Descriptive Statistics

Kosuke Imai

Department of Politics
Princeton University

Fall 2011

Variables and their Types

- Variables: characteristics about each unit which take different values across units
- Quantitative variables
  - Continuous: income, age, etc.
  - Discrete: number of siblings, etc.
  - Continuous variables may be measured discretely
- Qualitative (Categorical) variables
  - state, country, race, gender, party, etc.
  - Nominal vs. Ordinal
- In **R**,
  - quantitative variable as a `numeric` object
  - qualitative variable as a `factor` object
- Common data format:
  - comma separated values (csv)
  - other delimiter-separated values format; space, tab

Measurement Error

- Variables must be measured
- But often they are measured with error
- READING: FPP Chapter 6

- Random (chance) measurement error
- No bias: *on average*, there is no error
- Repeated measurements: "Law of large numbers" which we will cover later

- Systematic measurement error
  - government statistics in some countries
  - mis-reporting in surveys (social desirability bias)
  - interpersonal incompatibility in surveys

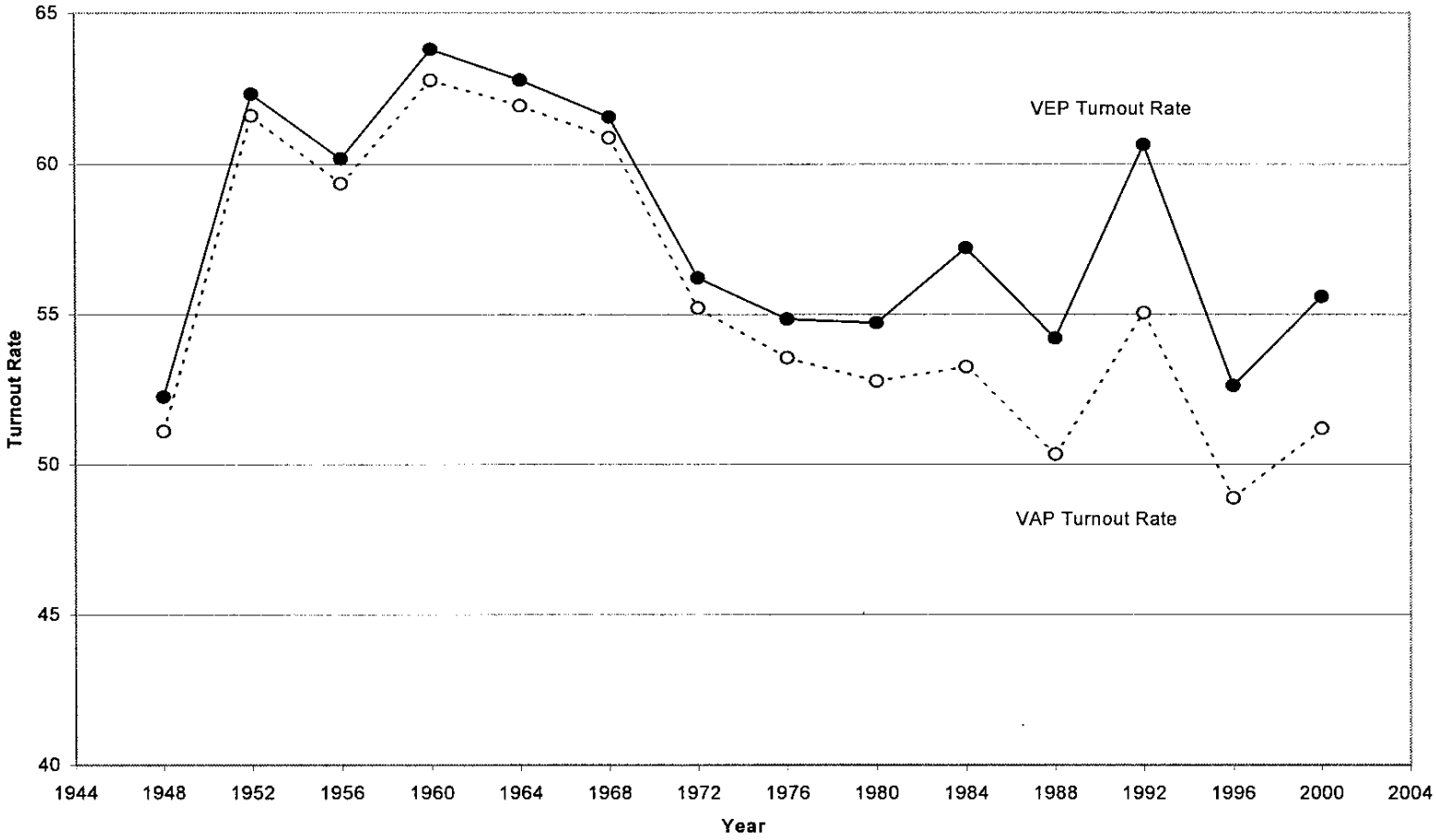
Measuring the US Turnout Rate

- Question: How do you measure turnout rate?
- Numerator: Total votes cast
  - Votes for highest office (president; governor or senator in midterm elections)
- Denomenator: VAP vs. VEP
  - VAP from Census (adjustments for non-Census years)
  - VEP $=$ VAP $+$ overseas voters $-$ ineligible voters
  - overseas voters: military personnel and civilians
  - ineligible voters: non-citizens, disenfranchised felons, mental incompetents, those who failed to meet states' residency requirement

VAP and VEP Turnout Rates are Different

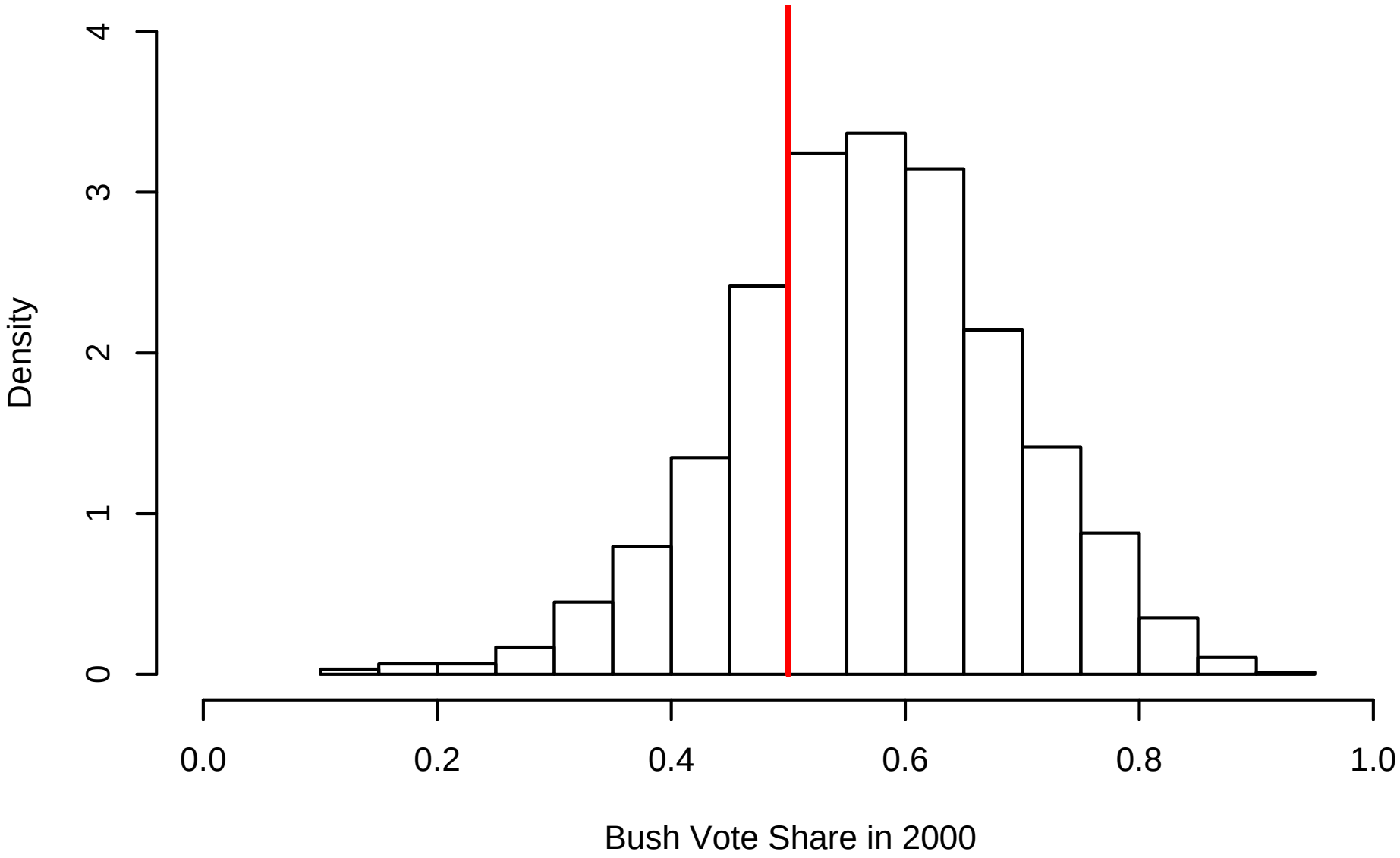
- Trend graph

FIGURE 1. National VAP and VEP Presidential Turnout Rates, 1948–2000



McDonald and Popkin (2001) *American Political Science Review*

Summarizing Quantitative Data



How to Read Histograms

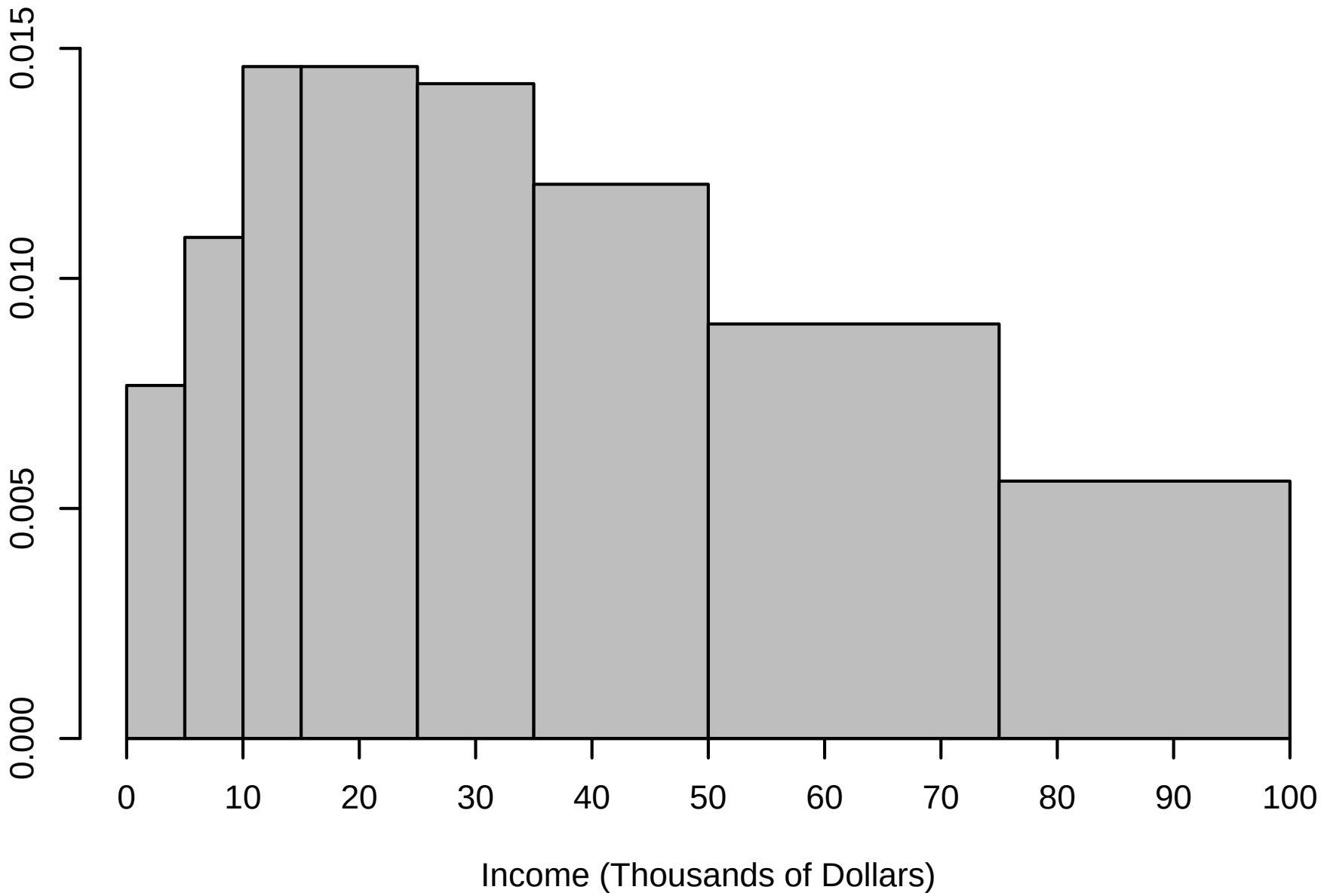
- READING: FPP Chapter 3
- What is density scale?
- The areas of the blocks sum to 1 or 100%
- Density scale $\neq$ Percentage (e.g., range $\neq [0, 1]$)
- The height of the blocks equals the percentage divided by *class interval* length
- In this case, "percent per vote share"
- Or more generally, "percentage per horizontal unit"

Interval Data

- U.S. Households Income in 2006:

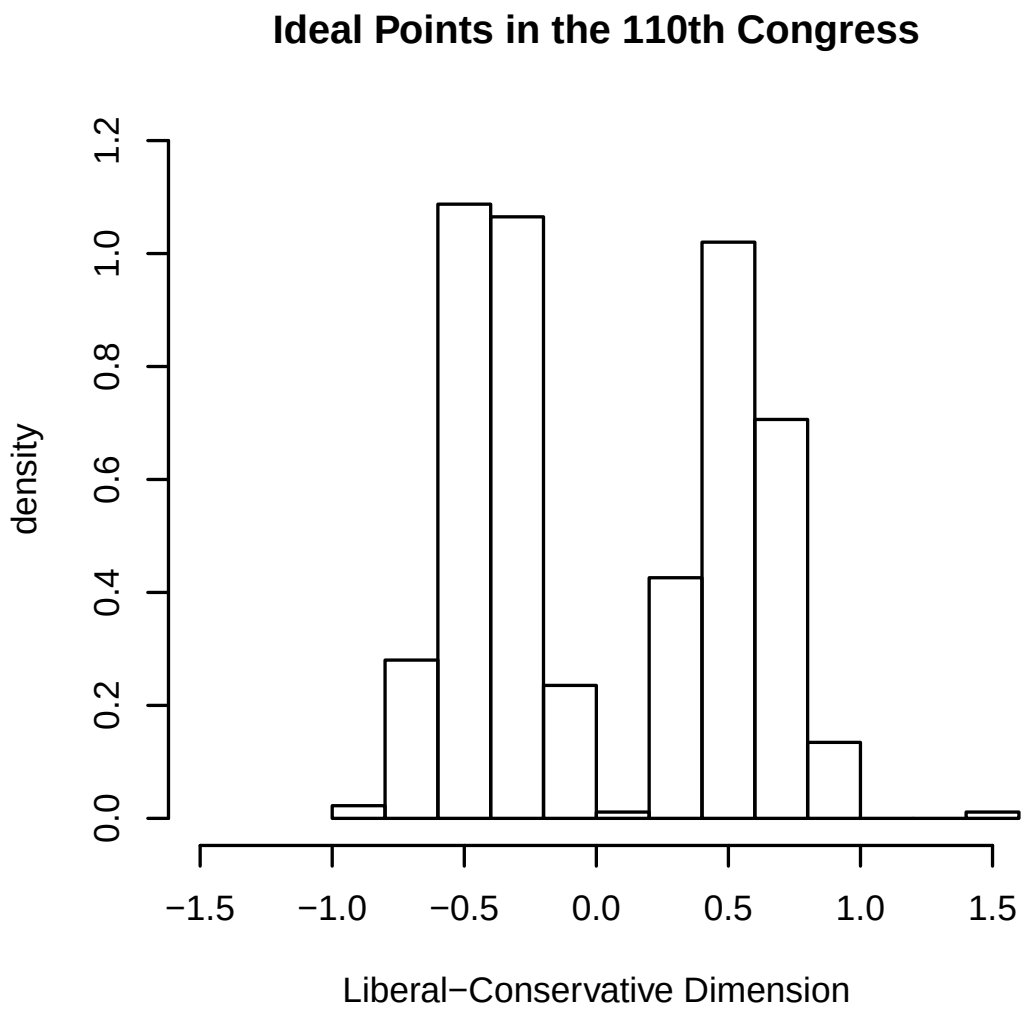
Household income	Percent	Percent (< $100K)
$0 - $4,999	3.1	3.8
$5,000 - $9,999	4.4	5.4
$10,000 - $14,999	5.9	7.3
$15,000 - $24,999	11.8	14.6
$25,000 - $34,999	11.5	14.2
$35,000 - $49,999	14.6	18.0
$50,000 - $74,999	18.2	22.5
$75,000 - $99,999	11.3	14.0
$\geq$ $100,000	19.1	

Histogram with Interval Data

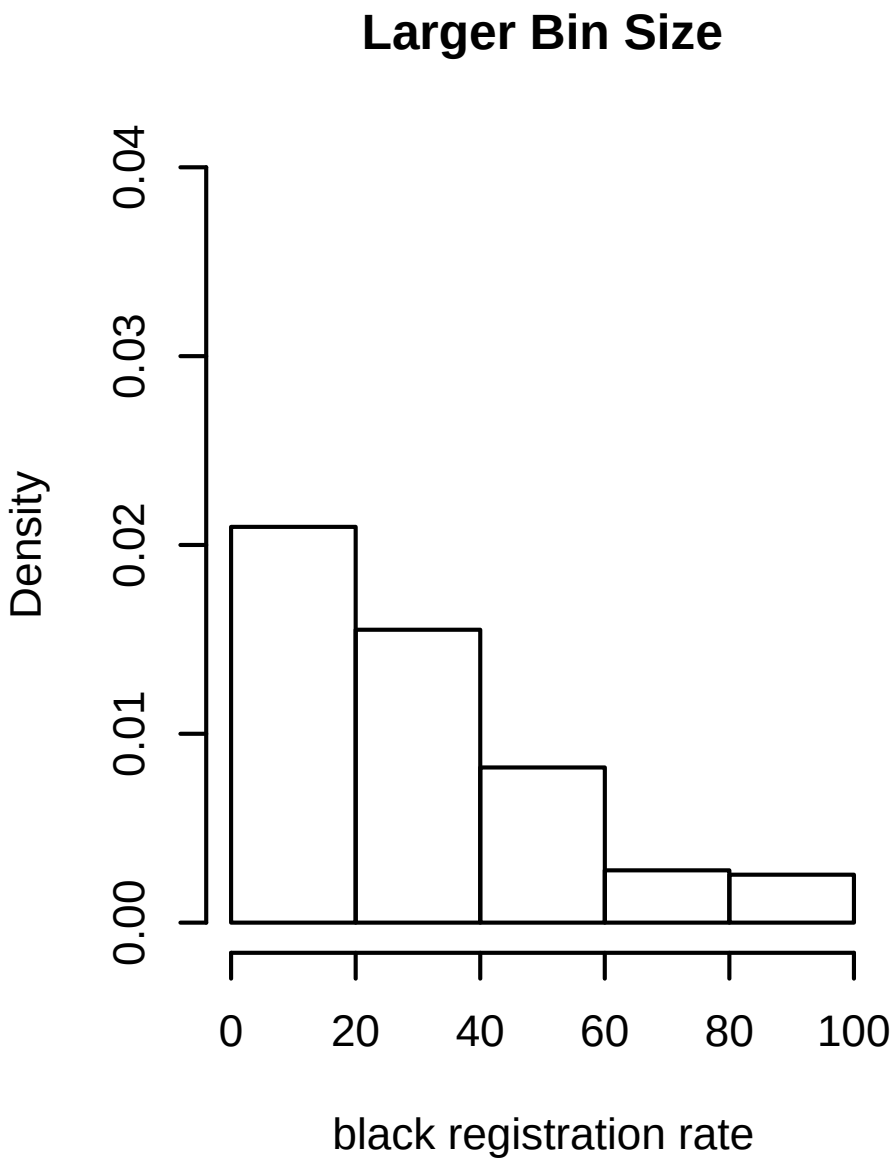


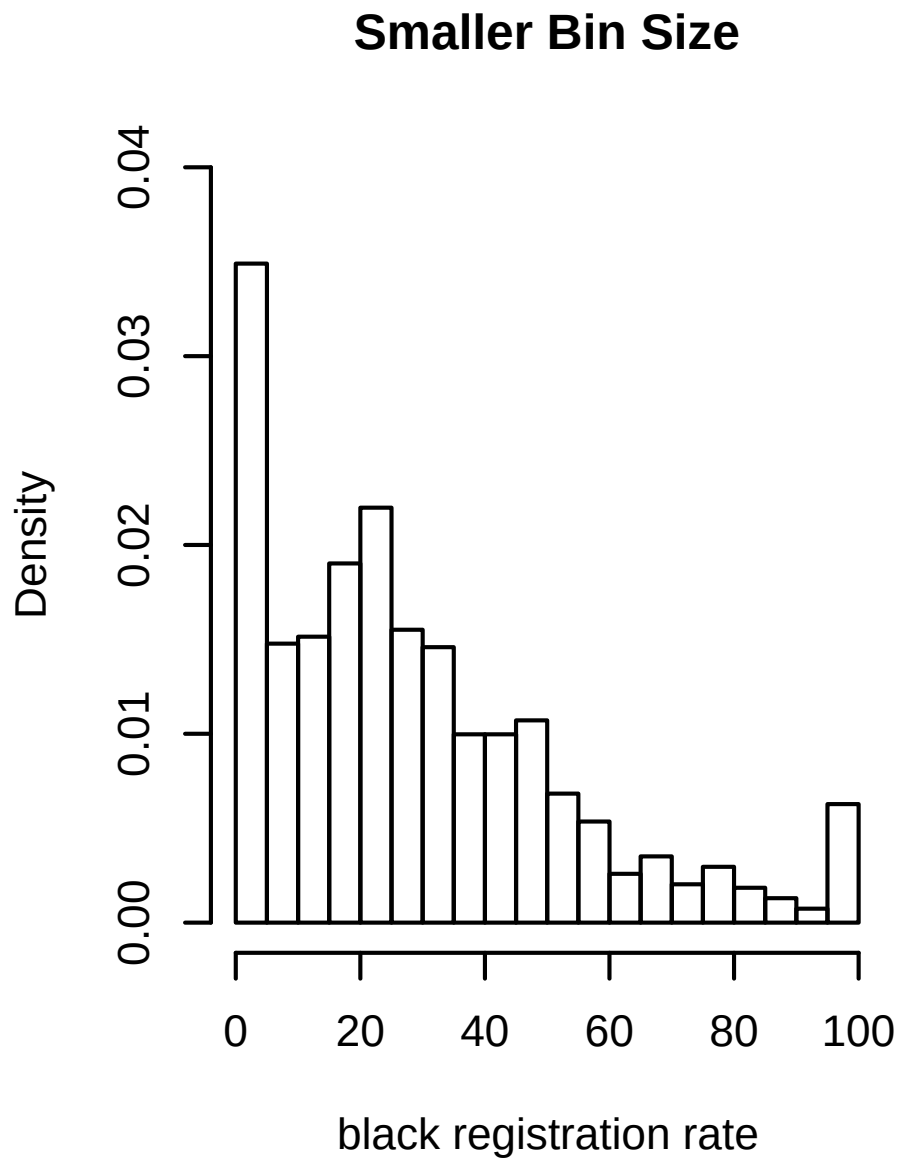
Ideal Points in Legislative Studies

- NOMINATE scores by Poole and Rosenthal

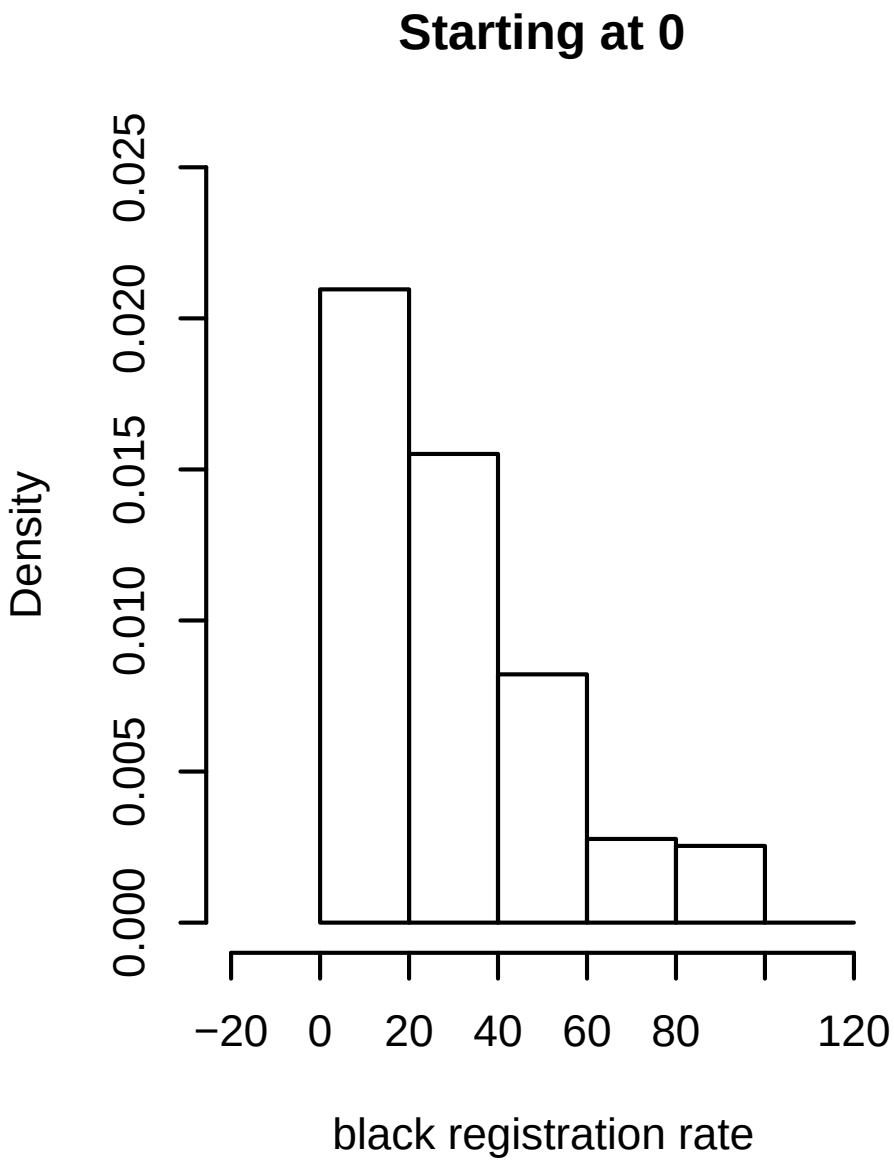


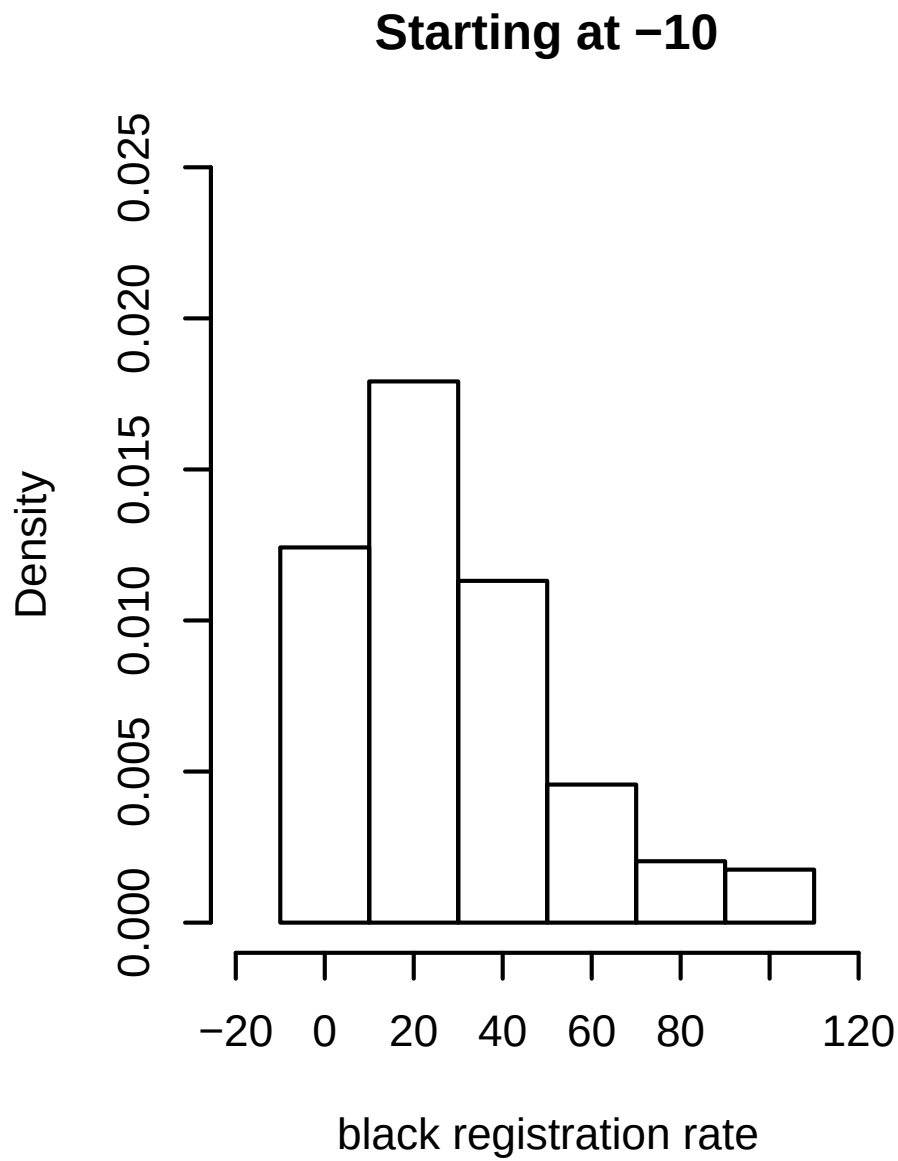
Sensitivity to the Choice of Bin Width





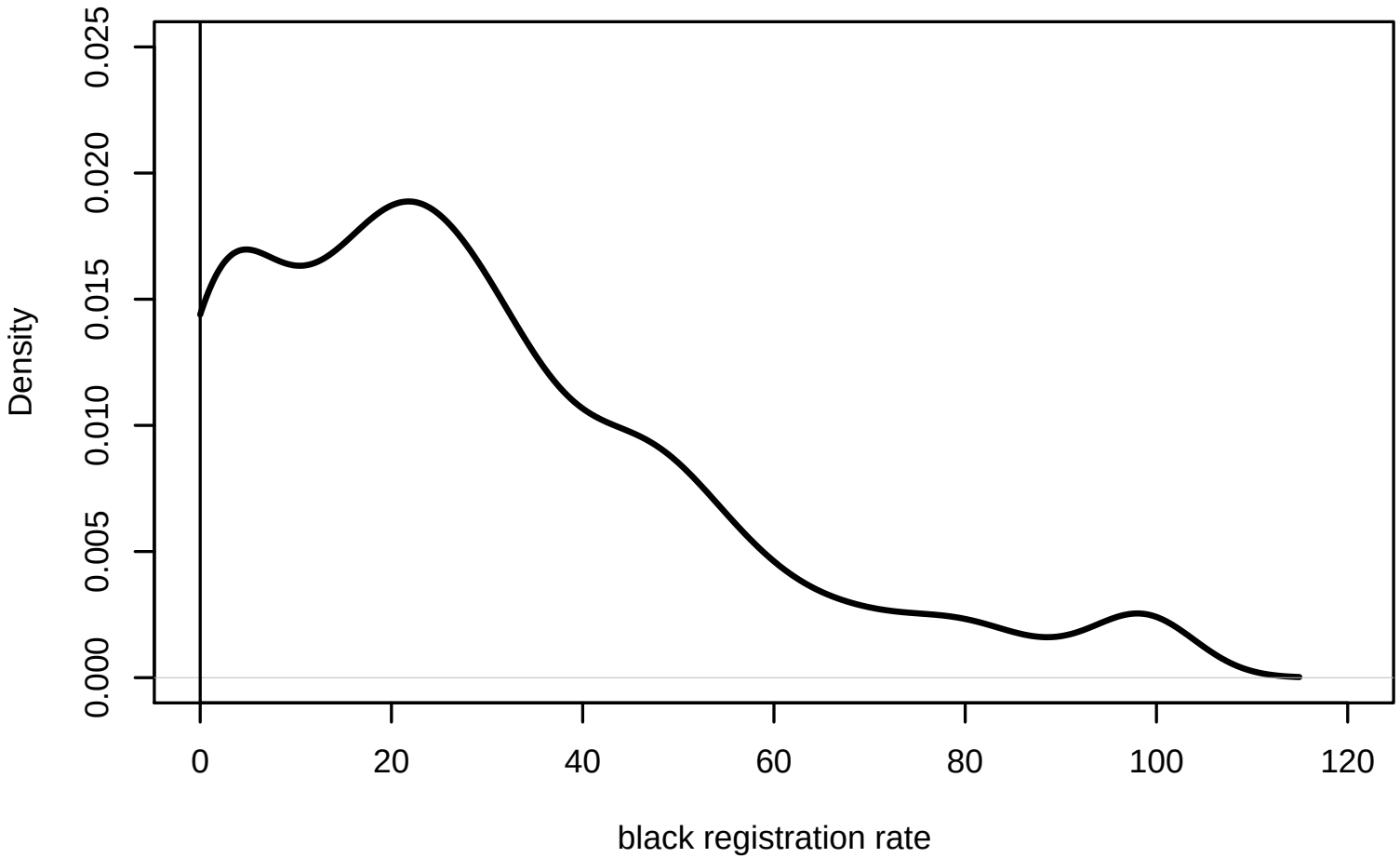
Sensitivity to the Choice of Origin





Smoothed Histogram

- Methods to "optimally" smooth histograms
- In **R**, `density()` will do it



Describing the Center of the Data

- READING: FPP Chapter 4
- (Sample) Mean or average:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

- (Sample) Median:

$$\tilde{X} = \begin{cases} X_{\left(\frac{n+1}{2}\right)} & \text{if } n \text{ is odd} \\ \frac{1}{2}X_{\left(\frac{n}{2}\right)} + \frac{1}{2}X_{\left(\frac{n}{2}+1\right)} & \text{if } n \text{ is even} \end{cases}$$

- Mean is not necessarily median
- Median is more robust to outliers than mean
- Example: data $= \{0,1,2,3,100\}$, mean $= 21.2$, median $= 2$

Median Justice in Supreme Court

- Estimated ideal points for Supreme court justices in 2008 by Martin and Quinn

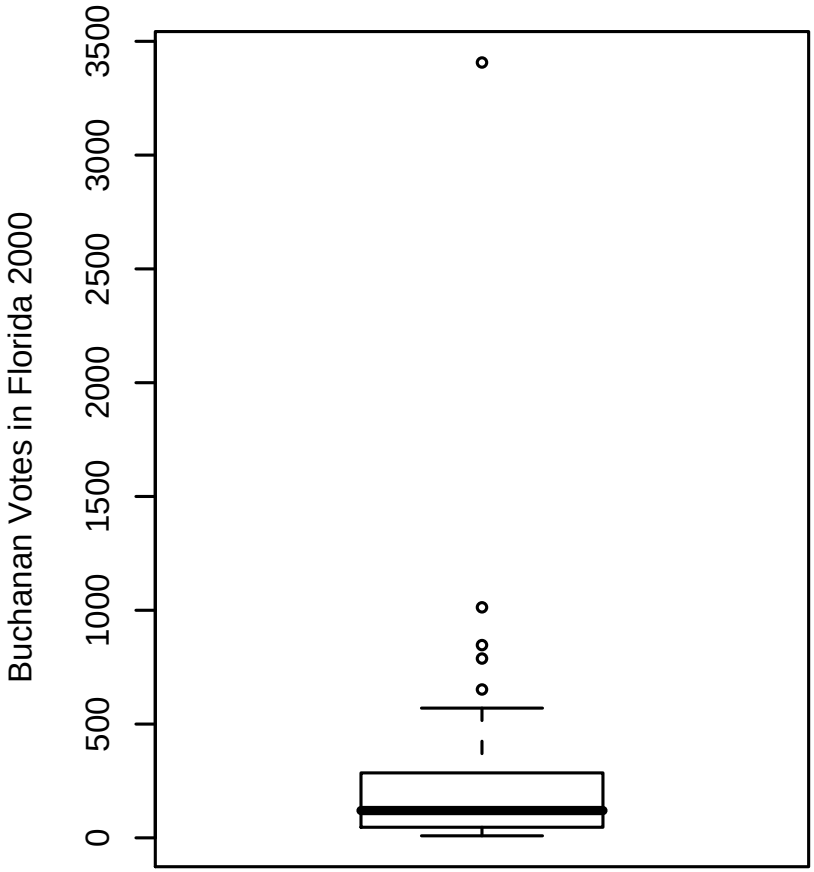
Alito	1.88
Breyer	−1.10
Ginsburg	−1.63
Kennedy	0.58
Roberts	1.69
Scalia	2.71
Souter	−1.50
Stevens	−2.51
Thomas	4.24

- Who was the median justice?

Describing the Spread of the Data

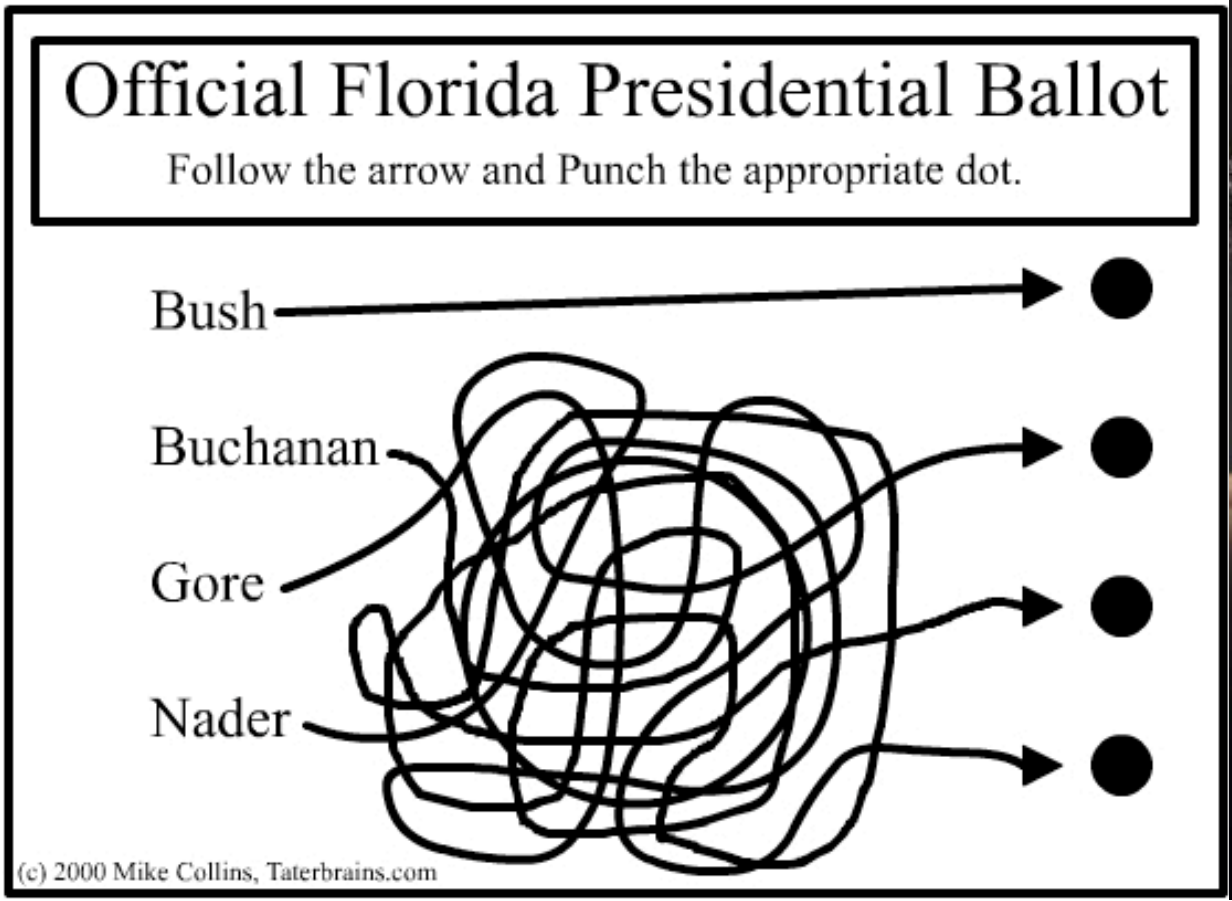
- Range: $[\min(X),\ \max(X)]$
- Quantile (quartile, quintile, percentile, etc.):
  - 25 percentile $=$ lower quartile
  - 50 percentile $=$ median
  - 75 percentile $=$ upper quartile
  - Interquartile Range (IQR): a measure of variability
- A definition of outliers: over 1.5 IQR above upper quartile or below lower quartile

Box Plot

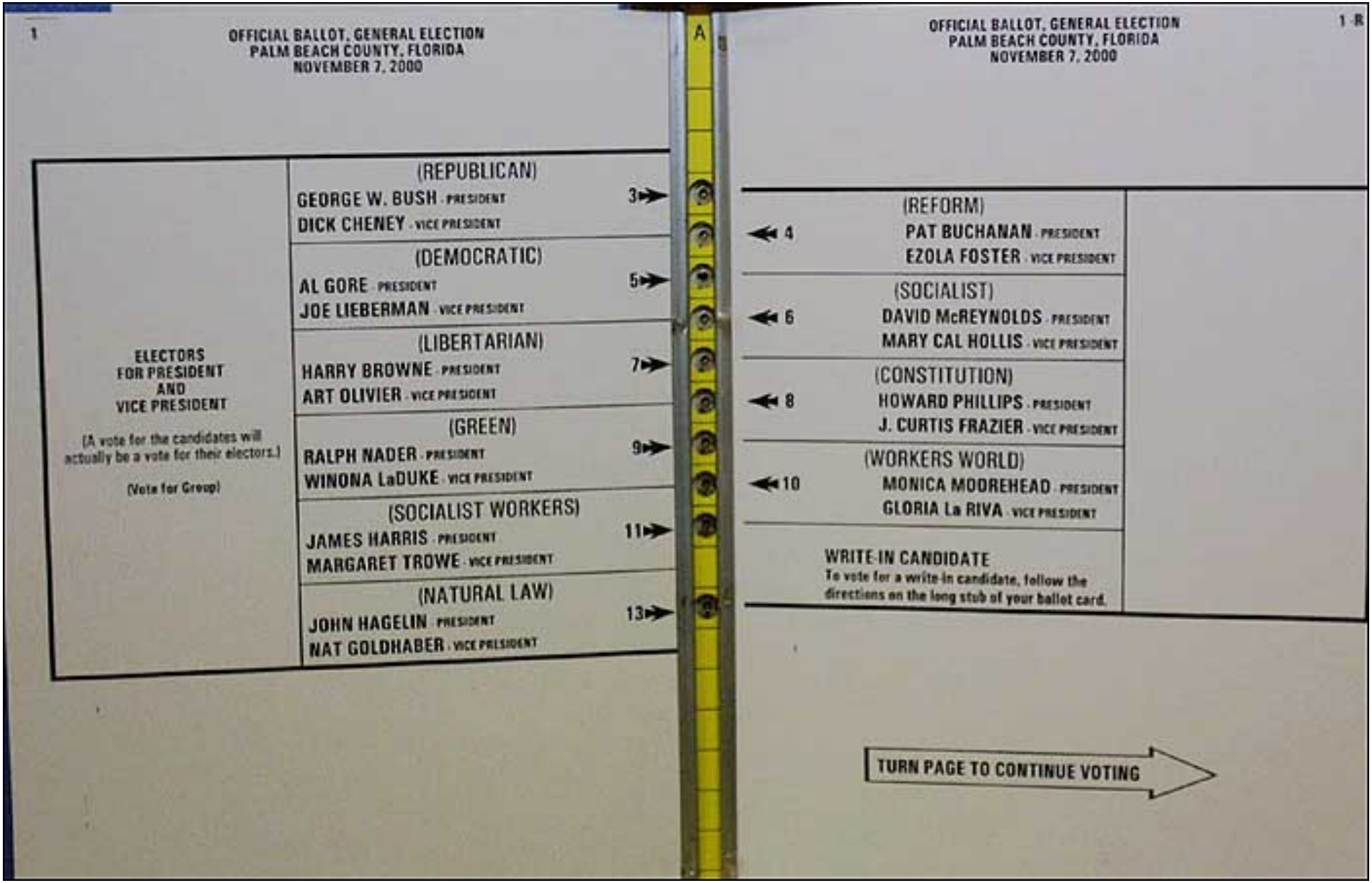


- BOX: 50% of the data from the lower to upper quartiles
- WHISKERS: 1.5 IQR lower than the lower quartile or higher than the upper quartile (but they do not exceed min/max)
- OUTLIERS: marked separately

Remembering Florida 2000



http://www.youtube.com/watch?v=GVd1LuadnDk


1
OFFICIAL BALLOT, GENERAL ELECTION
PALM BEACH COUNTY, FLORIDA
NOVEMBER 7, 2000
ELECTORS FOR PRESIDENT AND VICE PRESIDENT
(A vote for the candidates will actually be a vote for their electors.)
(Vote for Group)
(REPUBLICAN)
GEORGE W. BUSH - PRESIDENT
DICK CHENEY - VICE PRESIDENT
3
(DEMOCRATIC)
AL GORE - PRESIDENT
JOE LIEBERMAN - VICE PRESIDENT
5
(LIBERTARIAN)
HARRY BROWNE - PRESIDENT
ART OLIVIER - VICE PRESIDENT
7
(GREEN)
RALPH NADER - PRESIDENT
WINONA LaDUKE - VICE PRESIDENT
9
(SOCIALIST WORKERS)
JAMES HARRIS - PRESIDENT
MARGARET TROWE - VICE PRESIDENT
11
(NATURAL LAW)
JOHN HAGELIN - PRESIDENT
NAT GOLDHABER - VICE PRESIDENT
13
A
OFFICIAL BALLOT, GENERAL ELECTION
PALM BEACH COUNTY, FLORIDA
NOVEMBER 7, 2000
1-R
(REFORM)
4
PAT BUCHANAN - PRESIDENT
EZOLA FOSTER - VICE PRESIDENT
(SOCIALIST)
6
DAVID McREYNOLDS - PRESIDENT
MARY CAL HOLLIS - VICE PRESIDENT
(CONSTITUTION)
8
HOWARD PHILLIPS - PRESIDENT
J. CURTIS FRAZIER - VICE PRESIDENT
(WORKERS WORLD)
10
MONICA MOOREHEAD - PRESIDENT
GLORIA La RIVA - VICE PRESIDENT
WRITE-IN CANDIDATE
To vote for a write-in candidate, follow the directions on the long stub of your ballot card.
TURN PAGE TO CONTINUE VOTING

Standard Deviation

- On average, how far away are data points from their mean?
- Mathematically (and as implemented in **R**),

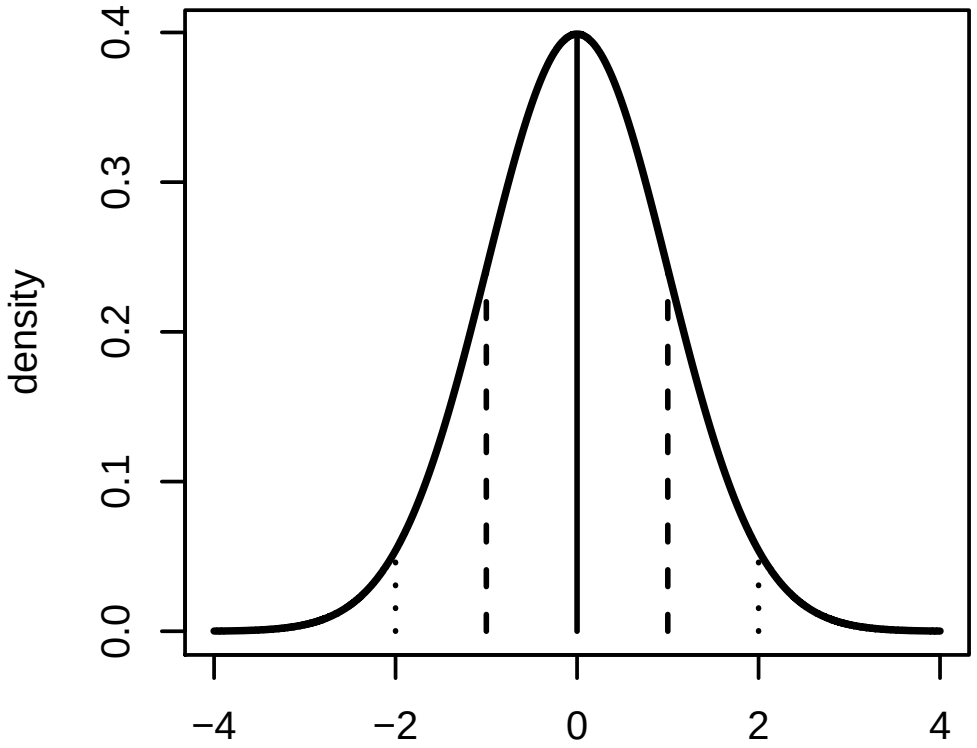
$$s_X = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

- Sometimes (as in FPP), $n-1$ is replaced with n
- Root-mean-square deviation from the mean
- Very few data points are more than 2 or 3 SDs away from the mean
- s_X^2 is called variance
- z-score for the ith observation:

$$z_i = \frac{x_i - \overline{x}}{s_X}$$

“Bell Shape” (Normal, Gaussian) Distribution

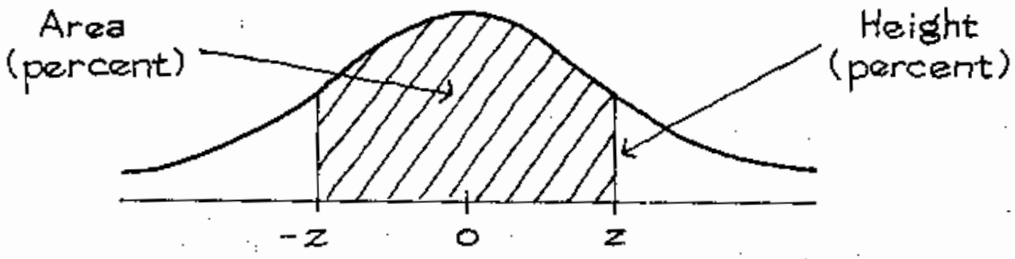
- READING: FPP Chapter 5
- Normal density curve:
  $\frac{1}{\sqrt{2\pi}} \exp(-x^2/2)$



- symmetric
- centered around zero
- standard deviation is one

- 68% of data points lie within 1 SD
- 95% of data points lie within 2 SDs
- 99.7% of data points lie within 3 SDs

The Standard Normal Table (FPP A–105)



A NORMAL TABLE

z	Height	Area	z	Height	Area	z	Height	Area
0.00	39.89	0	1.50	12.95	86.64	3.00	0.443	99.730
0.05	39.84	3.99	1.55	12.00	87.89	3.05	0.381	99.771
0.10	39.69	7.97	1.60	11.09	89.04	3.10	0.327	99.806
0.15	39.45	11.92	1.65	10.23	90.11	3.15	0.279	99.837
0.20	39.10	15.85	1.70	9.40	91.09	3.20	0.238	99.863

- To convert the normal with mean *a* and standard deviation *b*, subtract *a* and divide by *b*: z-score
- Use symmetry to get, for example, the area under the curve right of *z*

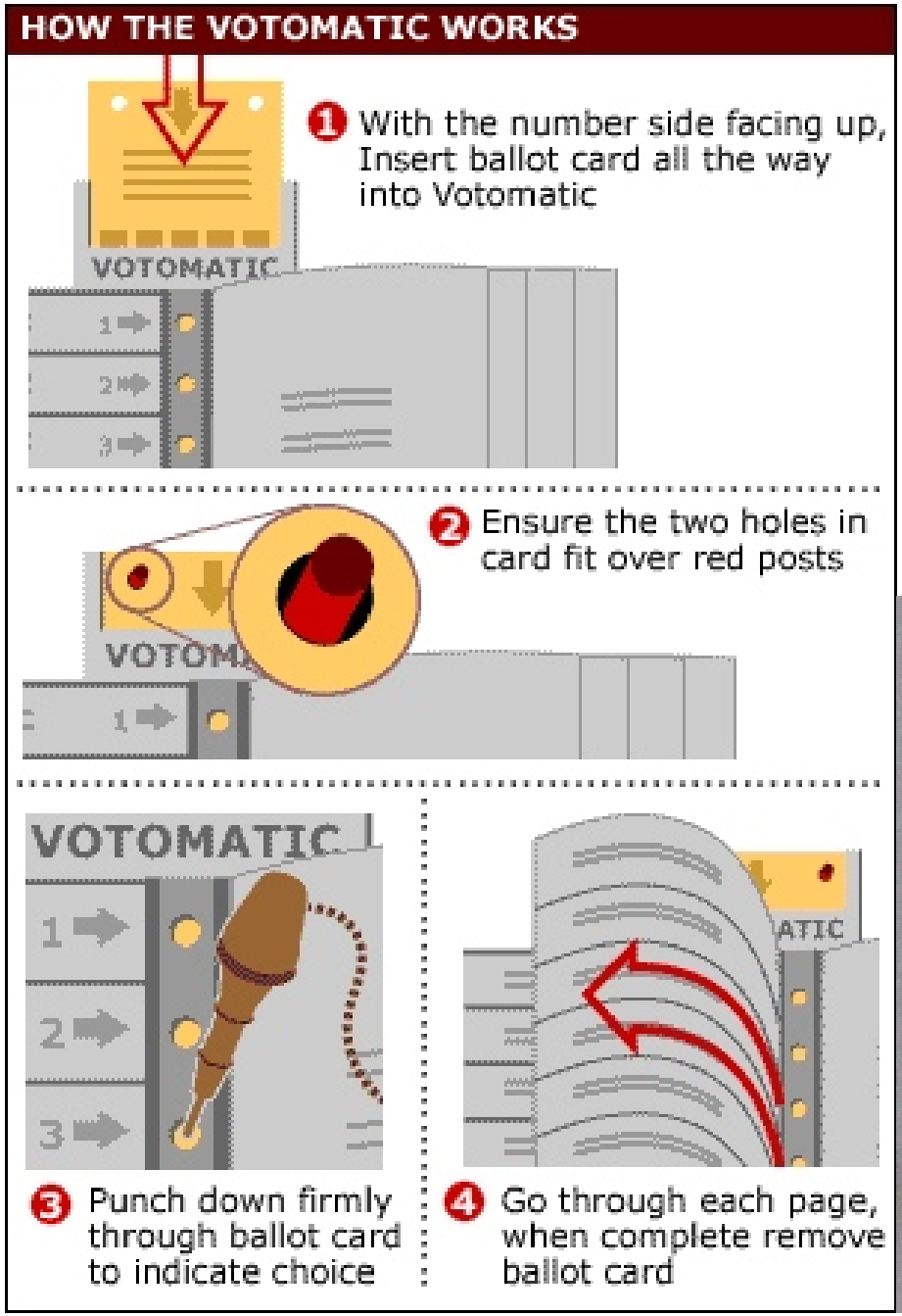
The Relationship between Two Variables

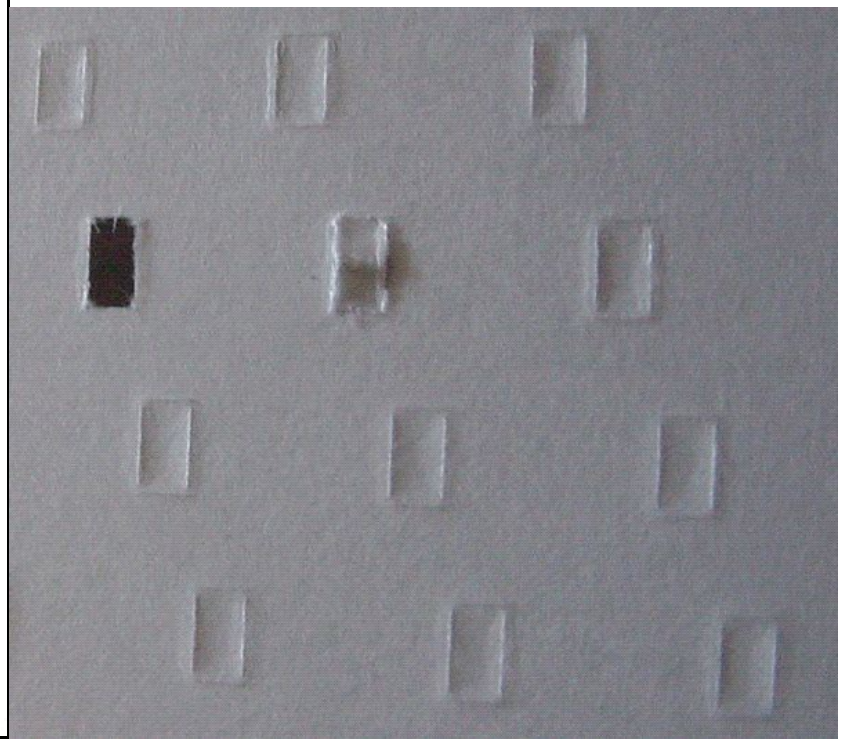
- Association (not causation!) between two variables
- Cross tabulation for categorical variables

		literacy requirement	
		Yes	No
polltax	Yes	109	369
	No	329	276

- Calculate the mean value for each category:

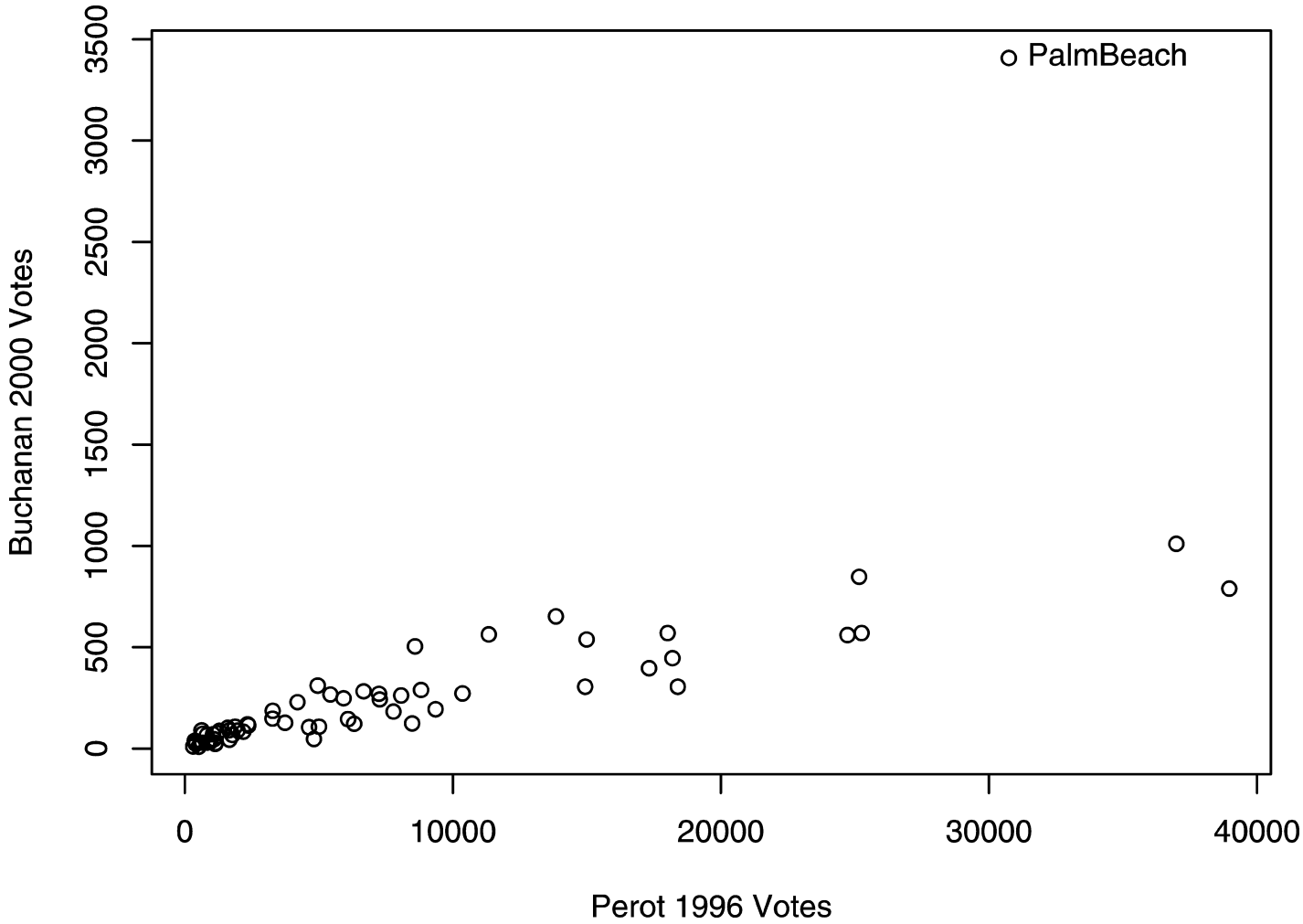
	Undervote	Overvote
Optical	0.3%	1.0%
Punch Card	1.5%	2.4%


HOW THE VOTOMATIC WORKS
1 With the number side facing up, Insert ballot card all the way into Votomatic
VOTOMATIC
1
2
3
2 Ensure the two holes in card fit over red posts
VOTOMATIC
1
2
3
3 Punch down firmly through ballot card to indicate choice
4 Go through each page, when complete remove ballot card



Scatter Plot for Two Quantitative Variables

- READING: FPP Chapter 7



Agresti and Presnell (2002) *Statistical Science*

Correlation

- READING: FPP Chapters 8 and 9

- On average, how do two variables move together?
- Positive (negative) correlation: When X is larger than its mean, Y is likely (unlikely) to be larger than its mean
- Positive (negative) correlation: data cloud slopes up (down)
- High correlation: data cluster tightly around a line

- Mathematical definition of correlation coefficient:

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left\{ \frac{(X_i - \overline{X})}{S_X} \times \frac{(Y_i - \overline{Y})}{S_Y} \right\}$$

- Sometimes (as in FPP) $n - 1$ is replaced with n

Properties of Correlation Coefficient

- Correlation is between -1 and 1
- Order does not matter: $\texttt{cor(x, y)} = \texttt{cor(y, x)}$
- Not affected by changes of scale:

$$\texttt{cor(x, y)} = \texttt{cor(ax + b, cy + d)}$$

  for any numbers $\texttt{a}$, $\texttt{b}$, $\texttt{c}$, and $\texttt{d}$
- Celsius vs. Fahrenheit; cm vs. inch; yen vs. dollar etc.
- Correlation size is relative to standard deviation
- Given the same correlation value, larger standard deviations make data cloud look more spread around a line
- Correlation only measures *linear* association

Correlation Doesn't Always Imply Causation

- The fact that two things tend to move together does not imply one causes the other
- If correlation always implies causation...
  1. Having telephone increases the chance of breast cancer death (correlation across 165 countries is 0.74)
  2. Population increase increases the chance of lung cancer death (correlation is 0.92 in the US over the last 50 years)
  3. Gaining weight makes you vote for Bush or Voting for Bush makes you fat (correlation between Bush's voteshare and self-reported obesity rates in states is 0.40)
- Predictive inference vs. Causal inference
- When does correlation imply causation?

Ecological Correlation

- Group-level correlation $\neq$ Individual-level correlation
- Robinson's example: literacy and race
  - Ecological correlation $= 0.946$;
  - Individual-level correlation $= 0.203$
- Another example: foreign born and literacy
  - Ecological correlation $= 0.118$
  - Individual-level correlation $= -0.619$
- Other examples:
  - Individual voting behavior and aggregate electoral data
  - Exposure and disease rates
- Ecological correlation tends to overstate the association
- Why researchers often use ecological correlation?: aggregate data are easier to obtain

The Source of Ecological Fallacy

- A numerical illustration
- Consider three districts with ten voters in each
  1. 2 Blacks (2/2), 8 Whites (6/8)
  2. 8 Blacks (1/8), 2 Whites (1/2)
  3. 3 Blacks (2/3), 7 Whites (6/7)
- Ecological correlation $= -0.99$
- Individual correlation $= -0.38$

- Contextual effects: blacks in the minority-majority districts vote less than blacks in the other districts
- Unobserved factors other than race: education, income etc.

Linear Regression Model

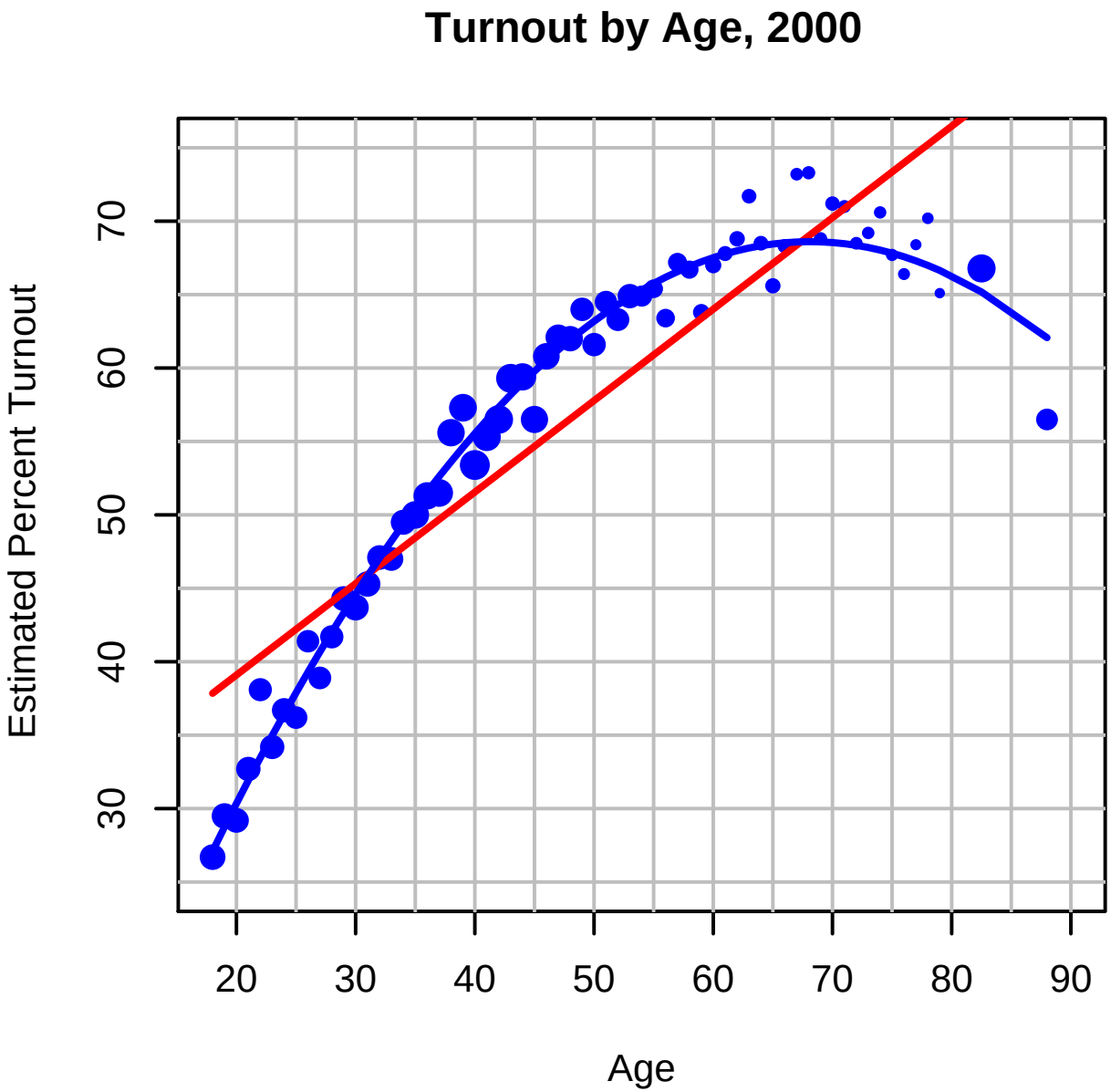
- READING: FPP Chapters 10 – 12
- A model for a linear relationship between two variables

$$Y_i = \alpha + \beta X_i + \epsilon_i$$

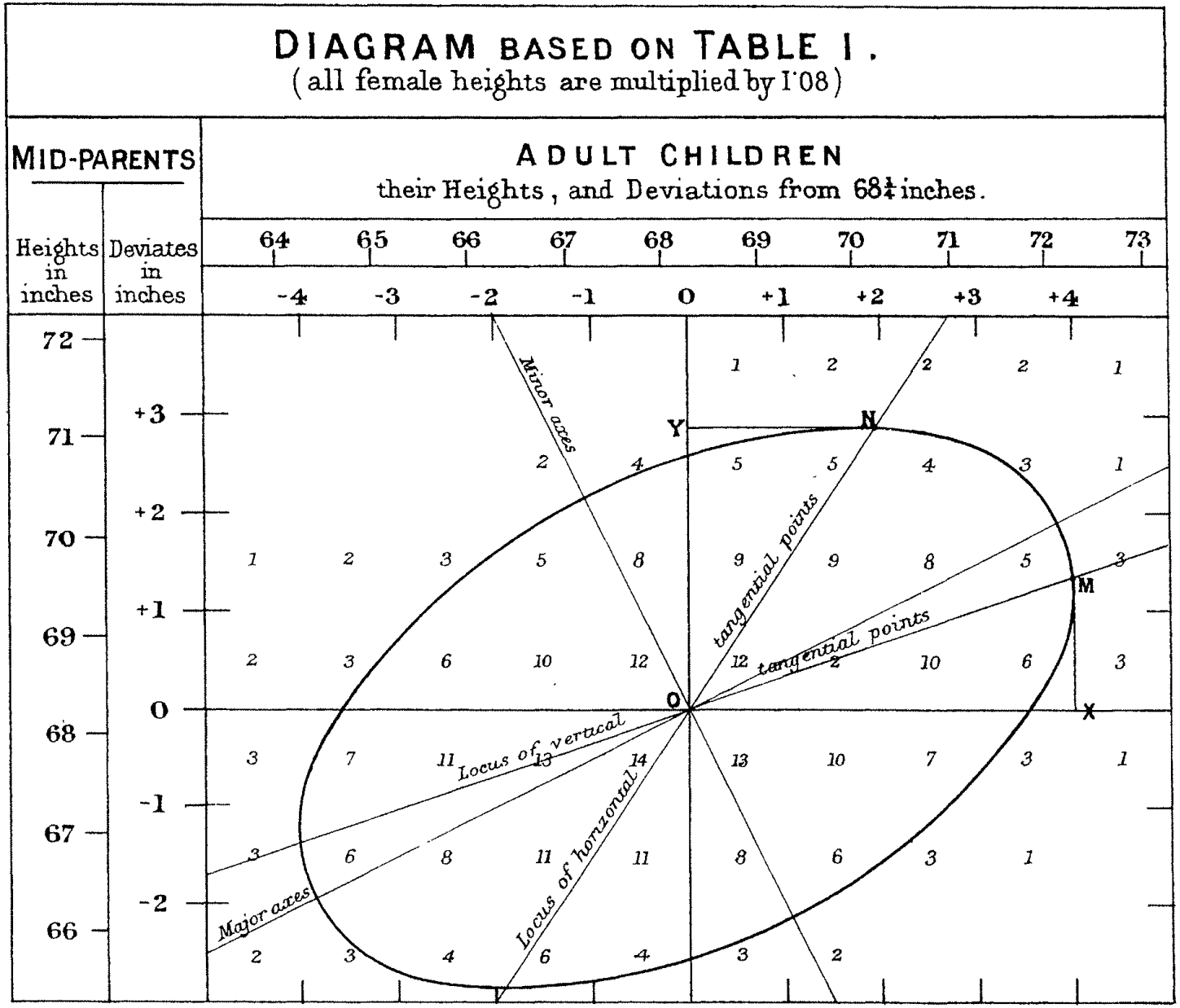
- X: Independent (explanatory) variable
- Y: Dependent (outcome, response) variable
- ϵ: error (disturbance) term

- Given a value of X, the model predicts the average of Y
- Abuse of regression: extrapolation, causal misinterpretation

Linear Model and Nonlinear Relationship



Galton (1886)'s Regression



The Regression Effect

- Also called regression towards mean
- "When mid-parents are taller (shorter) than mediocrity, their children tend to be shorter (taller) than they"
- Galton's calculation illustrated:

$$\frac{(8+5+3)}{(1+2+3+5+8+9+9+8+5+3)} = 0.30$$

$$\frac{(13+10+7+3+1)}{(14+13+10+7+3+1+3+7+11+13)} = 0.41$$

Analysis of Regression Effect

- If a student scores higher than average in the midterm, then her final score is likely to be less than the midterm score.
- A Regression Model: $Y_i = 15 + 0.8X_i + \epsilon_i$
  1. Y_i: Final exam score for student i (percent)
  2. X_i: Midterm exam score for student i (percent)
  3. ϵ_i is normally distributed with mean $= 0$ and SD $= 5$
- Two group of students: $X_1 = 60$ and $X_2 = 80$
- Which group of students is likely to do better in the final?
  - when compared with the other group
  - when compared with their own midterm score

Prediction and Prediction Error

- Model parameters: (α, β)
- Estimates: $(\hat{\alpha}, \hat{\beta})$
- Predicted (fitted) value given X_i:

$$\widehat{Y}_i = \hat{\alpha} + \hat{\beta} X_i$$

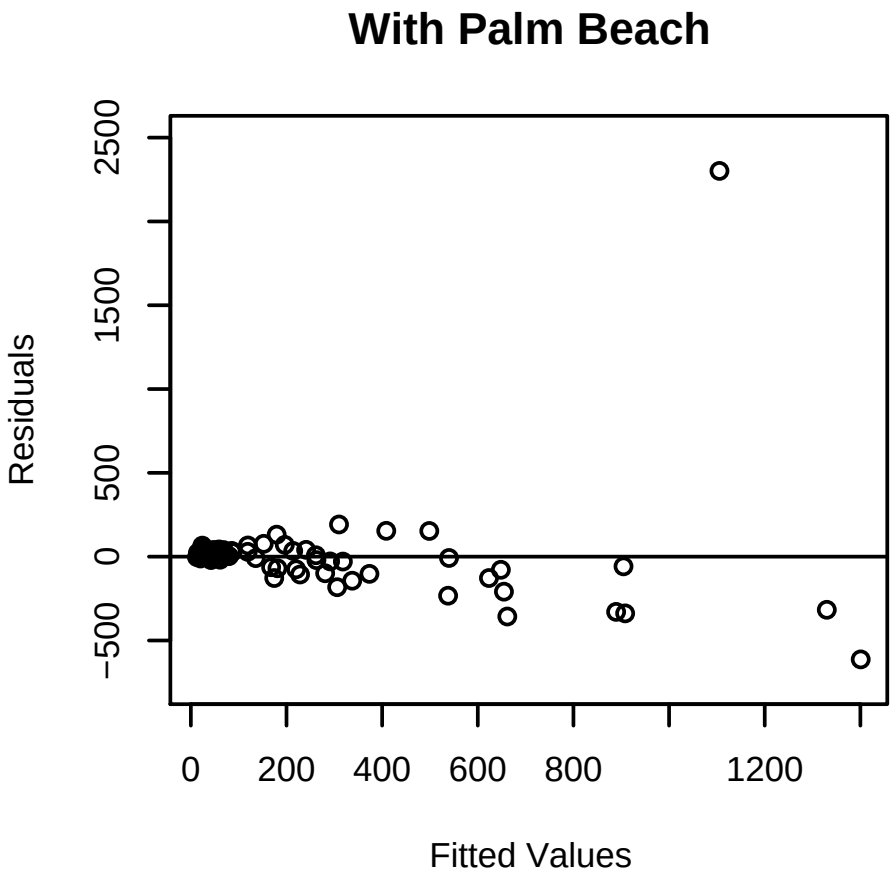
- Prediction error (residual) $=$ actual $-$ predicted:

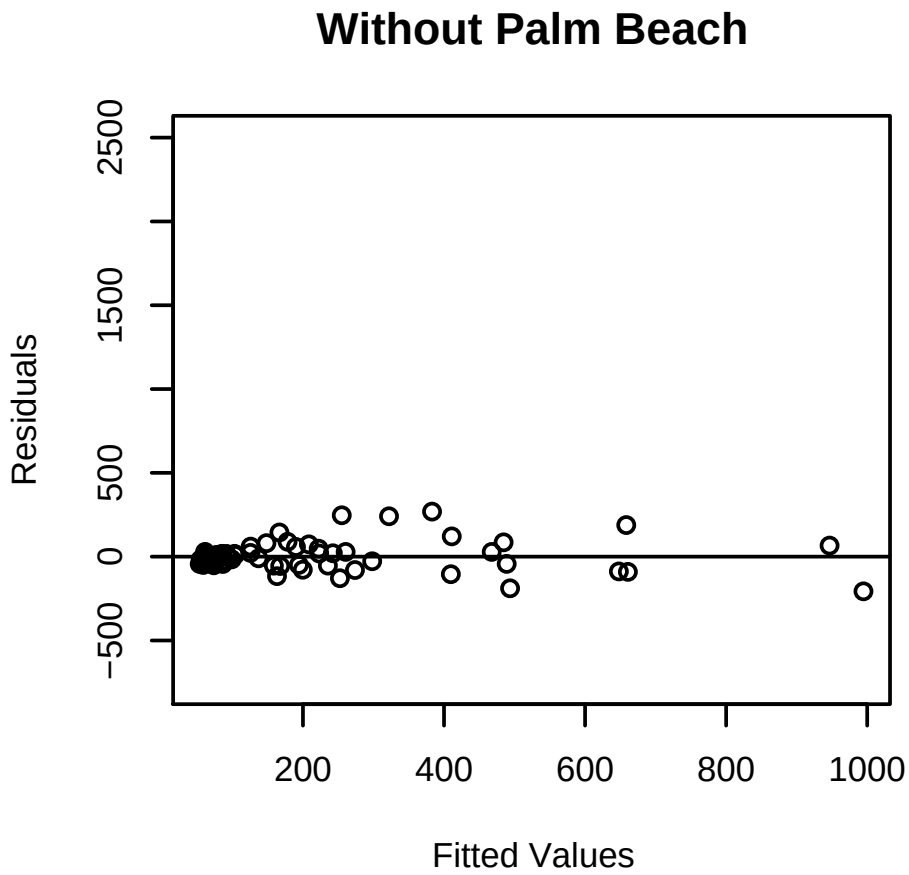
$$\hat{\epsilon}_i = Y_i - \widehat{Y}_i = Y_i - (\hat{\alpha} + \hat{\beta} X_i)$$

- r.m.s. of residuals $= \sqrt{1 - r^2} S_Y$
  - zero with perfect correlation $r = \pm 1$
  - S_Y with zero correlation

Residuals and Residual Plots

- Mean of residuals: $\frac{1}{n}\sum_{i=1}^{n}\hat{\epsilon}_i = 0$
- Does the regression model systematically overpredict or underpredict for some observations?
- Heteroskedastic errors?





The Method of Least Squares

- How do we obtain $\hat{\alpha}$ and $\hat{\beta}$?
- Minimize the sum of square errors (residual sum of squares):

$$SSE = \sum_{i=1}^{n} \hat{\epsilon}_i^2$$

which yields

$$\hat{\alpha} = \overline{Y} - \hat{\beta}\overline{X}, \quad \text{and} \quad \hat{\beta} = \frac{\sum_{i=1}^{n}(Y_i - \overline{Y})(X_i - \overline{X})}{\sum_{i=1}^{n}(X_i - \overline{X})^2}$$

- Equivalent to minimizing r.m.s. of residuals

Regression Line

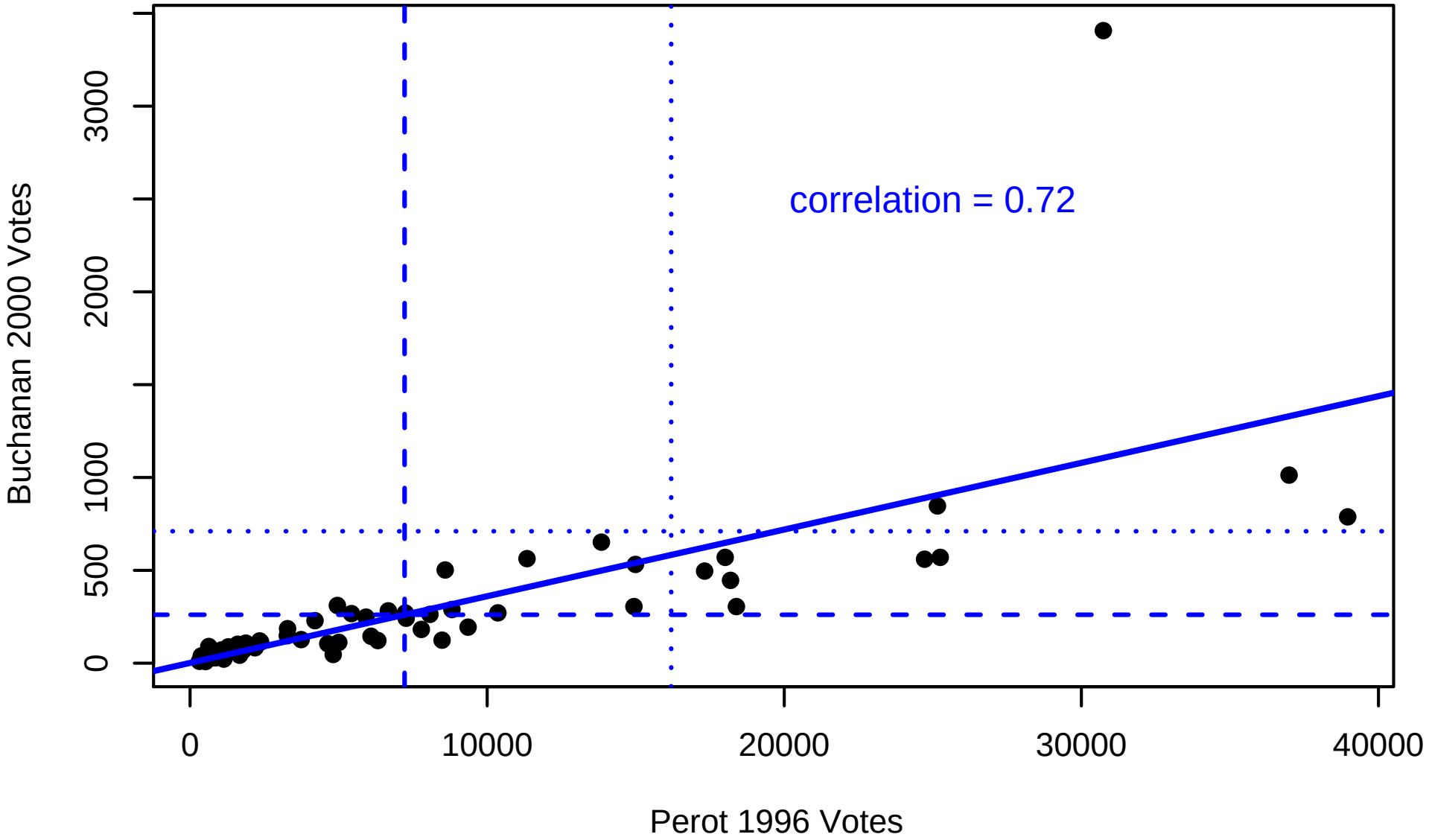
- Regression line goes through $(\overline{X}, \overline{Y})$
- Regression, standard deviation, and correlation:

$$\hat{\beta} = r \times \frac{S_Y}{S_X}$$

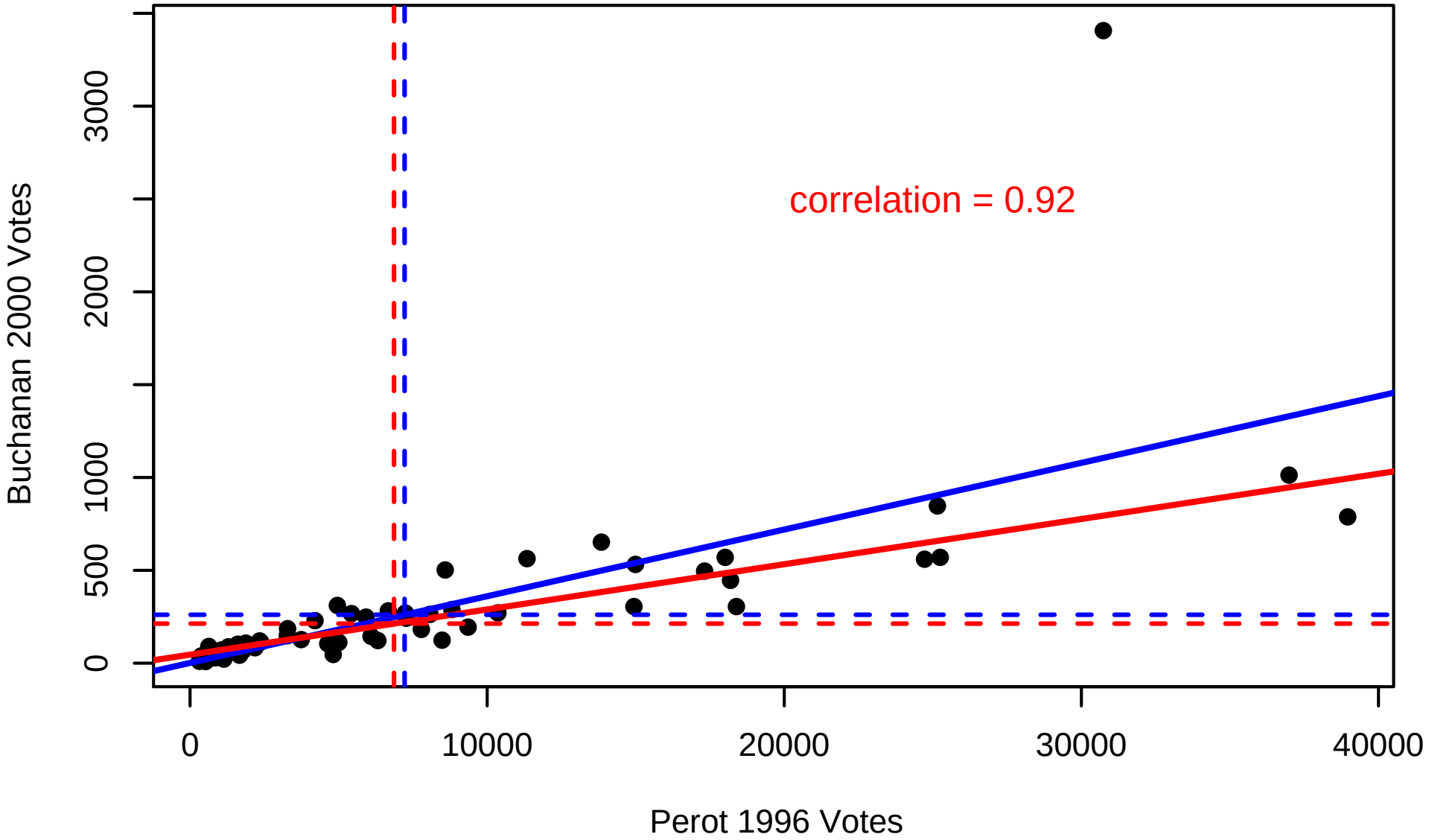
- One SD increase in $X \longrightarrow r$ SD increase in Y on average
- A practice problem: Correlation between midterm and final exams is 0.8. A student is 80 percentile on the midterm. Predict her percentile on the final exam using normal approximation.

- Regression line and correlation are based on averages $\Longrightarrow$ sensitive to outliers

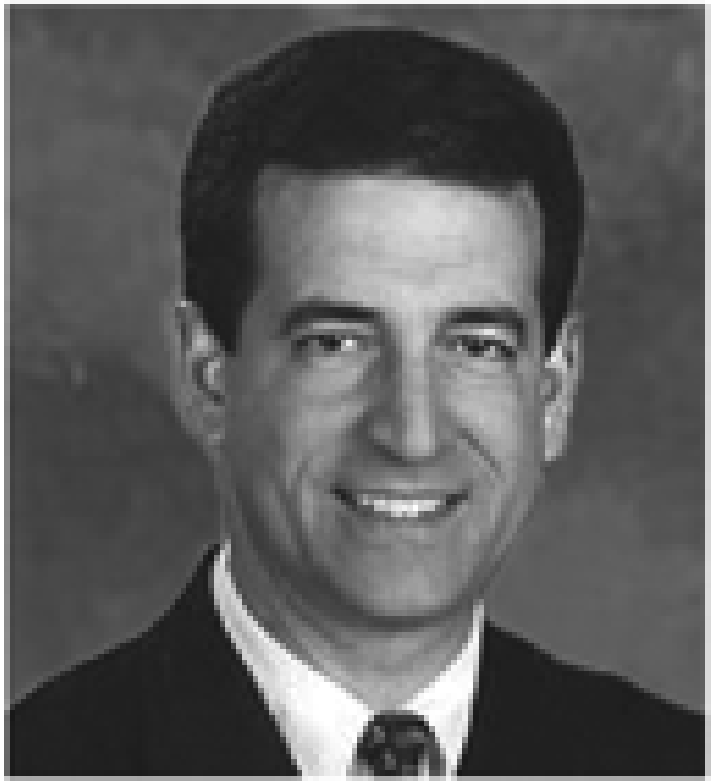
Regression, Correlation, and SD



Sensitivity to Outliers

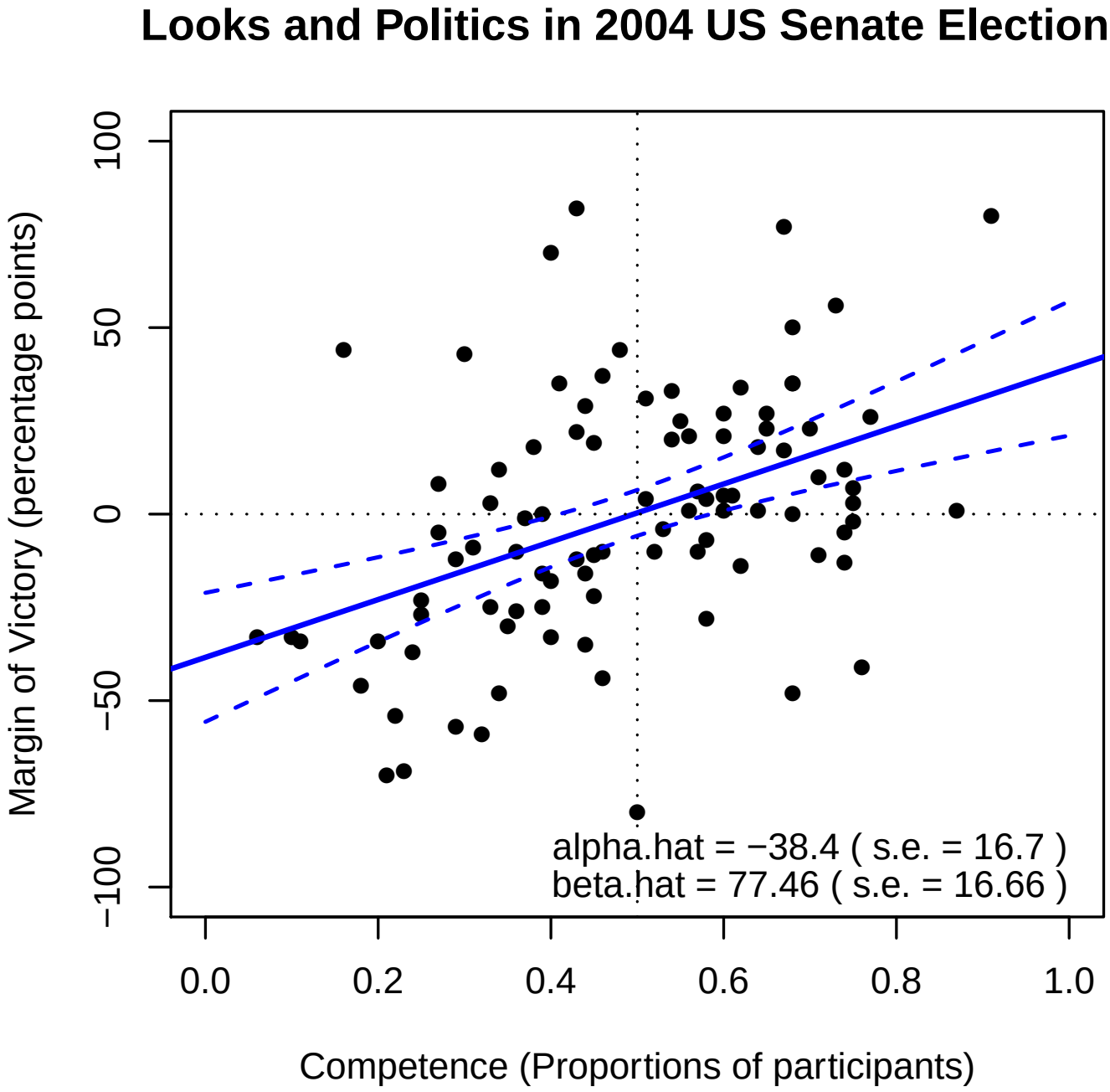


Looks and Politics



Which person is the more competent?

Todorov et al. Science


Looks and Politics in 2004 US Senate Elections
Margin of Victory (percentage points)
100
50
0
−50
−100
0.0
0.2
0.4
0.6
0.8
1.0
Competence (Proportions of participants)
alpha.hat = −38.4 (s.e. = 16.7)
beta.hat = 77.46 (s.e. = 16.66)

Another Example