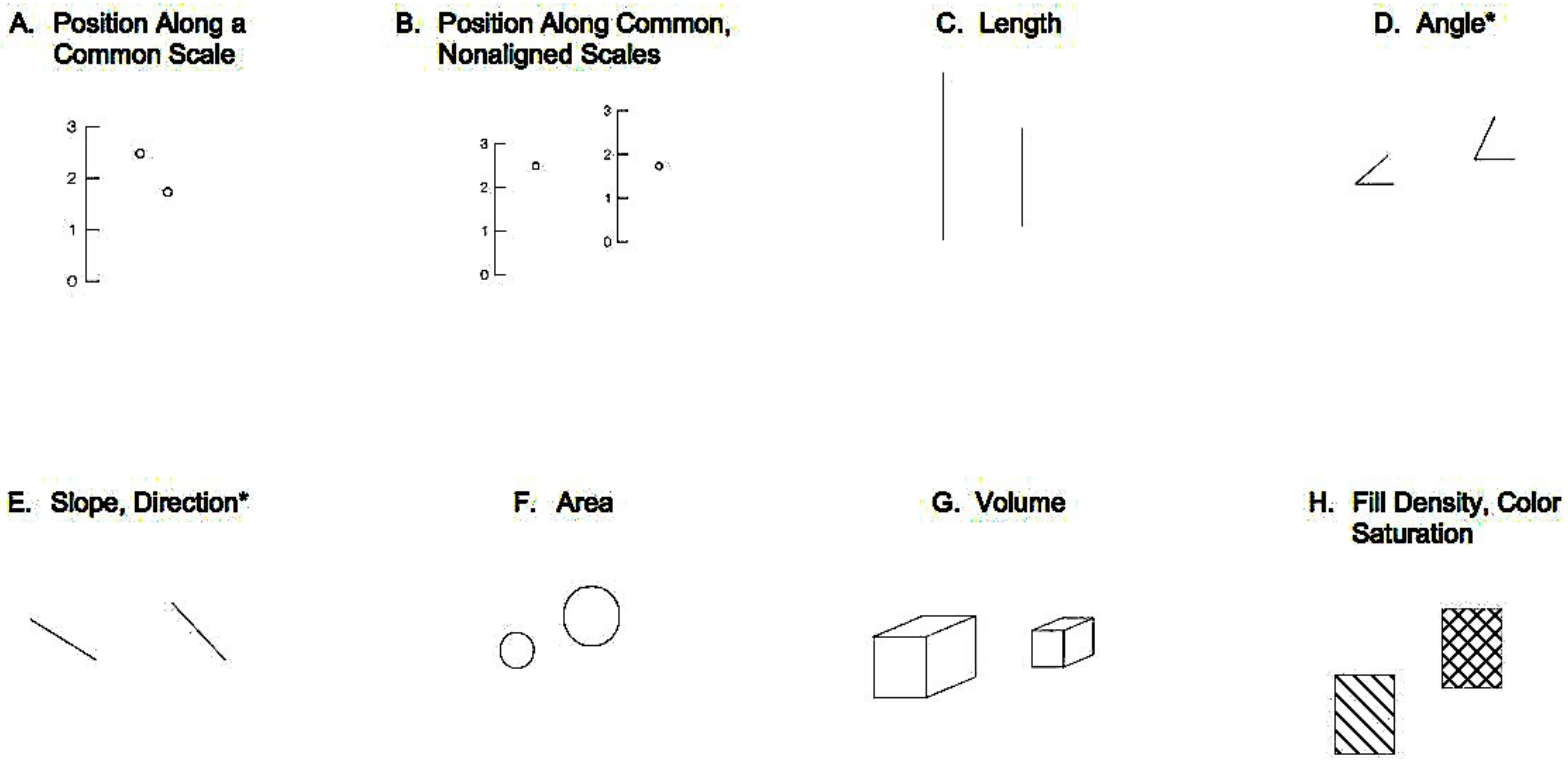


Cognition and graphics

- Not all graphs are created equal
- Humans perceive graphics with different levels of accuracy

Graphical perception tasks

Ordered from Most Accurate to Least Accurate



Cleveland (1994); Jacoby (2014)

Univariate Graphs

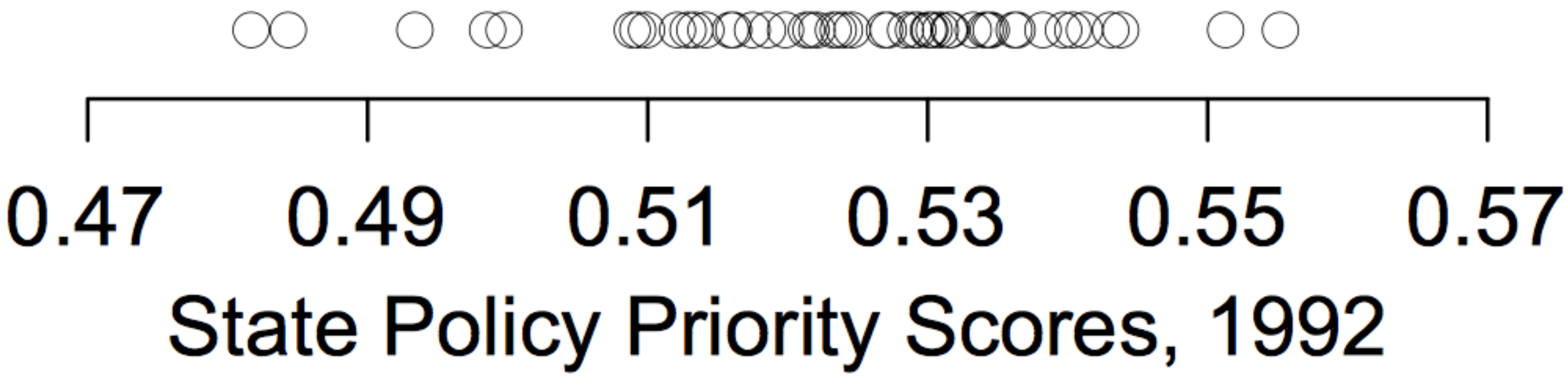
General Source: Jacoby (2014)

Univariate graphics

- A univariate graph is a **model** of your data
    - Highlight salient points without distorting features or imposing assumptions
    - Avoid excessive detail on individual points, highlight overall structure of data
- Many different univariate graphs: univariate scatterplots, histograms, smoothed histograms, quantile plots, box and whisker plots, dot plots

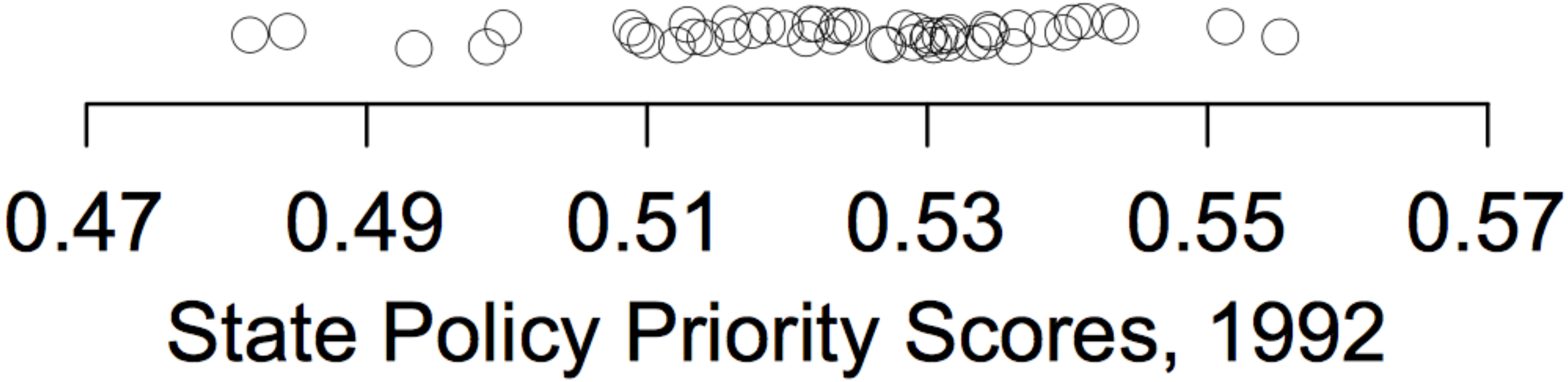
Univariate Scatterplots

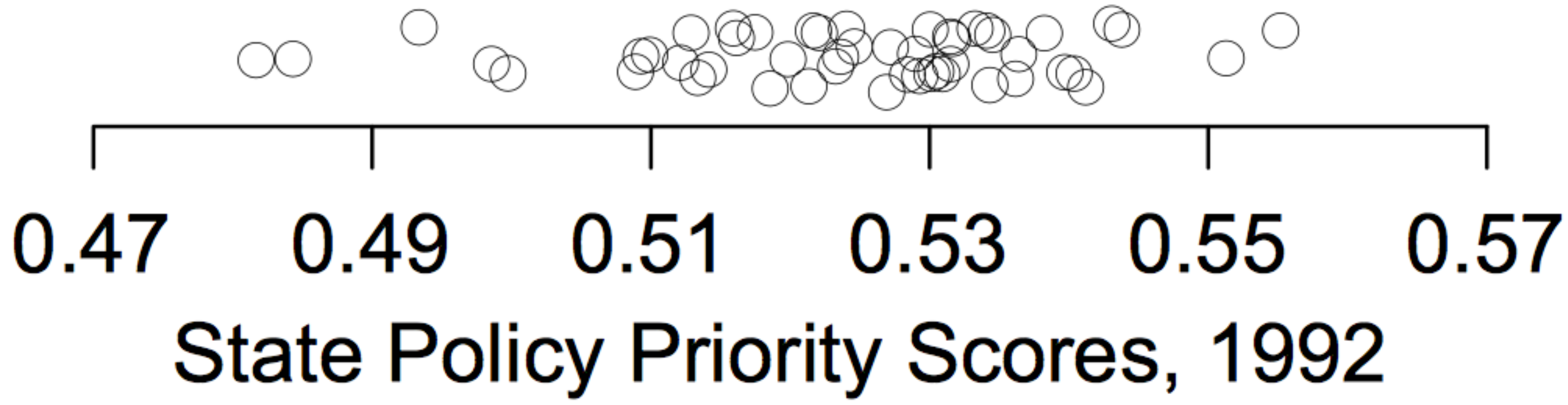
- Each observation plotted as a point along axis representing range of variable values

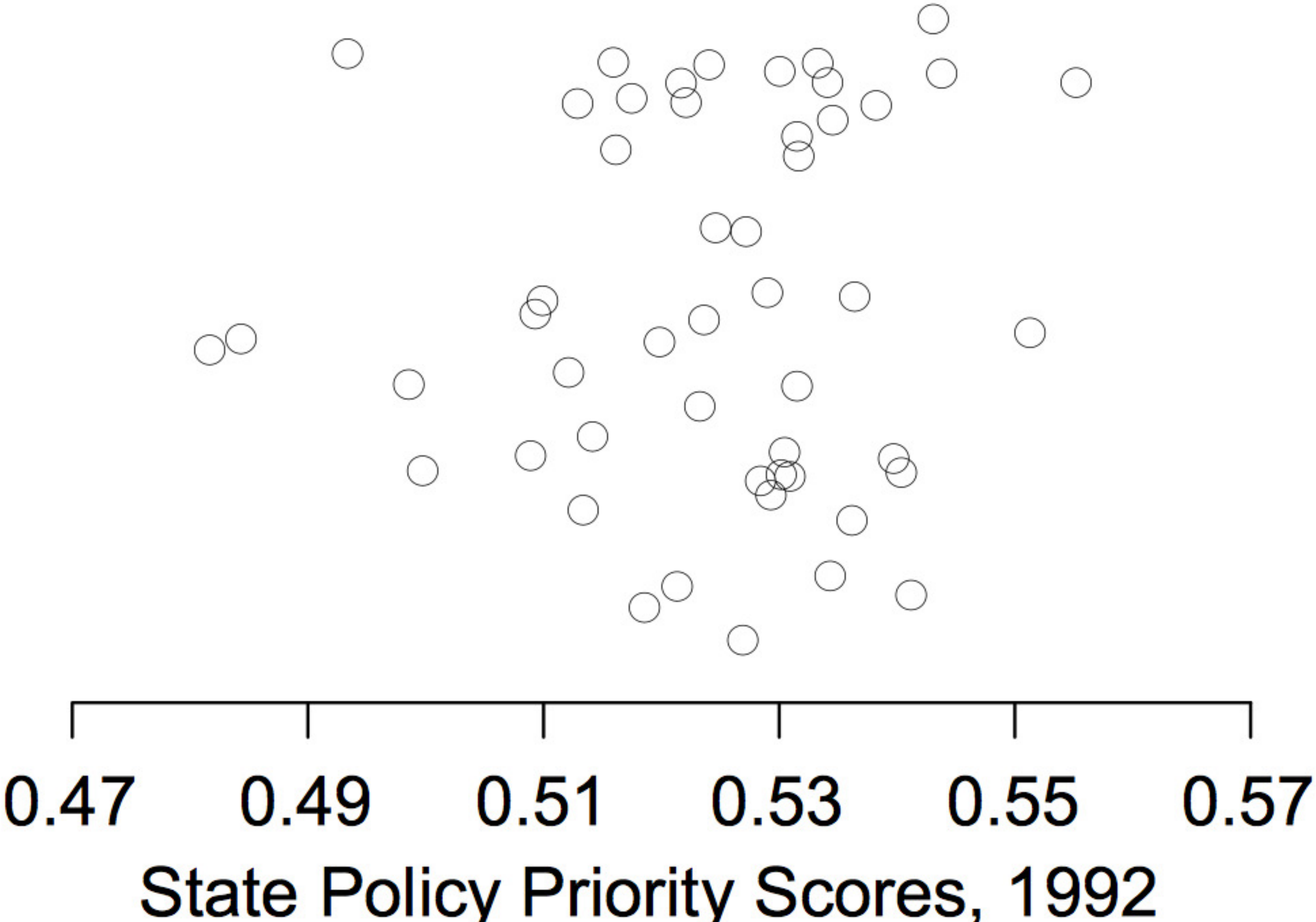

0.47
0.49
0.51
0.53
0.55
0.57
State Policy Priority Scores, 1992

Univariate Scatterplots

- Each observation plotted as a point along axis representing range of variable values
- We have an over plotting problem that is not clear how to fix. Often fix by attempting to jitter.
    - But unclear how far one should go.

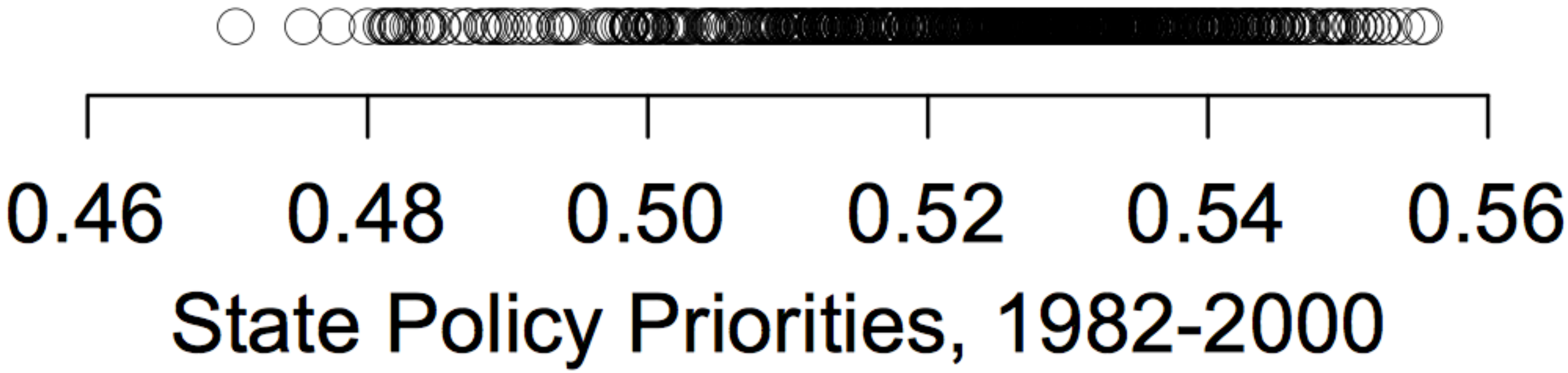

0.47
0.49
0.51
0.53
0.55
0.57
State Policy Priority Scores, 1992

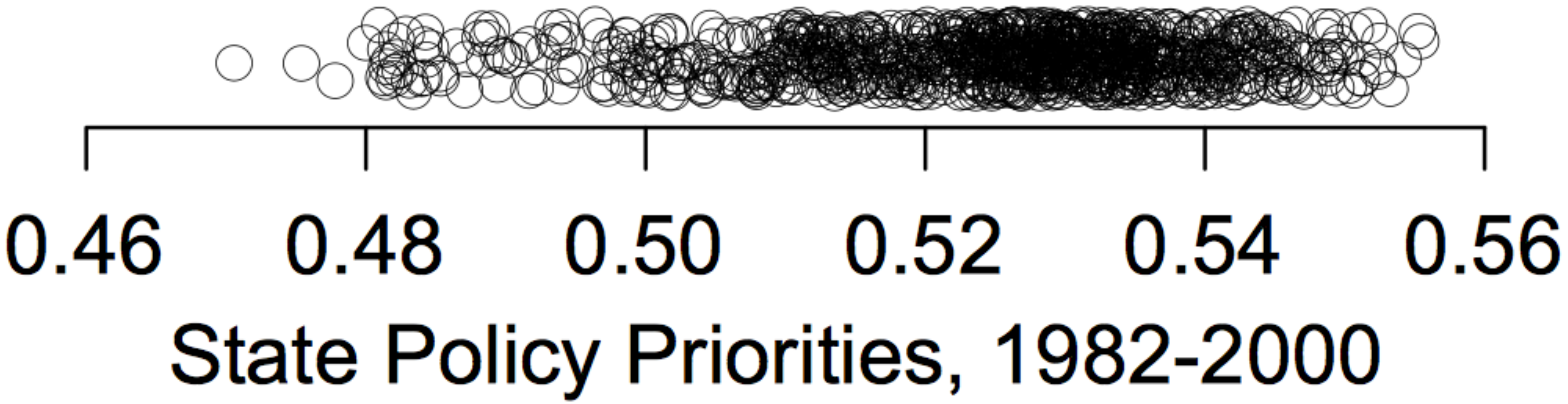

0.47
0.49
0.51
0.53
0.55
0.57
State Policy Priority Scores, 1992


0.47
0.49
0.51
0.53
0.55
0.57
State Policy Priority Scores, 1992

Univariate Scatterplots

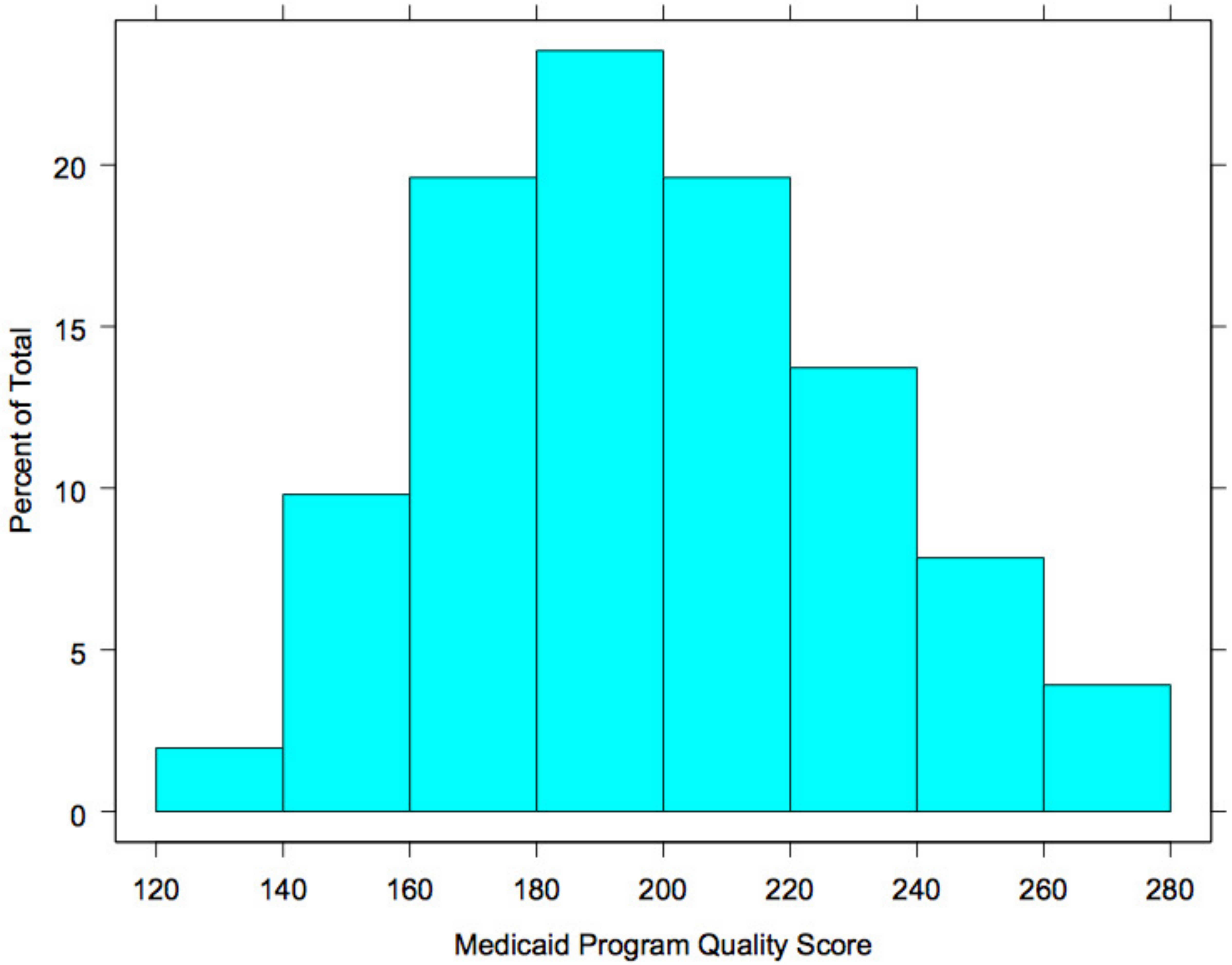
- Each observation plotted as a point along axis representing range of variable values
- Have an over plotting problem that is not clear how to fix. Often fix by attempting to jitter.
  - But unclear how far one should go.
  - Gets even worse with more data.


0.46
0.48
0.50
0.52
0.54
0.56
State Policy Priorities, 1982-2000

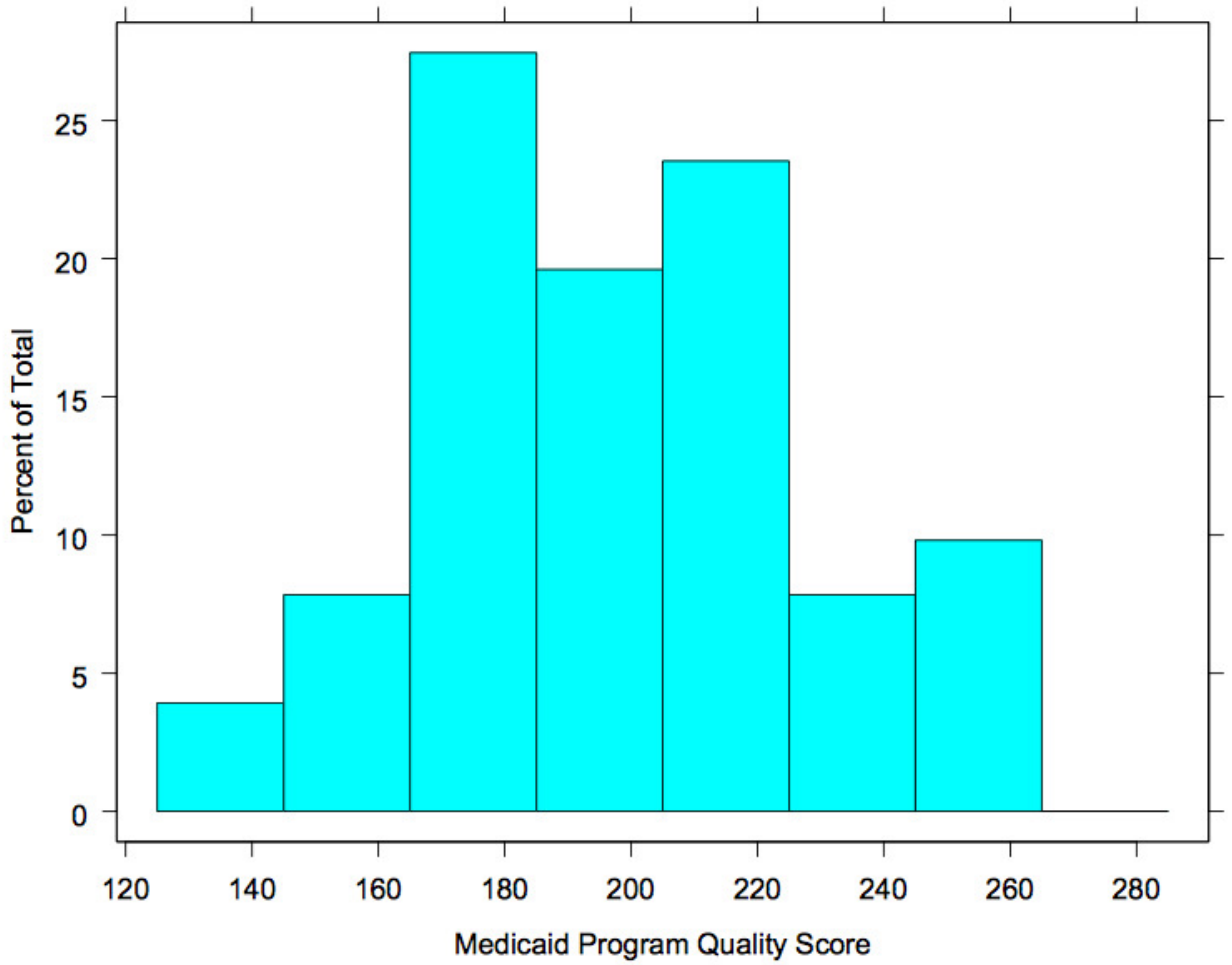

0.46
0.48
0.50
0.52
0.54
0.56
State Policy Priorities, 1982-2000

Histogram

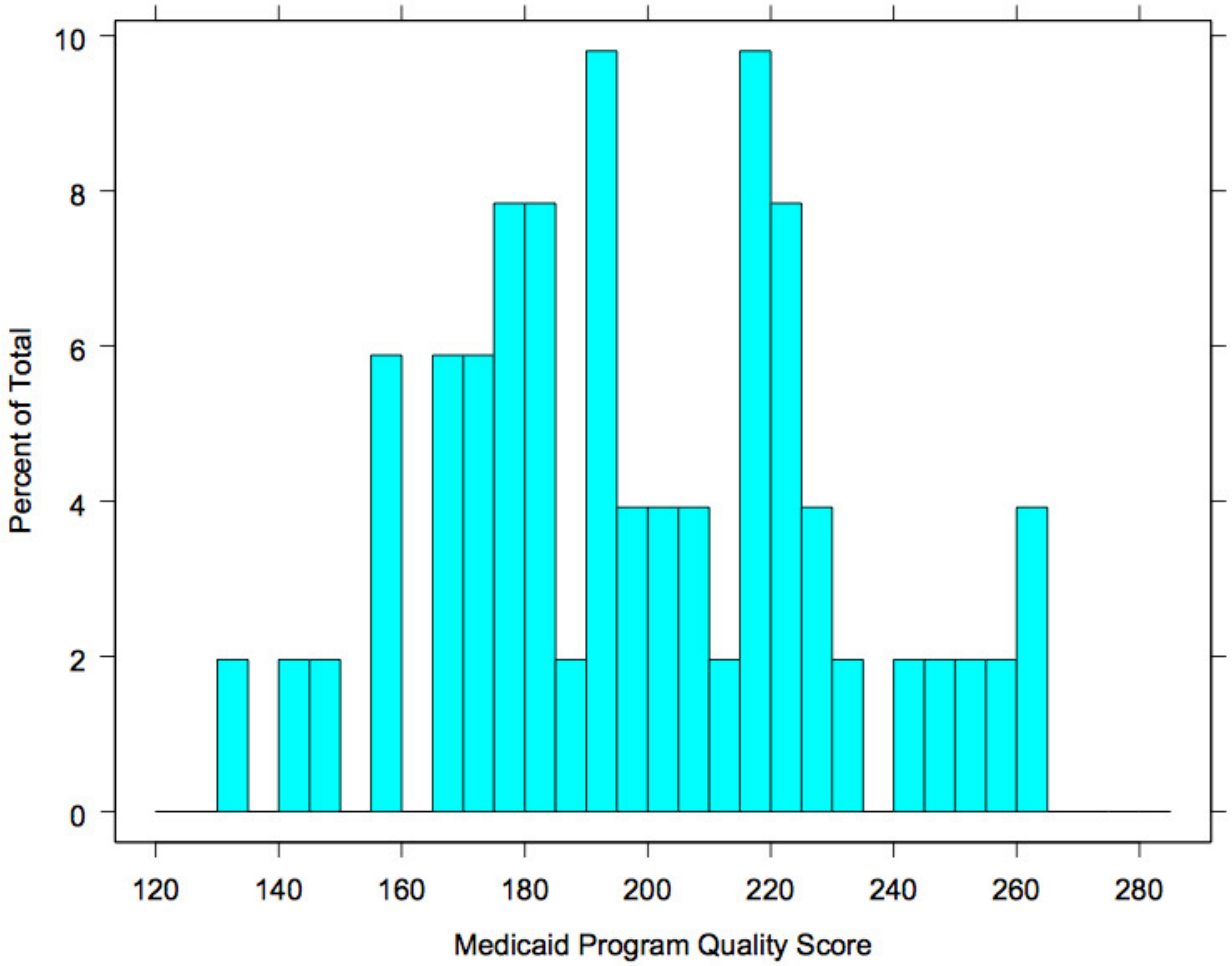
- A two-dimensional diagram, with one axis representing the range of the variable, and the other axis representing the data density a positions within the range
- Most common kind of univariate graph, provide a great deal of information about the distribution and data values
- If data are relatively continuous, observations are grouped into mutually exclusive and exhaustive categories, called “bins”
- Major issue with histograms: the binning process.



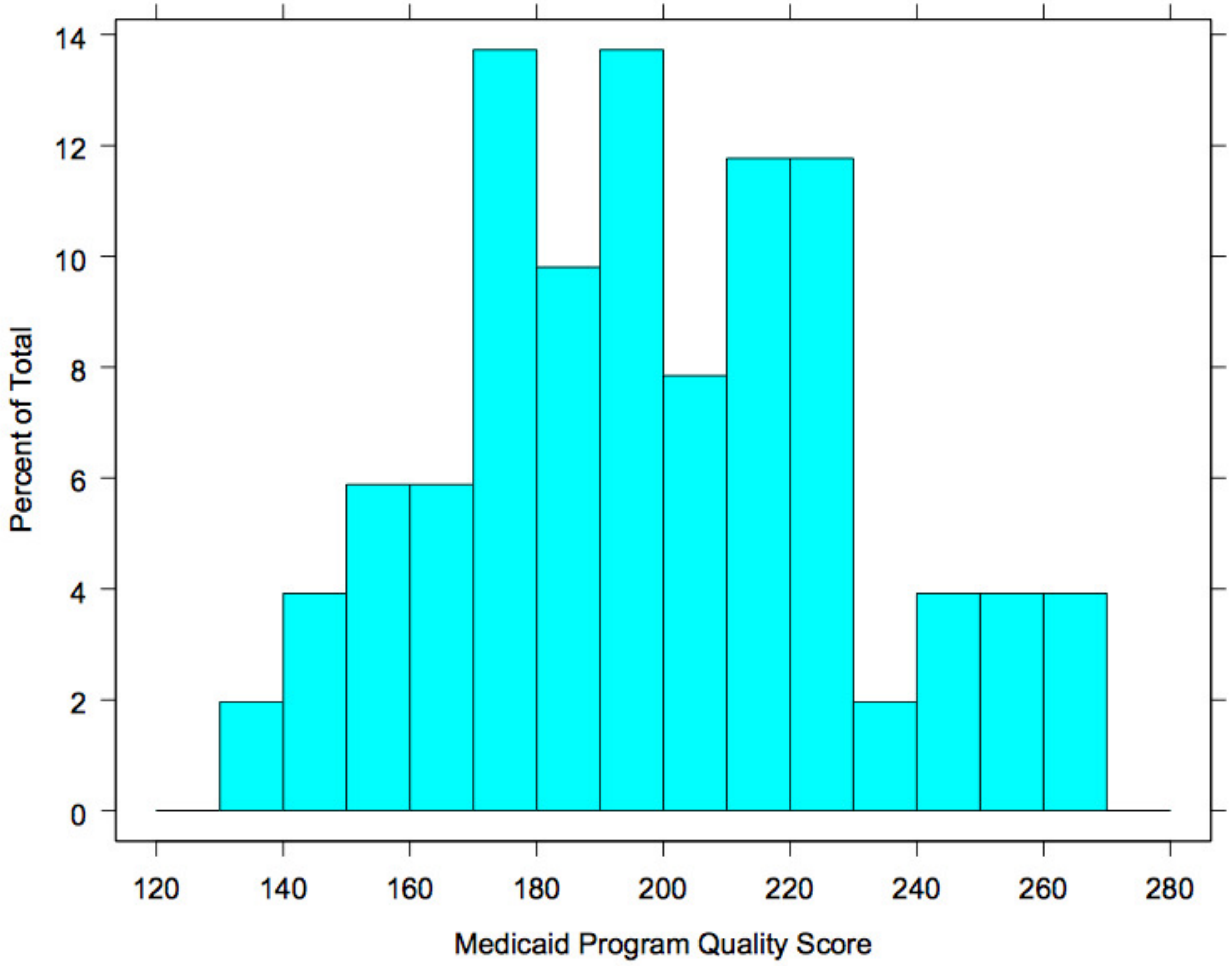
Note: Bin origin at 120, bin width is 20.



Note: Bin origin at 125, bin width is 20.



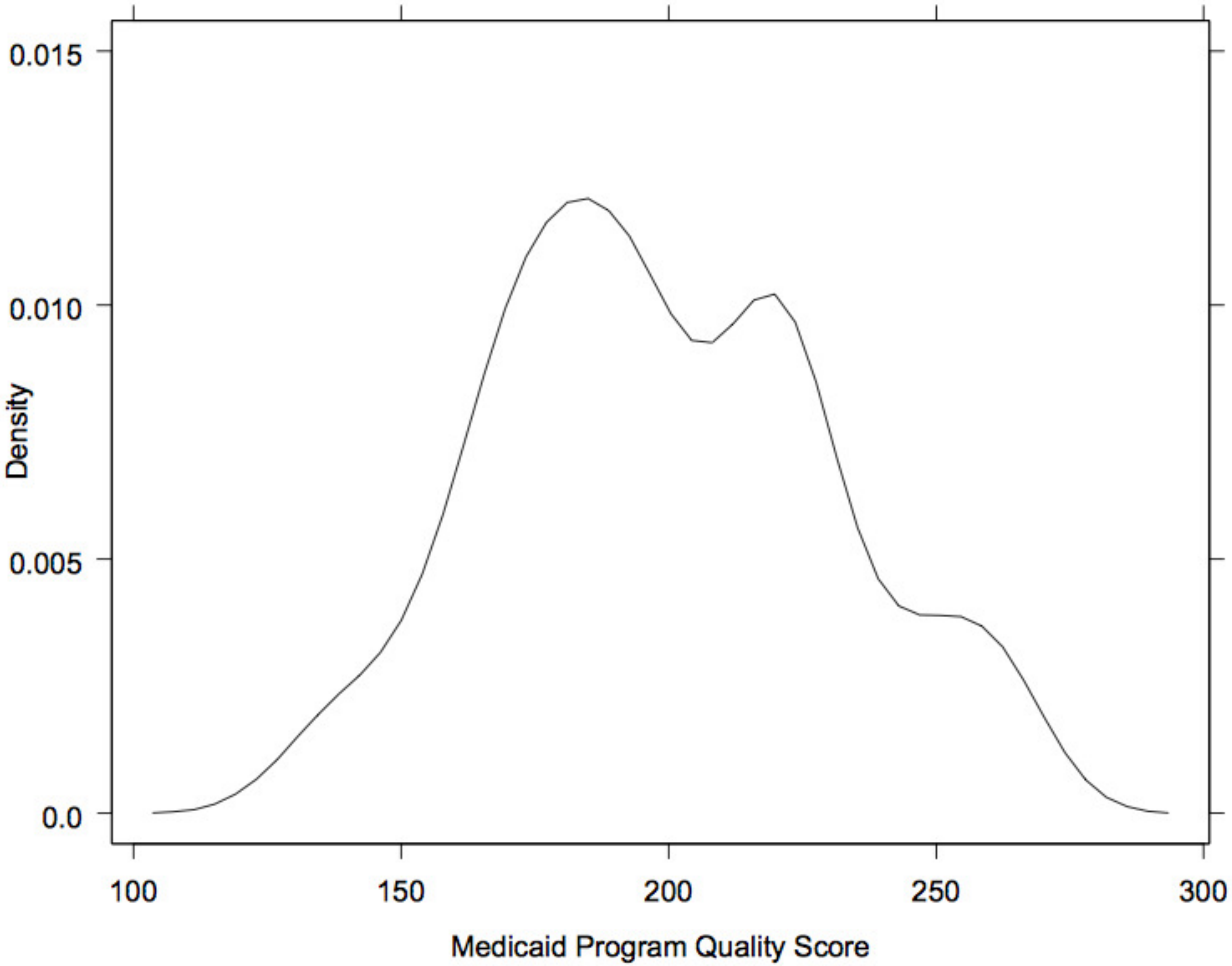
Note: Bin origin at 120, bin width is 5.



Note: Bin origin at 120, bin width is 10.

Smoothed histograms

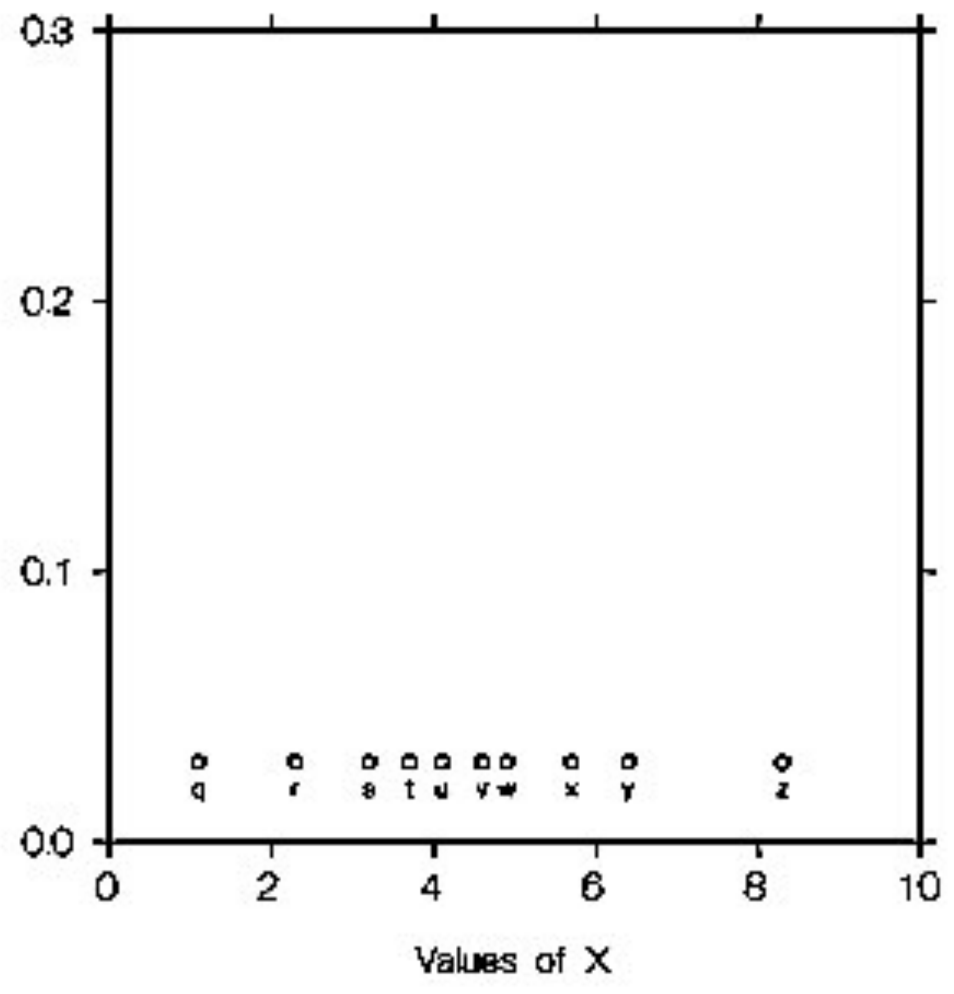
- Formally: ***kernel density plots***
- Shows local densities as a smooth, continuous function of the original data series


0.015
0.010
0.005
0.0
Density
100
150
200
250
300
Medicaid Program Quality Score

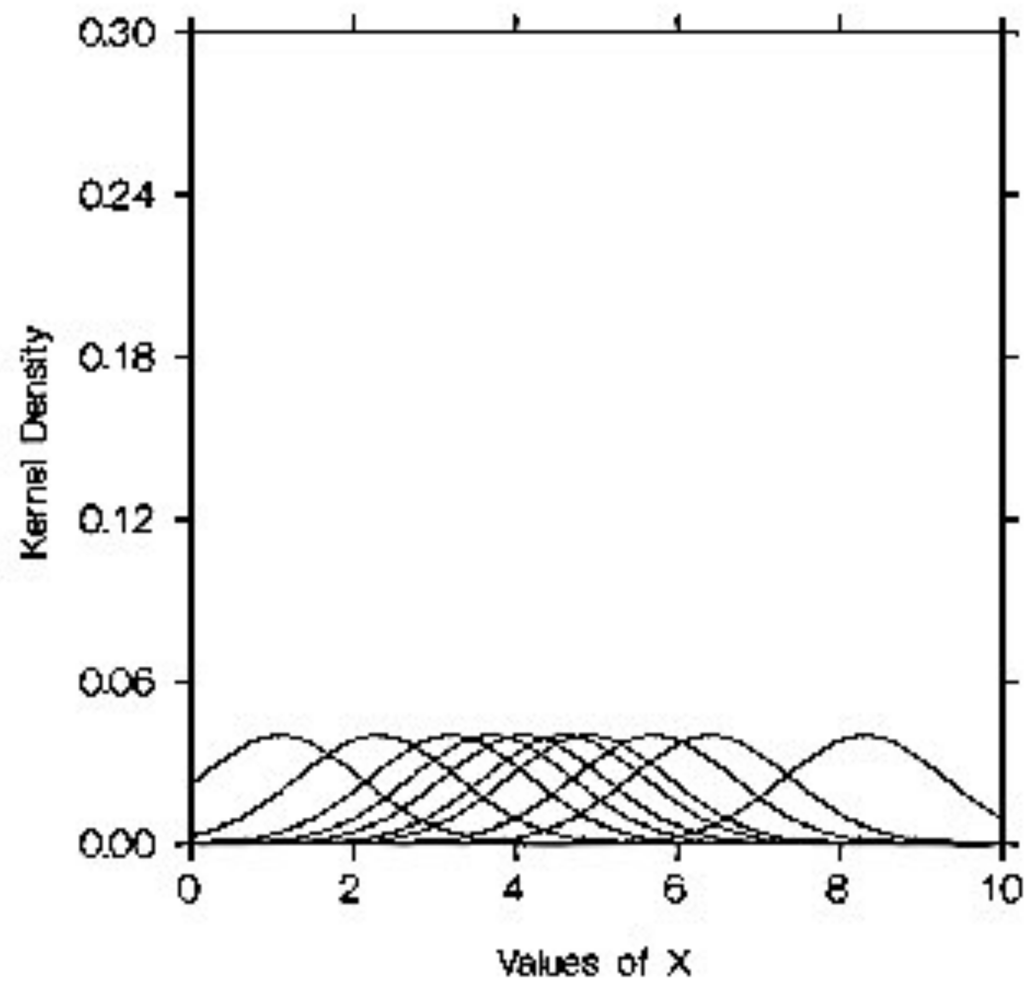
Smoothed Histograms

- Shows local densities as a smooth, continuous function of the original data series
- Looks better than a regular histogram, right?
- How do we make it?

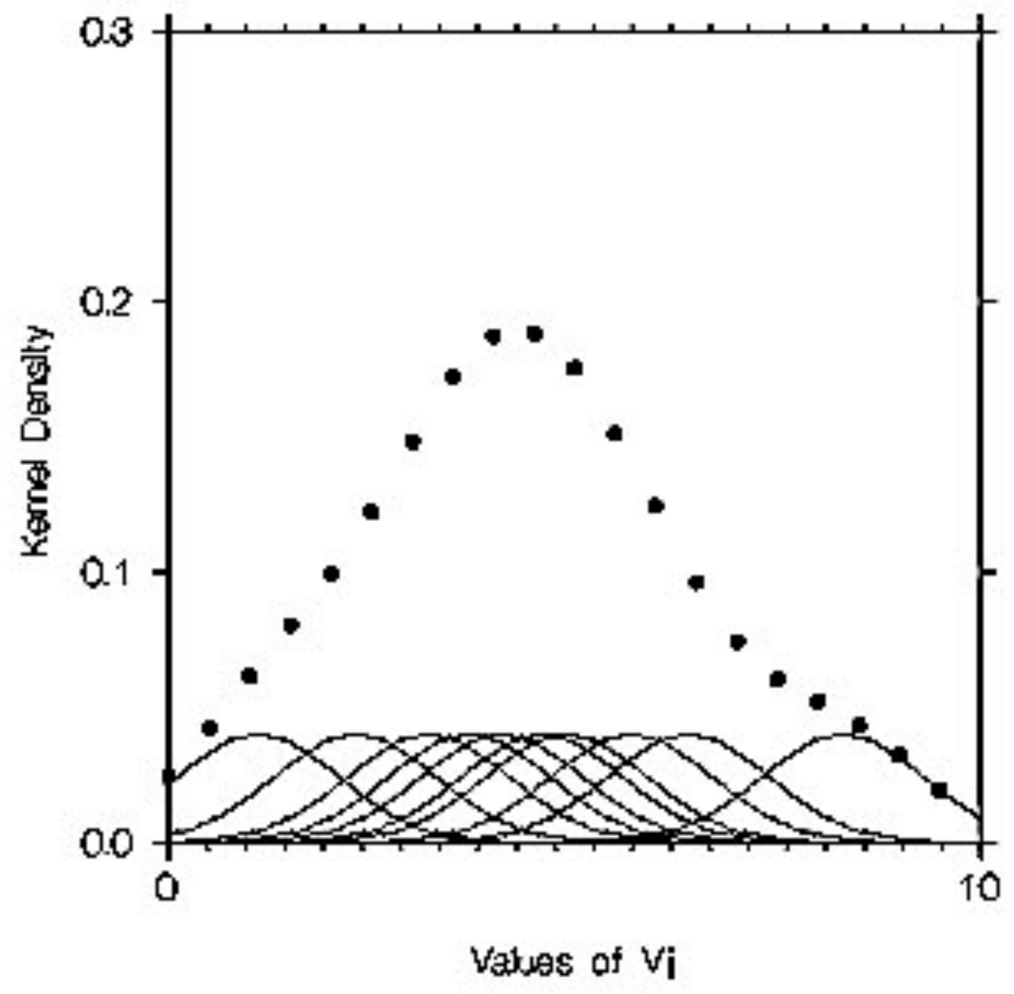
A. Univariate Scatterplot of 10 Data Points



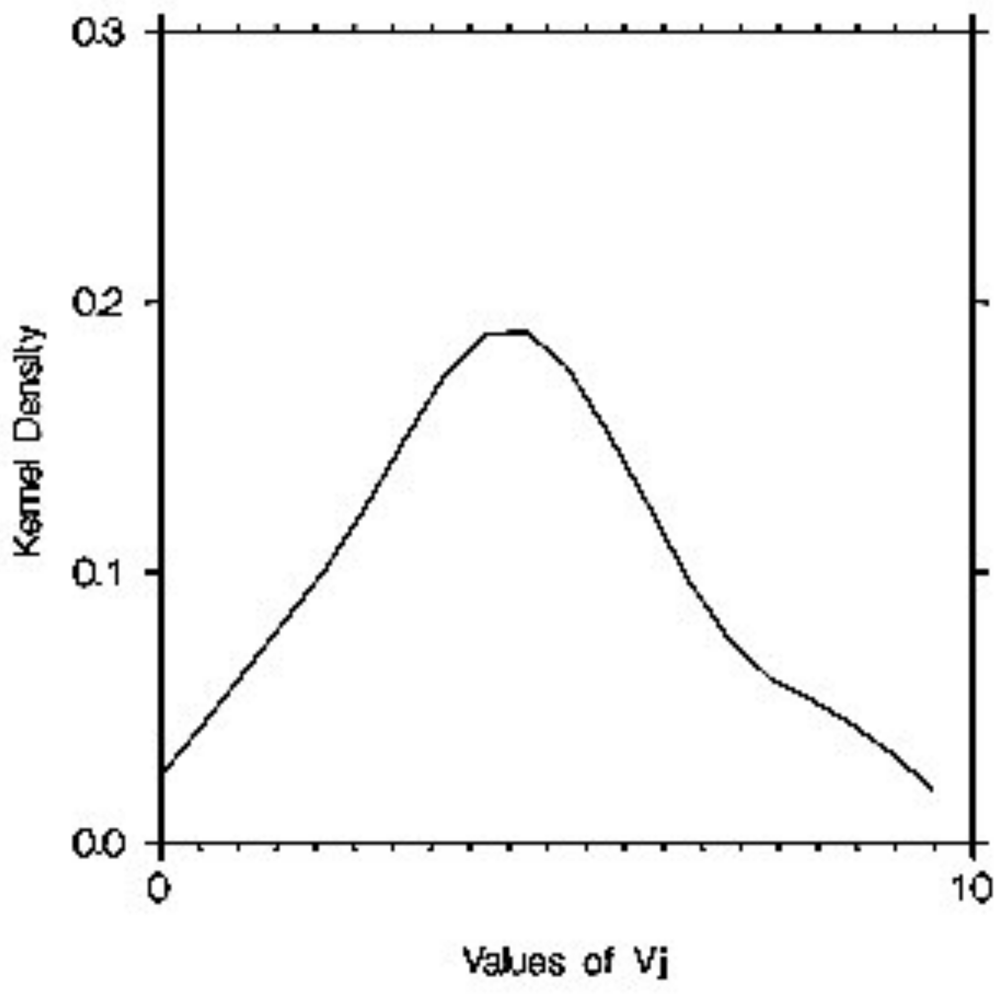
B. Data Shown as Kernel Densities



C. Summing Heights of Kernel Densities

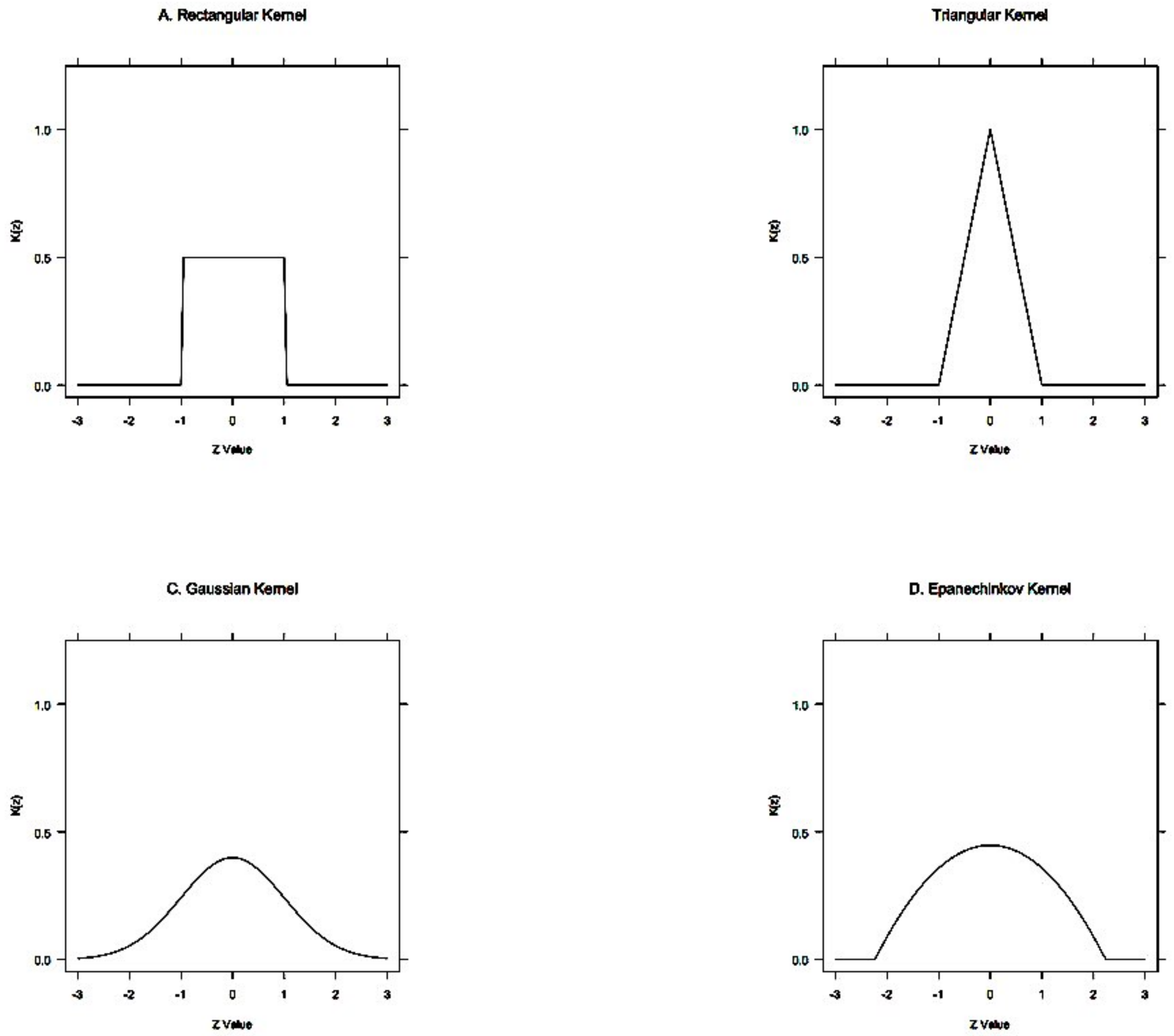


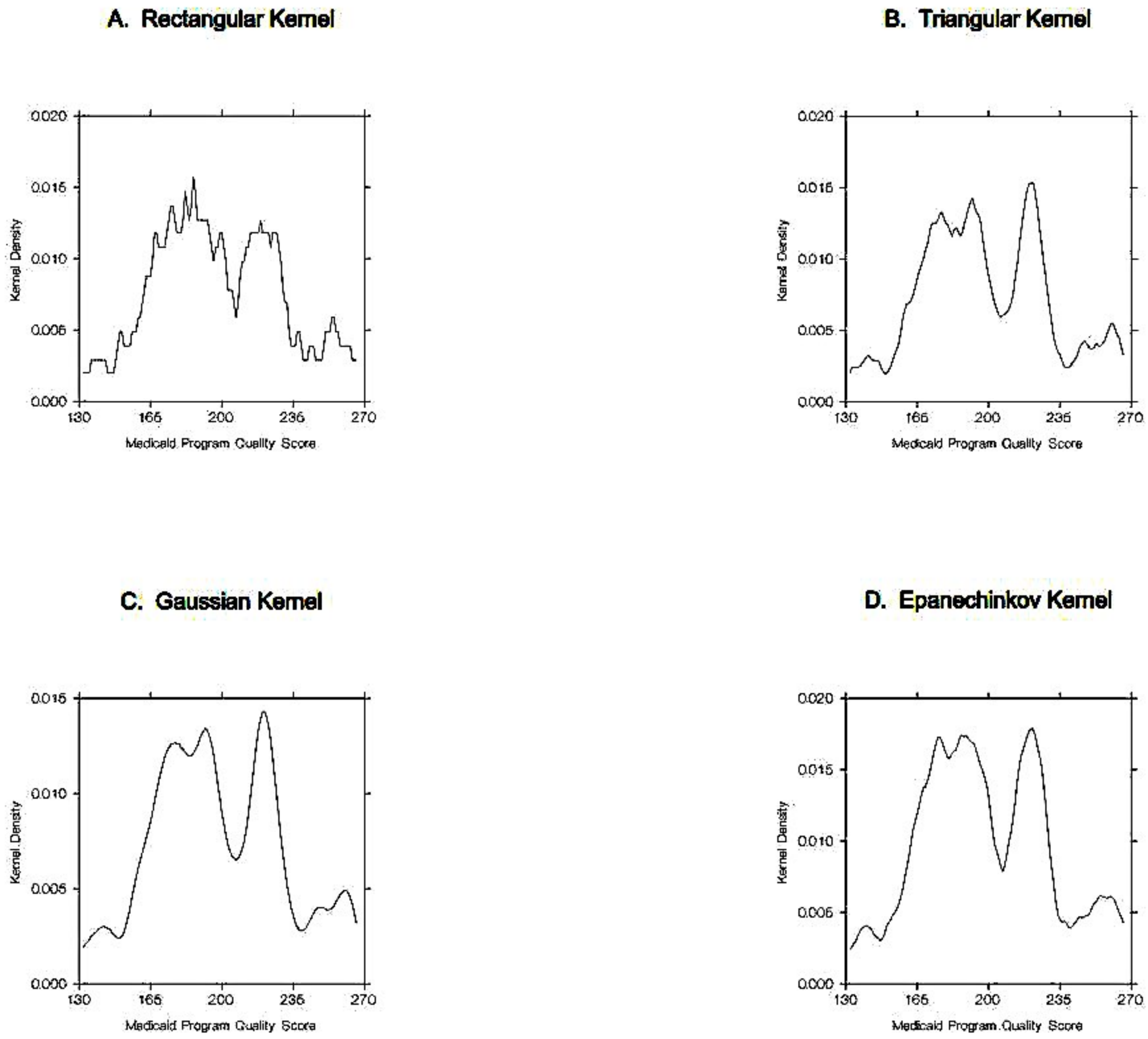
D. Smoothed Histogram



Smoothed Histograms

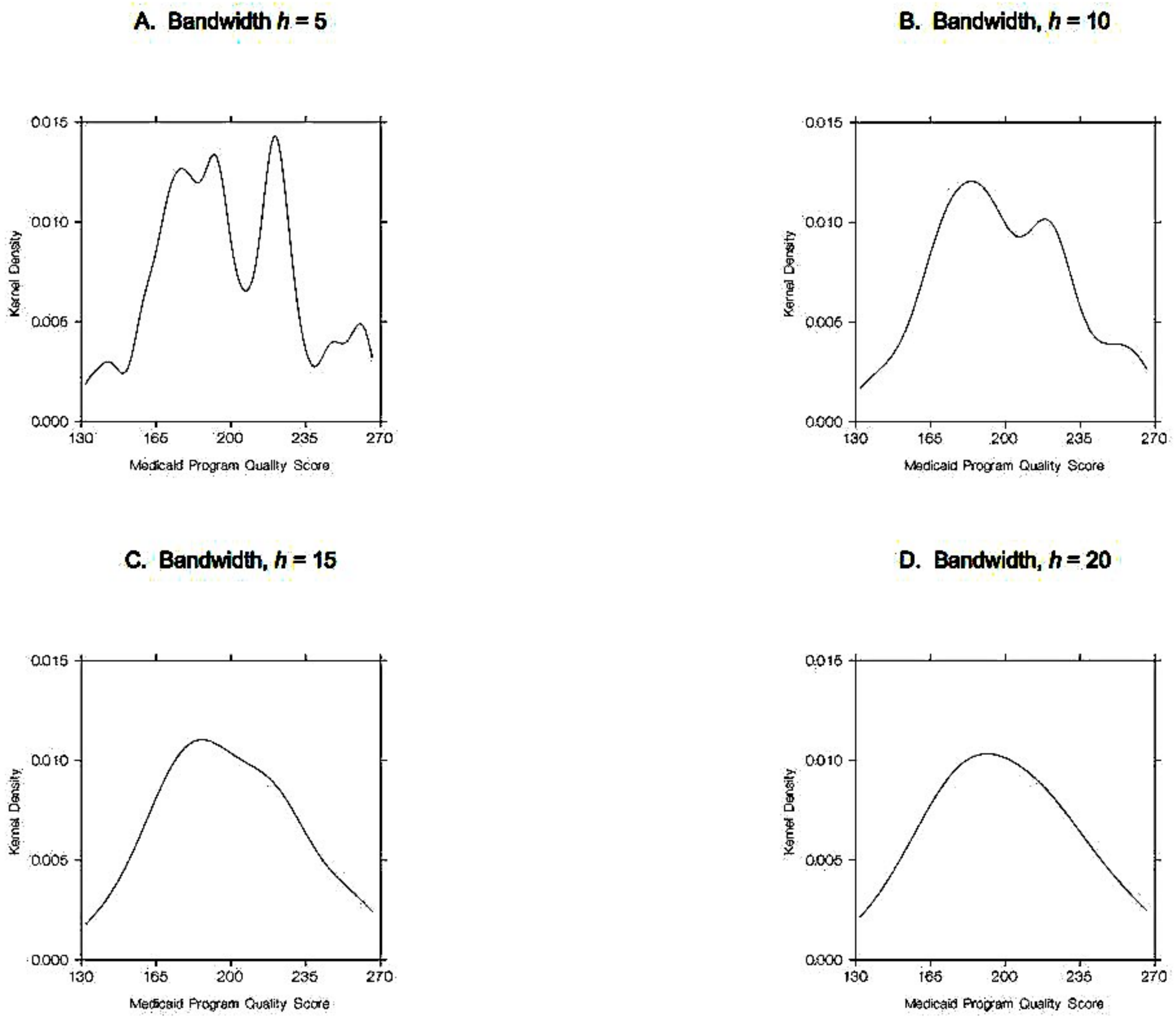
- Shows local densities as a smooth, continuous function of the original data series
- Woohoo! So we’ve fixed the problem! Yay creative statisticians!
  - Not so fast there, skippy.
- There are many different kernel functions we can use to smooth the data.


A. Rectangular Kernel
Triangular Kernel
C. Gaussian Kernel
D. Epanechinkov Kernel
K(z)
Z Value
1.0
0.5
0.0
-3
-2
-1
0
1
2
3


A. Rectangular Kernel
B. Triangular Kernel
C. Gaussian Kernel
D. Epanechinkov Kernel
Kernel Density
Medicaid Program Quality Score
0.000
0.005
0.010
0.015
0.020
130
165
200
235
270

Smoothed Histograms

- Shows local densities as a smooth, continuous function of the original data series
- There are many different kernel functions we can use to smooth the data.
- We also have to decide how much data those kernels look at (the *bandwidth*) . . . which means we still, fundamentally, have a kind binning problem.


A. Bandwidth h = 5
B. Bandwidth, h = 10
C. Bandwidth, h = 15
D. Bandwidth, h = 20
Kernel Density
0.015
0.010
0.005
0.000
130
165
200
235
270
Medicaid Program Quality Score

Smoothed Histograms

- Shows local densities as a smooth, continuous function of the original data series

- There are many different kernel functions we can use to smooth the data.

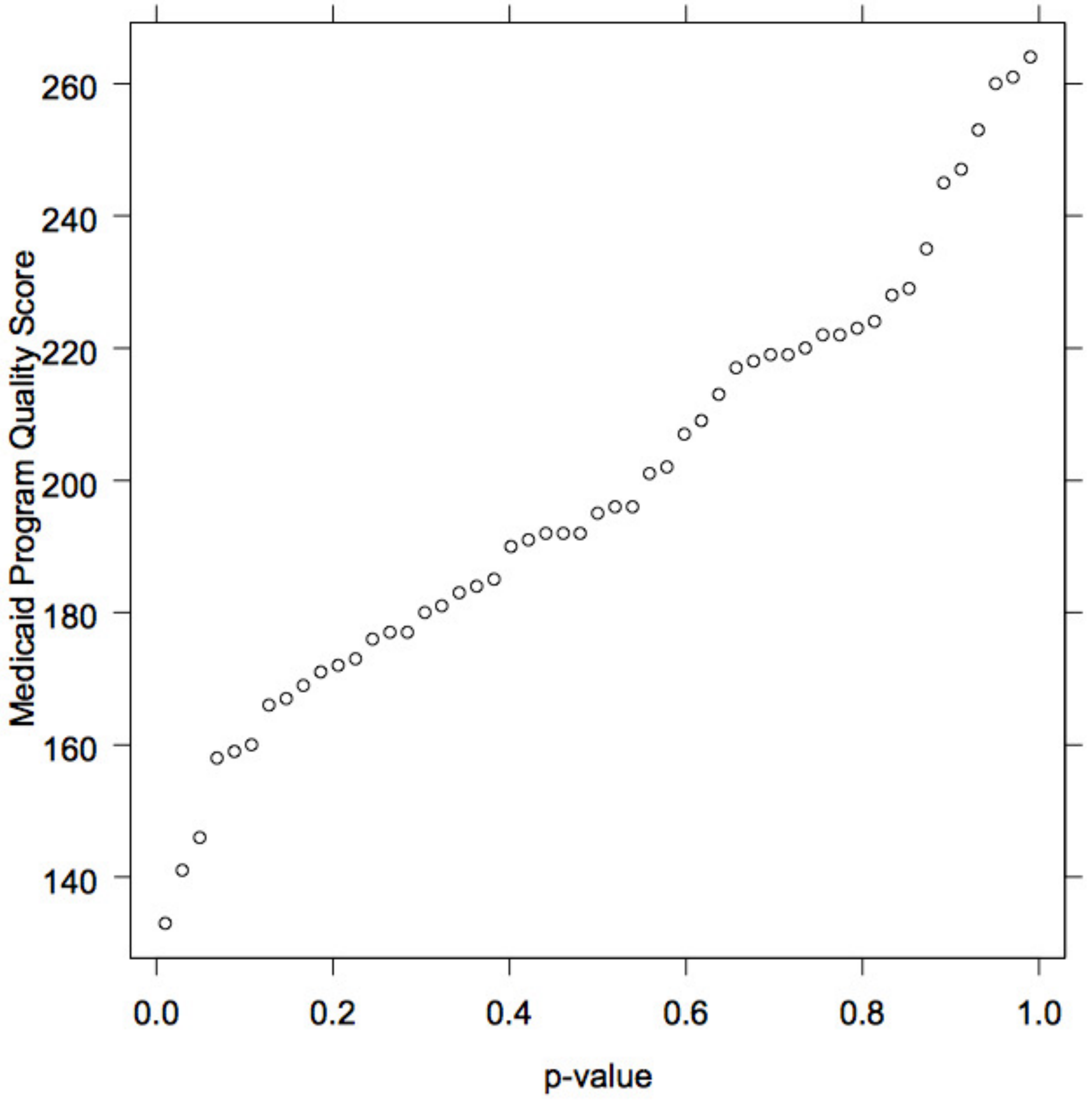
- We also have to decide how much data those kernels look at (the *bandwidth*) . . . which means we still, fundamentally, have a kind binning problem.

Assumptions and Models in Graphics

- OK, I give up. I’ll never make any graphs ever and never make any mistakes.
    - *No! Don’t do that!*
- Creation of graphics, like creation of OLS models, is always problematic. There are always assumptions floating around you can violate.
    - Purpose is to make sure to *understand* what trade-offs you’re making, and what the strengths and weaknesses of your graph are.
    - Nice parallel to actual statistics.

Quantile Plots

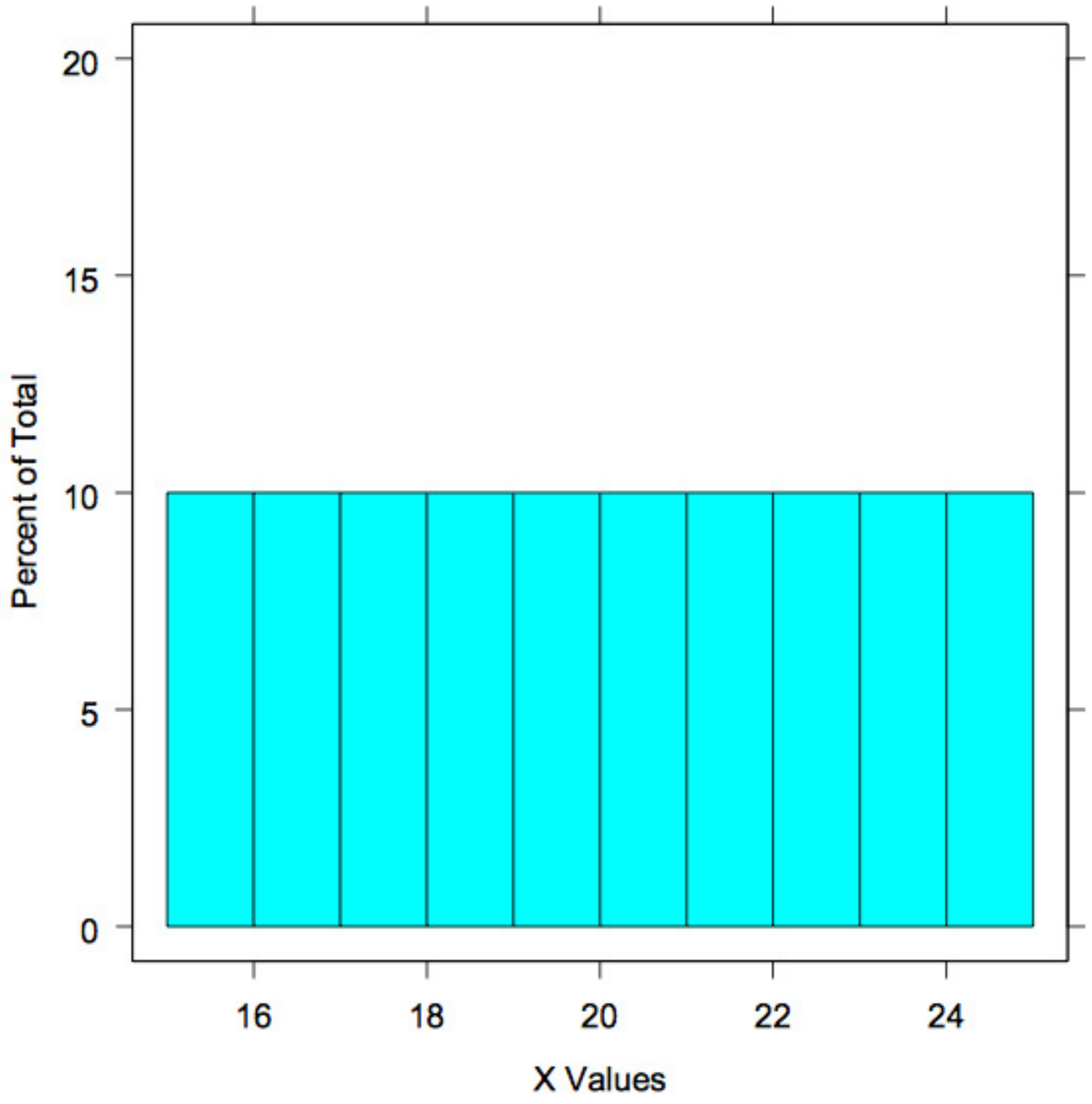
- Quantiles are the order statistics for a data distribution
  - A quantile plot is a two-dimensional display - a scatterplot of data values versus their position within the ordered distribution



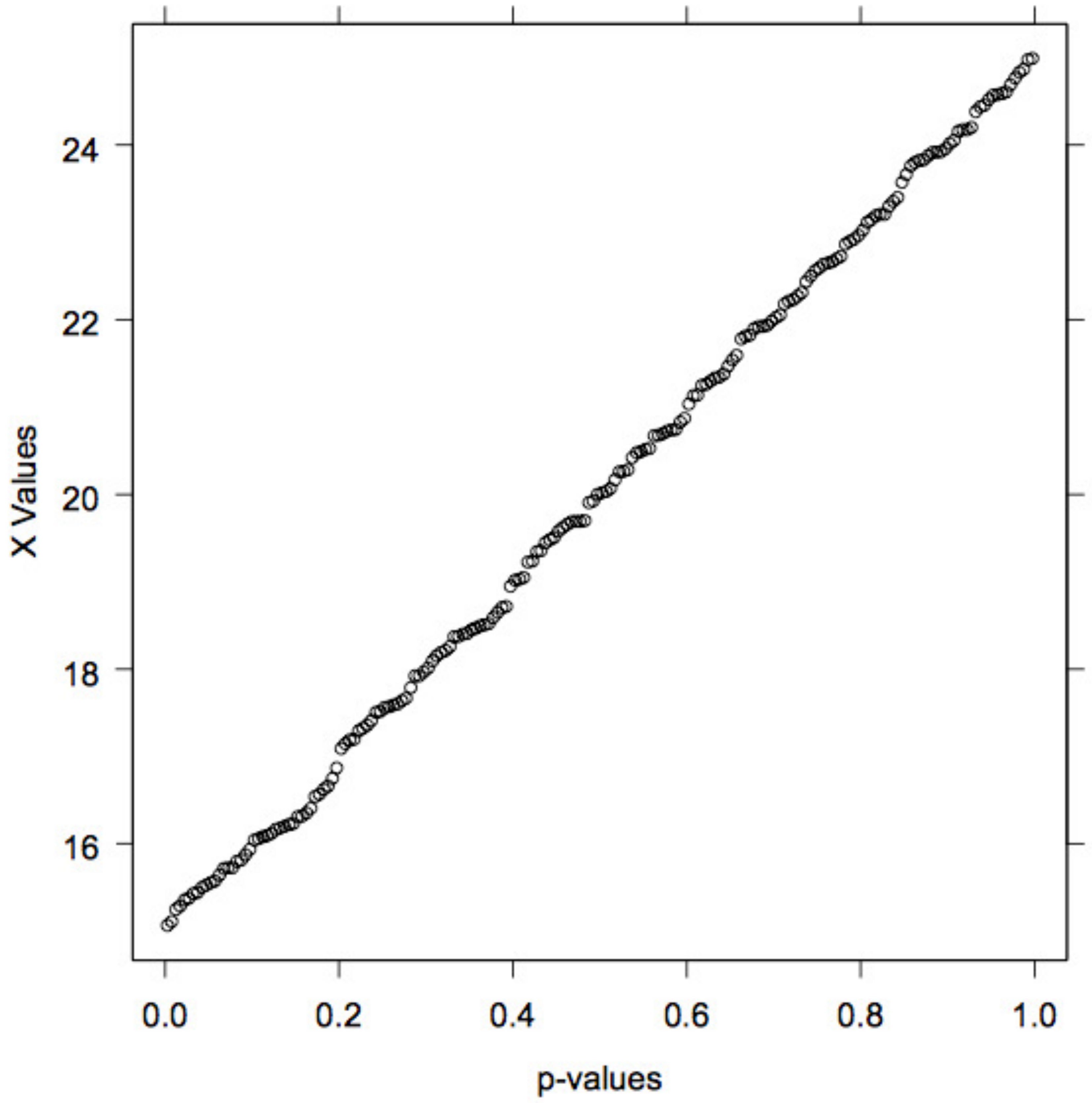
Medicaid Program Quality Score
140
160
180
200
220
240
260
0.0
0.2
0.4
0.6
0.8
1.0
p-value

Quantile Plots

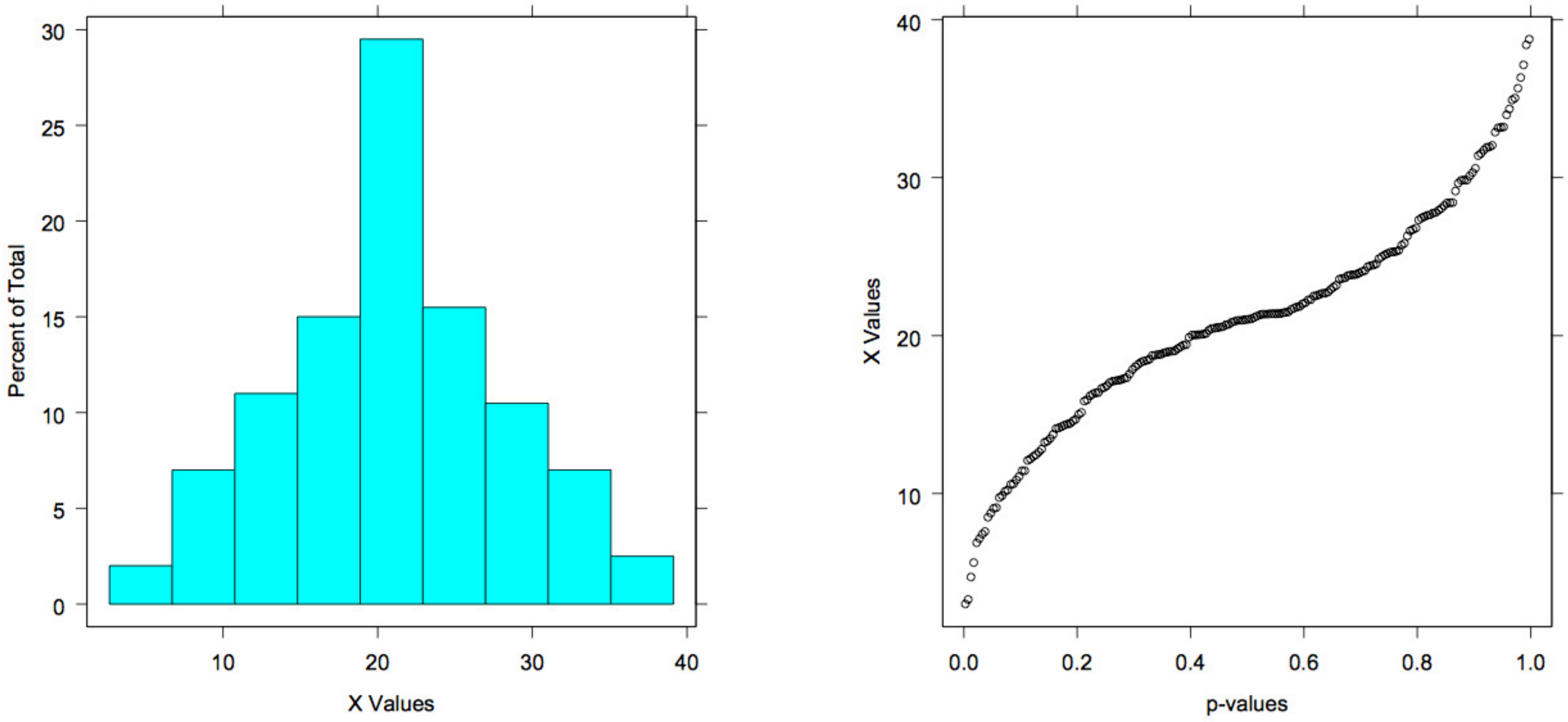
- Quantiles are the order statistics for a data distribution
  - A quantile plot is a two-dimensional display - a scatterplot of data values versus their position within the ordered distribution
- Quantile plots can solve many of the mathematical issues we have with histograms.
  - However, quantile plots are kind of tricky to interpret; we're not used to them. Compare to the following histograms:

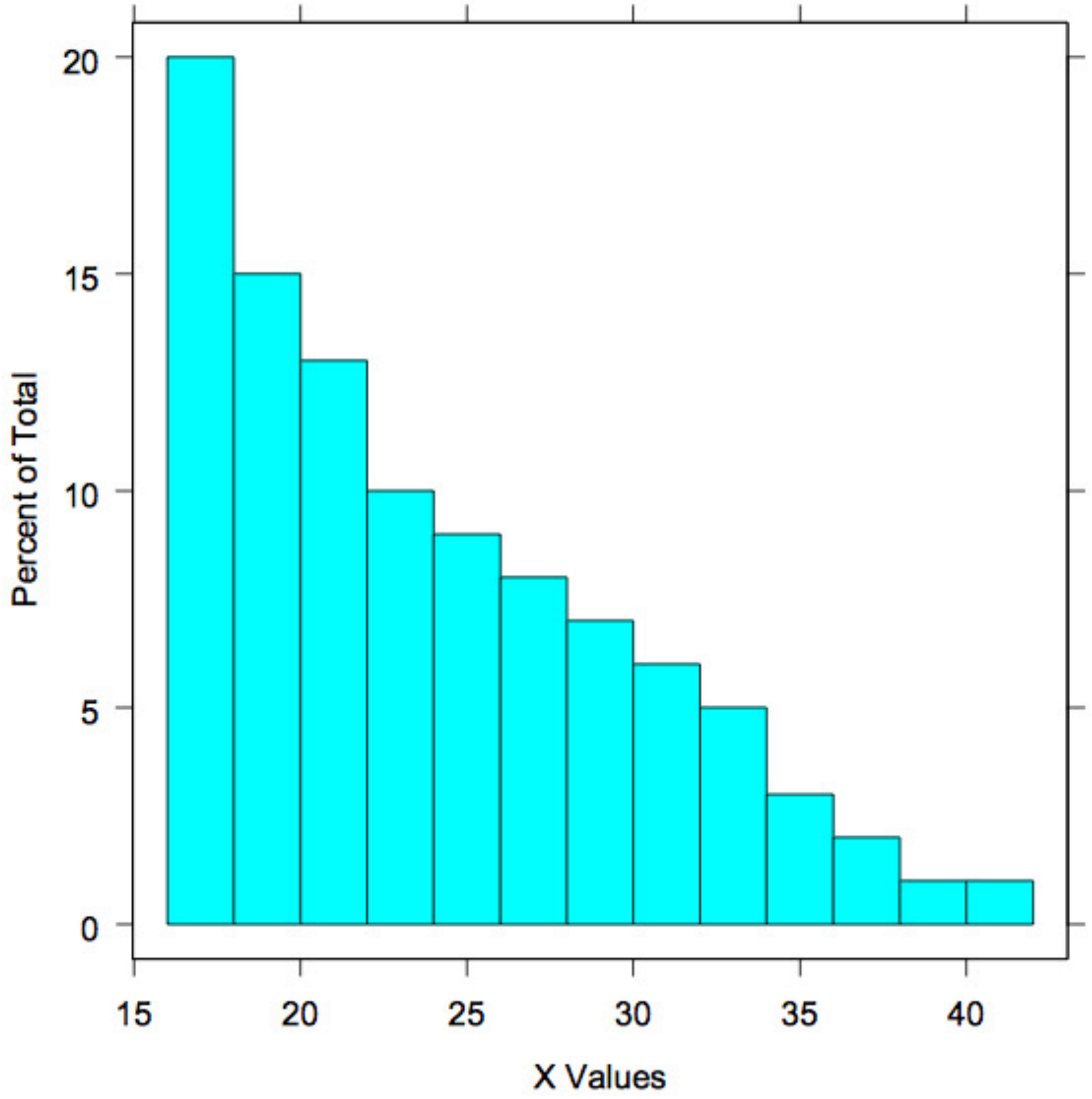


20
15
10
5
0
Percent of Total
16
18
20
22
24
X Values

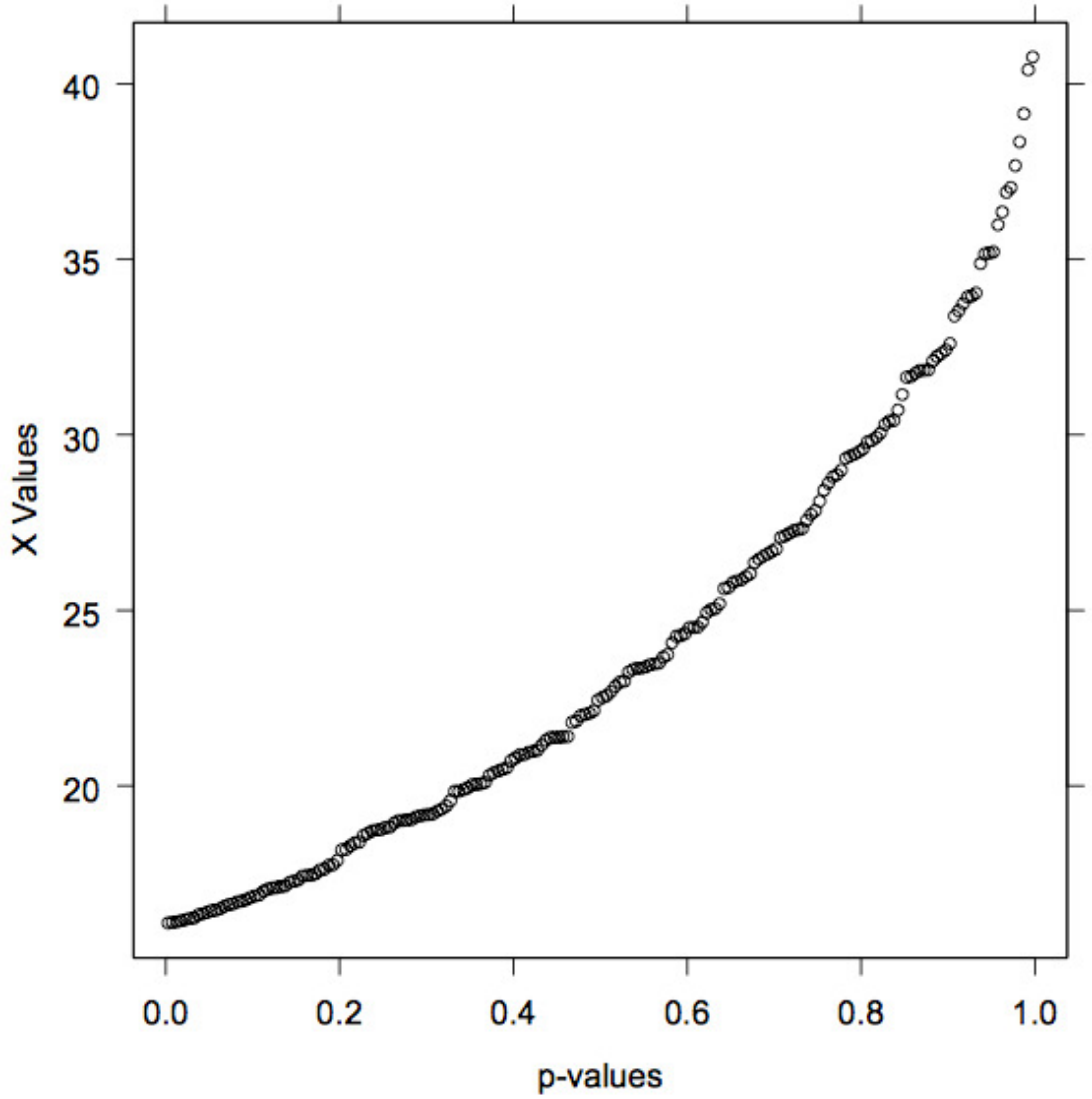


24
22
20
18
16
X Values
0.0
0.2
0.4
0.6
0.8
1.0
p-values

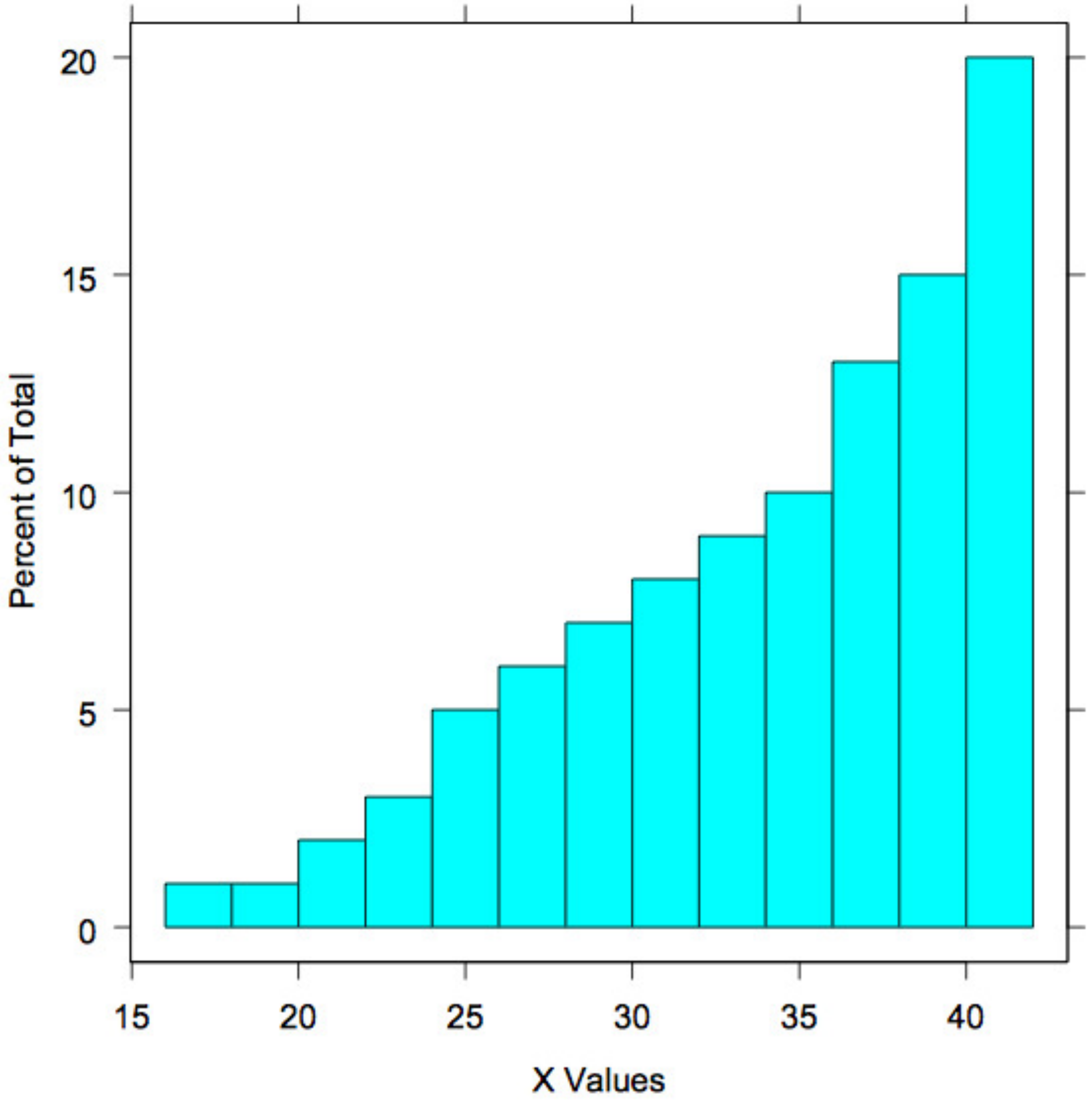

Percent of Total
0
5
10
15
20
25
30
10
20
30
40
X Values
X Values
10
20
30
40
0.0
0.2
0.4
0.6
0.8
1.0
p-values

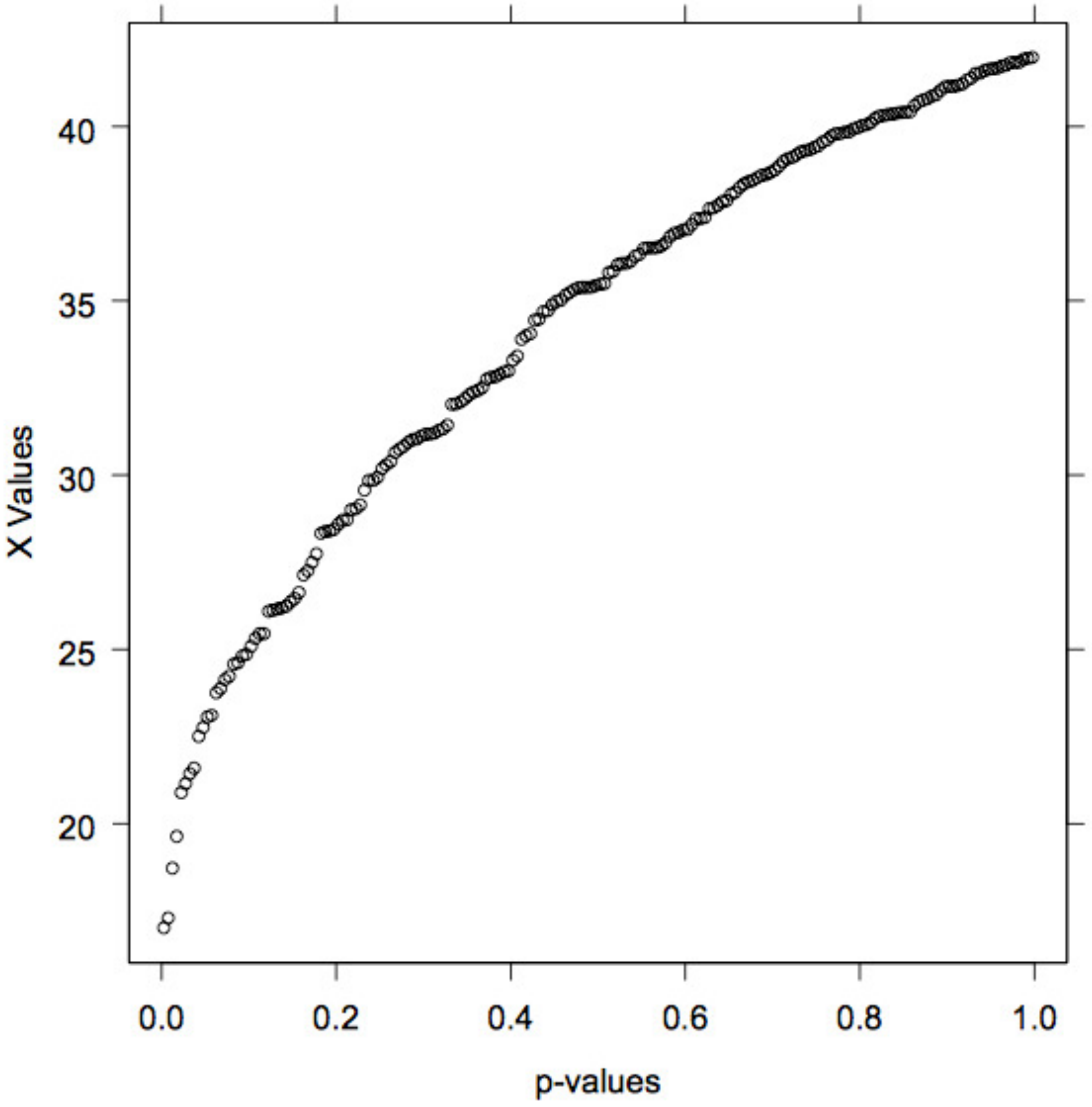


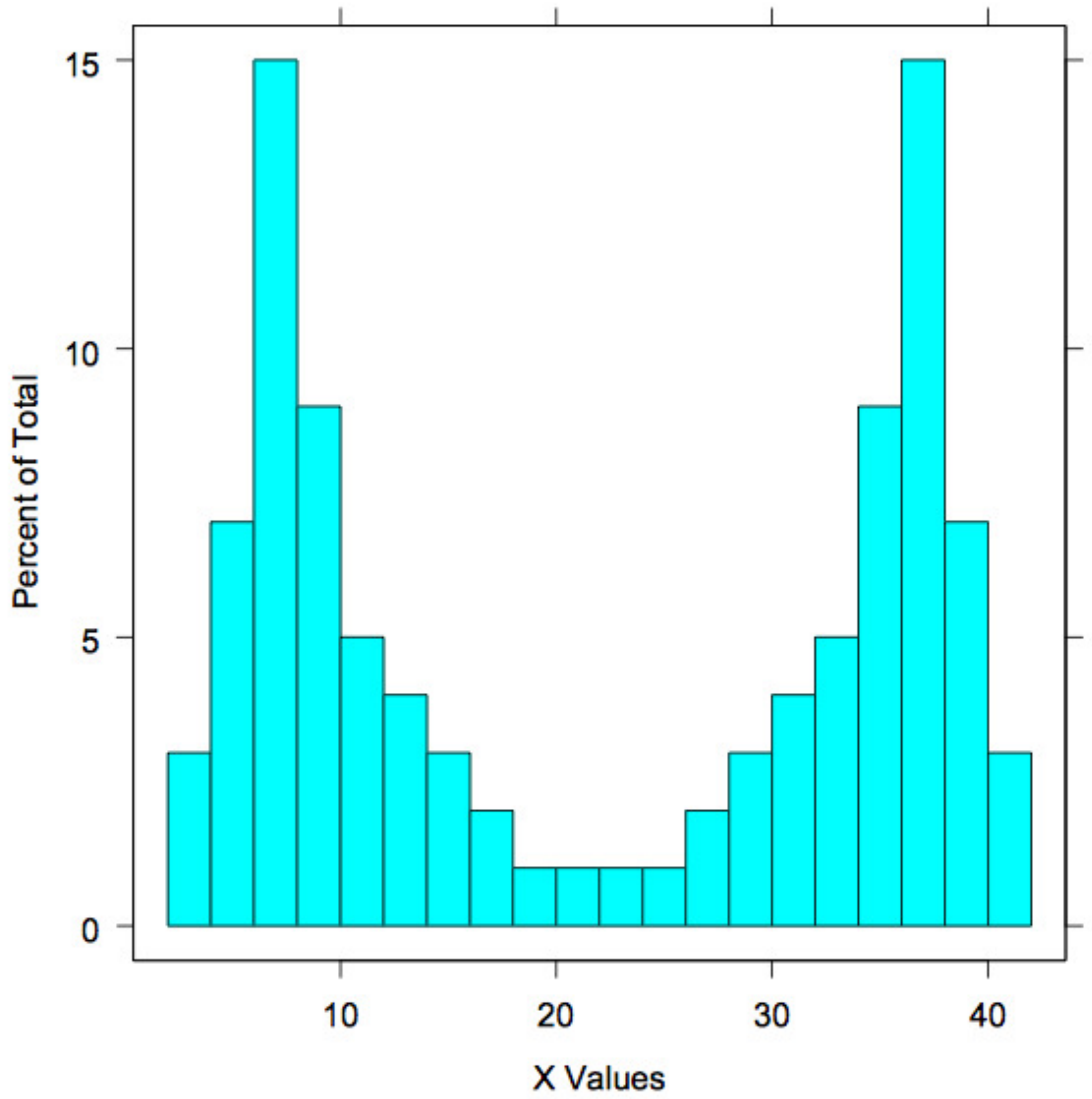
Percent of Total
20
15
10
5
0
15
20
25
30
35
40
X Values



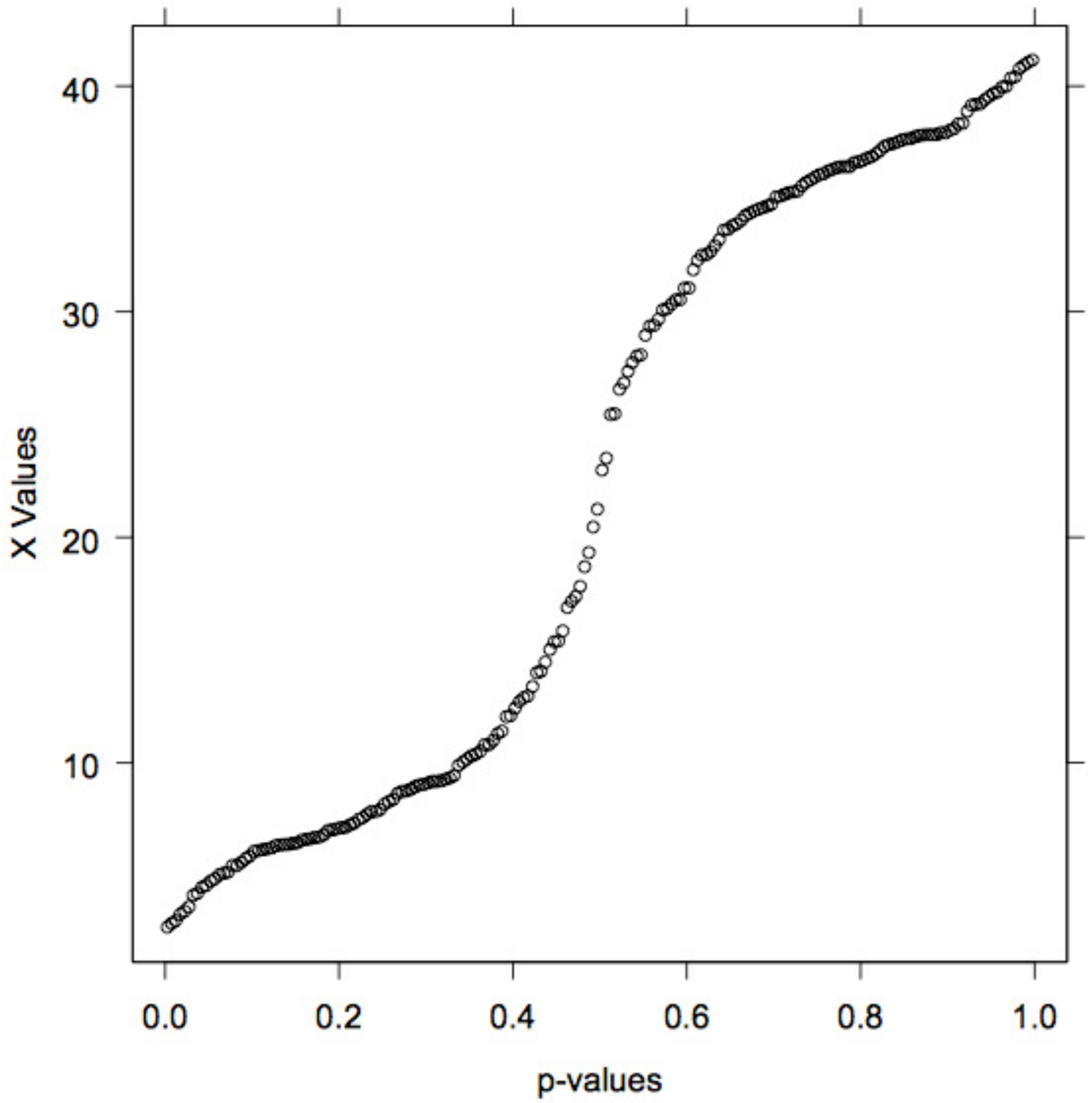
X Values
40
35
30
25
20
0.0
0.2
0.4
0.6
0.8
1.0
p-values


20
15
10
5
0
Percent of Total
15
20
25
30
35
40
X Values


40
35
30
25
20
X Values
0.0
0.2
0.4
0.6
0.8
1.0
p-values



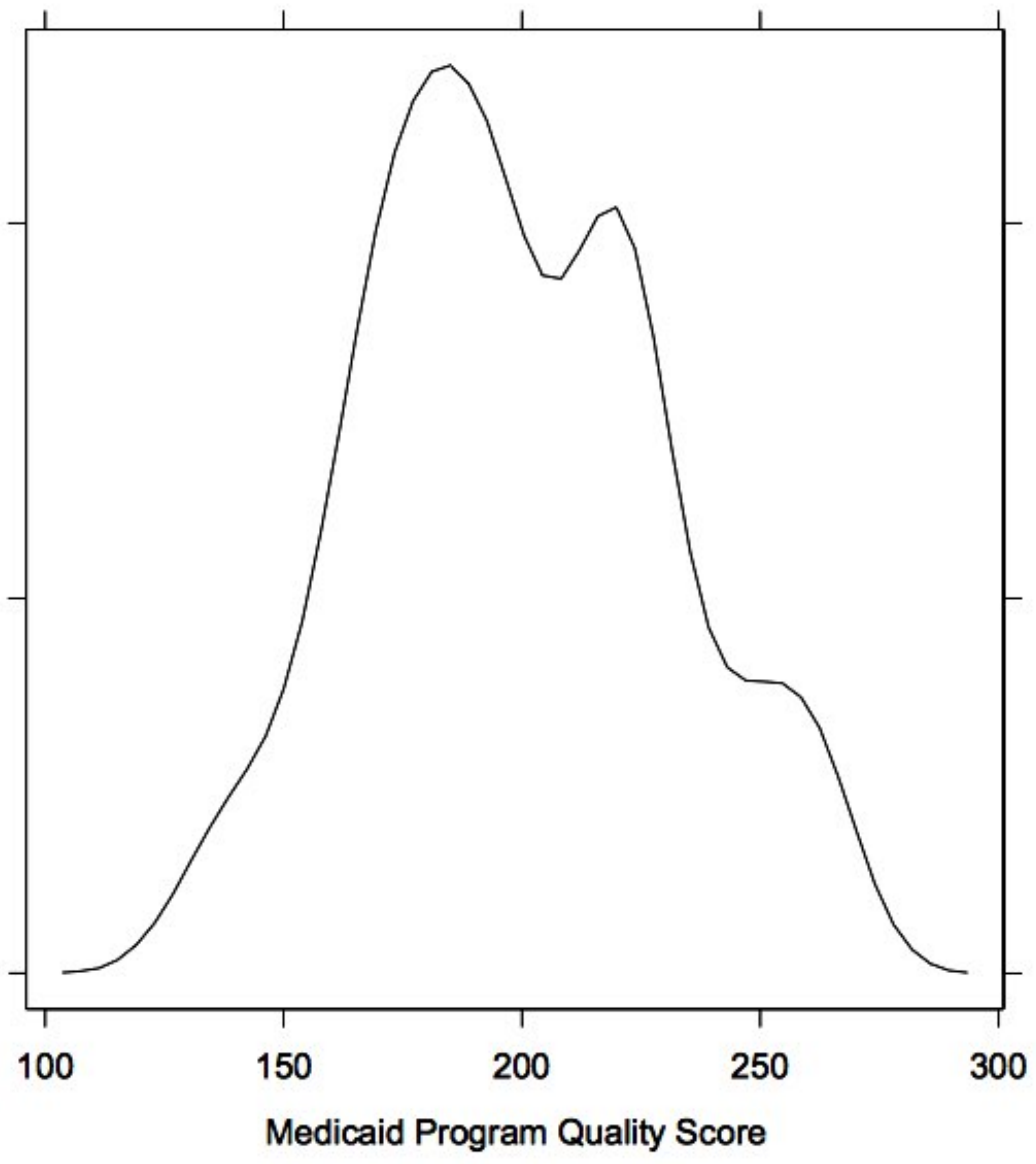
Percent of Total
15
10
5
0
10
20
30
40
X Values



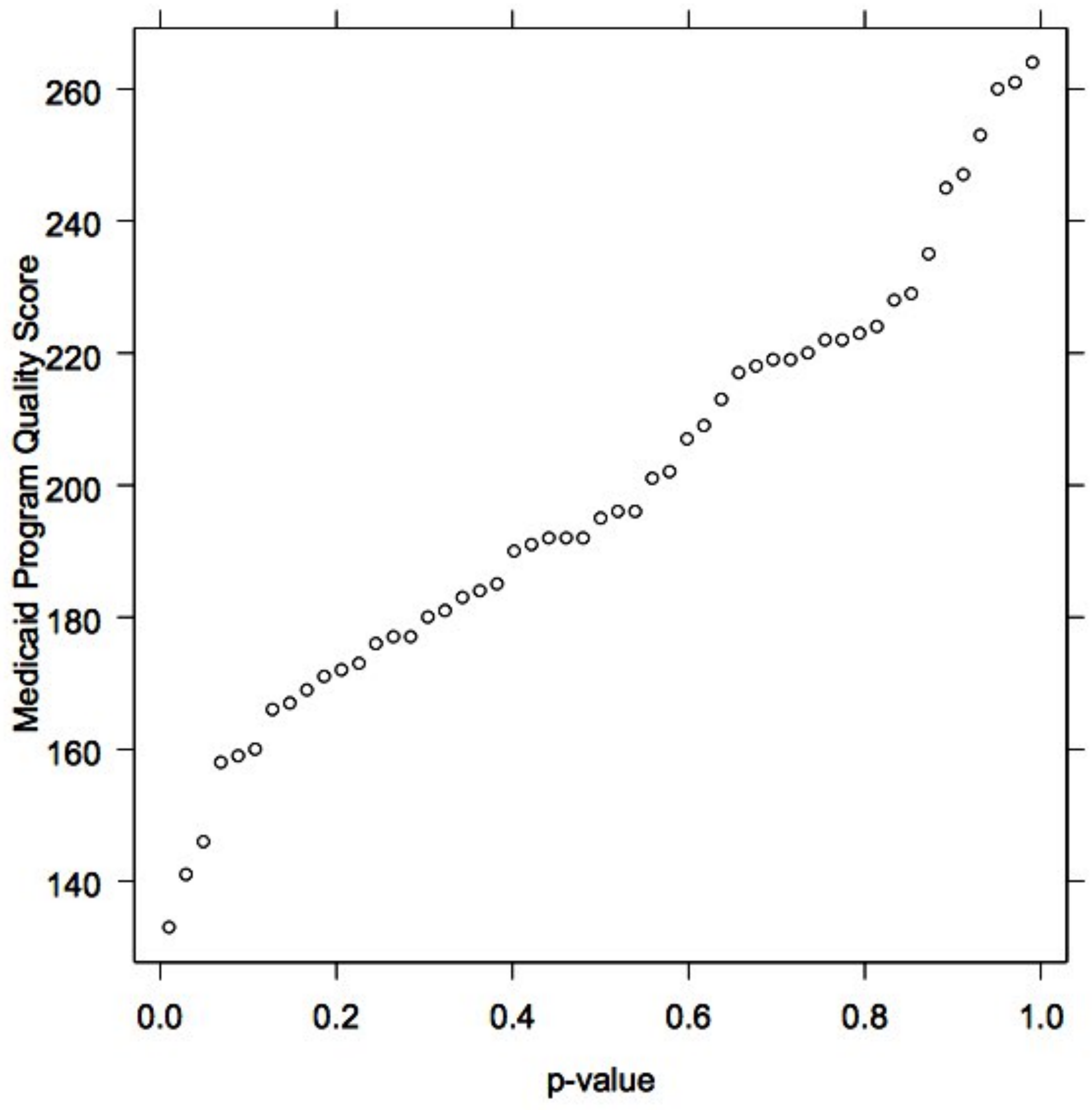
X Values
40
30
20
10
0.0
0.2
0.4
0.6
0.8
1.0
p-values

Quantile Plots

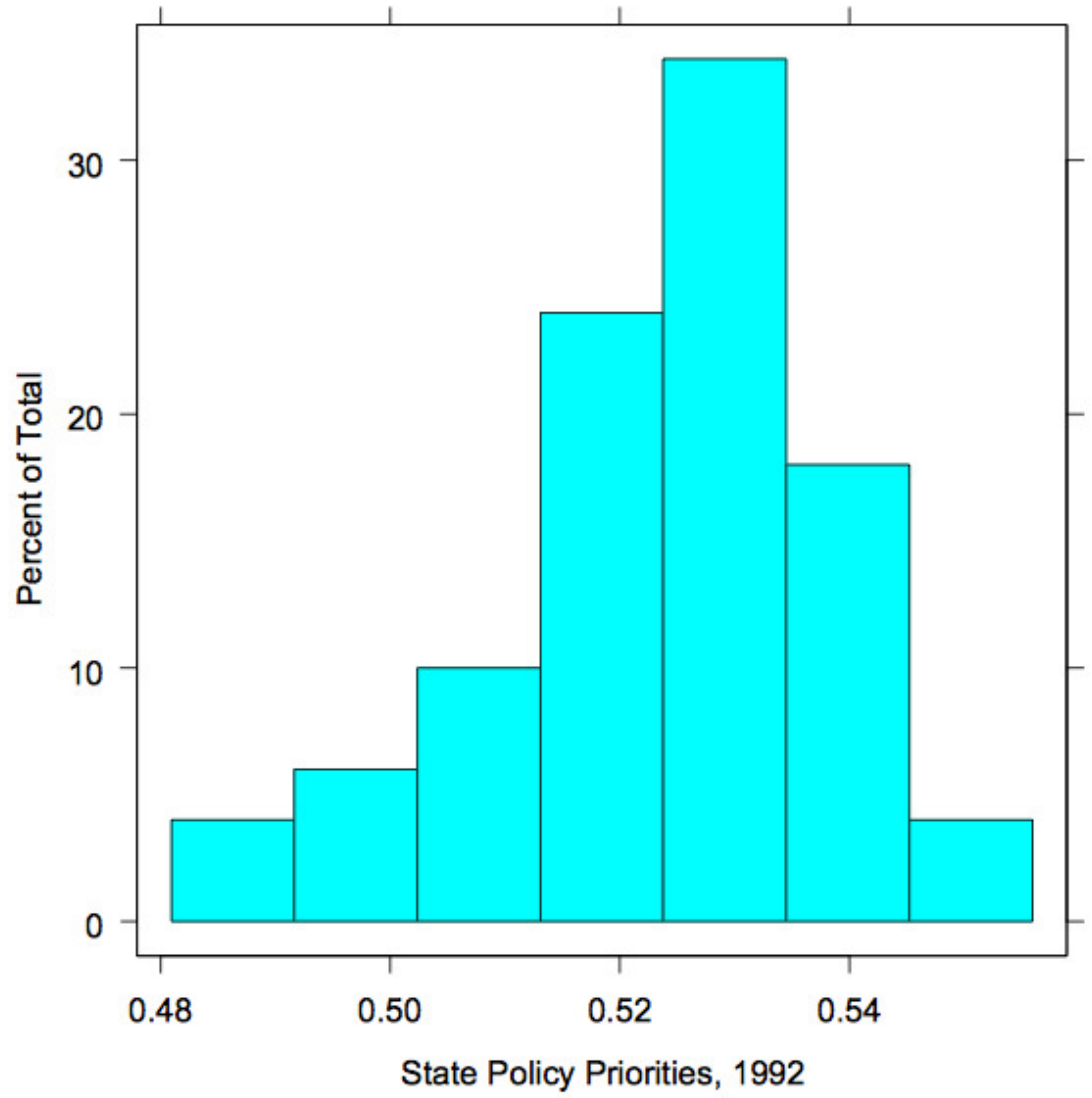
- Quantiles are the order statistics for a data distribution
    - A quantile plot is a two-dimensional display - a scatterplot of data values versus their position within the ordered distribution
- We're so used to using *area* or *height* to interpret numbers, that interpretation of quantile plots can be downright baffling and can hide variation other plots make obvious.
    - One can lose the forest through the trees.
    - (This is the purist's version of univariate statistics)



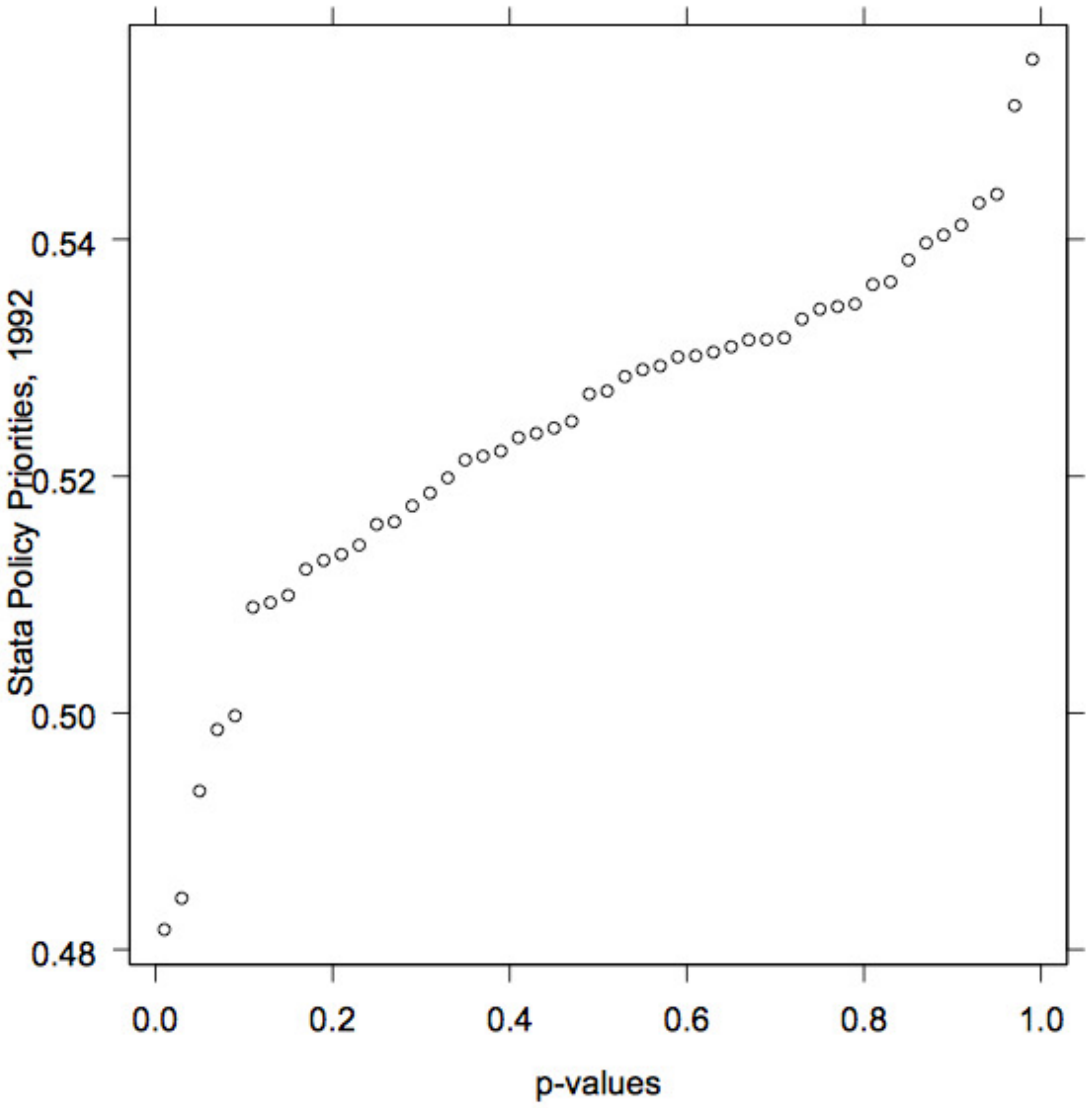
100
150
200
250
300
Medicaid Program Quality Score



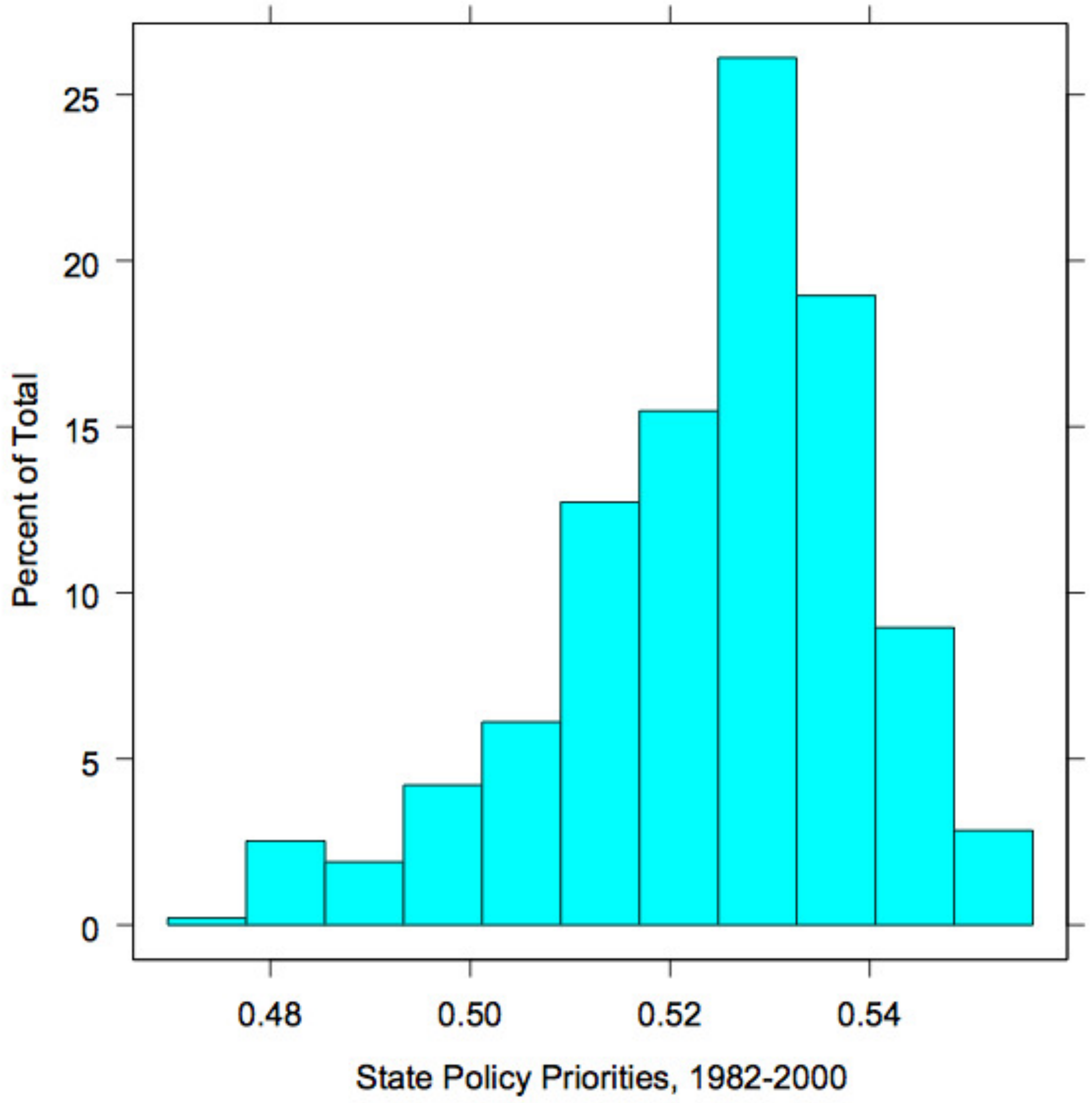
260
240
220
200
180
160
140
Medicaid Program Quality Score
0.0
0.2
0.4
0.6
0.8
1.0
p-value



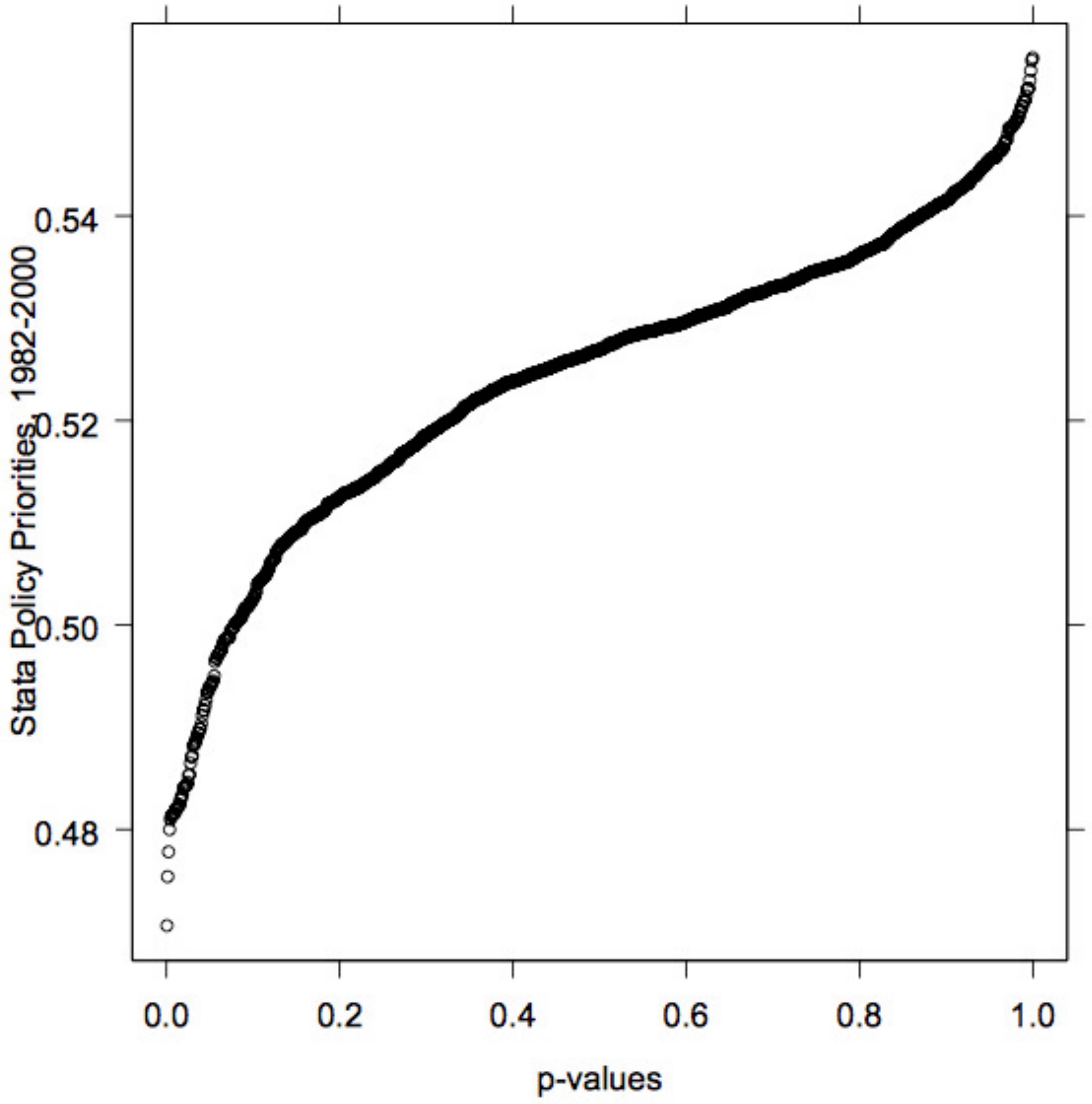
30
20
10
0
Percent of Total
0.48
0.50
0.52
0.54
State Policy Priorities, 1992



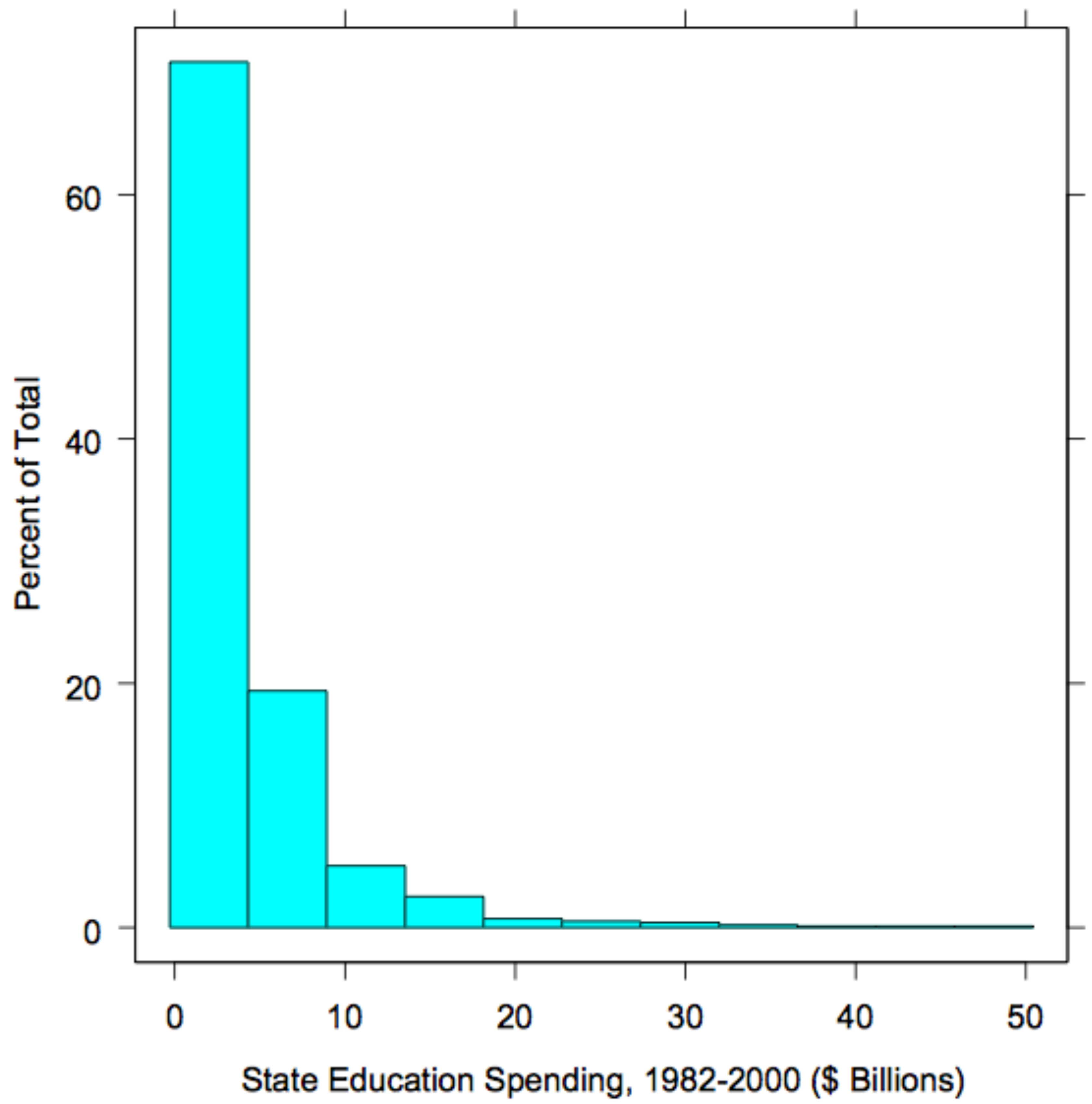
0.54
0.52
0.50
0.48
Stata Policy Priorities, 1992
0.0
0.2
0.4
0.6
0.8
1.0
p-values



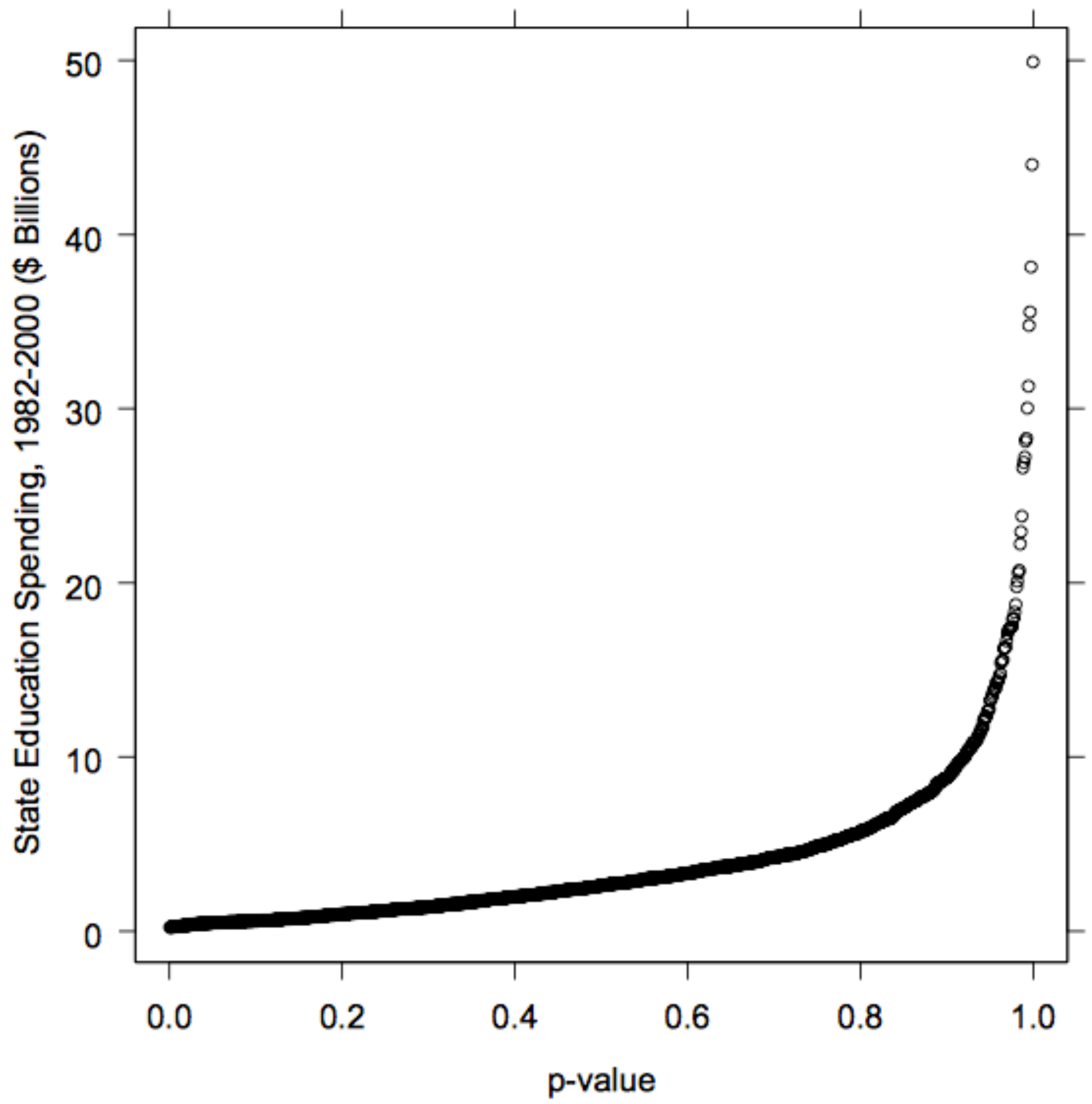
Percent of Total
25
20
15
10
5
0
0.48
0.50
0.52
0.54
State Policy Priorities, 1982-2000



Stata Policy Priorities, 1982-2000
0.54
0.52
0.50
0.48
0.0
0.2
0.4
0.6
0.8
1.0
p-values



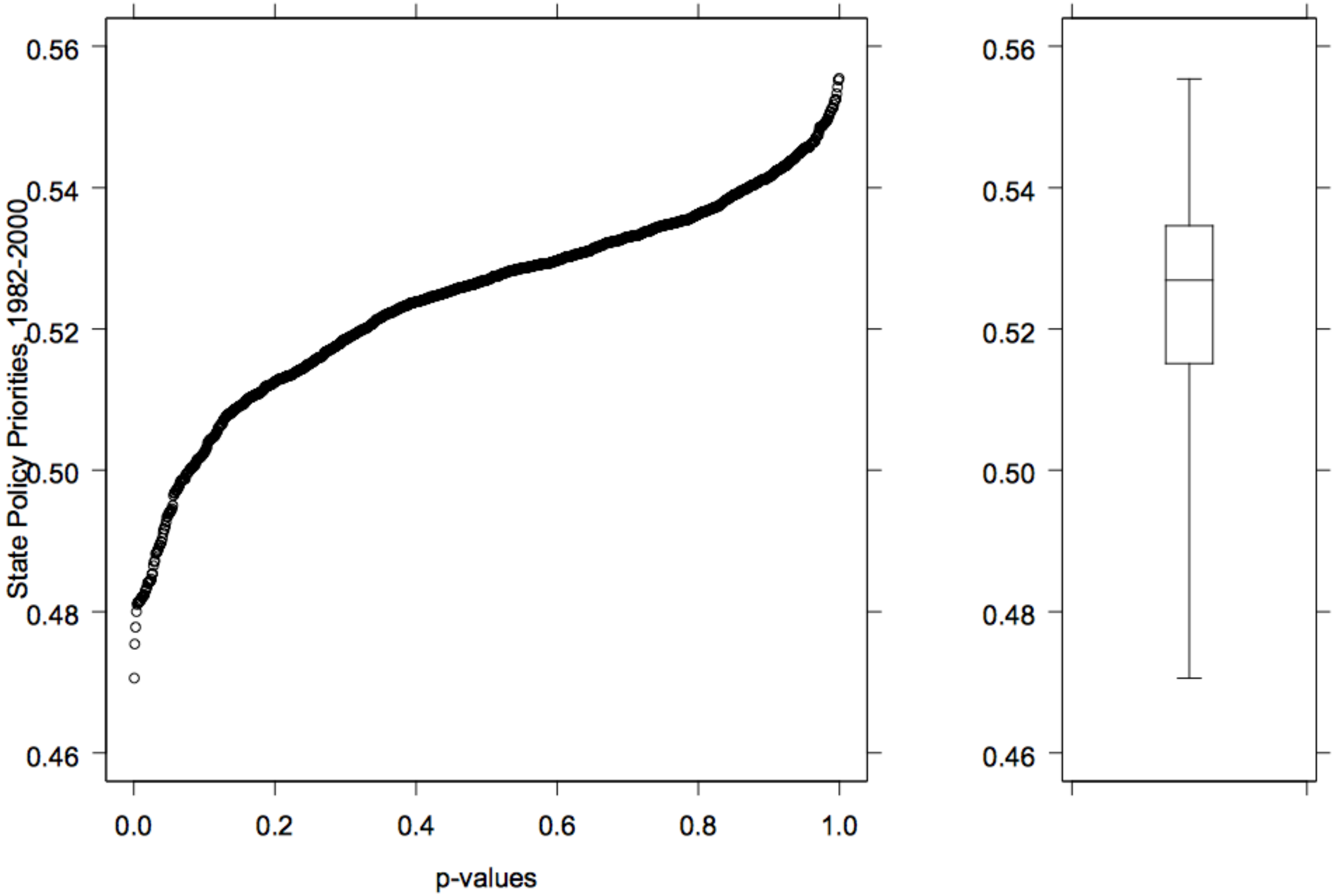
Percent of Total
0
20
40
60
0
10
20
30
40
50
State Education Spending, 1982-2000 ($ Billions)



State Education Spending, 1982-2000 ($ Billions)
0
10
20
30
40
50
0.0
0.2
0.4
0.6
0.8
1.0
p-value

Box Plots (Box and Whisker Plots)

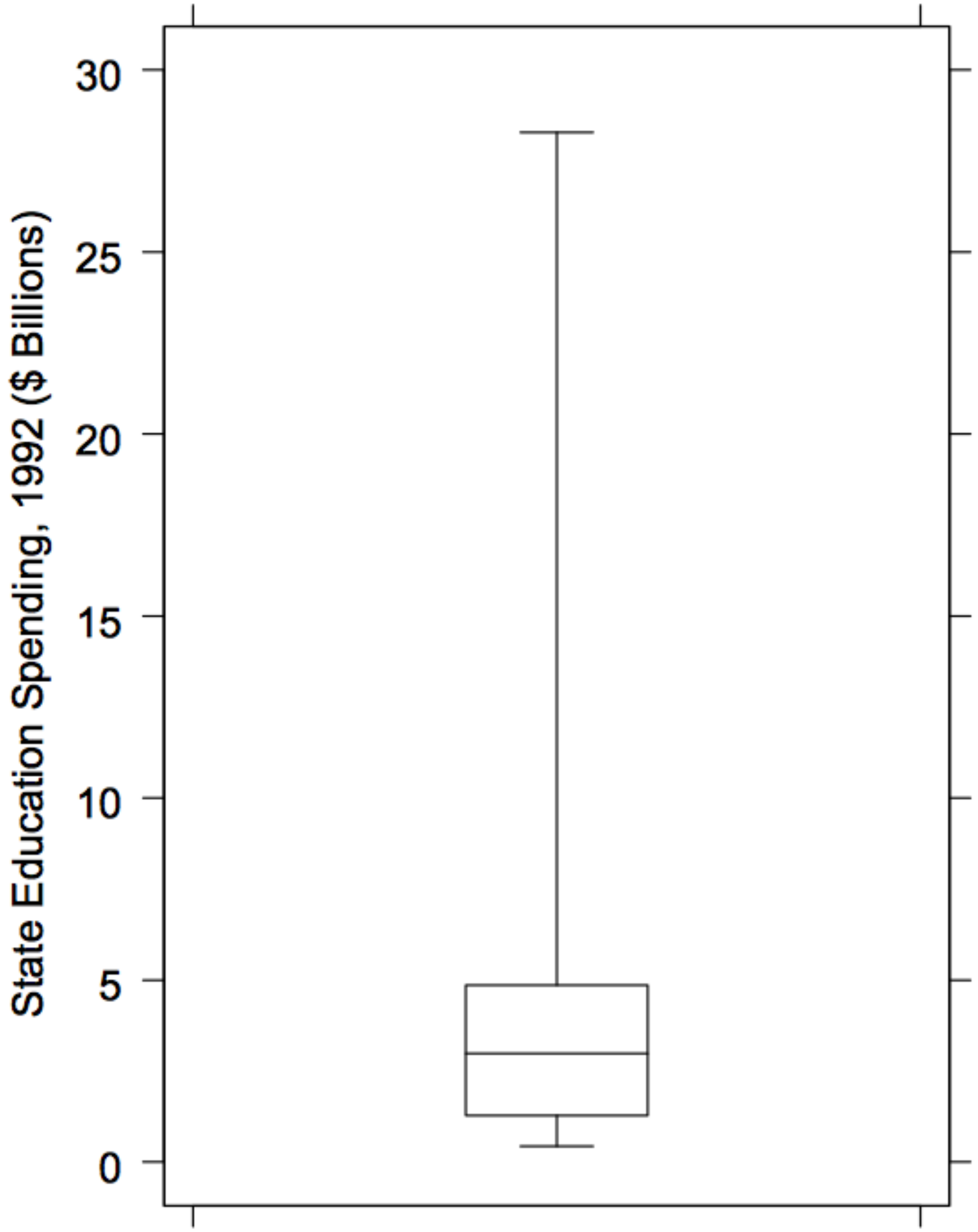
- Focuses on important quartiles.
    - Shows a box around 25%-75%, lines to 0% and 100%



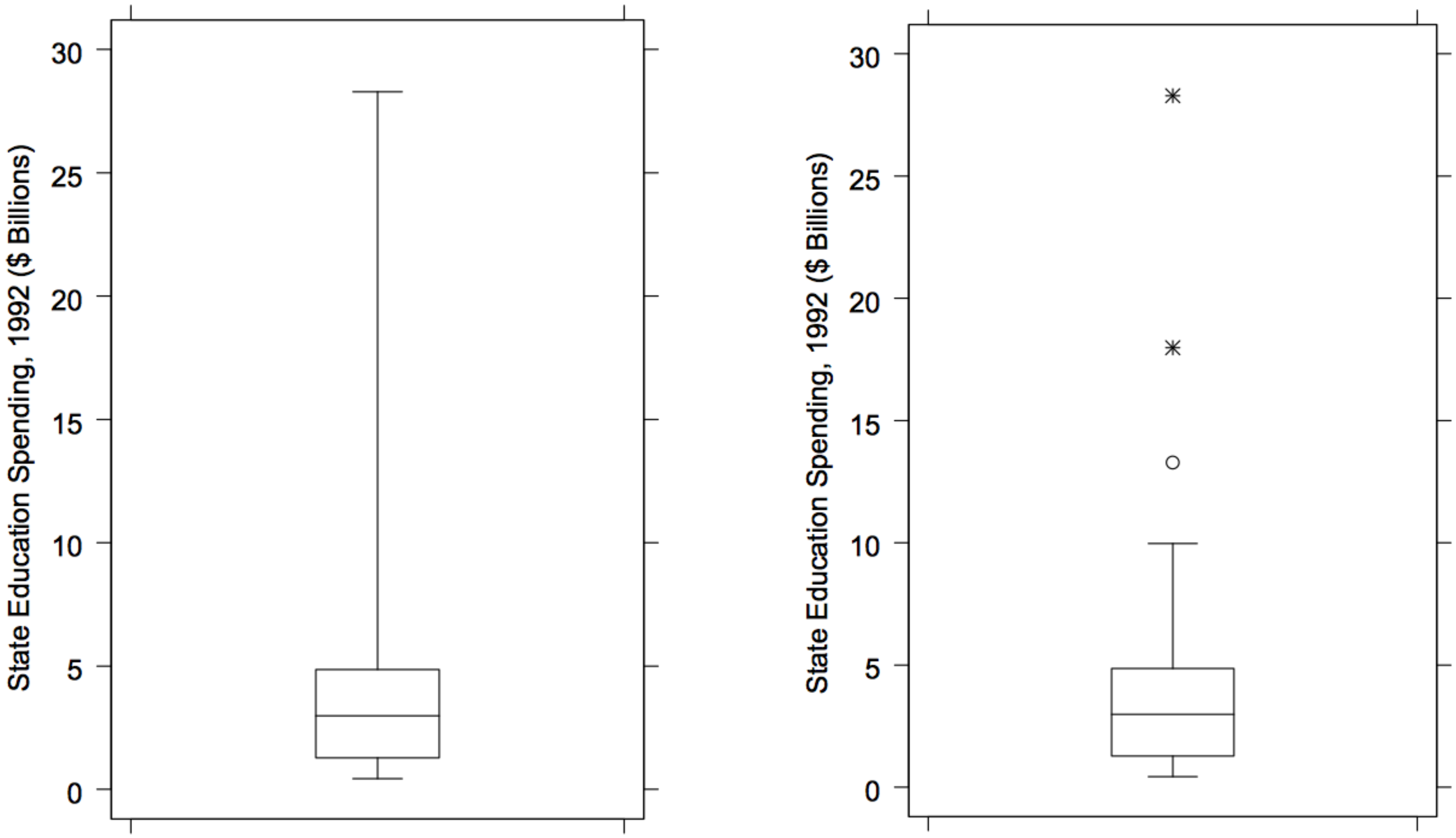
State Policy Priorities, 1982-2000
0.56
0.54
0.52
0.50
0.48
0.46
0.0
0.2
0.4
0.6
0.8
1.0
p-values

Box Plots (Box and Whisker Plots)

- Focuses on important quartiles.
  - Shows a box around 25%-75%, lines to 0% and 100%
- Whew. That's a relief.
  - Box plots, however, have an outlier problem.

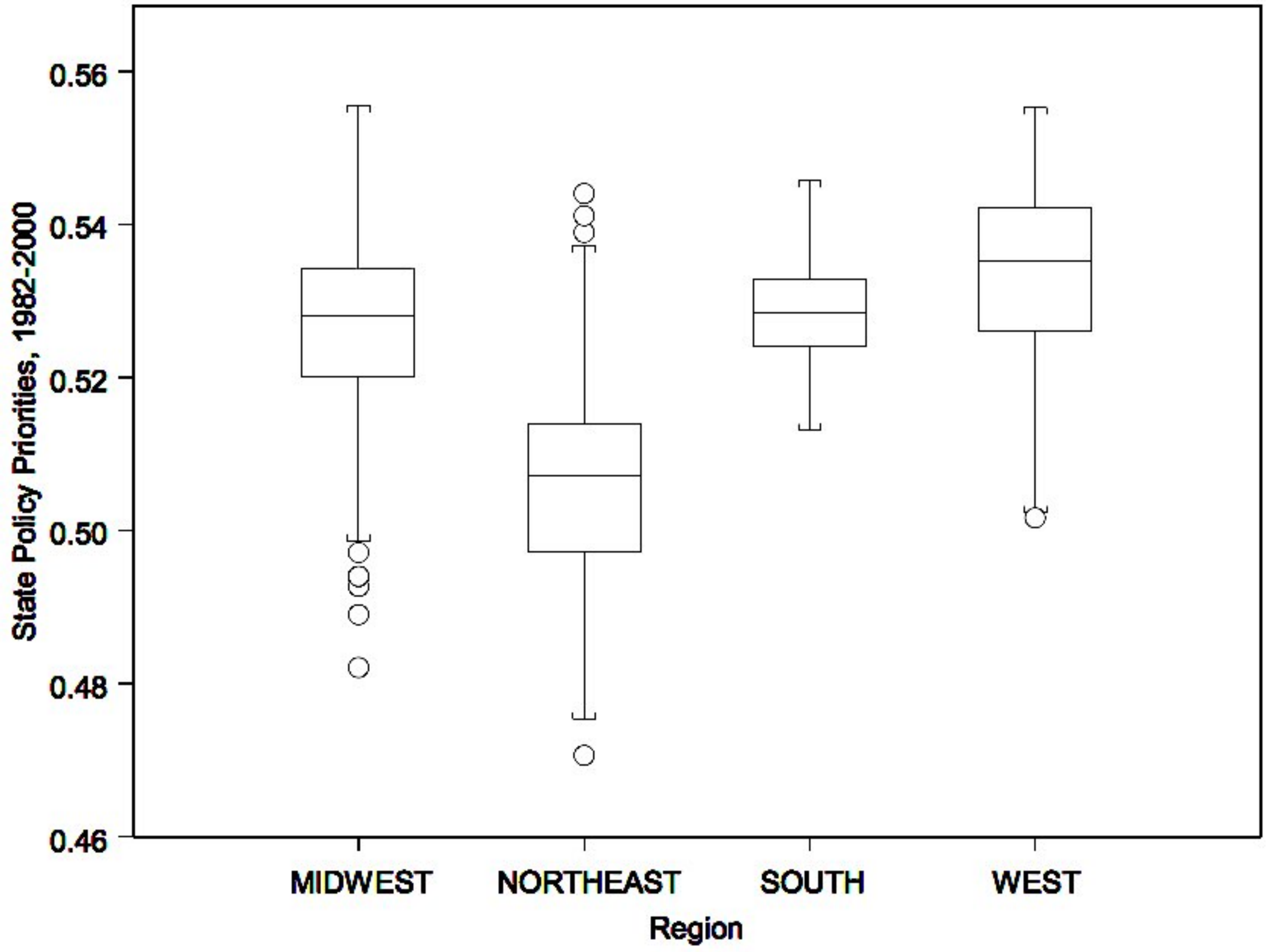

State Education Spending, 1992 ($ Billions)
30
25
20
15
10
5
0

?


State Education Spending, 1992 ($ Billions)
0
5
10
15
20
25
30
State Education Spending, 1992 ($ Billions)
0
5
10
15
20
25
30

Box Plots (Box and Whisker Plots)

- Focuses on important quartiles.
  - Shows a box around 25%-75%, lines to 0% and 100%
- Box plots, however, have an outlier problem.
  - They also lose individual data points and the “shape” of data.
- However, they lead to some very nice graphics, especially comparing different batches of data.

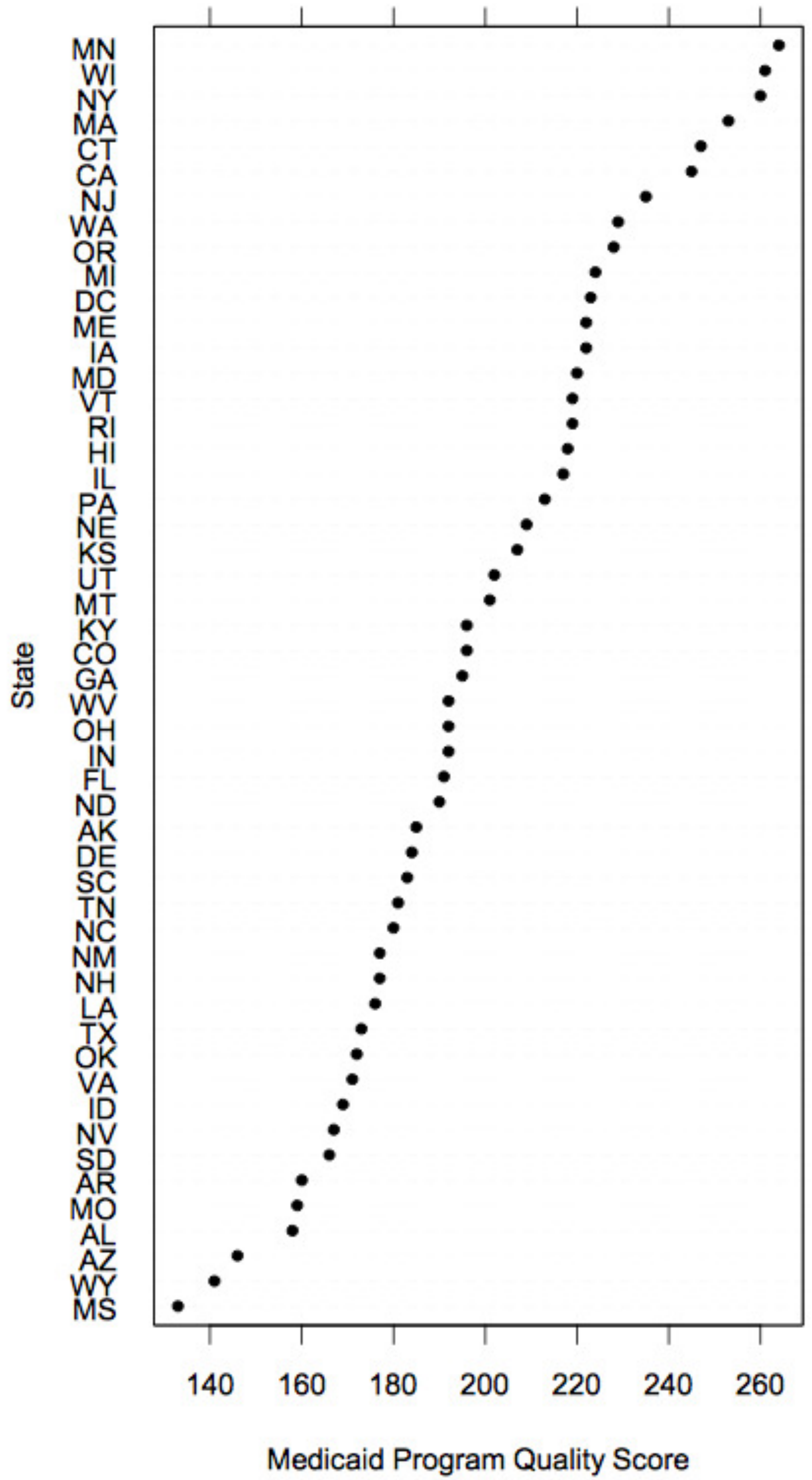

0.56
0.54
0.52
0.50
0.48
0.46
State Policy Priorities, 1982-2000
MIDWEST
NORTHEAST
SOUTH
WEST
Region

Box Plots (Box and Whisker Plots)

- Focuses on important quartiles.
  - Shows a box around 25%-75%, lines to 0% and 100%
- Box plots, however, have an outlier problem.
  - They also lose individual data points and the “shape” of data.
- Lead to some very nice graphics, especially comparing different batches of data.
  - But what do we do if we have lots of batches of data?

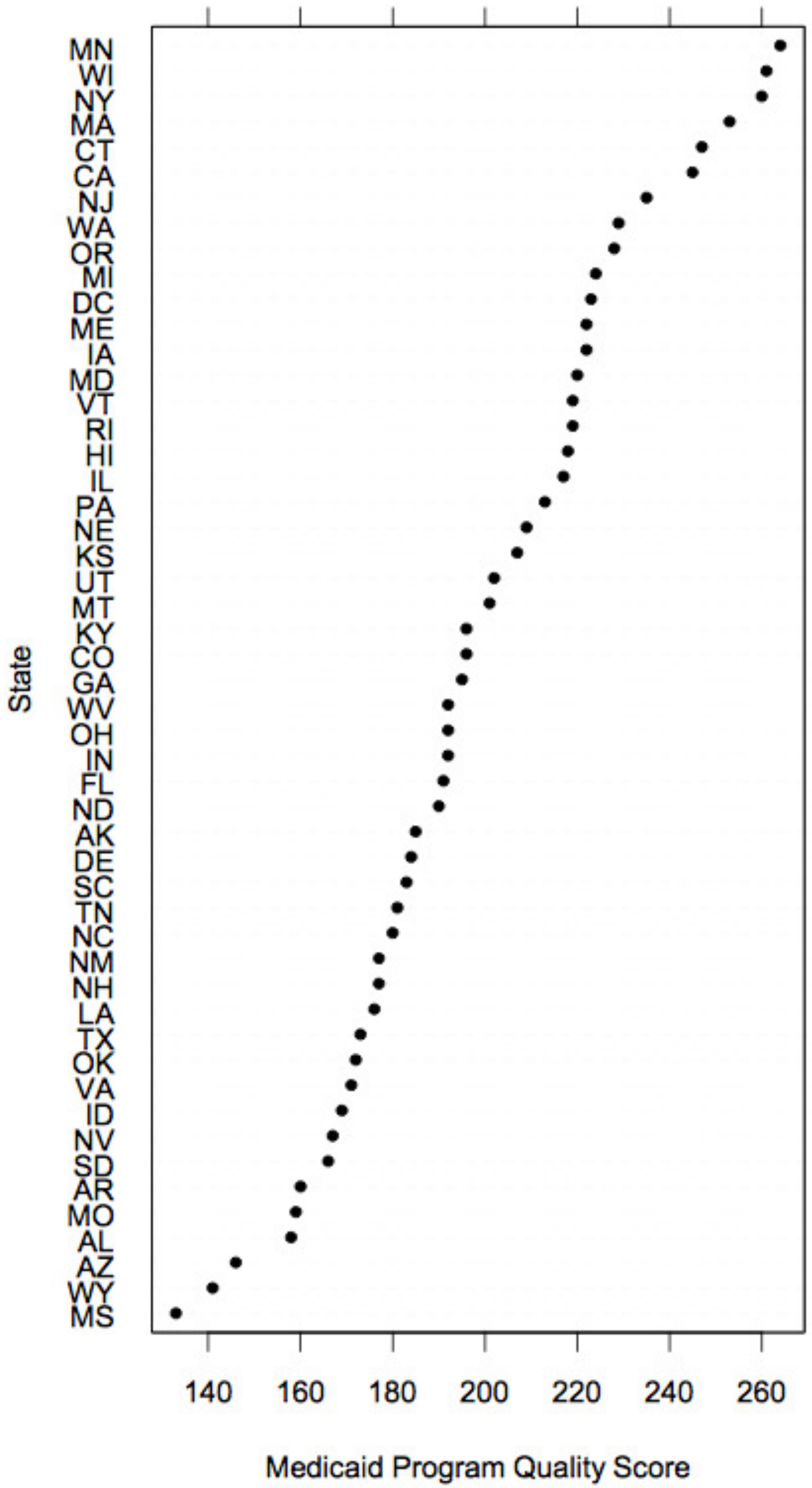
Dot Plots

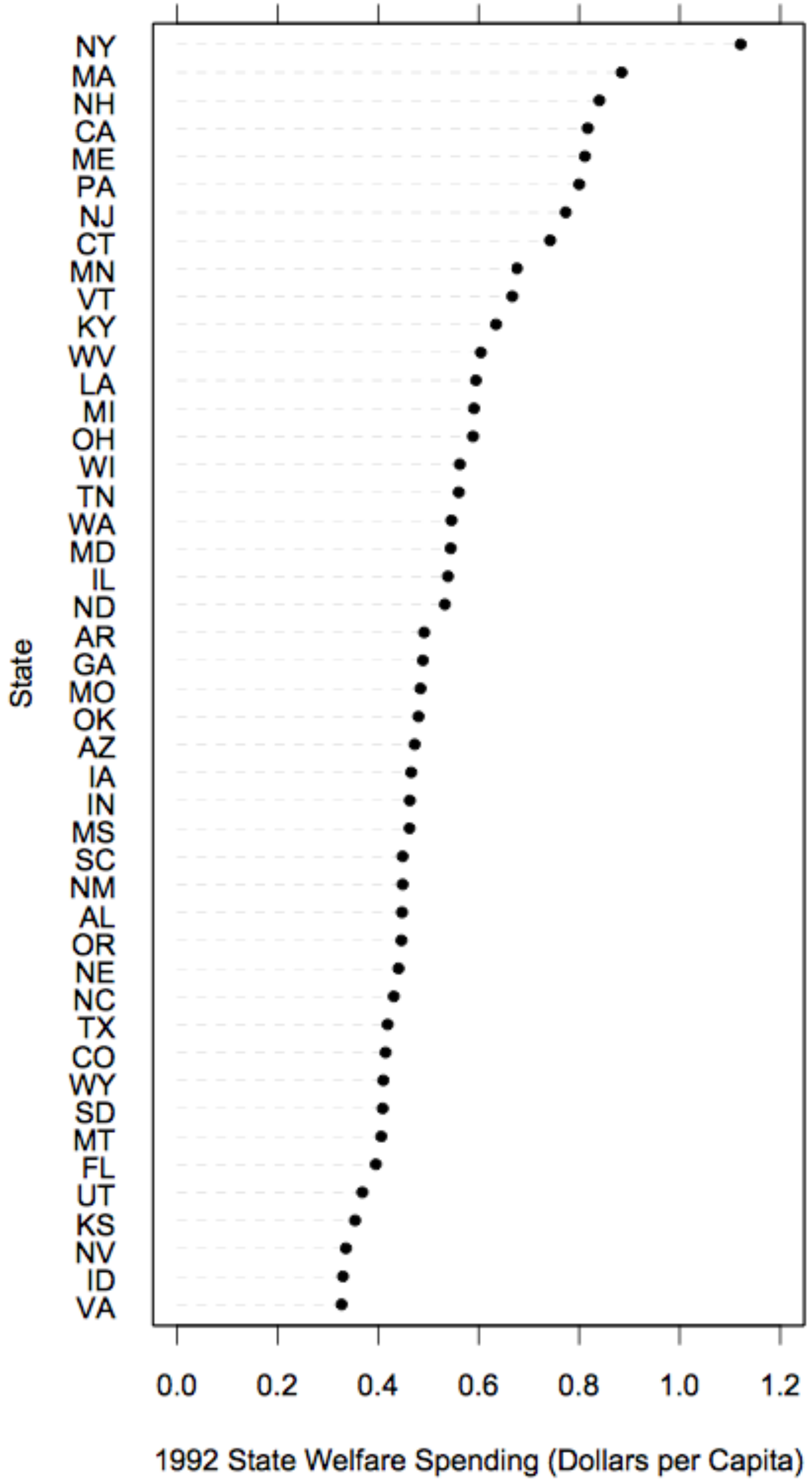
- Two-dimensional plot of data labels versus data values
  - Useful display when data values are associated with identifying information, such as a label or index number.
- Present a lot of information and are easy to interpret.


MN
WI
NY
MA
CT
CA
NJ
WA
OR
MI
DC
ME
IA
MD
VT
RI
HI
IL
PA
NE
KS
UT
MT
KY
CO
GA
WV
OH
IN
FL
ND
AK
DE
SC
TN
NC
NM
NH
LA
TX
OK
VA
ID
NV
SD
AR
MO
AL
AZ
WY
MS
State
140
160
180
200
220
240
260
Medicaid Program Quality Score

Dot Plots

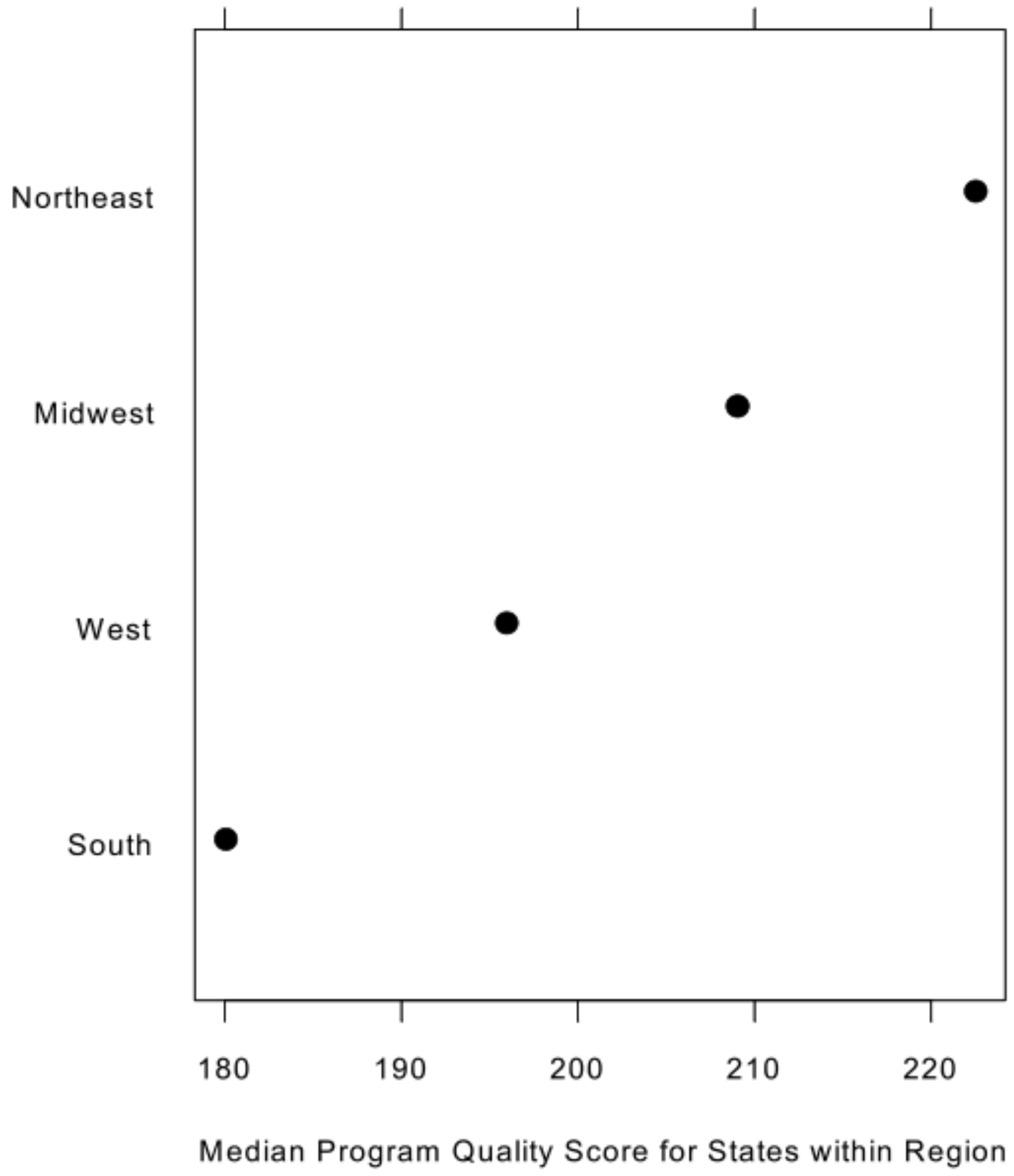
- Two-dimensional plot of data labels versus data values
    - Useful display when data values are associated with identifying information, such as a label or index number.
- Present a lot of information and are easy to interpret.
    - Note the ranking along the Y axis, which makes it very easy to compare.


State
MN
WI
NY
MA
CT
CA
NJ
WA
OR
MI
DC
ME
IA
MD
VT
RI
HI
IL
PA
NE
KS
UT
MT
KY
CO
GA
WV
OH
IN
FL
ND
AK
DE
SC
TN
NC
NM
NH
LA
TX
OK
VA
ID
NV
SD
AR
MO
AL
AZ
WY
MS
140
160
180
200
220
240
260
Medicaid Program Quality Score


State
NY
MA
NH
CA
ME
PA
NJ
CT
MN
VT
KY
WV
LA
MI
OH
WI
TN
WA
MD
IL
ND
AR
GA
MO
OK
AZ
IA
IN
MS
SC
NM
AL
OR
NE
NC
TX
CO
WY
SD
MT
FL
UT
KS
NV
ID
VA
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1992 State Welfare Spending (Dollars per Capita)

Dot Plots

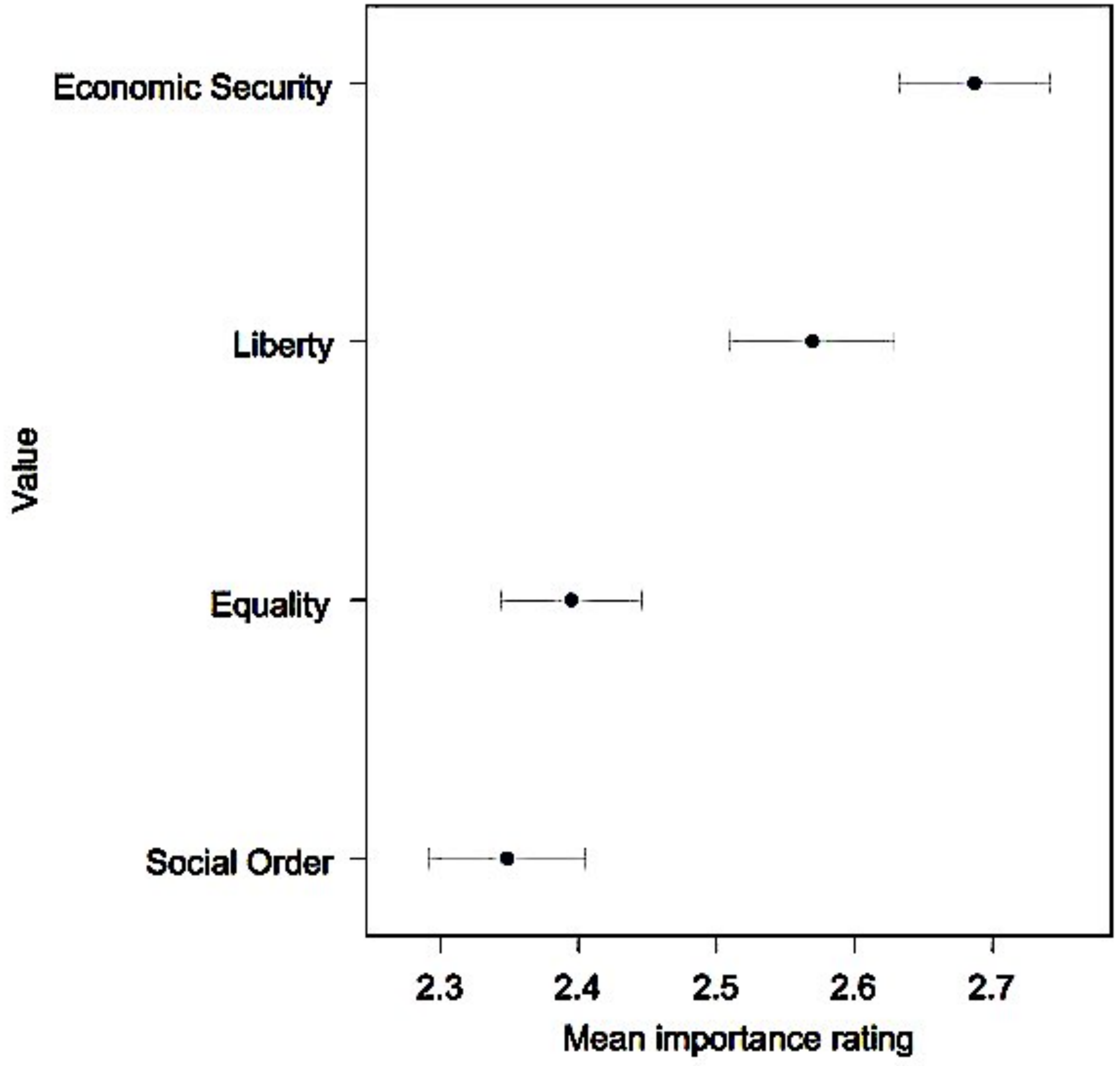
- Two-dimensional plot of data labels versus data values
  - Useful display when data values are associated with identifying information, such as a label or index number.
- Present a lot of information and are easy to interpret.
  - Can batch dot plots in useful ways to show data differently.


Northeast
Midwest
West
South
180
190
200
210
220
Median Program Quality Score for States within Region

Dot Plots

- Two-dimensional plot of data labels versus data values
    - Useful display when data values are associated with identifying information, such as a label or index number.
- Present a lot of information and are easy to interpret.
    - Can batch dot plots in useful ways to show data differently.
- Can also use dot plots in a variety of different ways, with different visual tweaks.

Value importance ratings in the American Public

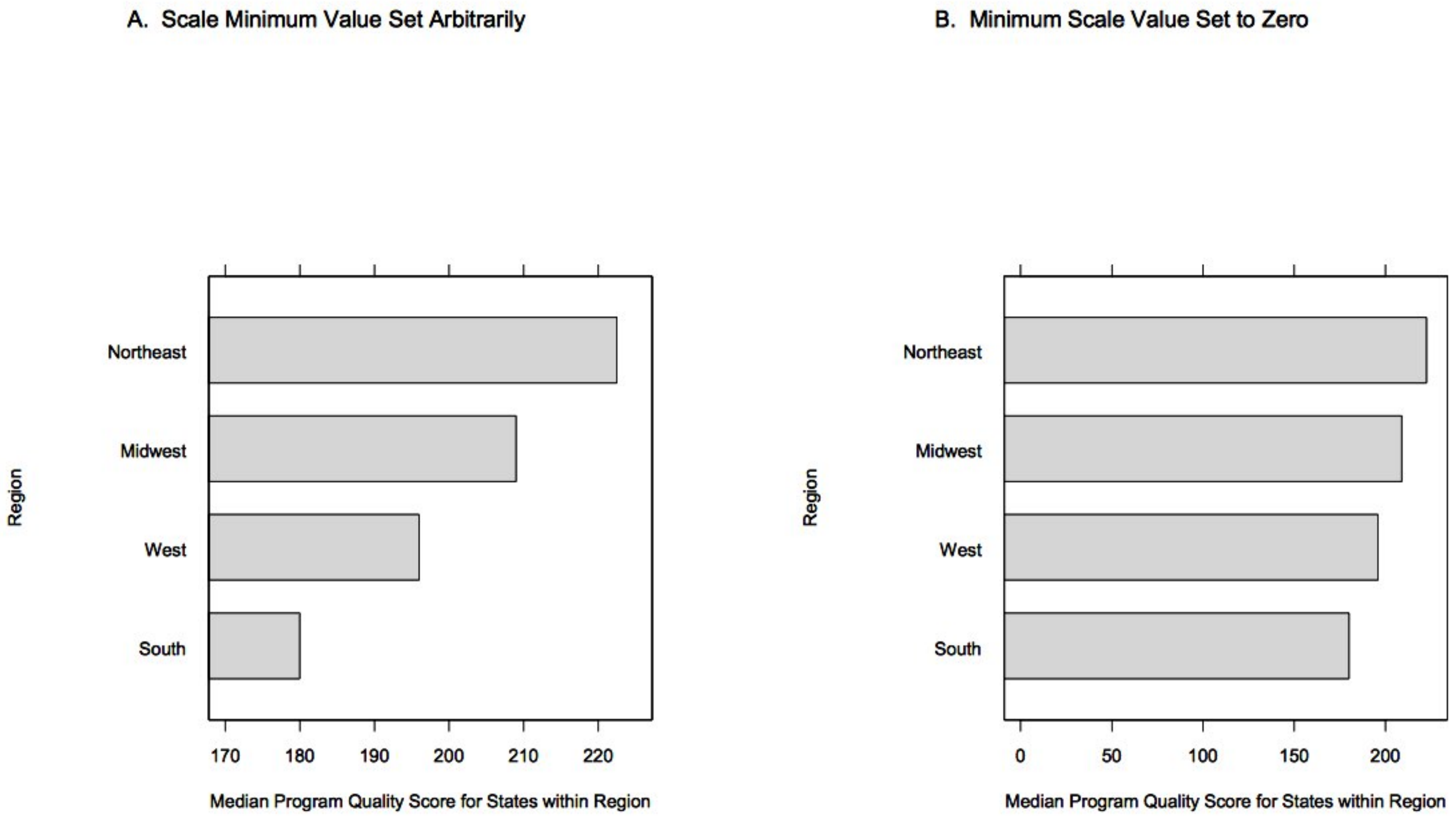


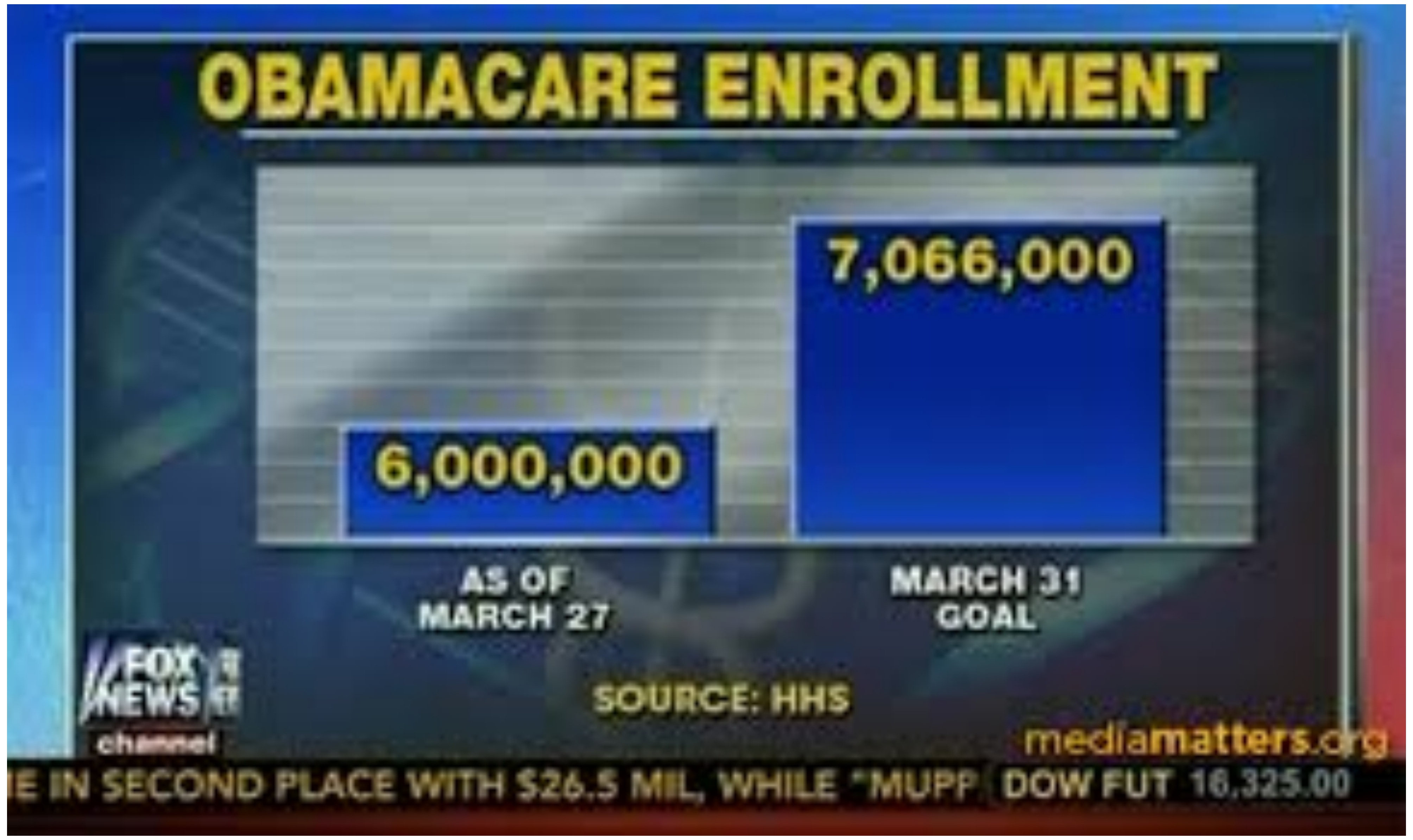
Note: The plotted points correspond to the mean importance rating assigned to each of the values, on a scale ranging from a minimum of zero to a maximum of three. The horizontal error bars show 95% confidence intervals for each of the means. Nonoverlapping confidence intervals imply that the means are reliably different from each other; that is, their difference is probably *not* merely due to sampling error.

Data Source: 1994 Multi-Investigator Study.

Dot Plots

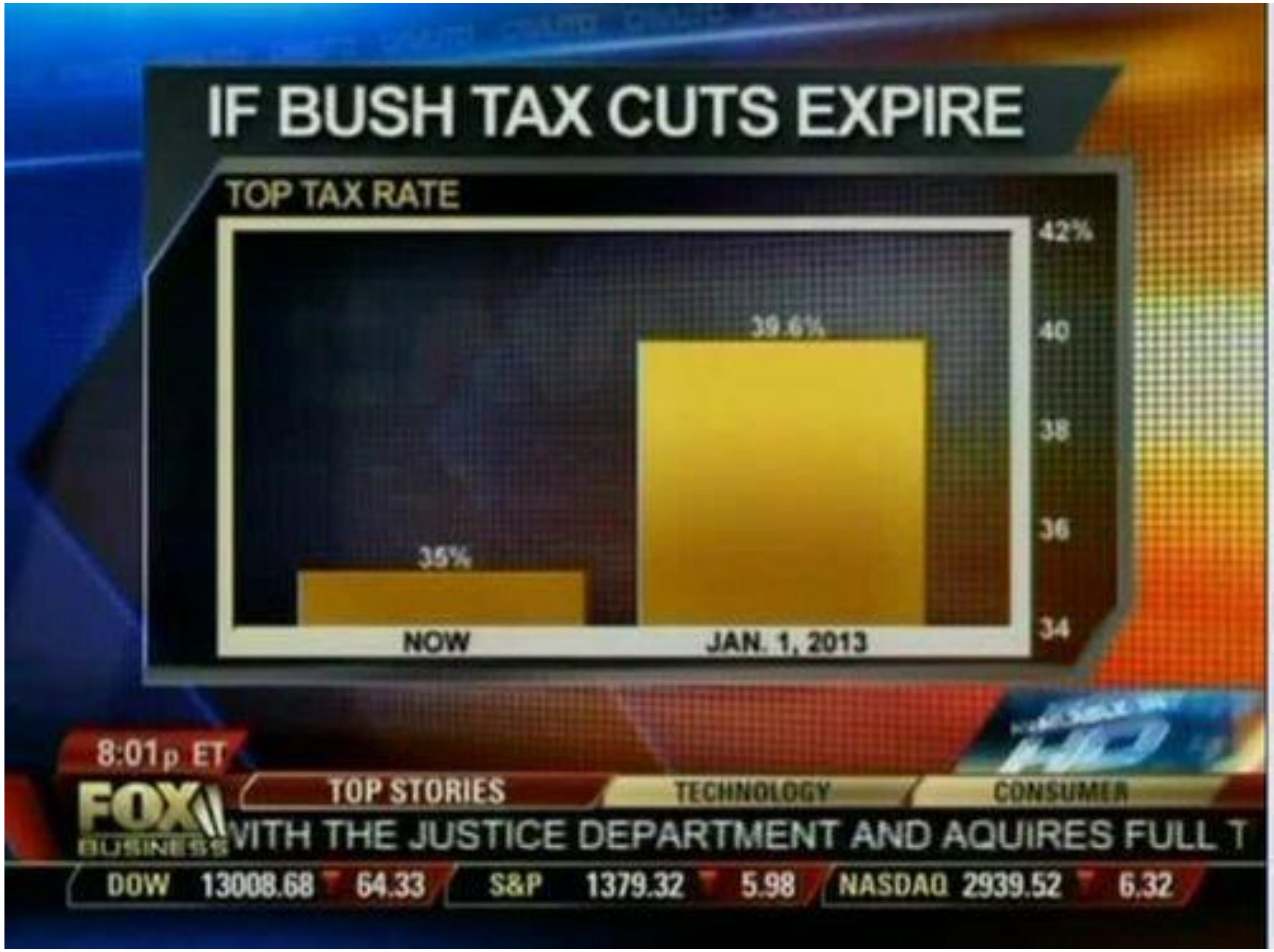
- Two-dimensional plot of data labels versus data values
  - Useful display when data values are associated with identifying information, such as a label or index number.
- Present a lot of information and are easy to interpret.
  - Can batch dot plots in useful ways to show data differently.
- Can also use dot plots in a variety of different ways, with different visual tweaks.
- The “zero” point of dot and bar plots should be paid attention to


A. Scale Minimum Value Set Arbitrarily
Northeast
Midwest
West
South
Region
170
180
190
200
210
220
Median Program Quality Score for States within Region
B. Minimum Scale Value Set to Zero
Northeast
Midwest
West
South
Region
0
50
100
150
200
Median Program Quality Score for States within Region



Original Graphic

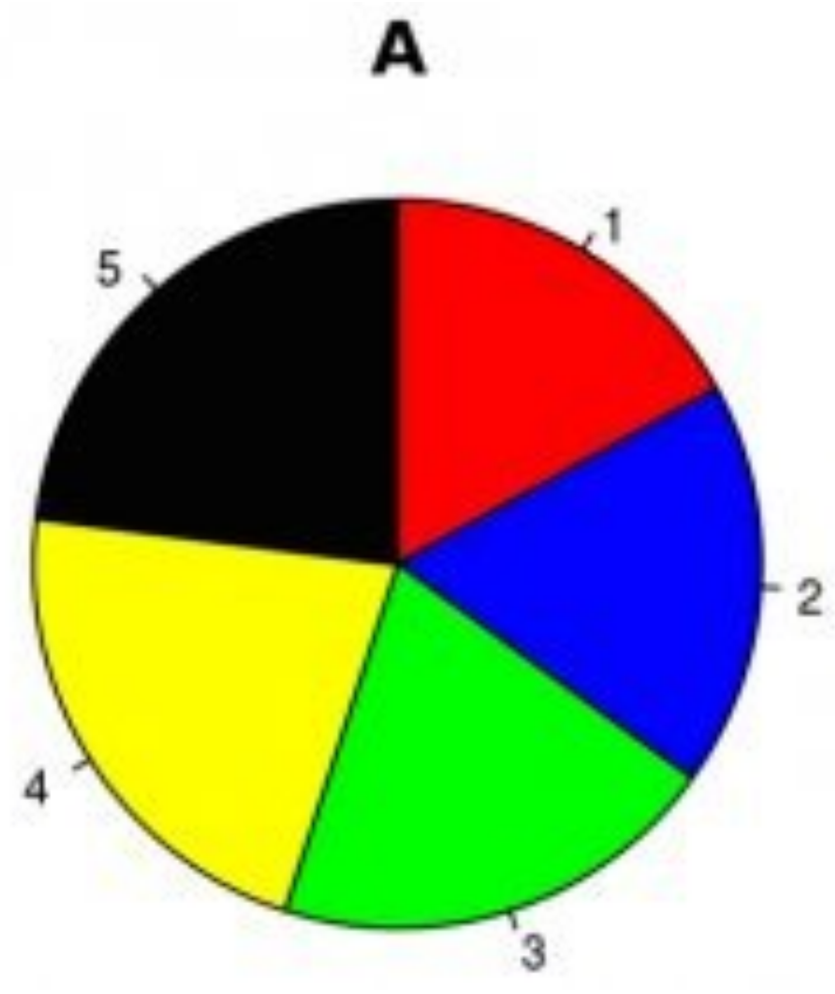
With Noted Zero Point


IF BUSH TAX CUTS EXPIRE
TOP TAX RATE
42%
40
38
36
34
35%
39.6%
NOW
JAN. 1, 2013
8:01p ET
TOP STORIES
TECHNOLOGY
CONSUMER
FOX BUSINESS
VITH THE JUSTICE DEPARTMENT AND AQUIRES FULL T
DOW 13008.68 64.33
S&P 1379.32 5.98
NASDAQ 2939.52 6.32

Bar and Pie Charts

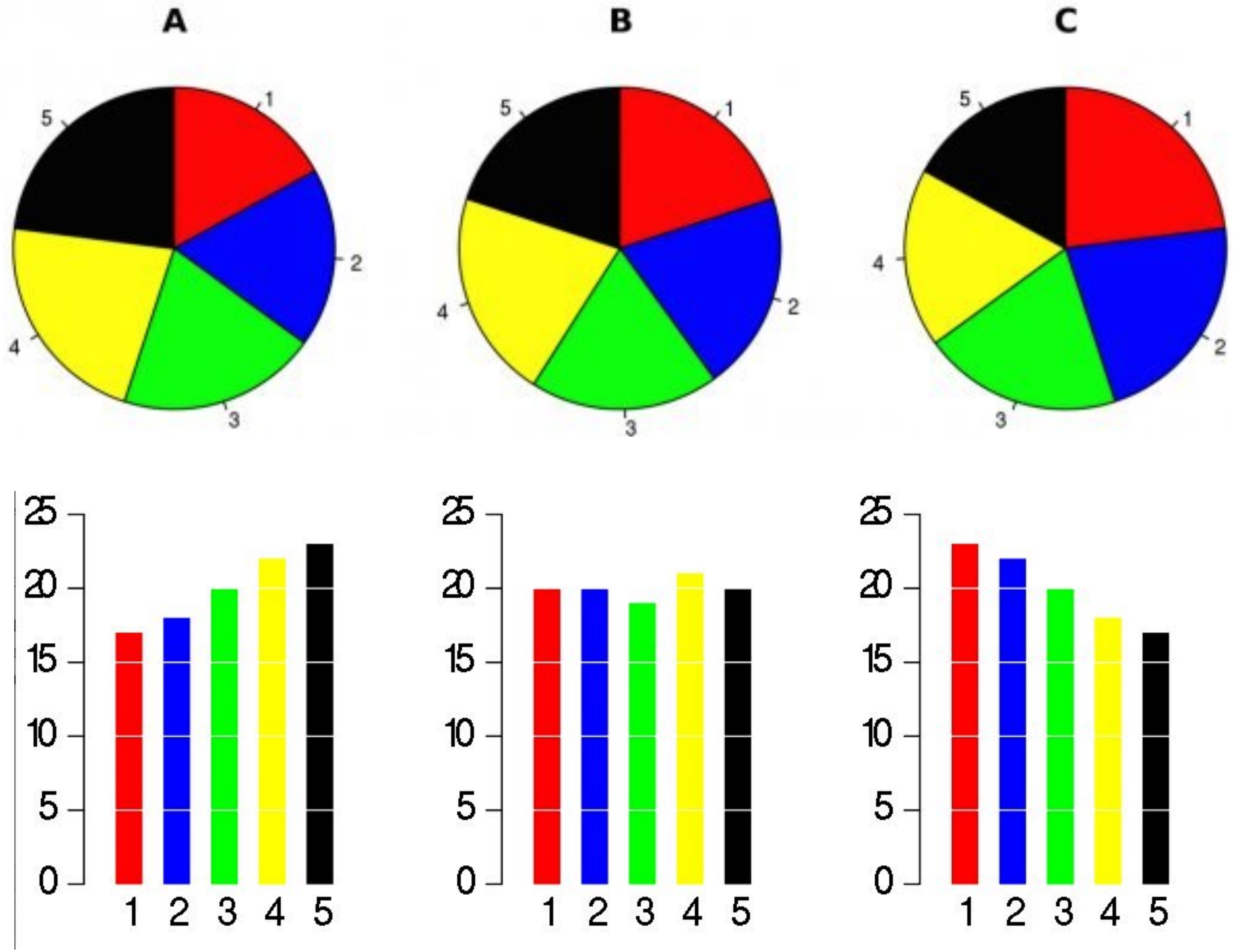
- For the love of all that is holy, please do not use a pie chart.
  - The human eye is particularly poor at interpreting area, especially in fine differences.
- Bar charts, and especially dot plots, represent a much better graph.
  - (Although bar charts, including the last, sometimes suggest “area” to the eye when they shouldn’t . . . use a dot plot when in doubt.)

Five candidates are running in an election. We poll the election at three separate times, and come up with the following poll results:



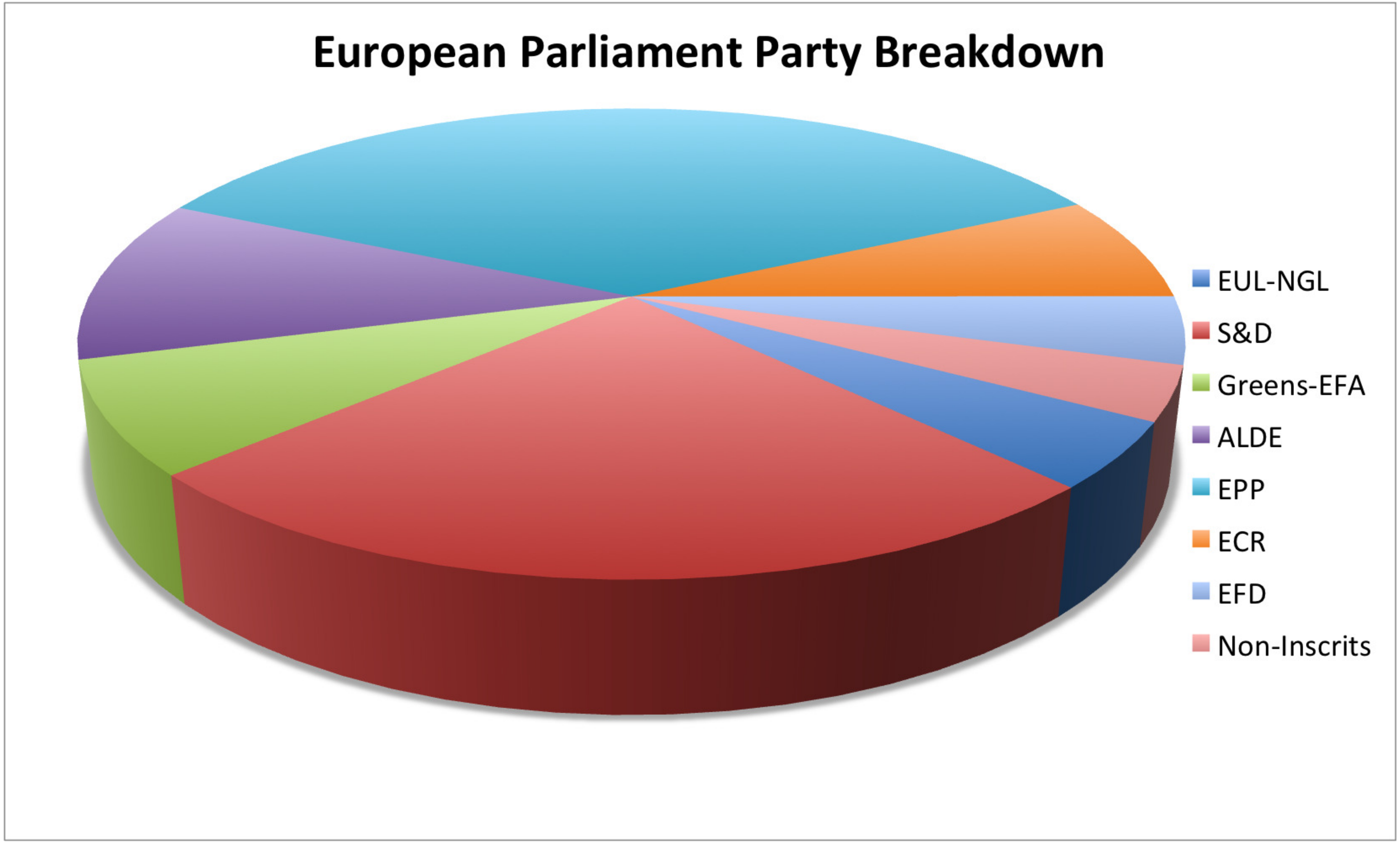
Which candidate is winning the election?

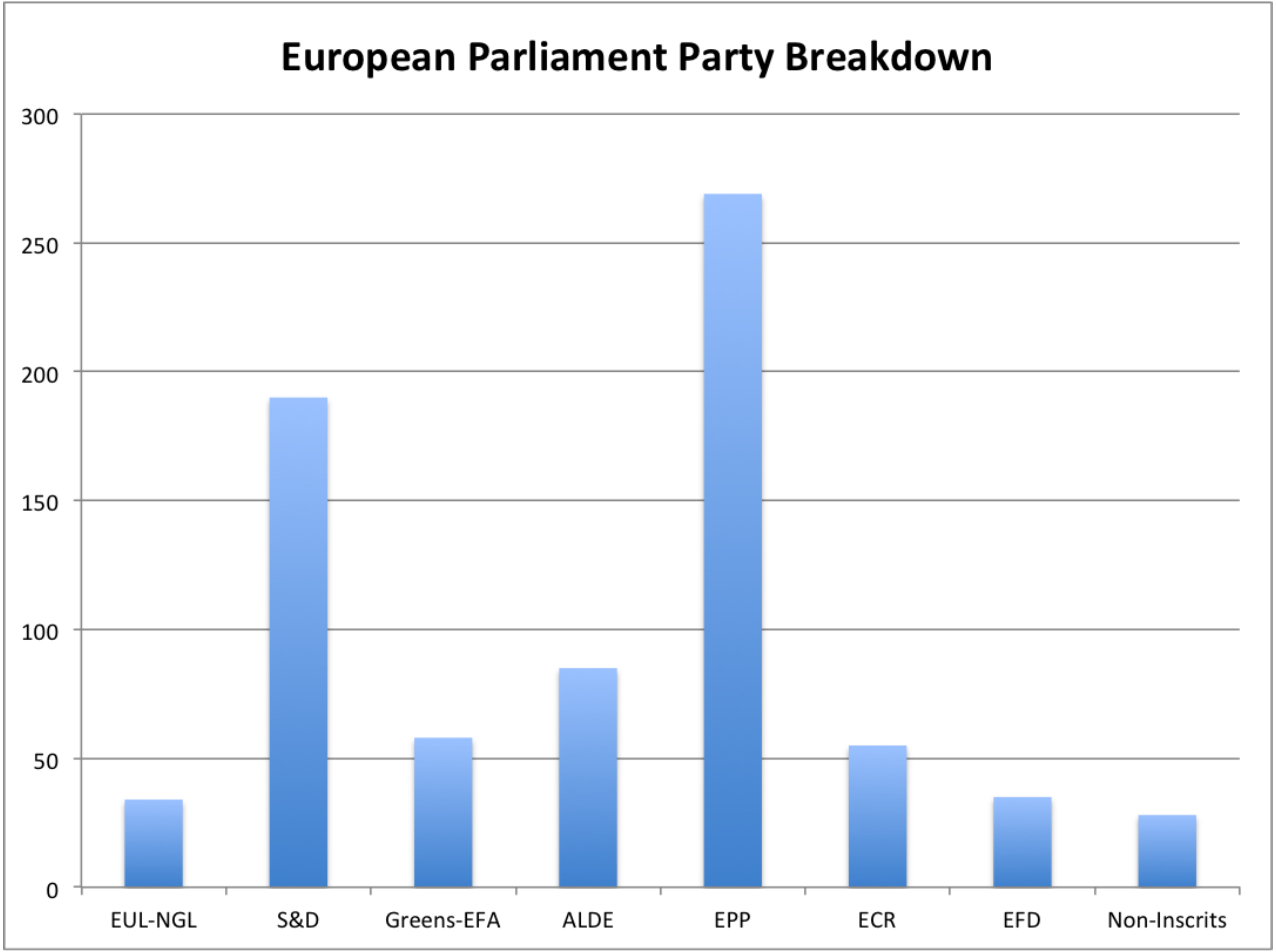
Source: http://www.businessinsider.com/pie-charts-are-the-worst-2013-6



A
B
C
1
2
3
4
5
25
20
15
10
5
0

Or, even better. What on earth is going on here?

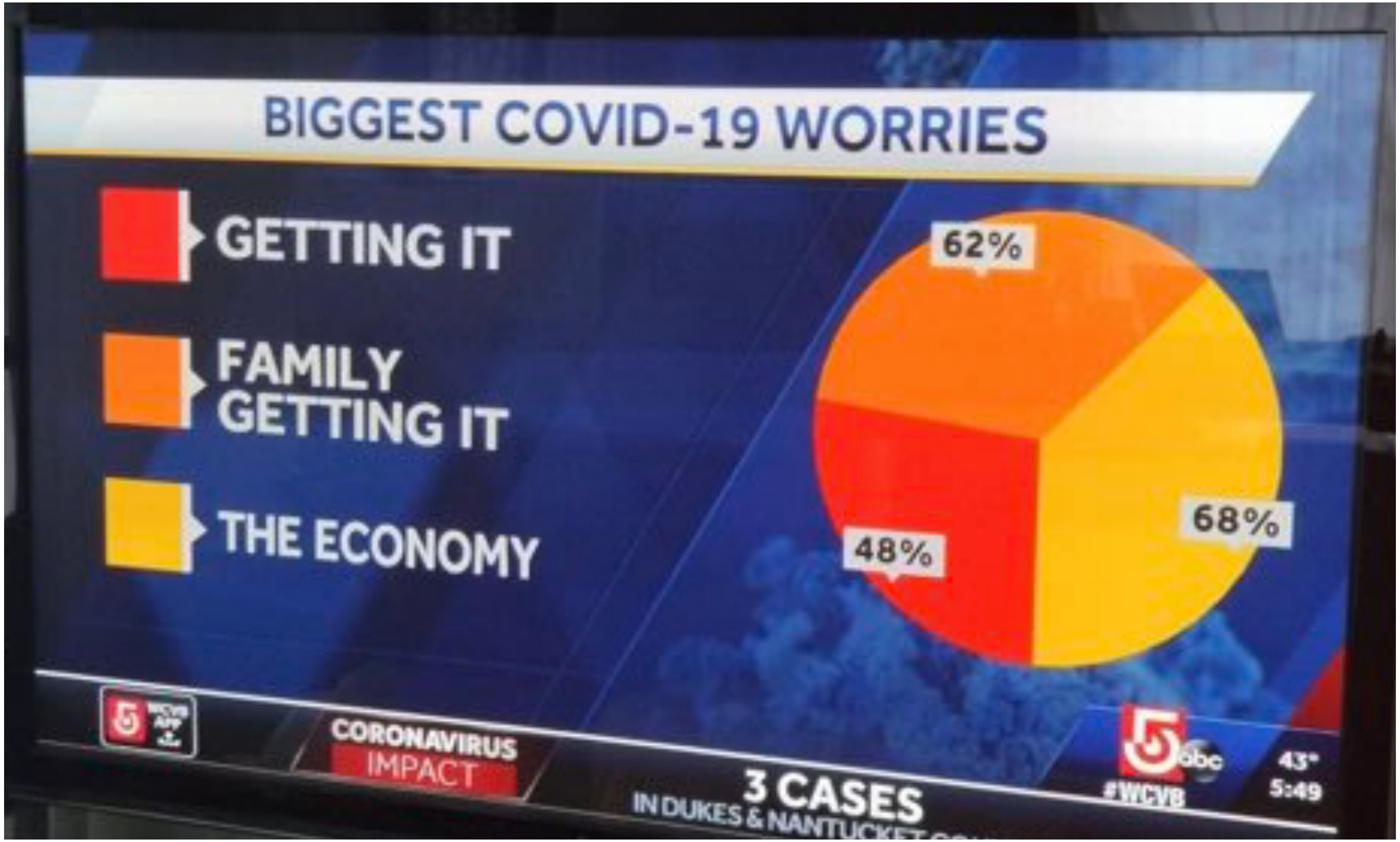



European Parliament Party Breakdown
300
250
200
150
100
50
0
EUL-NGL
S&D
Greens-EFA
ALDE
EPP
ECR
EFD
Non-Inscrits

The only good kind of pie chart

Wrapping up

- **Keep graphs simple.** Focus on your data.
- OK, so . . . how do I pick which one to use?
  - Depends on your data
    - As well as your own tastes
  - Different kinds of displays represent different data more effectively
    - *Understand* the choices you're making with your graphs
  - All bow to the histogram . . . it will probably continue to dominate univariate displays (for better or worse)


BIGGEST COVID-19 WORRIES
GETTING IT
FAMILY GETTING IT
THE ECONOMY
62%
48%
68%
CORONAVIRUS
IMPACT
3 CASES
#WCVB
43°
5:49