

Data Analysis: Interpretation and Pitfalls

Econ 102

Data Analysis

5	23	31	67	95	5	13	39
23	10	1	39	13	23	-3	56
31	-3	30	65	85	65	87	34
87	120	30	45	76	3	10	89
65	79	35	34	45	56	68	96
43	-34	39	23	34	44	40	-11
103	54	23	12	24	48	96	3

56 numbers, temperature

Need to summarize the information

What kind of information is valuable?

- range of values (difference between largest and smallest)
- most likely

Outliers; reduce their effect on overall impression

The word average has many meanings

5,3,8,10,8 ———▶ Observations; data set

Order them from lowest to highest

3,5,8,8,10

Mean = add all the observations, divide by the number of obs

$$34/5 = 6.8$$

Median: is the number which is in the exact middle of the data set.

Mode: **Mode is the number that appears the most often**

8

When is the mean a good indicator of the average?

Not much variation; Ames vs. LA

5,5,5,5,5 vs 2,1,200,3,4

Income: Mean of 5 vs 42

Median: 5 vs. 3 (50% point)

mean > median => wide dispersion, few have a lot

Mean < median ? grades

55	69	52.5	47	
34.5	36	50	61	
48	37	53	40	
33.5	15	42.5	37	
12	12	34	53	
58	60	57.5	67.5	
30	51	60	62.5	
70	33	2	21	
	mean =	43.58		
	median =	47.5		

Exactly 50% of students got below (above) 47.5

Data in economics

Cross section

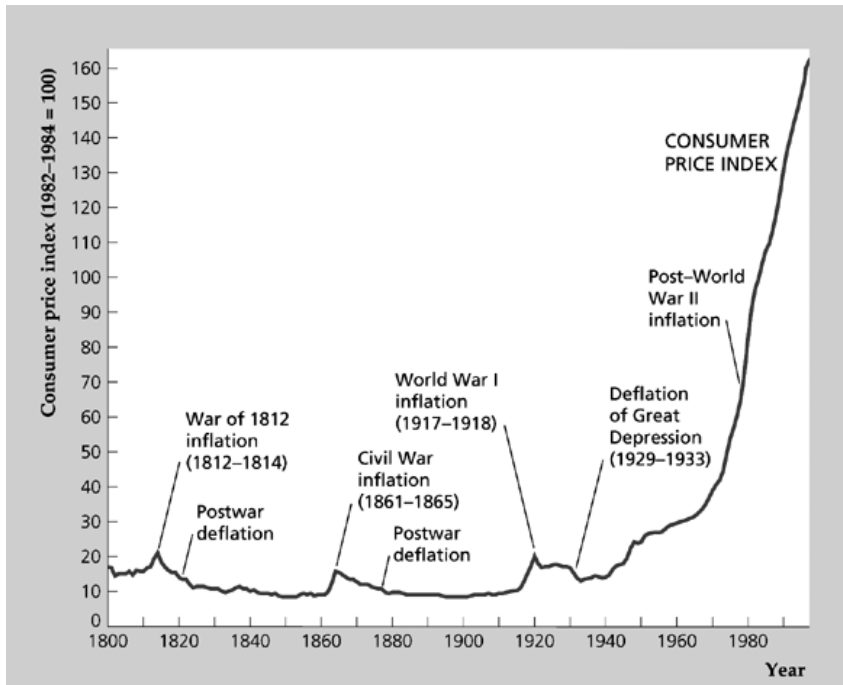
Population in 2002 across countries

Population 2002

Ranking Economy (thousands)

- 1 China 1,280,975
- 2 India 1,048,279
- 3 United States 288,369
- 4 Indonesia 211,716
- 5 Brazil 174,485
- 6 Pakistan 144,902
- 7 Russian Federation 144,071
- 8 Bangladesh 135,684
- 9 Nigeria 132,785
- 10 Japan 127,144
- 11 Mexico 100,021

Consumer prices in the United States, 1800-1998



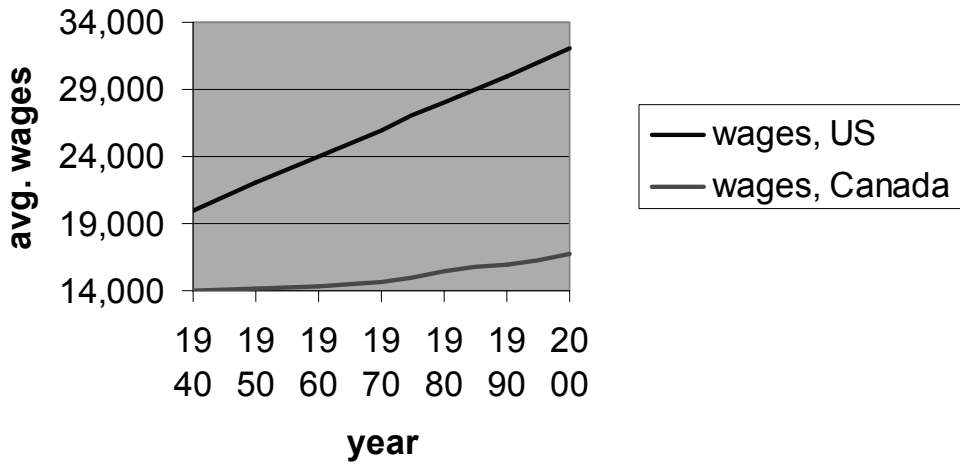
Time series: same country over time

Pitfalls of Data Analysis and Presentation

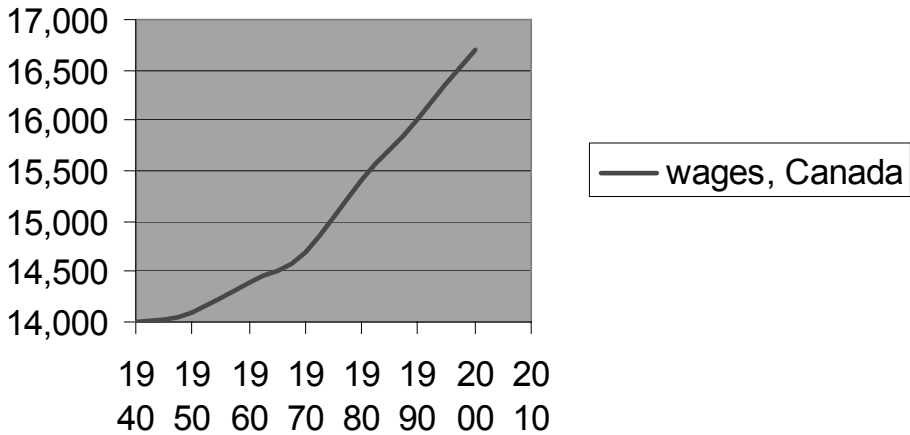
Scale matters

	wages, US	wages, Canada
1940	20,000	14,000
1950	22,000	14,100
1960	24,000	14,400
1970	26,000	14,700
1980	28,000	15,400
1990	30,000	16,000
2000	32,000	16,700

Wages over Time, Canada and US



wages, Canada



Growth rate:

$$\frac{(Y \text{ in } 2000 - Y \text{ in } 1999)}{Y \text{ in } 1999} * 100$$

Changes in the level of Y is growth rate of Y

What about changes in the growth rate of Y?

100 → 110 → 121

Pointers

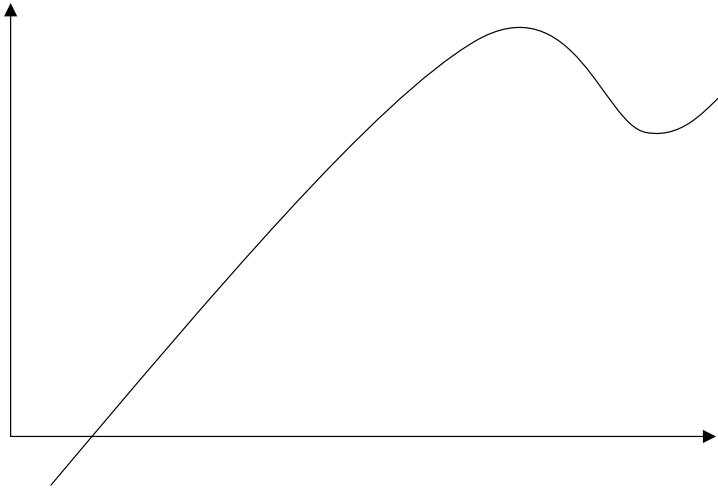
Deceptive: growth rates may hide actual levels

Growth over what length of time?

Few years esp. endpoint years may distort the picture

of people who visit Disneyland

enrollment of students at ISU



of accidents on Iowa's highways

of goats that die each year

Just because two variables are plotted together

does not mean they are related or that one causes the other

Establishing causation is tricky

does smoking *cause* cancer?

importance of ruling out a third contributory factor

pitfalls of interpreting causal relationships:
problem of *reverse* causation

more guns \Rightarrow more crime or more crime \Rightarrow more guns?

more banks \Rightarrow more growth or more growth \Rightarrow more banks?

Data Issues

Small sample bias

Are large samples necessarily good samples?

Surveying people outside Lowe's? At an ISU game?

Sample selection bias

Suppose 60% of pregnancy tests are positive. *Are 60% of women pregnant?*

Inference would have been correct if all women took pregnancy tests

Internet surveys; why aren't they "scientific"?

Sample must be representative for us to have faith

random sampling means *everyone has the same chance of being included in the survey*

Racist police officers and black crime statistics

Other problems with surveys

- who I think I am
 - costs nothing
 - survey influences respondents
1. Placebo
 2. Double blind trials