Lecture 3: Measures of spread and central tendency

Jacob M. Montgomery

Quantitative Political Methodology

Lecture 3

Class business

- Make sure you review online materials before lab
- Problem set 1 is due on Wed.

Facebook and survey

- Sign up for our Facebook group:
  `https://www.facebook.com/groups/1071702902960687/`
- Take the class survey! Can't assign teams until you all do.

`https://wustl.az1.qualtrics.com/jfe/form/SV_6rpSYD3xxmbRe5v`

Roadmap

Last time:

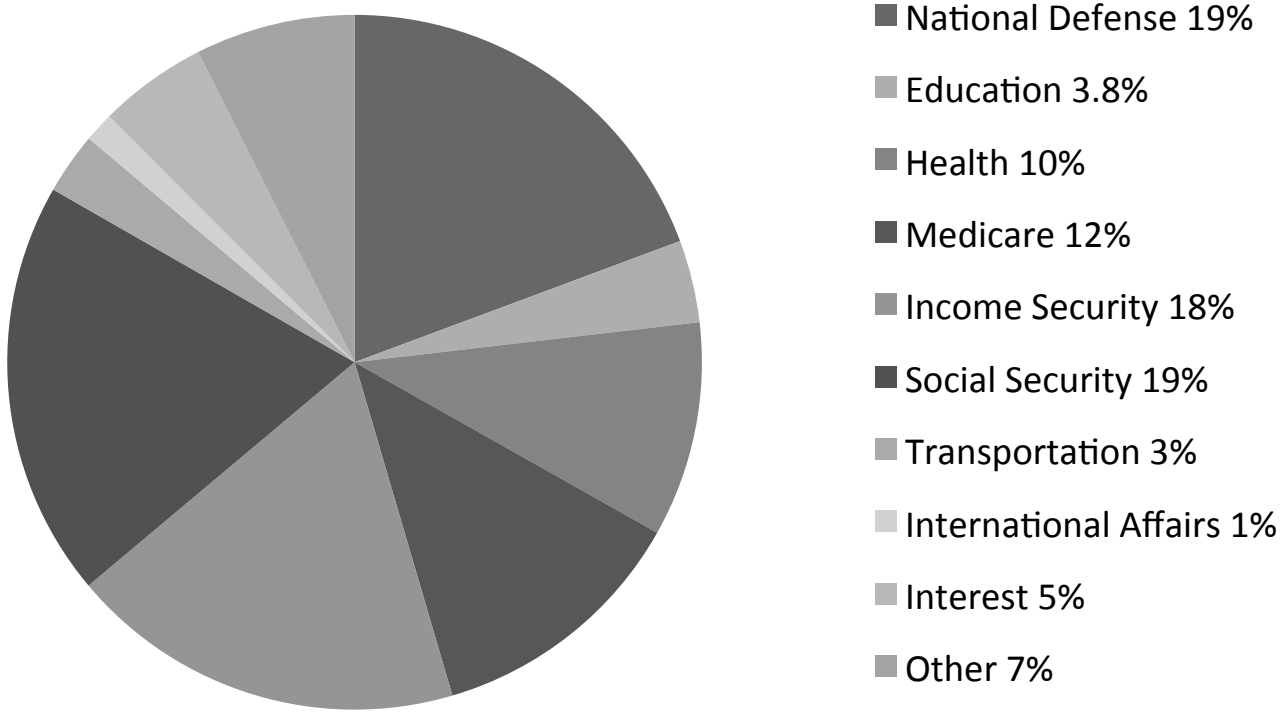
- Populations vs. samples
- Why sample?
- How to sample
- Problems in surveys

This time:

- Visualizing data
- Measures of central tendency and spread

Pie charts = mostly aweful

FIGURE 6.4a
Federal Spending by Category, 2010 (percent)

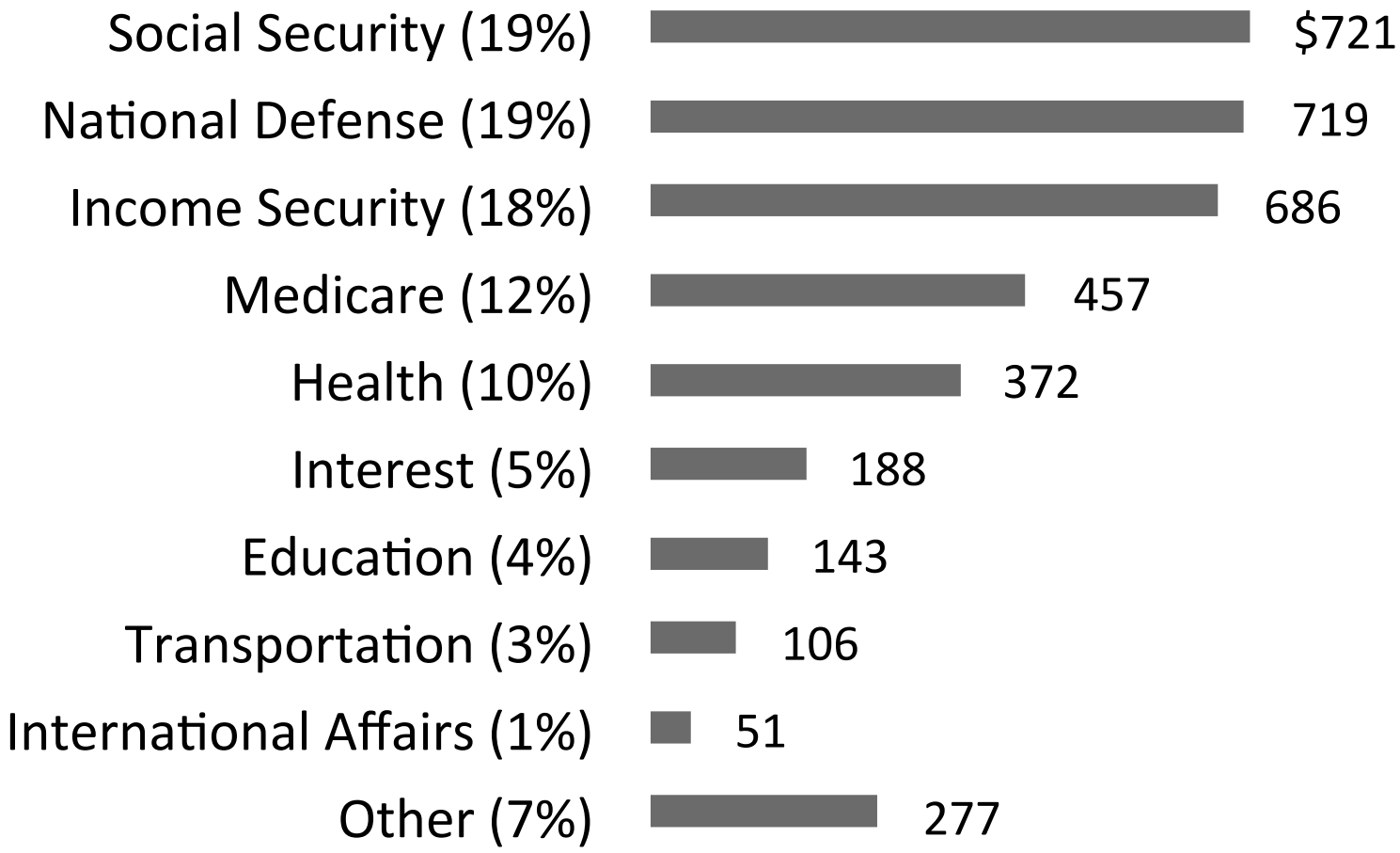


Source: *Budget of the U.S. Government, Fiscal Year 2011*

(Klass 2012)

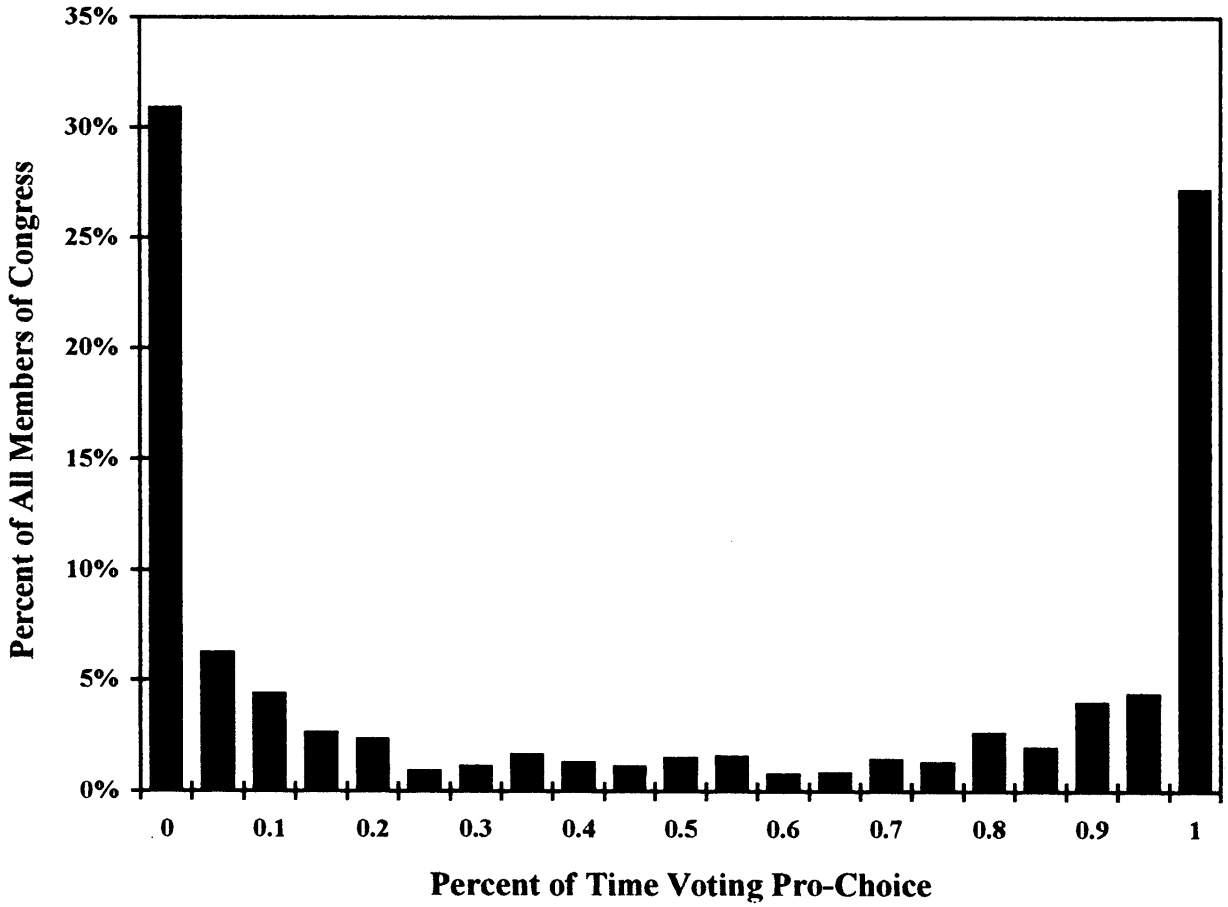
Bar charts – more useful

FIGURE 6.4b
Federal Expenditures by Function, 2010 (Billions)



Histograms – usually good

Figure 3. Distribution of Individual Legislators' Abortion Votes Over Entire Career



(Adams 1997)

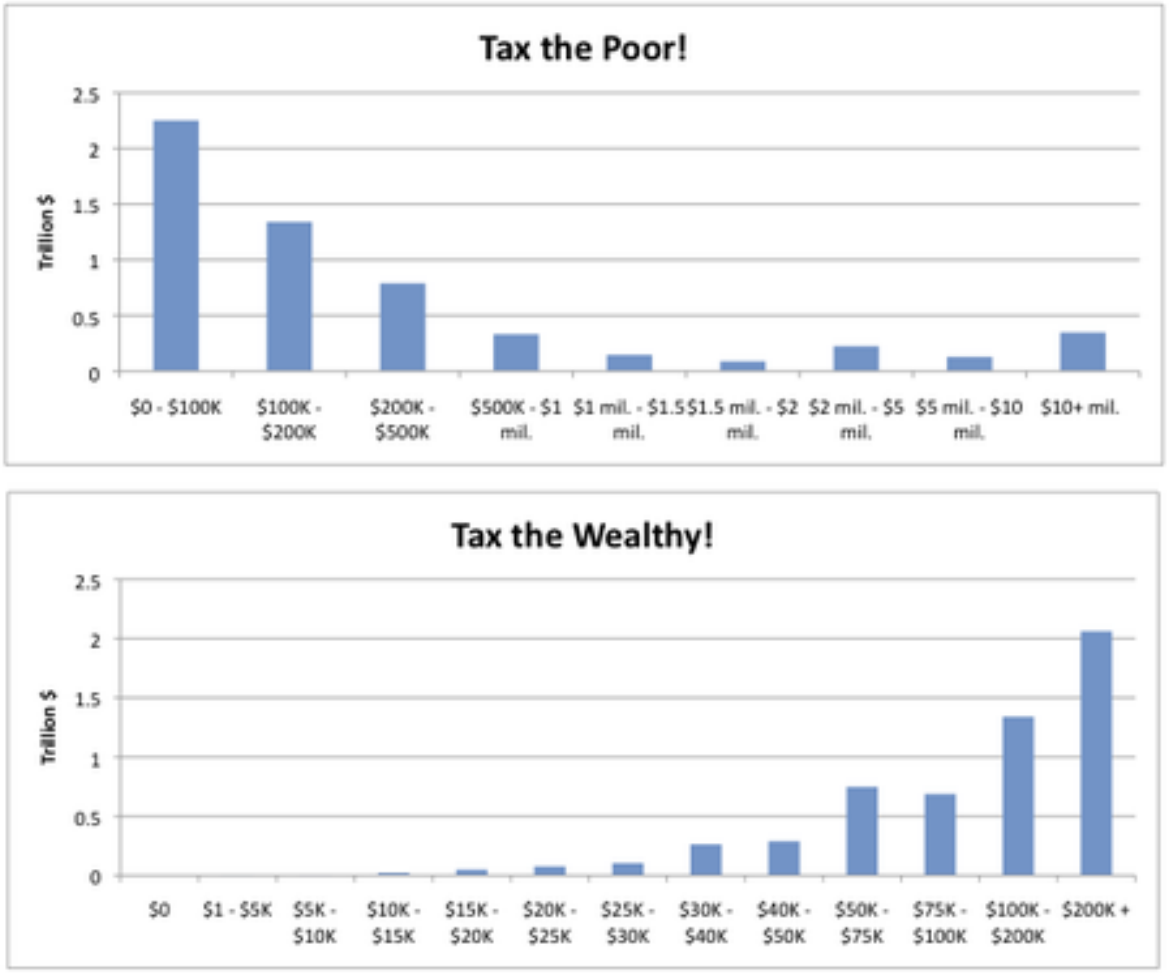
Making a histogram

	State	Median Income
1	Alabama	40489
2	Alaska	66953
3	Arizona	48745
4	Arkansas	37823
5	California	58931
6	Colorado	55430
7	Connecticut	67034
8	Delaware	56860
	...	
50	Wisconsin	49993
51	Wyoming	52664
52	Puerto Rico	18314

Making a histogram

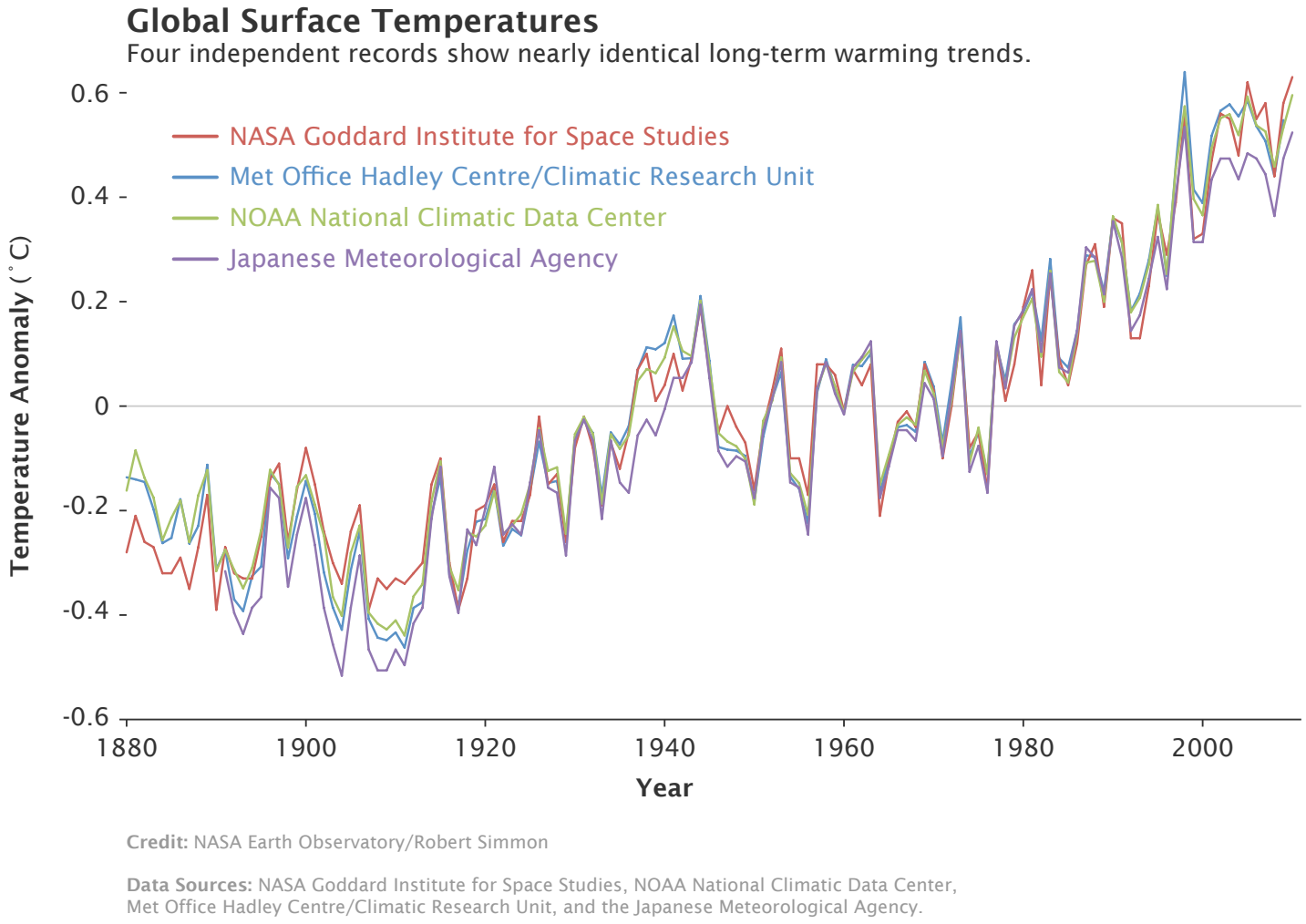
Class	Count
\$0-\$9,999	0
\$10,000-\$19,999	1
\$20,000-\$29,999	0
\$30,000-\$39,999	3
\$40,000-\$49,999	27
\$50,000-\$59,999	14
\$60,000-\$69,999	7
\$70,000-\$79,999	0

Binning can be evil



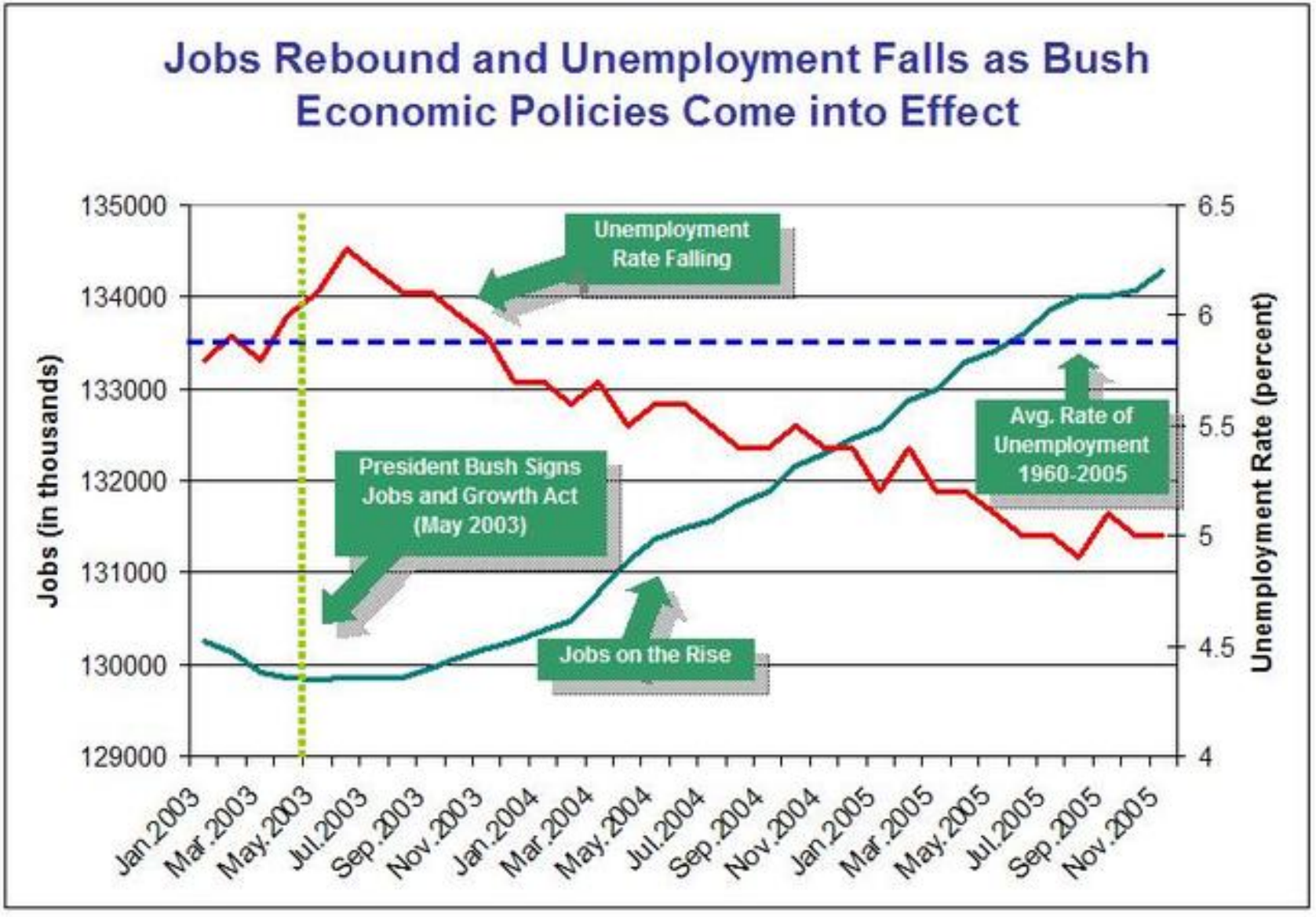
(Ken Schultz)

Line plots are often useful

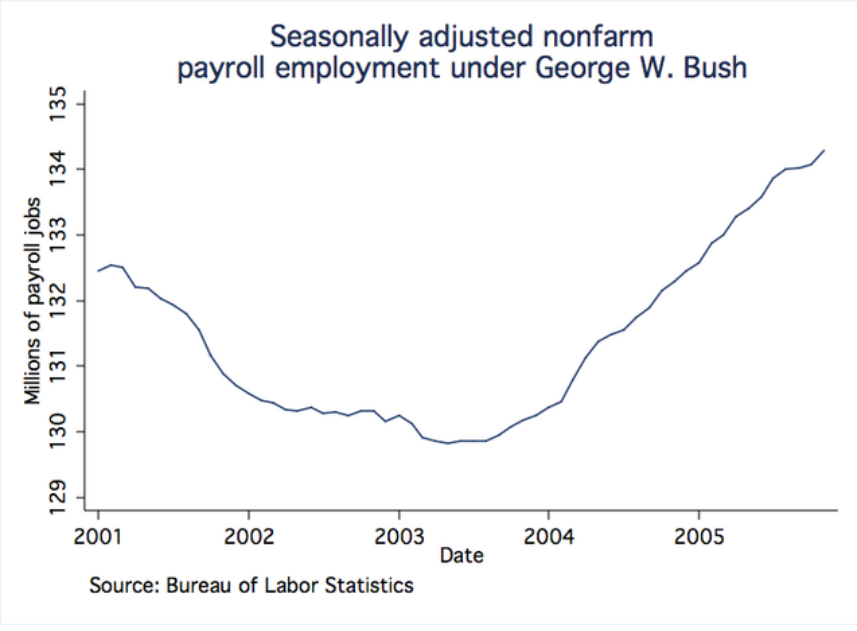


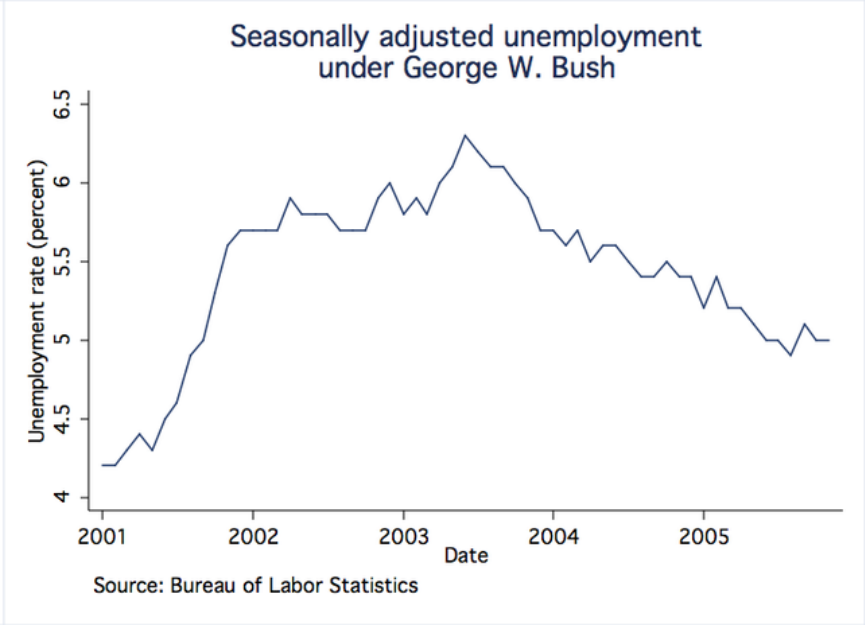
(NASA 2011)

Can also be misleading



Can also be misleading





(Nyhan 2005)

Looking at distributions of data

- General impressions (pattern, deviations)
- `Shape'' (symmetric, skewed,`bell-shaped," bimodal, etc.)
- `Center'' (`midpoint")
- "Spread" (range)
- `Outliers'' (`outside overall pattern")

Our first statistics – the mighty mean

$$n = \quad \text{Sample size}$$

$$y_1, y_2, \ldots, y_n = \quad \text{Observations}$$

Our first statistics – the mighty mean

$$
\begin{aligned}
n &= \quad \text{Sample size} \\
y_1, y_2, \ldots, y_n &= \quad \text{Observations}
\end{aligned}
$$

$$
\bar{y} = \frac{y_1 + y_2 + \ldots + y_n}{n} = \frac{1}{n} \sum_{i=1}^{n} y_i
$$

Our first statistics – the mighty mean

$$\begin{aligned} n &= \text{Sample size} \\ y_1, y_2, \ldots, y_n &= \text{Observations} \end{aligned}$$

$$\bar{y} = \frac{y_1 + y_2 + \ldots + y_n}{n} = \frac{1}{n}\sum_{i=1}^{n} y_i \equiv \frac{1}{n}\sum y_i$$

Test tip

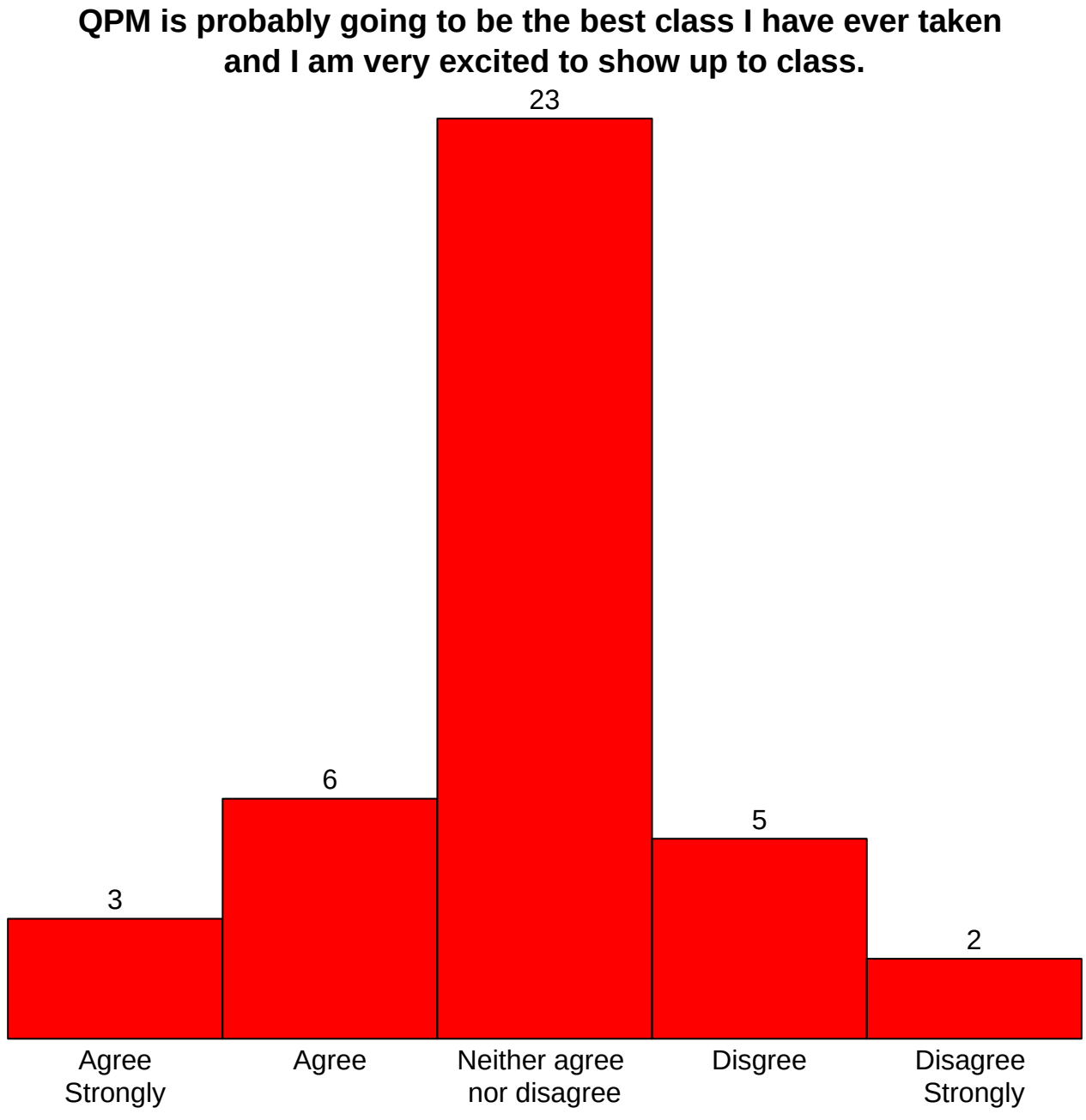
Grand mean of two samples $\bar{y} = \frac{n_1\bar{y}_1 + n_2\bar{y}_2}{n_1 + n_2}$

Test tip

Grand mean of two samples $\bar{y} = \frac{n_1\bar{y}_1+n_2\bar{y}_2}{n_1+n_2}$

What if I had j samples? $\bar{y} = \frac{n_1\bar{y}_1+n_2\bar{y}_2+...+n_j\bar{y}_j}{n_1+n_2+...+n_j}$

Finding the median of grouped data



Test tip: Find the median of grouped data

Category	Frequency	Percentage	Cumulative %
Agree Strongly	6	7.4	7.69
Agree	27	33.3	40.7
Neither agree nor disagree	18	22.2	63.0
Disagree	19	23.5	86.4
Disagree Strongly	11	13.6	100.00

Test tips: Standard deviation ($\sqrt{Var}$)

Standard deviation:

$$S = \sqrt{S^2} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}}$$

Test tips: Standard deviation ($\sqrt{Var}$)

Standard deviation:

$$S = \sqrt{S^2} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}}$$

- One standard deviation will be equal to something in original units (e.g., 2 inches)

Test tips: Standard deviation ($\sqrt{Var}$)

Standard deviation:

$$S = \sqrt{S^2} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}}$$

- One standard deviation will be equal to something in original units (e.g., 2 inches)
- $S \geq 0$ Why?

Test tips: Standard deviation ($\sqrt{Var}$)

Standard deviation:

$$S = \sqrt{S^2} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}}$$

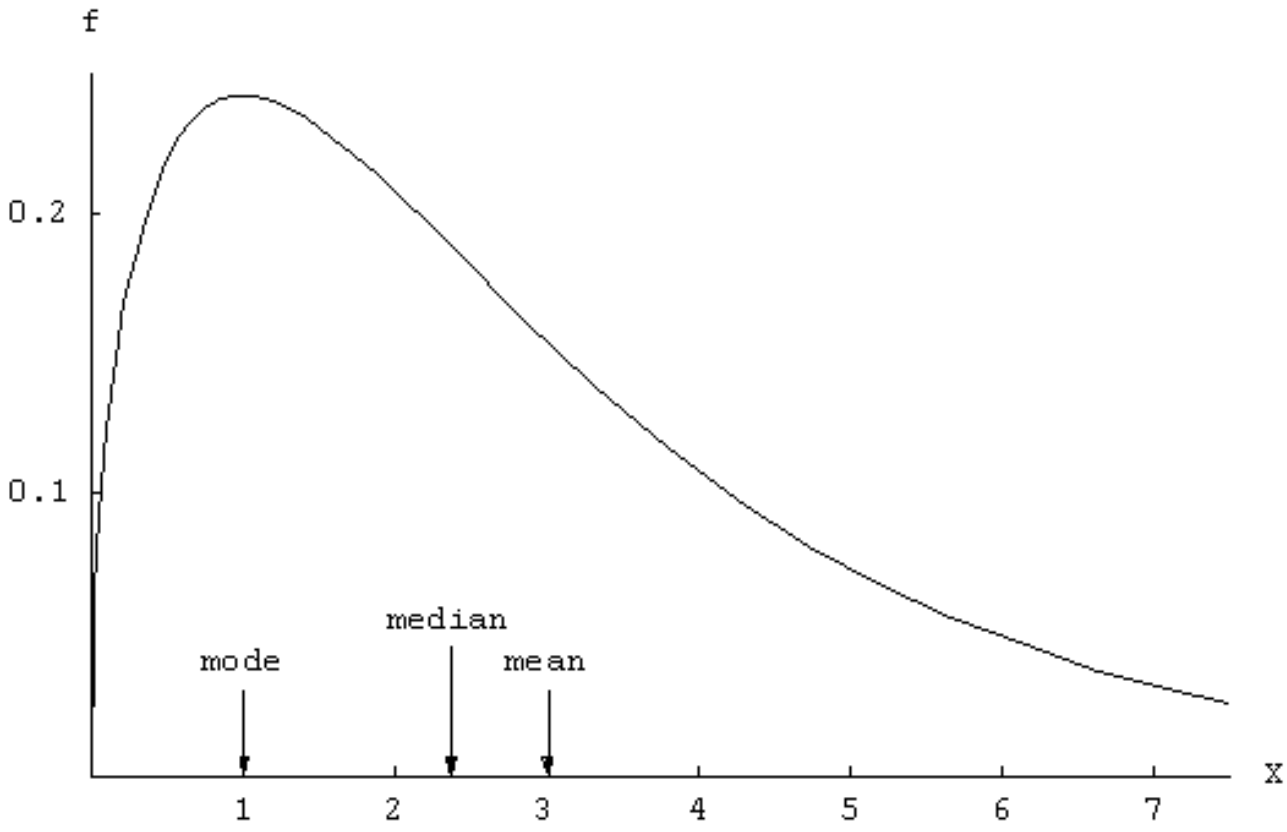
- One standard deviation will be equal to something in original units (e.g., 2 inches)
- $S \geq 0$ Why?
- $S = 0$ only if y is a constant.

Test tip on calculating variance

y_i	$(y_i - \bar{y})$	$(y_i - \bar{y})^2$
0	-5	25
4	-1	1
4	-1	1
5	0	0
7	2	4
10	5	25
$\sum y_i = 30$	$\sum(y_i - \bar{y}) = 0$	$\sum(y_i - \bar{y})^2 = 56$
$\bar{y} = 5$		

- $S^2 = \frac{56}{5} = 11.2$
- $S = \sqrt{11.2} = 3.3$

A little visulaization



Quantiles

- (55, 84, 65, 54, 61, 67, 80, 59, 81, 82)
- 54 55 59 61 65 67 80 81 82 84
- Median $= \frac{67-65}{2}$
- 50% of the data is at *or above*, and (100-50)% is at *or below*

Quantiles in R

It turns out that calculate percentiles is a more nuanced topic than I realized.

- R includes nine different methods for calculating percentiles
- We want to use

```
quantile(c(55, 84, 65, 54, 61, 67, 80, 59, 81, 82),
         probs=c(.25, .75),
         type=2)
```

```
## 25% 75%
##  59  81
```

Quantiles in R

It turns out that calculate percentiles is a more nuanced topic than I realized.

- R includes nine different methods for calculating percentiles
- We want to use

```
quantile(c(55, 84, 65, 54, 61, 67, 80, 59, 81, 82),
         probs=c(.25, .75),
         type=2)
```

```
## 25% 75%
##  59  81
```

- For more information:
tolstoy.newcastle.edu.au/R/e17/help/att-1067/Quartiles_in_R.pdf

Looking forward