Chapter 2

Describing Variables

2.1 Frequency Distributions for Discrete and Continuous Variables

2.2 Grouped and Cumulative Distributions

2.3 Graphing Frequency Distributions

Frequency Distributions

Frequency distribution: a table of outcomes (response categories) of a variable and the number of times [tally or count] each outcome is observed.

A frequency distribution shows the total number of persons responding to each of the variable's K categories.

Relative f.d. (= proportion): divide tally by total N of cases

Percentage f.d. shows proportions multiplied by 100%

Sum of all the percents = 100.0%

- **Tally (count) frequencies by hand or by calculator; or**
- **Use SPSS on GSS to tally frequencies & a print table**

ASTROSCI: Is Astrology Scientific?

GSS 2008: "Would you say that astrology is very scientific, sort of scientific, or not at all scientific?"

astrosci ASTROLOGY IS SCIENTIFIC

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 Very scientific	74	3.7	5.1	5.1
	2 Sort of scientific	434	21.5	30.1	35.2
	3 Not at all scientific	935	46.2	64.8	100.0
	Total	1443	71.3	100.0	
Missing	0 IAP	518	25.6		
	8 DONT KNOW	59	2.9		
	9 NO ANSWER	3	.1		
	Total	580	28.7		
Total		2023	100.0		

Calculating Relative Frequencies

Should you include or exclude cases with missing values when calculating a relative frequency distribution?

- SPSS "Percent" column includes all cases
- SPSS "Valid Percent" excludes any "Missing"

[0 = IAP; 8 = DK; 9 = NA]

For a variable with K categories, the valid N is the sum of the frequencies, f_i, across all K categories (where the subscript *i* indicates changing index values, from 1 to k) :

$$f_1 + f_2 + f_3 + \ldots + f_k = N$$

To find the **proportion** (relative frequency) in the *i*th category i, just divide f_i by valid N:

$$p_i = \frac{f_i}{N}$$

For ASTROSCI (exclude all Missing categories):

N = 74 + 434 + 935 = _____

p_1 = 74 / 1443 = _____

p_2 = 434 / 1443 = _____

p_3 = 935 / 1443 = _____

N = 1443 / 1443 = _____

Usually no more than four “significant digits” will be needed when calculating proportions; use rounding.

Calculating Percentages

To find the percentage in category i, multiply each p_i by 100%:

$$(p_i)(100\%) = \text{percent } \mathbf{i} = \mathbf{i}\%$$

$(p_1)(100\%) = (.0513)(100\%) =$ _____%

$(p_2)(100\%) = (.3008)(100\%) =$ _____%

$(p_3)(100\%) = (.6480)(100\%) =$ _____%

$(N)(100\%) = (1.0000)(100\%) =$ _____%

Percentages are typically rounded to the nearest tenth of one percent

See slide below on Rounding Rules

Grouped Distributions

Grouped data: continuous measures that have been collapsed into fewer categories

Measurement interval treats all cases that fall between the lower and upper limits as equal values

Use mutually exclusive & exhaustive limits:

- Each case falls into only one interval
- Every case is assigned somewhere

- *SSDA*: "Generally, between 6 and 20 intervals should be used…"

  Fewer than 10 intervals are preferable for simplicity

- Use SPSS RECODE to group adjacent categories together
- Label new category by the lower & upper limits of that interval

AGE in the 2008 GSS

Respondent's **AGE** is coded in years, 72 categories from 18 to 89 (and 10 cases with missing data, coded = 99).

Let's use these SPSS commands to collapse AGE into eight decades, by creating a new variable called **AGE10**:

```
COMPUTE age10 = age .

RECODE age10  (18 thru 19=1) (20 thru 29=2) (30 thru 39=3)
    (40 thru 49=4) (50 thru 59=5) (60 thru 69=6) (70 thru 79=7)
    (80 thru 89=8) (ELSE=SYSMIS) .

VARIABLE LABELS age10 'AGE IN DECADES' .

VALUE LABELS age10  1 '18-19' 2 '20-29' 3 '30-39' 4 '40-49'
     5 '50-59' 6 '60-69'  7 '70-79' 8 '80-89' .

FREQUENCIES VARIABLES = age age10 .
```

AGE Age of Respondent

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	18	6	.3	.3	.3
	19	31	1.5	1.5	1.8
	20	28	1.4	1.4	3.2
	***	***	***	***	***
	87	6	.3	.3	98.9
	88	1	.0	.0	99.0
	89+	21	1.0	1.0	100.0
	Total	2013	99.5	100.0	
Missing	99 NA	10	.5		
Total		4520	100.0		

*** 66 rows deleted here

AGE10 Age in Decades

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1.00 18-19	37	1.8	1.8	1.8
	2.00 20-29	322	15.9	16.0	17.8
	3.00 30-39	373	18.4	18.5	36.4
	4.00 40-49	381	18.8	18.9	55.3
	5.00 50-59	371	18.3	18.4	73.7
	6.00 60-69	272	13.4	13.5	87.2
	7.00 70-79	165	8.2	8.2	95.4
	8.00 80-89	92	4.5	4.6	100.0
	Total	2013	99.5	100.0	
Missing	System	10	.5		
Total		2023	100.0		

Which decade(s) has the most cases? _____

Which has the largest percentage? _____

Another Type of Grouped Data

Ordered frequency distributions may be tabled without collapsed any categories. Although each score doesn't involve a range from lower to upper limits, I also refer to such tabular displays as "grouped data" because each category represents numerous respondents:

NEWS HOW OFTEN DOES R READ NEWSPAPER	Frequency	Valid %
1 EVERYDAY	431	30.3
2 FEW TIMES A WEEK	300	21.1
3 ONCE A WEEK	297	20.9
4 LESS THAN ONCE WK	200	14.1
5 NEVER	191	13.5
Total	1419	100.0

Note the poor GSS practice of assigning higher numbers to lower-level activity! You should recode to reverse their order.

Cumulative Distributions

Cumulative frequency: for a given score or outcome of a variable, the total number of cases in the distribution at or below that value

Cumulation makes sense only for orderable discrete and continuous variables. Why should you never make a cumulative frequency distribution for a nonorderable discrete variable, such as race or state of residence?

Both cumulative frequency distributions and cumulative percentage distributions are created by adding the counts or the %s in the lower-valued categories

For an example, see the Cumulative Percent in the preceding AGE10 table

What % of 2008 GSS are < 60 years old? _____

Graphing Frequency Distributions

A **Graph** or **Diagram** visually summarizes the numbers in a frequency distribution or other table.

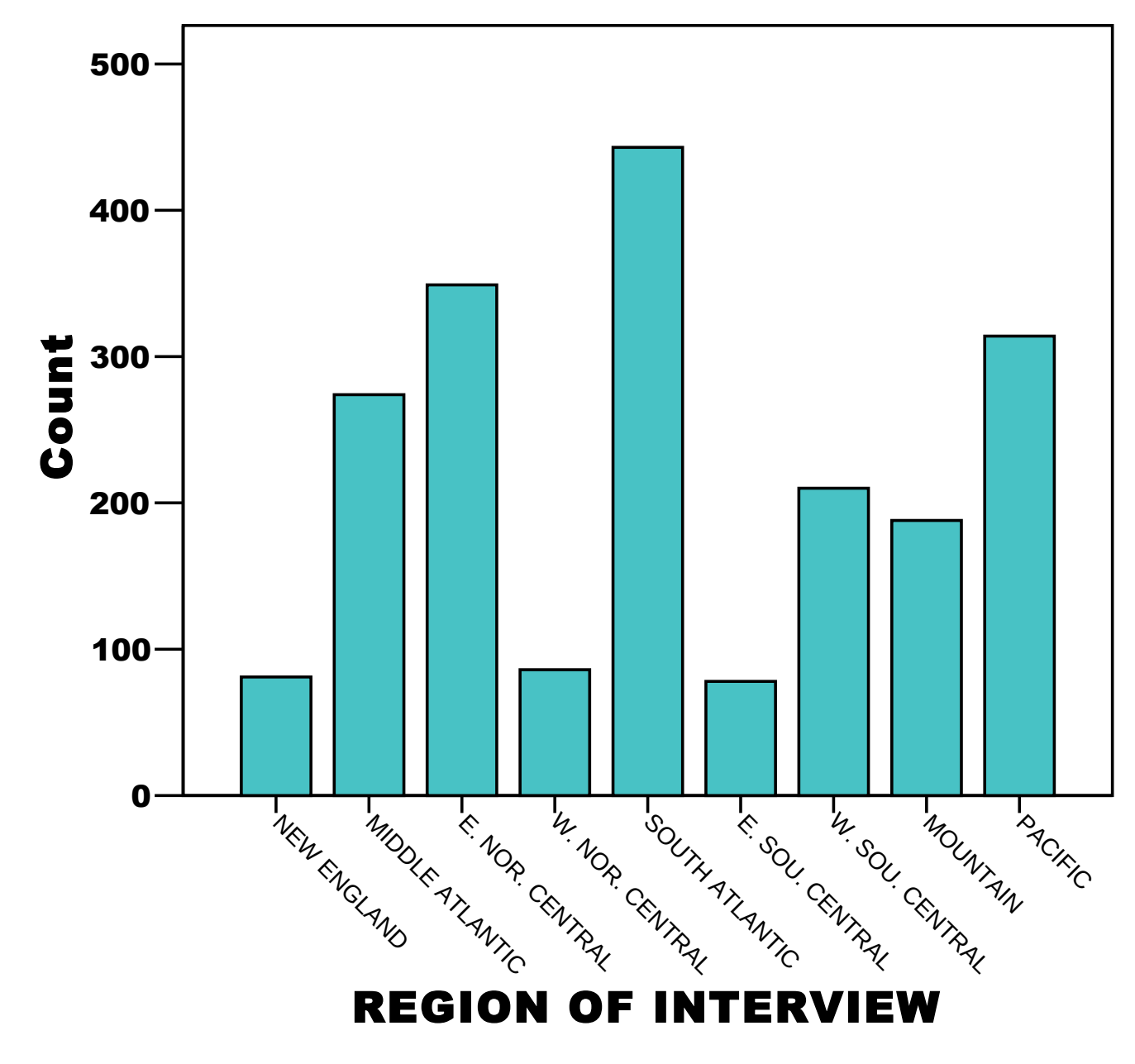
Three basic types of graphs:

BAR CHART	for nonordered discrete variables
HISTOGRAM	for ordered discrete variables
POLYGON	for continuous variables

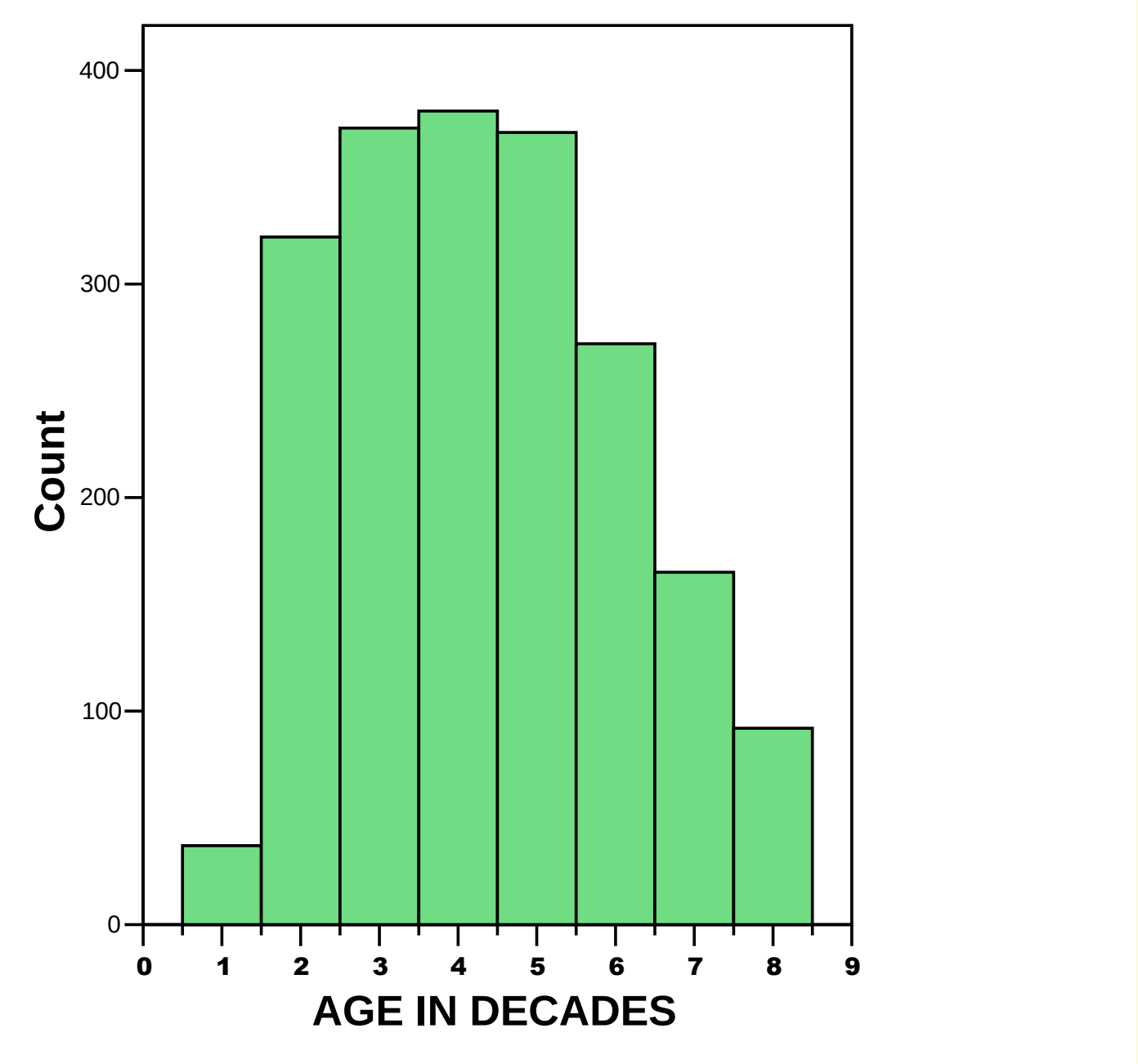
On the following slides, how do bar charts and histograms differ in the spaces between their bars? Why?

How does a histogram differ from a polygon?

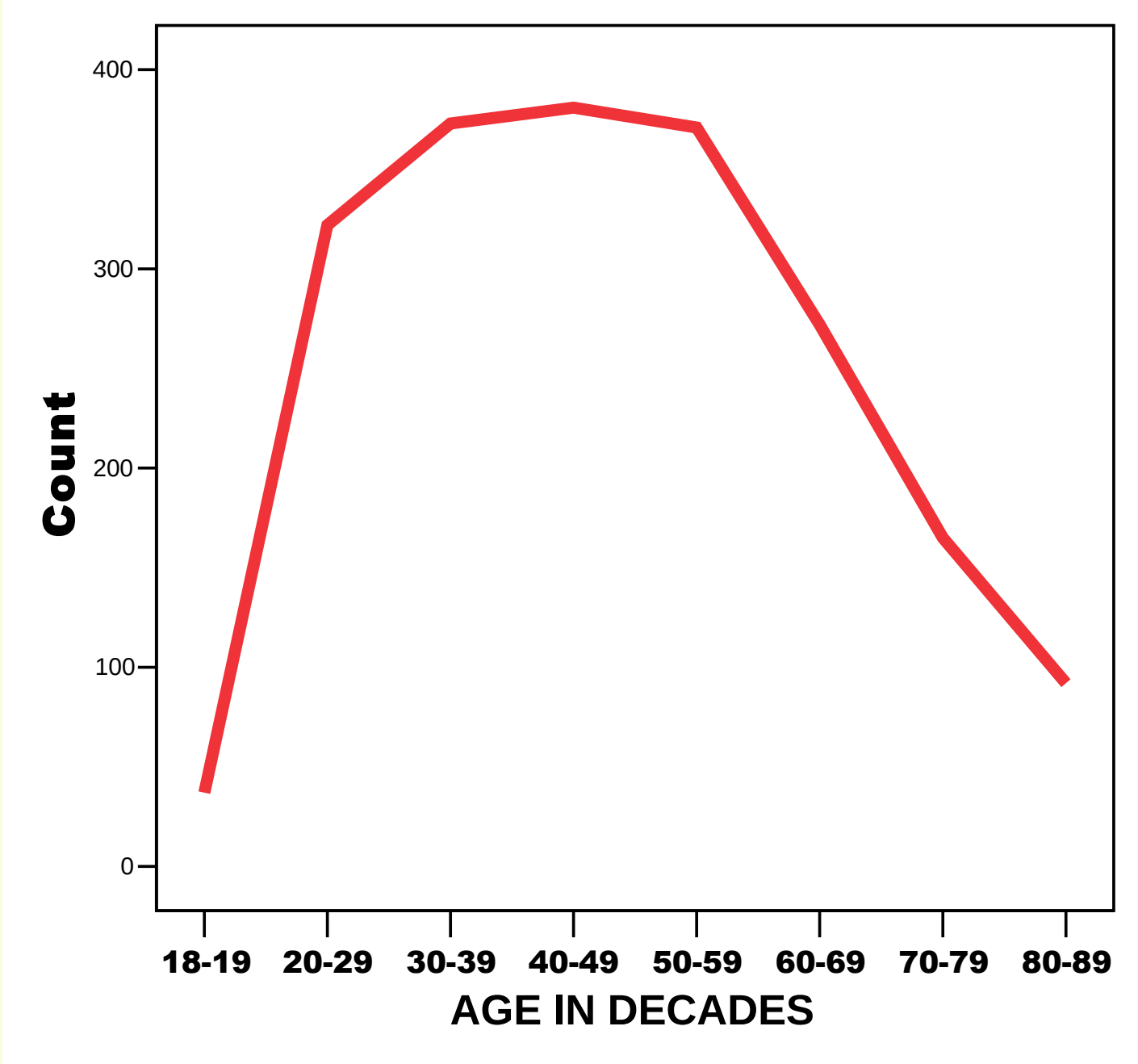
Bar Chart of REGION



Histogram of AGE10

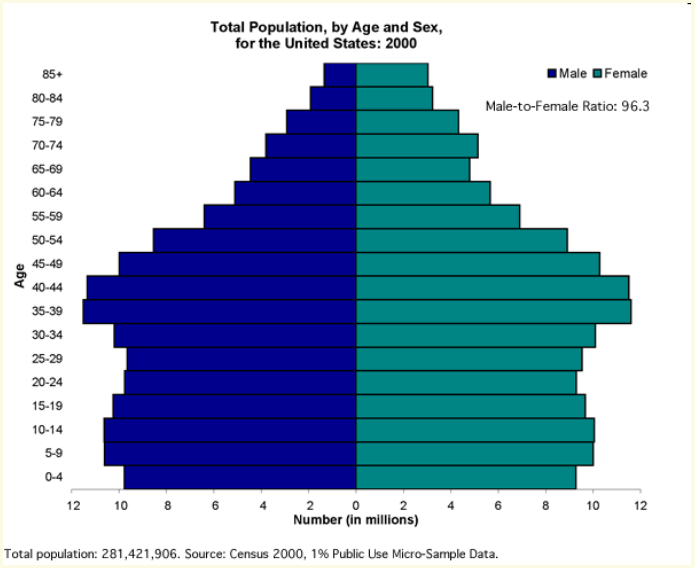


Polygon of AGE10

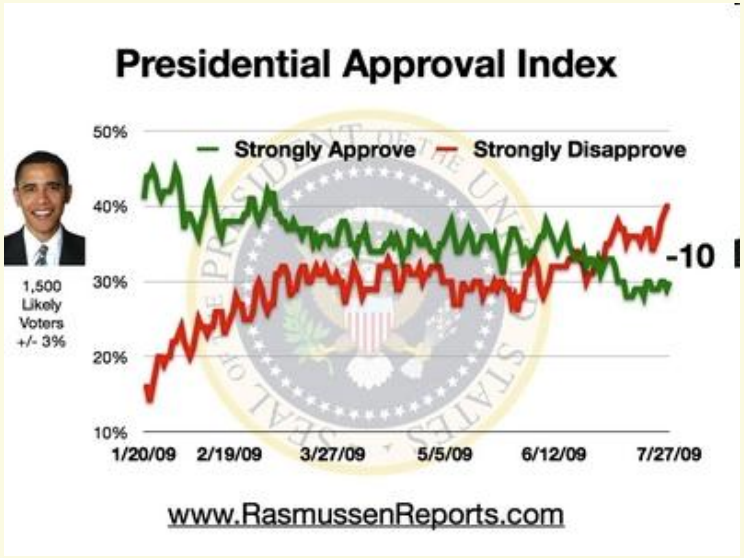


Variations on Basic Graphs

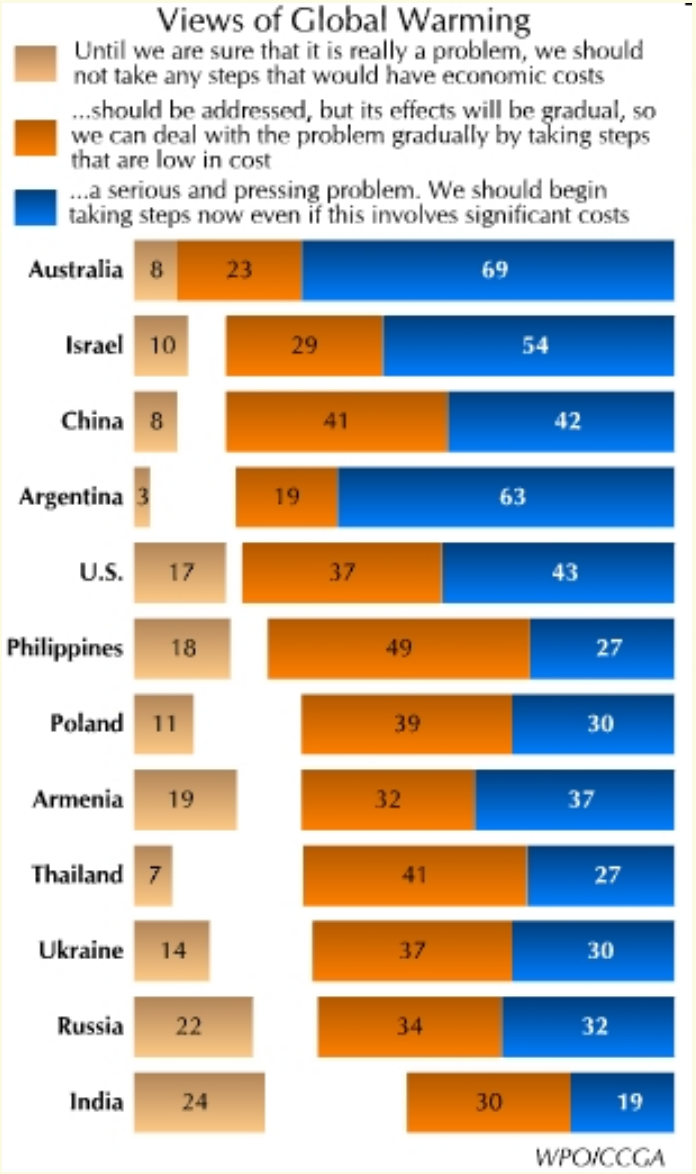
Two histograms: Age pyramids by sex



Two polygons: Approval over time



12 bar charts: Opinion by nation



ROUNDING RULES from Box 2.1

1. Round digits 1 to 4 down by leaving the digit to the left unchanged.
2. Round digits 6 to 9 up by increasing the digit to the left by 1.
3. Numbers ending in 5 are rounded alternately; the first number ending in 5 is rounded down, the second is rounded up, the third is rounded down, and so forth.
4. Never round past the original measurement interval.

Examples:

Unit of Measurement	Years (tenths)	Rounded No.
Years	36.6	37
Years	433.3	433
Decades	36.6	4
Decades	433.3	43
Centuries	36.6	0
Centuries	433.3	4