

Statistical Rules of Thumb

Gerald van Belle

Departments of Biostatistics and
Environmental Health

University of Washington,
Seattle, WA.

Outline

- A. Statistical rule of thumb
- B. Experimental or observational studies
- C. Covariation
- D. Sample size
- E. Presentation of results
- F. Recapitulation

Slides available on WEB at:

www.vanbelle.org

A. Statistical rule of thumb-1

A statistical rule of thumb is defined as a widely applicable guide to statistical practice—with sound theoretical basis. Characteristics include intuitive appeal, elegance, and transparency.

A. Statistical rule of thumb-2

1. Statistical rule as quick response
 - a. Typically committee meetings
 - b. Consultation sessions
 - c. Need to know the **basics**
 - d. May be preaching to the choir

A. Statistical rule of thumb-3

2. What are the basics?
 - a. Definition of statistical rule of thumb
 - b. Characteristics of rule of thumb
 - c. Substantive areas are idiosyncratic. In my case: environmental studies, epidemiology, statistical consulting

B. Randomized and Observational Studies-1

Rule 1.1 Distinguish between observational and randomized studies.

- a. Hohum!
- b. What's so nice about randomized studies?
- c. What's nice about observational studies?
- d. Gradation in observational studies.
- e. Some references.

B. Randomized and observational studies-2

a. Hohum!?

- * Epidemiology vs biostatistics
- * Selection issues
- * Missing data issues
- * Fragility of causal models



Arm Waving

B. Randomized and observational studies-3

a. Hohum!

b. What's so nice about randomized studies?

- * Provides a probability model

- * Rule 6.1 “Randomization puts systematic sources of variability into the error term.”
(DeLury)

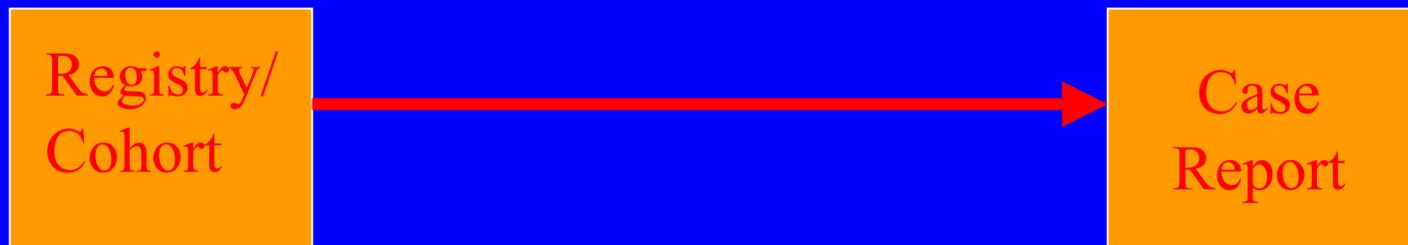
- * Lack of randomization leads to arm waving

B. Randomized and observational studies-4

- a. Hohum!
- b. What's so nice about randomized studies?
- c. **What's nice about observational studies?**
 - * Easier to carry out
 - * Majority of data; e.g. administrative data bases
 - * Ethical constraints on randomization
 - * Lots of data

B. Randomized and observational studies-5

- a. Hohum!
- b. What's so nice about randomized studies?
- c. What's nice about observational studies?
- d. Gradation in observational studies.



B. Randomized and observational studies-6

e. Some references.

Benson and Hartz (2000) NEJM

Concato, Shah and Horwitz (2000) NEJM

Copas and Li (1997) JRSS B

Copas and Shi (2000) BMJ

Hays (2001) Am. Scientist

(See May 2002 ROM on website)

C. Statistical rules of thumb for covariation-1

Rule 3.1: “Before choosing a measure of covariation determine the source of the data (sampling scheme), the nature of the variables, and the symmetry status of the measure.”

C. Statistical rules of thumb for covariation-2

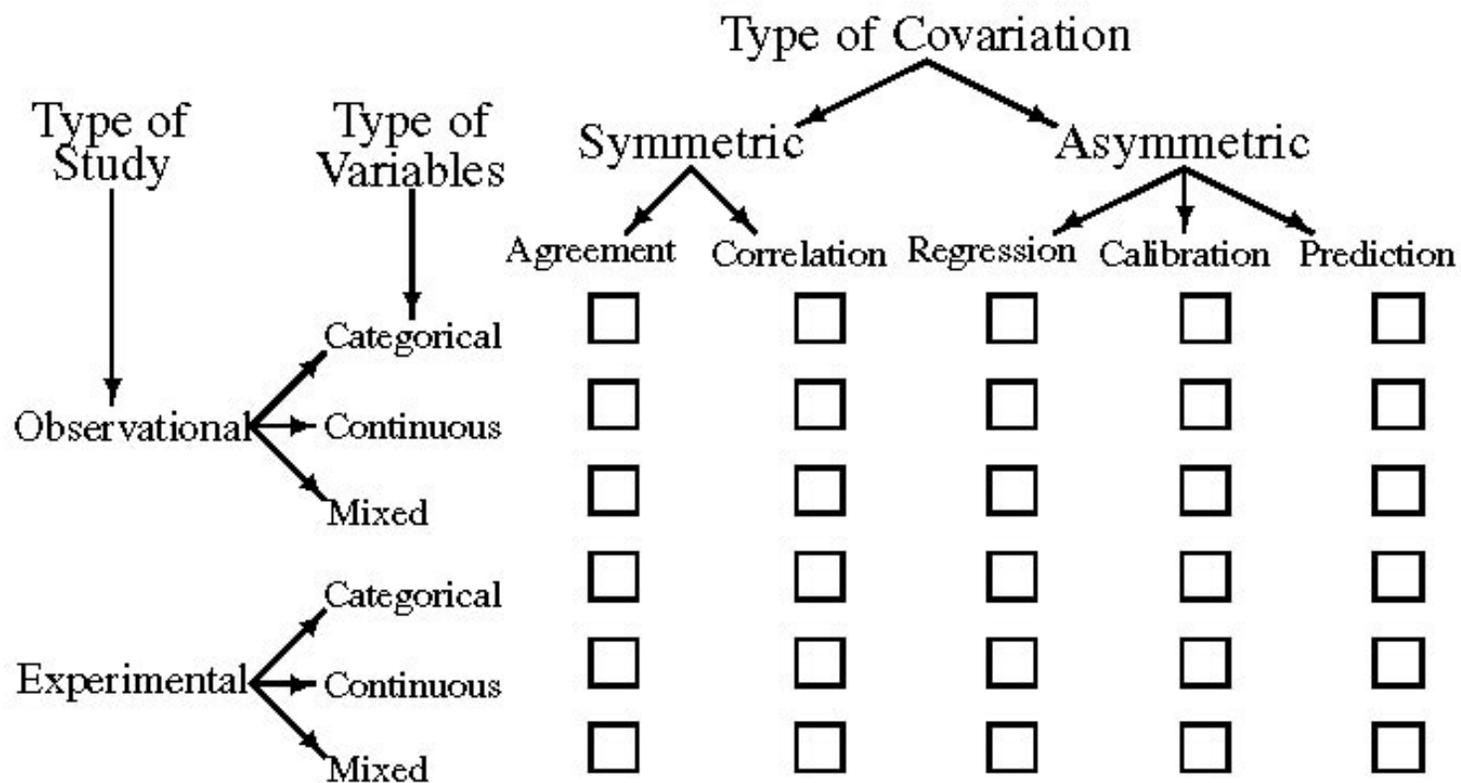


Fig. 3.1 Taxonomy of measures of covariation.

C. Statistical rules of thumb for covariation-3

Rule 3.2: “Do not summarize regression sampling schemes with correlations.”

Assume simple linear regression, Y on X

$$\frac{r_{regression}^2}{1 - r_{regression}^2} = \frac{r_{random}^2}{1 - r_{random}^2} \times \frac{S_{x,regression}^2}{S_{x,random}^2}$$

C. Statistical rules of thumb for covariation-5

	True r^2			
Ratio $s^2(\mathbf{x})/s^2(\text{true})$	0	0.10	0.20	0.30
0.25	0	0.03	0.06	0.10
1	0	0.10	0.20	0.30
4	0	0.31	0.50	0.63

C. Statistical rules of thumb for covariation-6

Rule 3: “Assess agreement by addressing accuracy, scale differential, and precision. Accuracy can be thought of as the lack of bias.”

C. Statistical rules of thumb for covariation-7

Model: Y_1 and Y_2 measured on the same “objects”

$$E(Y_1 - Y_2)^2 = (\mu_1 - \mu_2)^2 + (\sigma_1 - \sigma_2)^2 + 2(1 - \rho)\sigma_1\sigma_2$$



Total
Deviance

=



Bias

+



Scale
differential

+



Imprecision

Lin (1989)

C. Statistical rules of thumb for covariation-8

Model: Y_1 and Y_2 measured on the same “objects”

$$\frac{E(Y_1 - Y_2)^2}{2\sigma_1\sigma_2} = \frac{(\mu_1 - \mu_2)^2}{2\sigma_1\sigma_2} + \frac{(\sigma_1 - \sigma_2)^2}{2\sigma_1\sigma_2} + (1 - \rho)$$



Deviance

=



Bias

+



Scale
differential

+



Imprecision

D. Sample size responsibilities-1

$$n = 2 \frac{(z_{1-\alpha/2} + z_{1-\beta})^2}{\left(\frac{\mu_1 - \mu_2}{\sigma}\right)^2}$$

Investigator

Investigator

Statistician

D. Sample size responsibilities-2

$$n = \frac{16}{\left(\frac{\mu_1 - \mu_2}{\sigma}\right)^2}$$

Type I error = 0.05
Type II error = 0.20
(Power = 0.80)

Topic for discussion:
Treatment effect+
Variability=Effect Size

Two sample
(default)

D. Sample size responsibilities-3

Effect Size

$$\left(\frac{\mu_1 - \mu_2}{\sigma} \right) = \text{Effect Size}$$

$$\text{Effect Size} = \frac{4}{\sqrt{n}}$$

E. Presentation of results-1

Rule 7.1: When text, when tables, when graphs?

“Use sentence structure for displaying 2 to 5 numbers, tables for displaying more numerical information, and graphs for complex relationships.”

E. Presentation of results-2

Rule 7.1: When text, when tables, when graphs?

a. Illustration-version 1

“The blood type of the population of the United States is approximately 40%, 11%, 4% and 45% A, B, AB, and O, respectively.”

E. Presentation of results-3

Rule 7.1: When text, when tables, when graphs?

a. Illustration-version 2

“The blood type of the population of the United States is approximately 40% A, 11% B, 4% AB, and 45% O.”

E. Presentation of results-4

Rule 7.1: When text, when tables, when graphs?

a. Illustration-version 3

“The blood type of the population of the United States is approximately,

O 45%

A 40%

B 11%

AB 4%

E. Presentation of results-5

Rule 7.2: Table Structure

“Arrange rows and columns in meaningful way,
Limit the number of significant digits,
Make the table as self-contained as possible,
Use white space and lines to organize rows and columns,
Do not stint on table headings”

Rule 7.2: Table structure
Original table on right.
1. Different degrees of precision, due to different sources of data.
2. Ordering by alphabet;
Spanish version would look different.

Table 7.1 Number of Active Health Professionals According to Occupation in 1980: United States^a

Occupation	1980
Chiropractors	25,600
Dentists	121,240
Nutritionists/Dietiticians	32,000
Nurses, registered	1,272,900
Occupational therapists	25,000
Optometrists	22,330
Pharmacists	142,780
Physical Therapists	50,000
Physicians	427,122
Podiatrists	7,000
Speech therapists	50,000

^aFrom National Center for Health Statistics, 2000; Table 104 (Subcategories not included, Only data for 1980 is used). See MacKay et al. (2000).

Rule 7.2: Table structure

1. Reduced number of digits,
2. Rows arranged by frequency,
3. White space suggests similar groups

Table 7.2 Table 7.1 Re-arranged by Number in Occupational Category and Rounded to the Nearest 1000.

Occupation	1980 in 1000's
Nurses, registered	1,273
Physicians	427
Pharmacists	143
Dentists	121
Physical Therapists	50
Speech therapists	50
Nutritionists/Dietiticians	32
Chiropractors	26
Occupational therapists	25
Optometrists	22
Podiatrists	7

E. Presentation of results-6

Rule 7.2: Table Structure

Significant digits for frequencies

Table 7.5 Number of social activities in a two-week period among persons 70 years of age and older^a

Number of Activities	70-74 years	75-79 years	80-84 years	85 years and over
	%	%	%	%
Women				
0 activities	1.0	1.3	2.1	3.1
1-2 activities	6.8	10.5	11.9	19.2
3-4 activities	26.8	27.5	32.5	38.3
5-7 activities	65.4	60.7	53.5	39.4
Mean number	4.96	4.76	4.53	3.99
Men				
0 activities	1.9	1.7	2.9	5.3
1-2 activities	10.5	13.3	15.9	23.0
3-4 activities	26.3	30.3	36.7	35.9
5-7 activities	61.2	54.7	44.5	35.9
Mean number	4.75	4.54	4.17	3.96

^aFrom data for Figure 21 in Kramarov et al. (1999).

right just

← it's!

← it's!

6.1.0

6.2.0

Table 7.5 Reformatted

Number of activities	Age 70-74	Age 75-79	Age 75-79	Age 85+
Women				
	%	%	%	%
0	1	1	2	3
1-2	7	10	12	19
3-4	27	28	32	38
5-7	65	61	54	39
Mean	5.0	4.8	4.5	4.0
Men				
0	2	2	3	5
1-2	10	13	16	23
3-4	26	30	37	36
5-7	61	55	44	36
Mean	4.8	4.5	4.2	4.0

E. Presentation of results-7

Rule 7.3: Graph data

“When possible graph the data.”

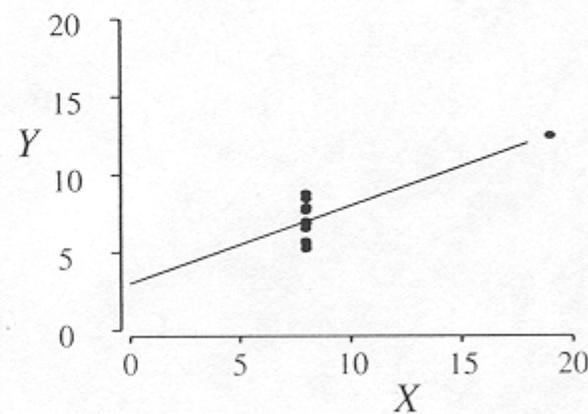
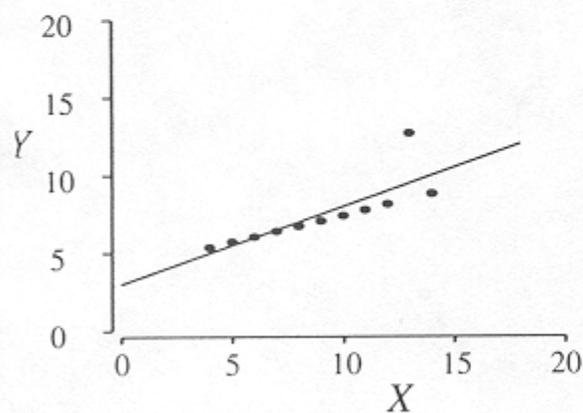
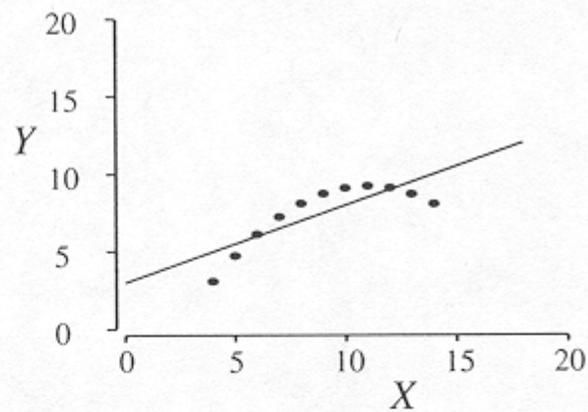
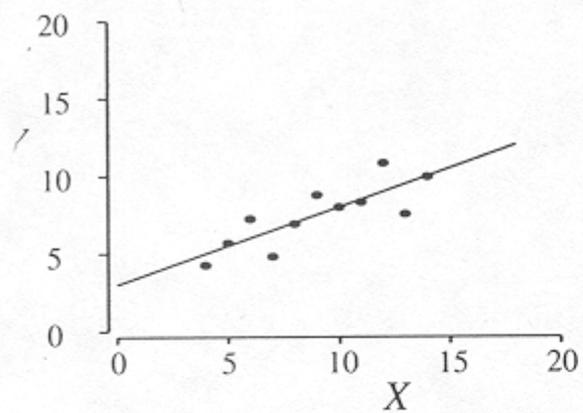


Fig. 7.1 Regression data with the property that for every set the means are equal, the regression lines are $Y = 3 + 0.5X$, the standard errors of estimate of slope are 0.118, and the correlation coefficients are 0.82. Data from Anscombe (1977).

The following three tables and figures are from a great paper:

Gelman, A., Pasarica, C. and Dohdia, R. (2002). Let's practice what we preach: Turning tables into graphs. *The American Statistician*, 56: 121-130.

Profession	Frequency of recent citations	1996 total employed (1,000)	Relative frequency
Lawyers	8101	880	9.2
Economists	1201	148	8.1
Architects	1097	160	6.9
Physicians	3989	667	6.0
Statisticians	34	14	2.4
Psychologists	479	245	2.0
Dentists	165	137	1.2
Teachers (not university)	3938	4724	0.8
Engineers	934	1960	0.5
Accountants	628	1538	0.4
Computer programmers	91	561	0.2
Total	20,657	11,034	1.9

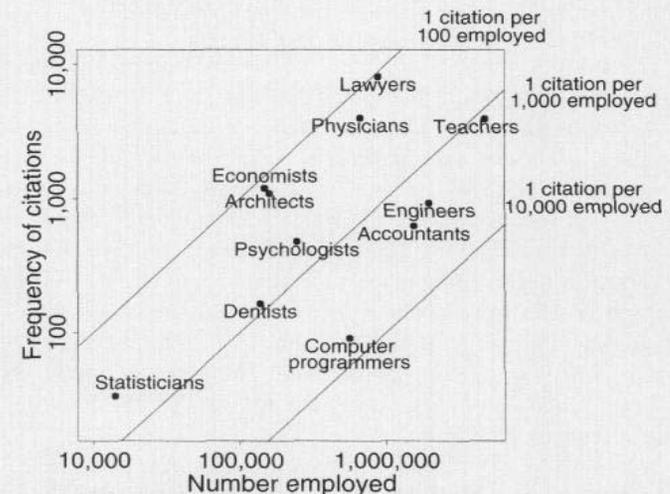
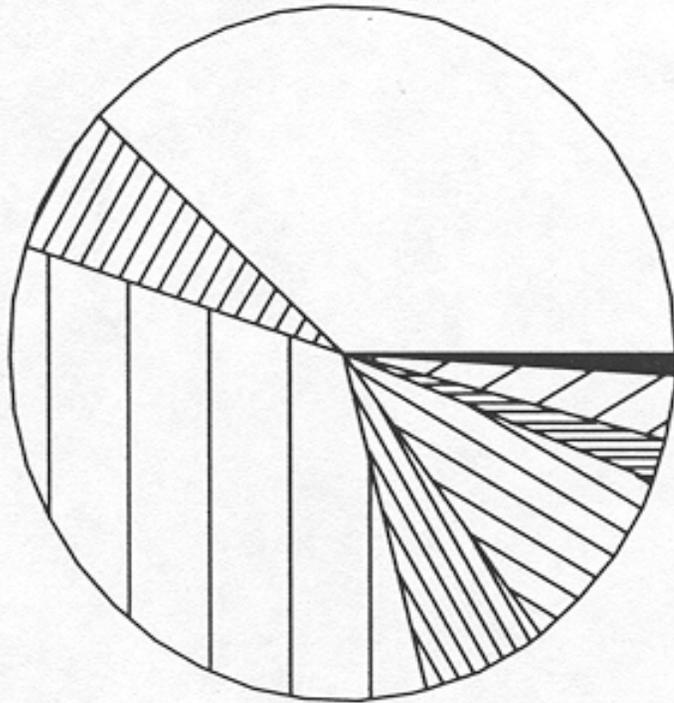


Figure 2. Top panel: Table from Ellenberg (2000) displays counts and rates of citations of various professions from the New York Times database. Bottom panel: Display as a figure shows the relative positions of the different professions much more clearly. The log-log display allows comparison across several orders of magnitude. The x and y axes are on the same scale, so that any 45° line indicates a constant relative frequency.

E. Presentation of results-8

Rule 7.4: Pie Charts

“Never use a pie chart. Present a simple list of percentages, or whatever constitutes the divisions of the pie chart.”



<i>Blood Type</i>	<i>Rh+</i>	<i>Rh-</i>	<i>Total</i>
<i>O</i>	<i>38</i>	<i>7</i>	<i>45</i>
<i>A</i>	<i>34</i>	<i>6</i>	<i>40</i>
<i>B</i>	<i>9</i>	<i>2</i>	<i>11</i>
<i>AB</i>	<i>3</i>	<i>1</i>	<i>4</i>
<i>Total</i>	<i>84</i>	<i>16</i>	<i>100</i>



Fig. 7.2 Pie chart of frequencies of blood type and Rh classification for the population of the United States. Percentages in the right-hand panel form the basis for the pie chart.

E. Presentation of results-9

Rule 7.5: Bar Charts

“Always think of alternatives to a bar graph.”

Table 7.5 Number of social activities in a two-week period among persons 70 years of age and older^a

Number of Activities	70-74 years	75-79 years	80-84 years	85 years and over
	%	%	%	%
Women				
0 activities	1.0	1.3	2.1	3.1
1-2 activities	6.8	10.5	11.9	19.2
3-4 activities	26.8	27.5	32.5	38.3
5-7 activities	65.4	60.7	53.5	39.4
Mean number	4.96	4.76	4.53	3.99
Men				
0 activities	1.9	1.7	2.9	5.3
1-2 activities	10.5	13.3	15.9	23.0
3-4 activities	26.3	30.3	36.7	35.9
5-7 activities	61.2	54.7	44.5	35.9
Mean number	4.75	4.54	4.17	3.96

^aFrom data for Figure 21 in Kramarov et al. (1999).

right just

← it's!

← it's!

6.1.0

6.2.0

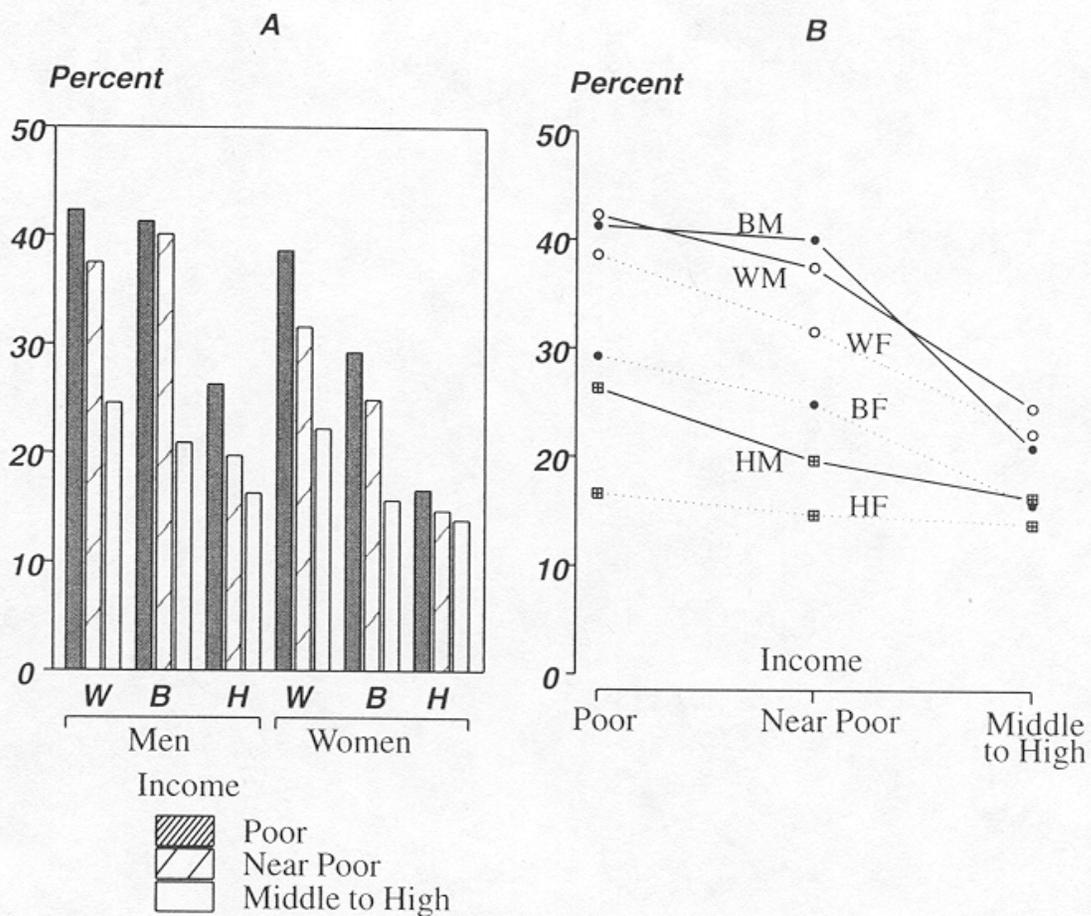


Fig. 7.3 Cigarette smoking (%) in 1995 among adults 18 years of age and over by family income, sex, race, and Hispanic origin for the United States. Adjusted for age. From data for Figure 36, Pamuk et al. (1998). Panel A: original format; Panel B: new format.

E. Presentation of results-10

Rule 7.6: Stacked bar charts

“There are much more effective ways of showing data structure than stacked bar charts.”

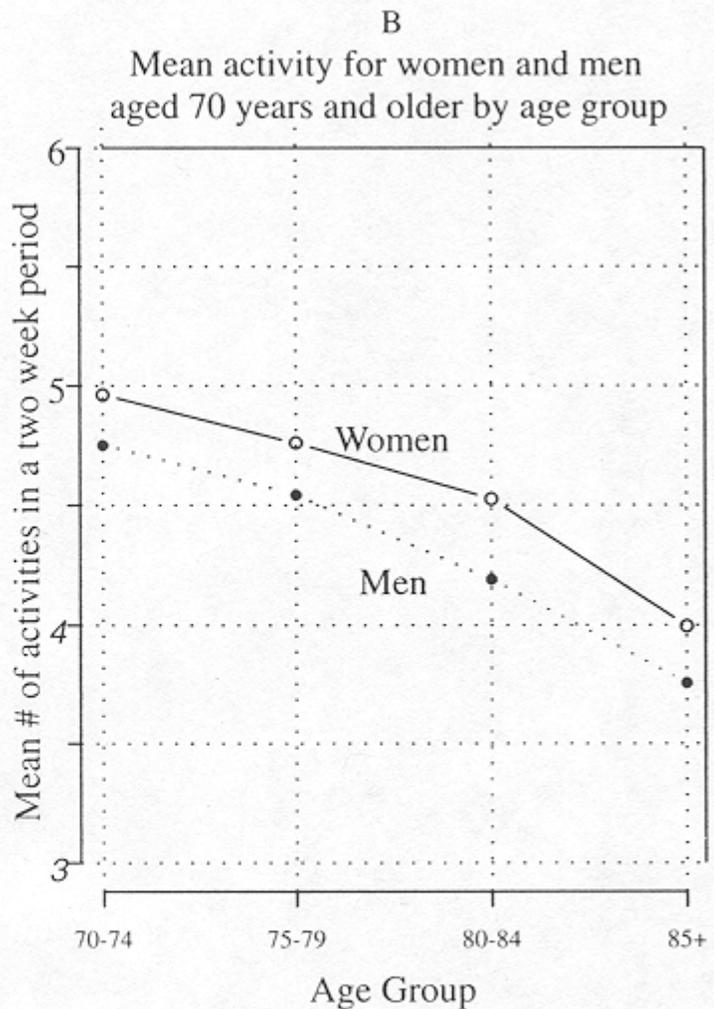
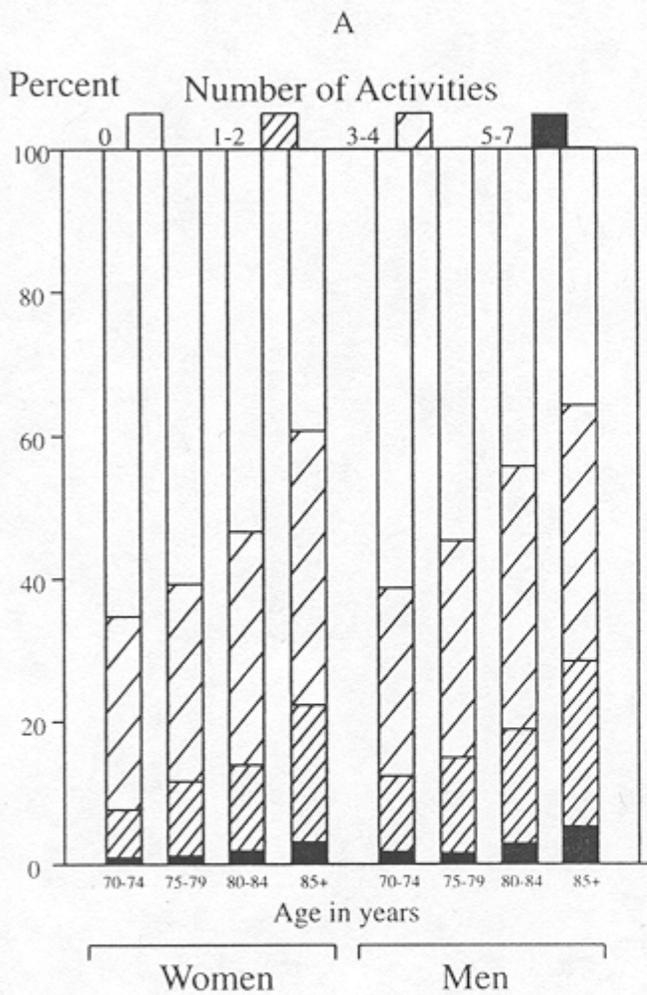


Fig. 7.4 Number of social activities in 1995 in a 2-week period among persons 70 years of age and older. From data for Figure 21 in Kramarov et al. (1999). Panel A: original format; Panel B: new format.

E. Presentation of results-11

Rule 7.7: Three-dimensional bar graphs

“Never use three-dimensional bar graphs.”

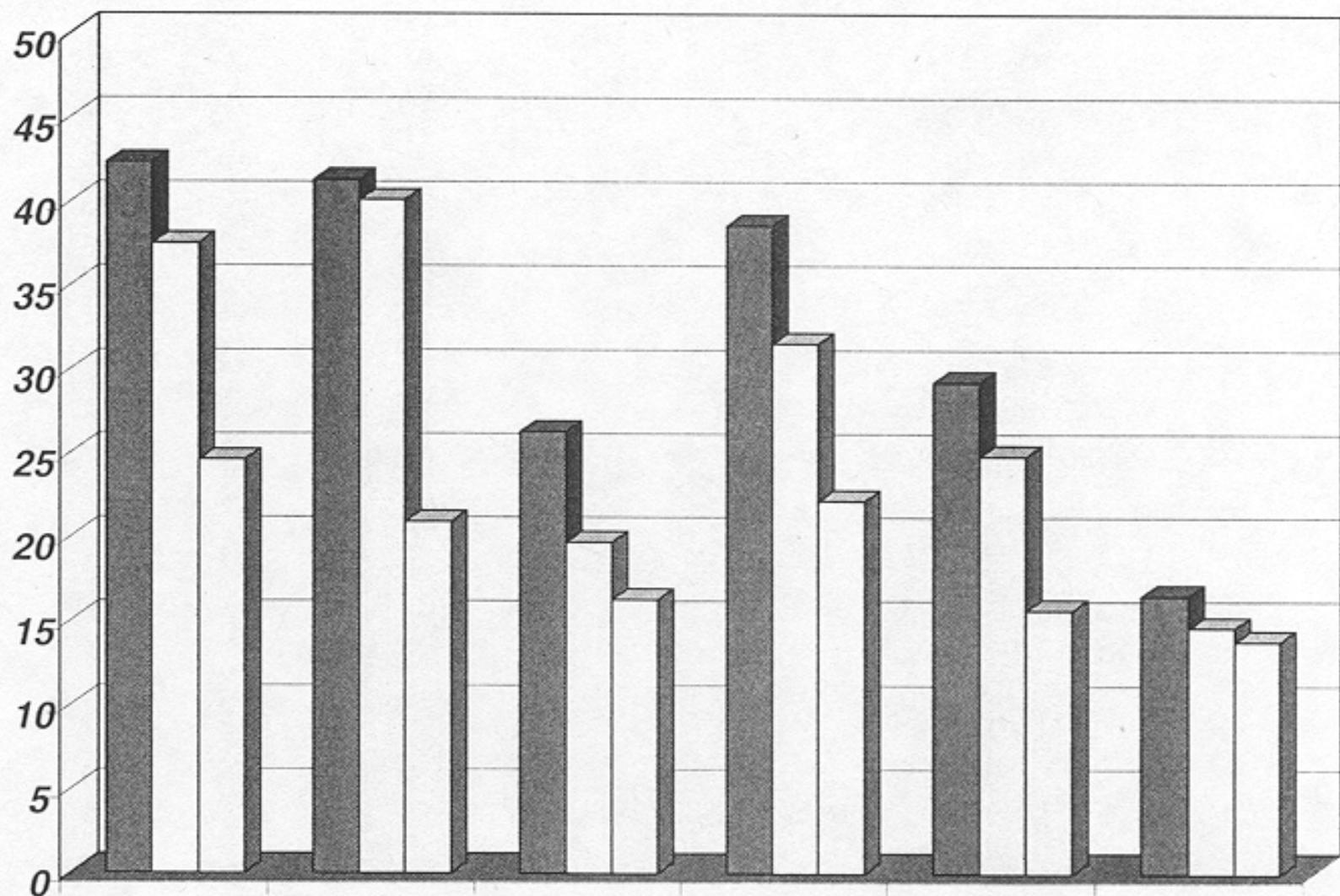


Fig. 7.5 Distortion introduced by adding extra dimension to a bar graph. Same data, and graph, as in Figure 6.3.

E. Presentation of results-12

Rule 7.8: Longitudinal data

“In the case of longitudinal data identify both cross-sectional and longitudinal patterns.”

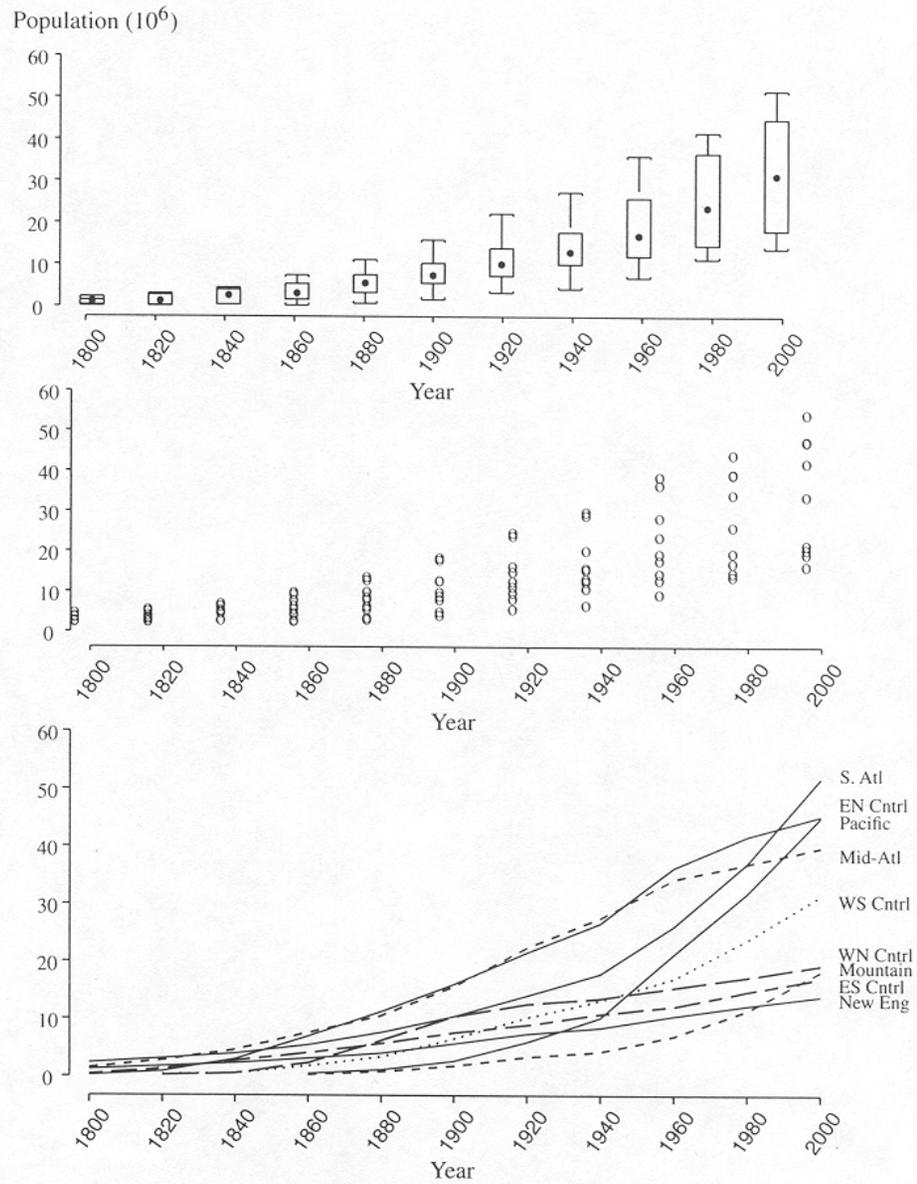


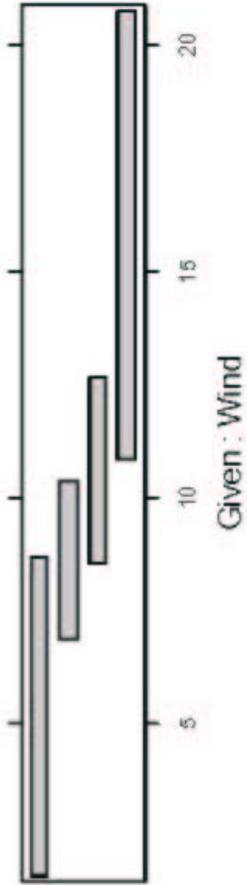
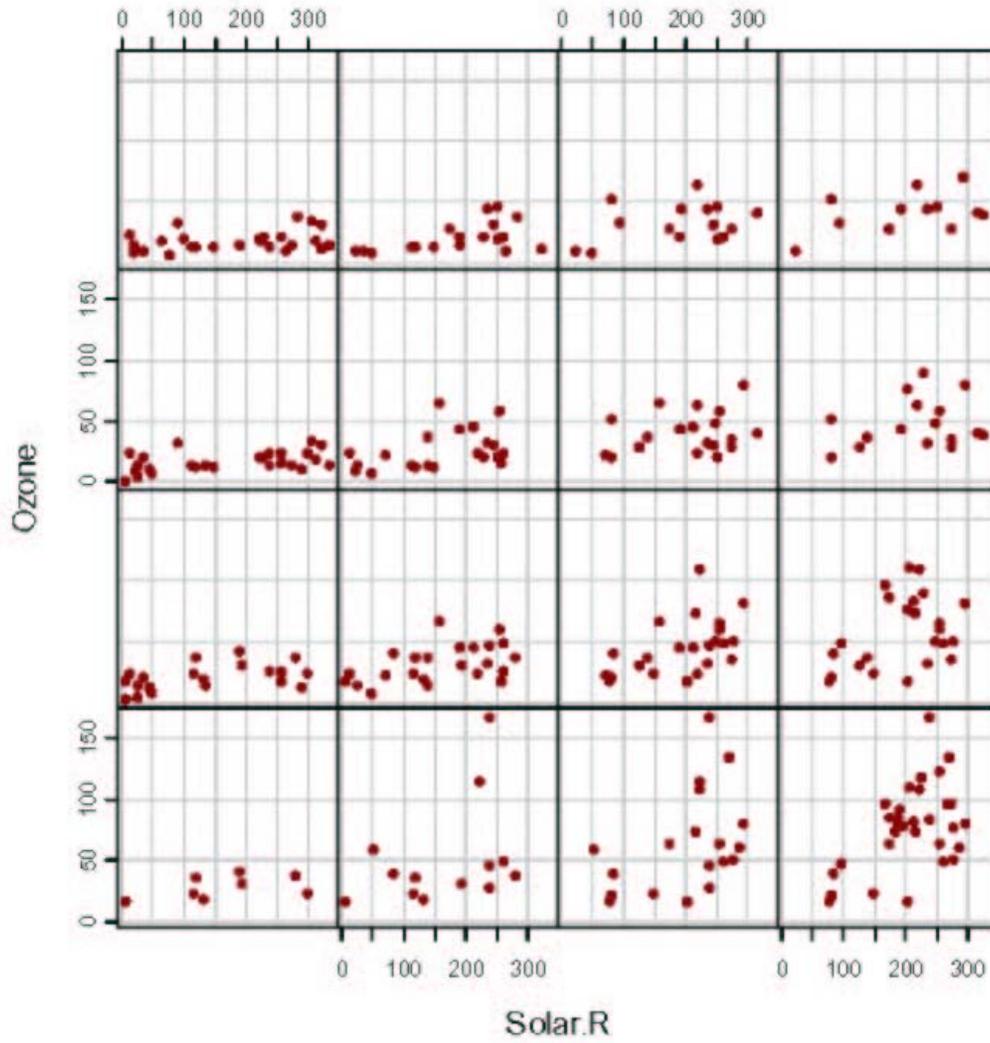
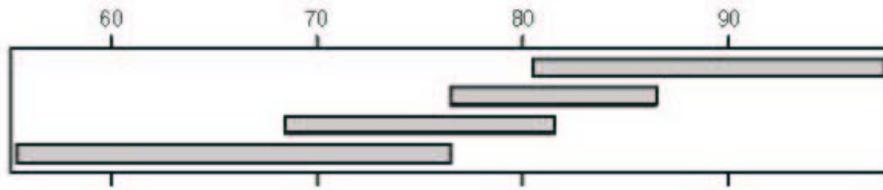
Fig. 7.6 US population by region and year from 1800 to 2000.

E. Presentation of results-13

Rule 7.9: High dimensional data

“Three key aspects of presenting high dimensional data are: rendering, manipulation, and linking. *Rendering* determines what is to be plotted, *manipulation* determines the structure of the relationships, and *linking* determines what information will be shared between plots or sections of the graph.”

Given : Temp



Given : Wind

E. Presentation of results-14

- 1. Chew words carefully**
- 2. Watch table manners**
- 3. Avoid pies**
- 4. Stay away from bars**

F. Summary and recapitulation

1. Concept of statistical rule of thumb
2. Observation studies are fragile
3. Covariation needs to be described correctly
4. Logic to presentation of results

G. Resources

1. Gerald van Belle, (2002). *Statistical Rules of Thumb*, John Wiley and Sons, New York, NY.
2. WEB sites: evolving rapidly
4. Statistical books and journals: Audience recommendations
4. My choices at <http://www.vanbelle.org>
5. General purpose books and journals on science
6. Colleagues. Find a consultants' consultant

H. Acknowledgments

1. DeLury (University of Toronto)
2. Steve Millard, Jim Hughes, Michael Levin
3. Paul Crane
4. Biostatistics, statistics, and epidemiology colleagues
5. Consultees who sharpened my skills