Chapter 2

Organizing and Summarizing Data

After collecting a sample, statistical data is often first analyzed in a descriptive manner. In particular, qualitative or quantitative (discrete or continuous) data is described in both a graphical and tabular form.

2.1 Organizing Qualitative Data

Distribution tables, bar graphs and Pareto charts are discussed in this section.

In addition to bar graphs, text discusses other graphs, including pie graphs, which are *not* covered in this workbook. Although not covered in workbook, this other (simple) material on graphs could appear on either a quiz or homework.

Exercise 2.1 (Organizing Qualitative Data)

1. *Patient health.* Health of twenty patients in a high blood pressure study are:

   good, good, fair, poor, bad, poor, great, fair, good, good,
   good, fair, fair, fair, good, poor, poor, bad, good, good.

   Distribution table, bar graph and Pareto chart for this data is given below.

category	frequency	relative frequency
bad	2	$\frac{2}{20} = 0.10$
poor	4	$\frac{4}{20} = 0.20$
fair	5	$\frac{5}{20} = 0.25$
good	8	$\frac{8}{20} = 0.40$
great	1	$\frac{1}{20} = 0.05$
total	20	1.0

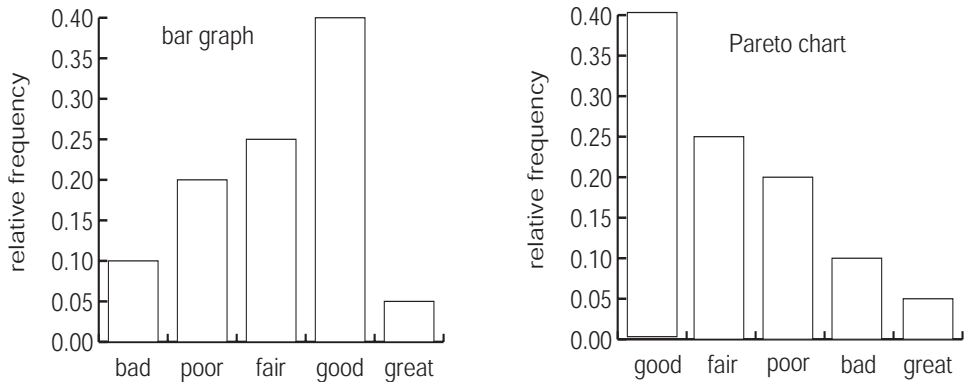

Figure 2.1 (Bar Graph and Pareto Chart for Patient Health)

StatCrunch: For bar graph on left, relabel var1 as Health, var2 as Frequency; type health types in Health column, type frequencies in Frequency column. Click on Graphics, Bar Plot, with summary, Categories: Health, Counts: Frequency, Next, Type: Relative Frequency, Order by: Worksheet, then Create Graph! For Pareto chart on right, same as bar graph, except Order by: Count Descending. Once done, click Data, Save data "2.1 Health Bar Graph".

(a) *Review.* This qualitative data is (choose one)
nominal / **ordinal** / **interval** / **ratio** because this data is ordered.

(b) Of 20 patients, (circle one) **2** / **4** / **5** / **8** are in fair health.

(c) *Frequency* means (choose *one or more*!)
number / **count** / **proportion** / **percentage** / **category**

(d) *Relative frequency* means (choose *two*!)

i. number in a particular category
ii. count in a particular category
iii. proportion of observations that fall into each category
iv. percentage of observations that fall into each category
v. proportion of observations that fall into at least two categories

(e) Height of each vertical bar in bar graph corresponds to the relative frequency for each category. For example, vertical bar for "good" category has a height (or relative frequency) of (choose one) **0.30** / **0.35** / **0.40**.

(f) Adding all heights of vertical bars in five categories together, we get (choose one) **0.40** / **0.75** / **1.00**.

(g) **True** / **False**. Pareto chart is a bar graph where bars are arranged left to right in decreasing order.

(h) **True** / **False**. Another possible variation of a bar graph for qualitative data would be to draw bar graph where y–axis is frequency, rather than relative frequency.

(i) **True** / **False** Although height of each vertical bar gives relative frequency of a particular category of health occurring for twenty patients, *width* of each vertical bar has *no* meaning.

2. *Comparing patient health.* Health of twenty patients in a high blood pressure study are compared in 2001 and 2009.

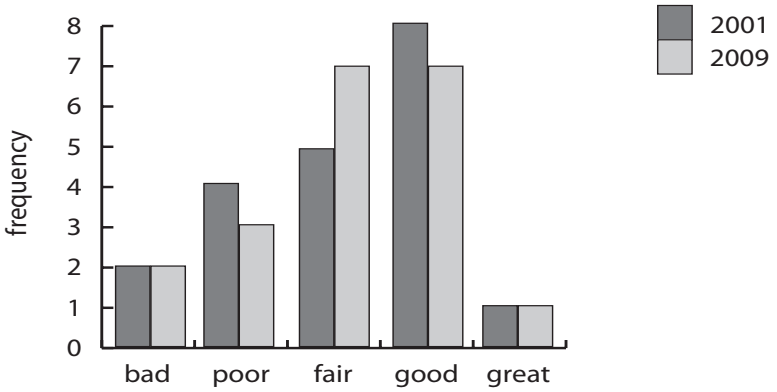

Figure 2.2 (Side–by–side bar graph)

(a) Number of patients in poor health in 2009: **3** / **4** / **5** / **8**.

(b) What category of health most improved from 2001 to 2009?
bad / **poor** / **fair** / **good** / **great**

(c) *Percentage* of patients in fair health in 2009: **25%** / **30%** / **35%** / **40%**.

2.2 Organizing Quantitative Data: The Popular Displays

Histograms, graphs of quantitative (discrete or continuous) data, are discussed. Stem-and-leaf plots of small discrete data sets are also discussed.

Exercise 2.2 (Organizing Quantitative Data: The Popular Displays)

1. *Histograms for Discrete Quantitative Data.*

(a) *Number of Burgers Made.* Distribution table and histogram for number of burgers made in some fixed time period are given below.

3, 3, 3, 4, 4, 4, 4, 4, 4, 4,
4, 4, 4, 4, 4, 4, 4, 4, 5, 5

burgers made	frequency	relative frequency
3	3	$\frac{3}{20} = 0.15$
4	15	0.75
5	2	0.10

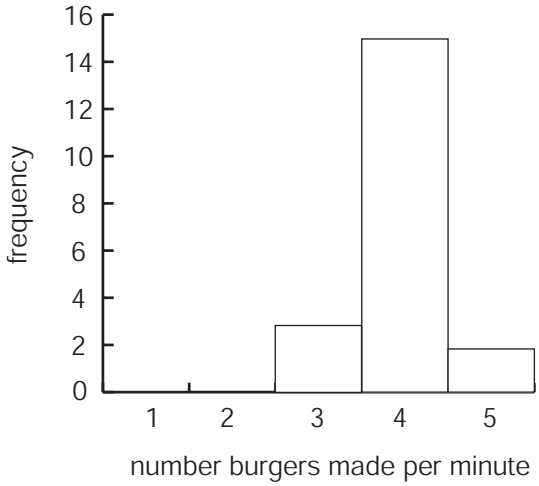

Figure 2.3 (Discrete Data Histogram for Number Burgers Made)

If rectangles replaced by vertical dots, for discrete data, histogram becomes a *dot plot*.

i. *Review*. Burgers data is quantitative data and
(choose one) **discrete** / **continuous**
because we *count* the number of burgers made.

ii. Most frequent number burgers made is (choose one) **3** / **4** / **5**.

iii. Relative frequency 5 burgers made is **0.05** / **0.10** / **0.15** / **0.20**.

iv. Sum of heights of rectangles in histogram equal **17** / **18** / **19** / **20**.

v. Rectangles in histogram **touch one another** / **separated by gaps**.

(b) *Patient Ages.* Distribution table and histogram for patient ages, treated as discrete data, are given below.

32, 37, 39, 40, 41, 41, 41, 42, 42, 43,
44, 45, 45, 45, 46, 47, 47, 49, 50, 51

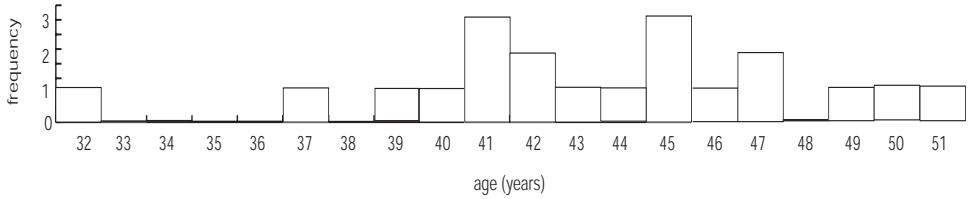

Figure 2.4 (Discrete Data Histogram for Patient Ages)

patient age	frequency	relative frequency
32	1	$\frac{1}{20} = 0.05$
33	0	0.00
34	0	0.00
35	0	0.00
36	0	0.00
37	1	0.05
38	0	0.00
39	1	0.05
40	1	0.05
41	3	0.15
42	2	0.10
43	1	0.05
44	1	0.05
45	3	0.15
46	1	0.05
47	2	0.10
48	0	0.00
49	1	0.05
50	1	0.05
51	1	0.05

i. *Review.* Patient age data is quantitative data and
(choose one) **discrete** / **continuous** because
a patient age *always* exists between any two different patient ages.

ii. Number of classes is (circle one) **17** / **18** / **19** / **20**.

iii. Most frequent age is (choose *two*!) **41** / **43** / **45** / **47** years.

iv. Sum of heights of rectangles in histogram equal **17** / **18** / **19** / **20**.

v. Discrete data histogram (choose one) **appropriate** / **inappropriate**
in this case because data continuous, not discrete, and too dispersed.

2. *Histograms for Continuous Quantitative Data.*

(a) *Patient Ages.* Distribution table and histogram for patient ages, treated as continuous data, are given below.

32, 37, 39, 40, 41, 41, 41, 42, 42, 43,
44, 45, 45, 45, 46, 47, 47, 49, 50, 51

class	frequency	relative frequency
30 to 34	1	$\frac{1}{20} = 0.05$
35 to 39	2	0.10
40 to 44	8	0.40
45 to 49	7	0.35
50 to 54	2	0.10

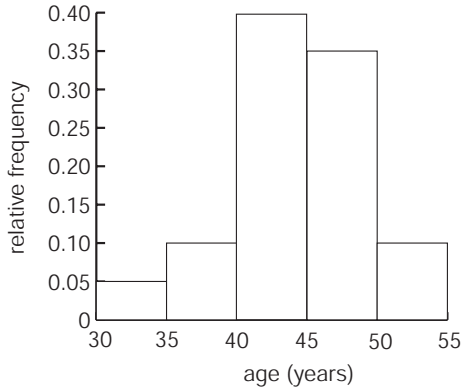

Figure 2.5 (Histogram for Patient Ages)

StatCrunch: Relabel var1 as Age, type 20 ages into Age column. Click on Graphics, Histogram, select Age, Relative Frequency, Start bins at: 30, Binwidth: 5, then Create Graph! Once done, click Data, Save data "2.2 Age Histogram".

i. Number of classes is (circle one) **3** / **4** / **5** / **6**.

ii. First class is (circle one) **30 to 34** / **35 to 39** / **40 to 44**.

iii. *Lower* class limit of first class is (circle one) **30** / **34** / **35** / **39**.

iv. *Upper* class limit of first class is (circle one) **30** / **34** / **35** / **39**.

v. *Width* of first class is $35 - 30 =$ (circle one) **3** / **4** / **5** / **6** years.

vi. *Class width* equals (choose two)
difference between consecutive lower class limits
difference between consecutive upper class limits
difference between upper and lower class limits

vii. Number of patients in 30-34 age class is (circle one) **1** / **2** / **3** / **4**.

viii. Percentage of patients in 30-34 age class: **5%** / **10%** / **35%** / **40%**.

(b) *pH levels.* Consider distribution table and histogram of 28 pH levels of soil data below.

4.3	5	5.9	6.5	7.6	7.7	7.7	8.2	8.3	9.5
10.4	10.4	10.5	10.8	11.5	12	12	12.3	12.6	12.6
13	13.1	13.2	13.5	13.6	14.1	14.1	15.1		

class	frequency	relative frequency
4–5.9	3	$\frac{3}{28} \approx 0.107$
6–7.9	4	0.143
8–9.9	3	0.107
10–11.9	5	0.179
12–13.9	_____	_____
14–15.9	_____	_____

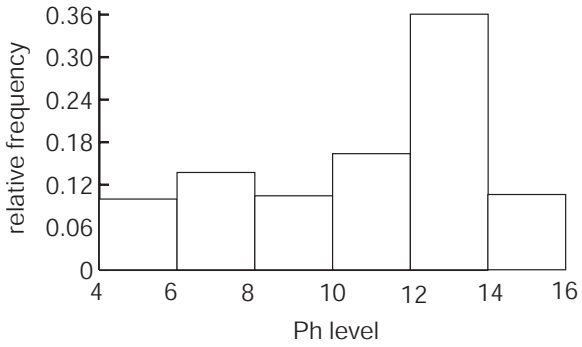

Figure 2.6 (Histogram for pH Level Data)

StatCrunch: Relabel var1 as pH, type 28 pH levels into pH column. Click on Graph, Histogram, select pH, Relative Frequency, Start bins at: 4, Binwidth: 2, then Create Graph! Once done, click Data, Save data "2.3 pH data".

i. Fill in blanks in distribution table. Hint: 28 readings total.

ii. Number of classes is (circle one) **3** / **4** / **5** / **6**.

iii. Width of each class is (circle one) **2** / **3** / **4** / **5** pH.

iv. Most frequent pH reading is
8 − 9.9 / **10 − 11.9** / **12 − 13.9** / **14 − 15.9**.

3. *Stem-and-Leaf Plots.*

(a) *Patient Ages.* Stem-and-leaf plot for patient ages is given below.

32, 37, 39, 40, 41, 41, 41, 42, 42, 43,
44, 45, 45, 45, 46, 47, 47, 49, 50, 51

3	2	7	9*													
4	0	1	1	1**	2	2	3	4	5	5	5	6	7	7	9	stem: 10s
5	0	1														leaf: 1s

i. Starred number, **9***, represents age (circle one) **39** / **93** / **9**.
Double–starred number, 1**, represents age (circle one) **41** / **14** / **1**.

ii. Numbers left of double line (in first column) are called **stems** / **leaves**; numbers to right are called (circle one) **stems** / **leaves**.

iii. Starred number 9* is a leaf of stem (circle one) **3** / **4** / **5**.

iv. **True** / **False** Note to right of stem-and-leaf plot specifies numbers used as stems are "tens" (or "10s") and numbers used as leaves are "ones" (or "1s"). So, for instance, stem "3" represents 3x10 = 30 and leaf "2" represents 1x2 = 2.

v. Stem–and–leaf plot is ordered where, in first stem, for example, 32 is followed by (circle one) **37** / **39** / **40**.

vi. Stem–and–leaf plot is useful in identifying "center" of data, or, where "most" data values are located. In this case, this is **30**s / **40**s / **50**s.

(b) *Split stem-and-leaf plots: patient ages.* Sometimes, to spread data out, stems are *split* as, for example, in following table.

3	2							
3	7	9						
4	0	1	1	1	2	2	3	4
4	5	5	5	6	7	7	9	
5	0	1						
5								

stem: 10s
leaf: 1s

i. **True** / **False** Low stem 3 contains one half of leaves, 0, 1, 2, 3 or 4; high stem 3 contains other half of leaves, 5, 6, 7, 8 and 9.

ii. **True** / **False** Stem-and-leaf plots can have stems split not only twice, but also three or more times. Splitting each stem three times might, say, consist of a low stem which contains leaves 0, 1 and 2; a middle stem with leaves 3, 4, 5 and 6 and a high stem with leaves 7, 8 and 9.

iii. A stem and leaf plot with 10s as stems can be split at most (circle one) **5** / **7** / **10** / **100** times.

iv. **True** / **False** Although no one "best" way of constructing a stem–and–leaf plot, most stem–and–leaf plots consist of 5 to 20 stems.

(c) *Back–to–back plots.* Back–to–back stem and leaf plot compares "body temperature (before taking tablets)" to "body temperature (after taking tablets)".

body						4	97L								body
temperature						8	97H								temperature
before							98L	1							after
tablet					8	5	98H	5							tablet
						0	99L								
						9	99H	6	8						
			3	2	0	0	100L	0	0	1	1	2	3	4	stem: 1s
	9	9	8	7	7	6	100H	7	8	8	8	9	9		leaf: 0.1s
				1	1	0	101L	0	1	4					(after tablet)
						5	101H								outlier: 96.2

StatCrunch: Relabel var1 as Before Temperature, var2 as After Temperature, type temperatures into two columns. Click on Graph, Stem and Leaf, select both Before Temperature and After Temperature, then Create Graph! Two separate graphs are created, not one back-to-back plots. Once done, click Data, Save data "2.2 Body Temperature Stem and Leaf".

i. Before and after body temperatures appear centered around **90** / **100** / **110** Also, "shape" of both plots seem about the same.

ii. Before body temperatures **higher than** / **about same as** / **lower than** after body temperatures.

iii. Based on back–to–back stem and leaf plot, it (circle one) **does** / **does not** appear as though drug is lowering high body temperature.

iv. Stem–and–leaf plots are most useful for (circle one)

A. qualitative data

B. small quantitative (either discrete or continuous) data sets

C. large quantitative data sets

4. *Shapes of Histograms for Continuous Quantitative Data.*
Common shapes of continuous *quantitative* histograms are given below. Identify previous quantitative histograms as one of these shapes.

It does not make sense to talk about shapes of *qualitative* nominal histograms because order of data is arbitrary.

(a) Figure 2.3 Discrete Data Histogram for Number Burgers Made is roughly **uniform** / **bell-shaped** / **right-skewed** / **left-skewed** / **none**.

(b) Figure 2.4 Discrete Data Histogram for Patient Ages is roughly **uniform** / **bell-shaped** / **right-skewed** / **left-skewed** / **none**.

(c) Figure 2.5 Histogram for Patient Ages is roughly **uniform** / **bell-shaped** / **right-skewed** / **left-skewed** / **none**.

(d) Figure 2.6 Histogram for pH Level Data is roughly **uniform** / **bell-shaped** / **right-skewed** / **left-skewed** / **none**.

(e) Stem-and-leaf and split stem-and-leaf of patient ages are both roughly **uniform** / **bell-shaped** / **right-skewed** / **left-skewed** / **none**.

(f) Back-to-back stem-and-leaf of blood pressure data are both roughly **uniform** / **bell-shaped** / **right-skewed** / **left-skewed** / **none**.

Rotate stem-and-leaf 90^o counterclockwise so higher temperatures to right, then check shape.

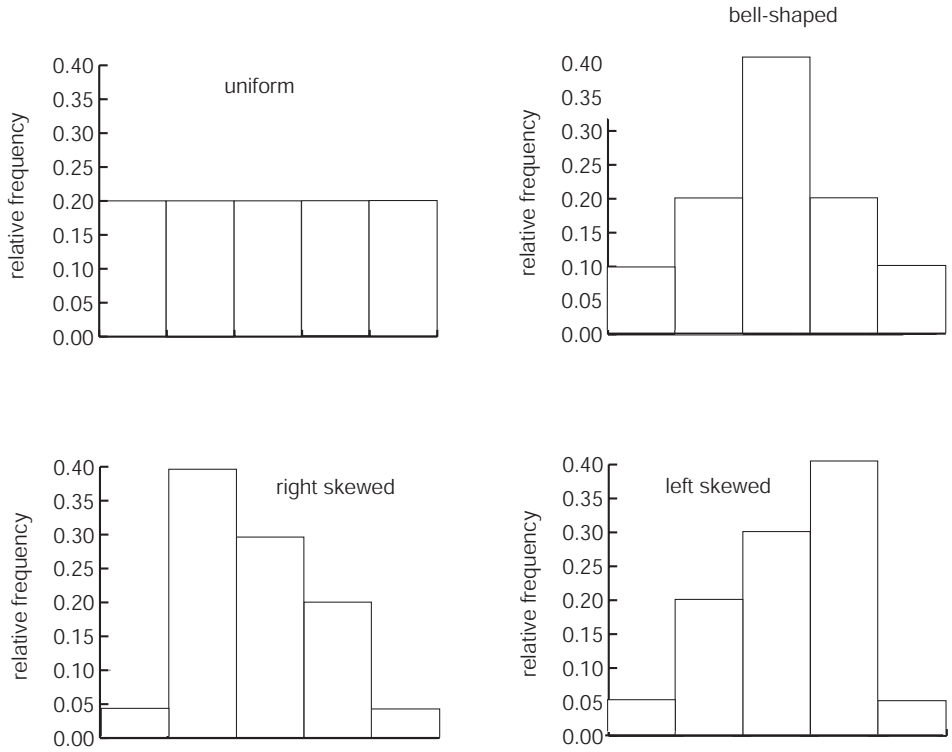

Figure 2.7 (Shapes of Histograms)

2.3 Additional Displays of Quantitative Data

One additional display of quantitative data is an *ogive*, which is a graph of *cumulative* frequency or *cumulative* relative frequency data.

In addition to ogives (pronounced "o-jives"), text discusses other graphs, including frequency polygons and time series, which are *not* covered in this workbook. Although not covered in workbook, this other (simple) material on graphs could appear on either a quiz or homework.

Exercise 2.3 (Additional Displays of Quantitative Data)

1. *Ogive (cumulative relative frequency graph): patient ages.* Consider both cumulative distribution table and ogive for age of patient data below.

32, 37, 39, 40, 41, 41, 41, 42, 42, 43,
44, 45, 45, 45, 46, 47, 47, 49, 50, 51

class	frequency	relative frequency	cumulative relative frequency
30-34	1	0.05	0.05
35-39	2	0.10	$0.05 + 0.10 = 0.15$
40-44	8	0.40	$0.15 + 0.40 = 0.55$
45-49	7	0.35	0.90
50-54	2	0.10	1.00

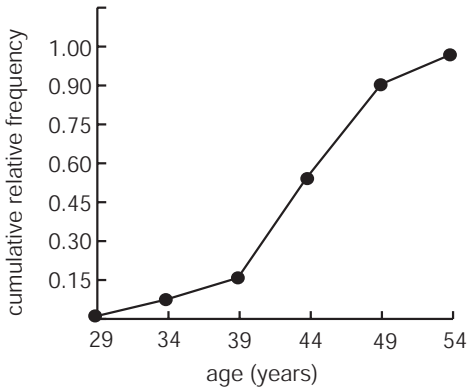

Figure 2.8 (Ogive for Age Data)

(a) Cumulative relative frequency below age 35 is **0.05** / **0.10** / **0.15**.

(b) Cumulative relative frequency below age 40 is **0.05** / **0.10** / **0.15**.

(c) Cumulative relative frequency below age 55 is **0.70** / **0.95** / **1.00**.

(d) **True** / **False**. Cumulative relative frequency for a class is equal to relative frequency for that class plus all of previous classes.

2. *Ogive: patient ages again.* Consider cumulative distribution table below.

class	frequency	cumulative relative frequency
30-34	2	0.10
35-39	5	0.35
40-44	6	0.65
45-49	3	0.80
50-54	4	1.00

(a) Cumulative relative frequency below age 40 is **0.35** / **0.45** / **0.55** / **0.65**.

(b) Cumulative relative frequency *above* age 39 is **0.35** / **0.45** / **0.55** / **0.65**.

2.4 Graphical Misrepresentations of Data

Data can be misrepresented in different ways, often related to improper manipulation of scale.

Exercise 2.4 (Graphical Misrepresentations of Data)

1. *Unequal widths.*

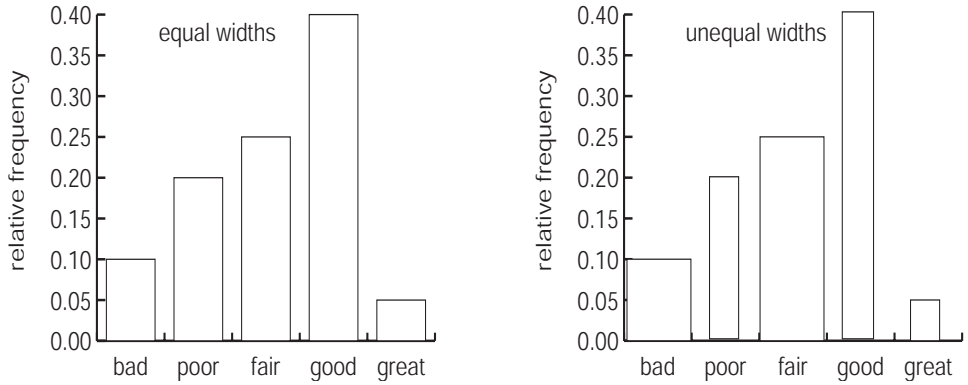

Figure 2.9 (Unequal widths)

Bar graph on right possibly misleading because it seems "bad" and "fair" health occur **less frequently than** / **as frequently as** / **more frequently than** other categories.

2. *Truncated and adjusted scale.*

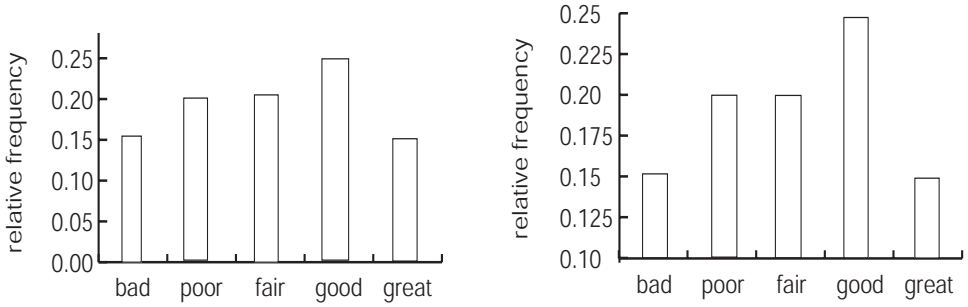

Figure 2.10 (Truncated and adjusted scale)

Bar graph on right possibly misleading because it seems **greater** / **same** / **lesser** difference between categories.