BIOS 662 Fall 2018

Descriptive Statistics

David Couper, Ph.D.

david_couper@unc.edu

or

couper@bios.unc.edu

https://sakai.unc.edu/portal

Descriptive Statistics

- **Types of variables**
- Measures of location
- Measures of spread, shape
- Data displays

Types of Variables

- Definition 3.1. A *variable* is a quantity that may vary from object to object
- Definition 3.2. A *sample* or *data set* is a collection of values of one or more variables.
- Types of variables
  - Quantitative variable – intrinsically numeric
    e.g. age, height, counts
  - Qualitative (categorical) – intrinsically non-numeric
    e.g. gender, state, country

Types of Variables

- Qualitative (categorical) - intrinsically non-numeric
  - Binary, dichotomous

    e.g., alive/dead, female/male
  - Ordinal - natural ordering

    e.g., diagnosis (certain, probable, unlikely, ...)

    e.g., attitude (strongly agree, agree, neutral, ...)
  - Nominal - no natural ordering

    e.g., religion, race
- In recording qualitative data (and using them in analyses), numeric values may be assigned
- Some "values" may have special meaning, such as missing, N/A, unknown

Descriptive Statistics

- Types of variables
- **Measures of location**
- Measures of spread, shape
- Data displays

Definition 3.10. A *statistic* is a numerical characteristic of a sample

Measures of Location

- (Arithmetic) Mean
- Percentiles
- Median
- Mode
- Geometric mean

Arithmetic Mean

- Data:

$$x_1, x_2, \ldots, x_n$$

- Mean:

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- Example

  Duration of hospital stay in days:

$$x_1 = 5,\ x_2 = 10,\ x_3 = 6,\ x_4 = 11$$

  Mean: $\bar{x} = \frac{1}{4}(5 + 10 + 6 + 11) = \frac{32}{4} = 8$

Reporting of Decimals

- Report mean with one more significant digit than the observations
- Example:

  If x is measured in whole numbers and $\bar{x} = 6.345$, report $\bar{x} = 6.3$

Properties of the Mean

- Let c be any constant
- If

$$y_i = x_i + c \quad \text{for} \quad i = 1, 2, 3, \ldots, n,$$

then

$$\bar{y} = \bar{x} + c$$

- If

$$y_i = cx_i \quad \text{for} \quad i = 1, 2, 3, \ldots, n,$$

then

$$\bar{y} = c\bar{x}$$

Properties of the Mean – Example

- A sample of birth weights in a hospital found

$$\bar{y} = 3166.9 \text{ grams}$$

- 1 oz = 28.35 g
- Therefore the mean in ounces is

$$\bar{x} = \frac{\bar{y}}{28.35} = 111.7$$

Order Statistics

- Data: $x_1, x_2, \ldots, x_n$
- Order data from smallest to largest

$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$$

- $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ are *order statistics*
- Note

$$x_{(1)} = \min\{x_1, x_2, \ldots, x_n\}$$
$$x_{(n)} = \max\{x_1, x_2, \ldots, x_n\}$$

- $(1), (2), \ldots, (n)$ are the *ranks* of the observations
- The textbook defines a *half rank* such that the value at the half rank $(i + 1/2)$ is $(x_{(i)} + x_{(i+1)})/2$

Example

- Duration of hospital stay in days:

$$x_1 = 5,\ x_2 = 10,\ x_3 = 6,\ x_4 = 11$$

- Order statistics:

$$x_{(1)} = 5,\ x_{(2)} = 6,\ x_{(3)} = 10,\ x_{(4)} = 11$$

$$x_{(2.5)} = (6+10)/2 = 8$$

Percentiles

- Intuitive definition: the *x percentile* is such that *x%* of the observations are less than that value
- Also known as sample *quantile*

Percentiles: Text Definition

- The $(p \times 100)^{\text{th}}$ percentile of a sample of size n is

$$\hat{\zeta}_p = \begin{cases} y_{(np+p)} & \text{if } np+p \text{ is an integer} \\ \{y_{(\lfloor np+p \rfloor)} + y_{(\lceil np+p \rceil)}\}/2 & \text{otherwise} \end{cases}$$

  for $0 < p < 1$

- Note:

  $\lfloor y \rfloor$ is the greatest integer $\leq y$; (the *floor* function)

  $\lceil y \rceil$ is the smallest integer $\geq y$; (the *ceiling* function)

- Compare with Definition 3.11 of text: Pth *percentile* is the value with rank $(P/100)(1+n)$. If this rank is not an integer, it is rounded to the nearest half rank.

Percentiles: General Form

- General form (Hyndman and Fan, *Am Stat* 1996)

$$\hat{\zeta}_p = (1-\gamma)y_{(j)} + \gamma y_{(j+1)}$$

  where $j = \lfloor pn + m \rfloor$ for some $m \in \mathbb{R}$ and $0 \leq \gamma \leq 1$.

- Let $g = pn + m - j$
- If $m = p$ then $j = \lfloor pn + p \rfloor$ and we set

$$\gamma = \begin{cases} 0 & \text{if } g = 0 \\ \\ 1/2 & \text{if } g > 0 \end{cases}$$

  we recover the text definition

Percentiles: Software

- SAS PROC UNIVARIATE: 5 definitions of percentile
- R: 9 definitions

R "quantile()" Function

```
> ?quantile
quantile                package:stats                R Documentation

Sample Quantiles

Description:

     The generic function 'quantile' produces sample quantiles
     corresponding to the given probabilities. The smallest observation
     corresponds to a probability of 0 and the largest to a probability
     of 1.

Usage:

     quantile(x, ...)

     ## Default S3 method:
     quantile(x, probs = seq(0, 1, 0.25), na.rm = FALSE,
              names = TRUE, type = 7, ...)
```

```
Arguments:

      x: numeric vectors whose sample quantiles are wanted.

  probs: numeric vector of probabilities with values in [0,1].

  na.rm: logical; if true, any 'NA' and 'NaN''s are removed from 'x'
         before the quantiles are computed.

  names: logical; if true, the result has a 'names' attribute.  Set to
         'FALSE' for speedup with many 'probs'.

   type: an integer between 1 and 9 selecting one of the nine quantile
         algorithms detailed below to be used.

    ...: further arguments passed to or from other methods.

Types:

    'quantile' returns estimates of underlying distribution quantiles
    based on one or two order statistics from the supplied elements in
    'x' at probabilities in 'probs'.  One of the nine quantile
    algorithms discussed in Hyndman and Fan (1996), selected by
    'type', is employed.
```

Percentiles: Class Definition

- The $(p \times 100)^{\text{th}}$ percentile of a sample:

$$\hat{\xi}_p = \begin{cases} y_{(\lfloor np \rfloor + 1)} & \text{if } np \text{ is not an integer} \\ \{y_{(np)} + y_{(np+1)}\}/2 & \text{if } np \text{ is an integer} \end{cases}$$

  for $0 < p < 1$

- Definition 2 of R / Hyndman and Fan: $m = 0$, so $j = \lfloor pn \rfloor$, $g = pn - j$, and

$$\gamma = \begin{cases} 1 & \text{if } g > 0 \\ 1/2 & \text{if } g = 0 \end{cases}$$

- Definition 5 of SAS

Example

- Suppose $n = 278$ and we want the 75th percentile

$$np = 278 \times 0.75 = 208.5$$

so

$$\hat{\xi}_{0.75} = x_{(209)}$$

- R

```
> x <- 1:278
> quantile(x,0.75,type=2)
75%
209
```

Example: SAS

```
data;
  infile "C:\BIOS662\2018fall\percentile.txt";
  input x;

proc univariate; var x; run;
```

```
The UNIVARIATE Procedure
Variable:  x

Quantiles (Definition 5)

Quantile      Estimate

75% Q3           209.0
50% Median       139.5
25% Q1            70.0
10%               28.0
5%                14.0
1%                 3.0
0% Min             1.0
```

Median

- The sample median is the 50th percentile

$$\hat{\xi}_{0.5} = \begin{cases} y_{(\frac{n+1}{2})} & \text{if } n \text{ is odd} \\ \{y_{(n/2)} + y_{(n/2+1)}\}/2 & \text{if } n \text{ is even} \end{cases}$$

  for $0 < p < 1$

- Example
  - Duration of hospital stay in days:

$$x_1 = 5,\ x_2 = 10,\ x_3 = 6,\ x_4 = 11$$

  - Median:

$$\hat{\xi}_{0.5} = \{x_{(2)} + x_{(3)}\}/2 = (6+10)/2 = 8$$

Mode

- The mode is the most frequently occurring value in the data set
- Example:

  If $x_1 = 5,\ x_2 = 11,\ x_3 = 6,\ x_4 = 11$

  then the mode is 11

Geometric Mean

- Data: $x_1, x_2, \ldots, x_n > 0$
- The geometric mean of x is

$$\bar{x}_g = (x_1 \cdot x_2 \cdots x_n)^{1/n}$$

- Let $y_i = \log(x_i)$ for $i = 1, 2, \ldots, n$. Then

$$\bar{x}_g = \exp(\bar{y})$$

- $\bar{x}_g$ is used when data are of the form c^k
- Example:
  Suppose $x_1 = 10$ and $x_2 = 0.1$
  Then $\bar{x}_g = 1$

Comments

- The mean is the most often used measure
- The median is better if there are influential observations (it is more robust to extreme values)
- The mode is rarely used (exception: nominal data)

Example

- Duration of hospital stay in days:

$$x_1 = 5,\ x_2 = 10,\ x_3 = 6,\ x_4 = 11$$

$$\hat{\xi}_{0.5} = \bar{x} = 8, \quad \bar{x}_g = 7.6$$

- Alter last observation:

$$x_1 = 5,\ x_2 = 10,\ x_3 = 6,\ x_4 = 50$$

$$\hat{\xi}_{0.5} = 8, \quad \bar{x} = 17.7, \quad \bar{x}_g = 11.1$$

Descriptive Statistics

- Types of variables
- Measures of location
- **Measures of spread, shape**
- Data Displays

Measures of Spread, Shape

- Range
- Variance and standard deviation
- Interquartile range
- Skewness, kurtosis

Range

- Range:

$$r_a = x_{(n)} - x_{(1)}$$

- Easy to calculate
- Sensitive to unusual observations (outliers)
- Usually, the larger n is, the larger r_a

Sample Variance and Standard Deviation

- Want to measure deviation from mean
- Sample variance

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)$$

- Sample standard deviation

$$s = \sqrt{s^2}$$

Sample Variance and Standard Deviation

- An alternative form of the sample variance is

$$s_1^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

- Can show s^2 is unbiased for population variance σ^2, however

$$E(s_1^2) = \sigma^2 - \frac{\sigma^2}{n}$$

- van Belle *et al.* argue for s^2 based on degrees of freedom (d.f.) (Note 3.5)

Sample Standard Deviation

- The units of s are the same as the units of x_i
- If s is large, the data are spread over a wide range
- The textbook recommends reporting the standard deviation with two more significant digits than the original observations

Properties of the Standard Deviation

- If c is a constant and

$$y_i = x_i + c,$$

then

$$s_y = s_x$$

- If

$$y_i = cx_i$$

then

$$s_y = cs_x$$

Some Approximations

- The interval $\bar{x} \pm s$ usually contains approximately 68% of the observations
- The interval $\bar{x} \pm 2s$ usually contains approximately 95% of the observations
- Approximate s by

$$s \approx \frac{\hat{\xi}_{0.75} - \hat{\xi}_{0.25}}{1.35}$$

- Note

$$\hat{\xi}_{0.75} - \hat{\xi}_{0.25}$$

  is called the *interquartile range*

Comments

- $\hat{\xi}_{0.25}$ is the lower quartile
- $\hat{\xi}_{0.75}$ is the upper quartile
- Epidemiologists often use quantiles to categorize a continuous variable
- Splitting a variable at $\hat{\xi}_{0.25}$, $\hat{\xi}_{0.5}$, $\hat{\xi}_{0.75}$ yields four (roughly) equally-sized groups
- Statisticians use quartile to refer to one of the cut-points
- Epidemiologists usually use it to refer to the categories
- Tertiles (for three groups) or quintiles (for five groups) are also often used

Symmetry and Skewness

- Informally, define *symmetry* to indicate having a uniform or even distribution about the mean
- If a distribution is symmetric,

$$\text{mean} = \text{median}$$

- Data that are not symmetric are said to be *skewed*
- *Skewness* is a measure of the degree to which a data set is skewed

Skewness

- Define rth sample moment about the mean

$$m_r = \frac{\sum_i (y_i - \bar{y})^r}{n} \quad \text{for } r = 1, 2, 3, \ldots$$

- Text definition of sample skewness:

$$a_3 = \frac{m_3}{(m_2)^{3/2}} = \frac{\sum_i (y_i - \bar{y})^3 / n}{\{\sum_i (y_i - \bar{y})^2 / n\}^{3/2}} = \sqrt{n} \frac{\sum_i (y_i - \bar{y})^3}{\{\sum_i (y_i - \bar{y})^2\}^{3/2}}$$

- Typo in text on page 51
- In SAS PROC UNIVARIATE need to use the option VARDEF=N to obtain a_3

Interpretation?

- Text:

  "skewed to the right if the mean is greater than the mode"

  "Values of $a_3 > 0$ indicate ... skewness to the right"

- However, for $\{0, 2, 2, 3, 4\}$

  $\bar{x} = 2.2$

  mode $= 2$

  skewness $= -0.37$

Alternative Definitions

- Another definition of skewness:

$$b_3 = \frac{n\sqrt{n-1}}{n-2} \frac{\sum_i (y_i - \bar{y})^3}{\{\sum_i (y_i - \bar{y})^2\}^{3/2}}$$

- b_3 is the default in SAS

- Many more definitions; cf. Joanes and Gill (JRSS D 1998)

Kurtosis

- *Kurtosis* is a measure of the flatness or peakedness of a distribution; degree of archedness; thickness of tails
- Text definition of *sample* kurtosis:

$$a_4 = \frac{m_4}{(m_2)^2} = \frac{\sum_i (y_i - \bar{y})^4 / n}{\{\sum_i (y_i - \bar{y})^2 / n\}^2} = n \frac{\sum_i (y_i - \bar{y})^4}{\{\sum_i (y_i - \bar{y})^2\}^2}$$

- Typo in text on page 51

Kurtosis: SAS

- In SAS PROC UNIVARIATE need to use the option VARDEF=N to obtain a_4

$$b_4 = \frac{1}{n}\sum\left(\frac{y_i - \bar{y}}{s_1}\right)^4 - 3$$

i.e.,

$$b_4 = \frac{\sum(y_i - \bar{y})^4/n}{s_1^4} - 3$$

i.e.,

$$b_4 = \frac{m_4}{(m_2)^2} - 3 = a_4 - 3$$

- Why minus 3?

Descriptive Statistics

- Types of variables
- Measures of location
- Measures of spread, shape
- **Data displays**

Data displays

- Simplest form is a line listing – essentially a tabular display with each line containing the data for a single person (or other unit of observation)
- A *frequency table* gives the frequency of observations within a set of ordered intervals
  - Intervals should be mutually exclusive and exhaustive
  - 8 to 10 intervals is usually sufficient
  - With the exception of the end intervals, the length of the intervals should be constant

Frequency Table – Example: Table 3.6

Blood Pressure	Native Japanese	Generation	
		1st	2nd
≤ 104	218	4	23
106-114	272	23	132
116-124	337	49	290
126-134	362	33	347
136-144	302	41	346
146-154	261	38	202
156-164	166	23	109
> 164	314	52	112
Total	2232	263	1561

Frequency Tables

- Table on previous slide an example of an *empirical frequency distribution*
- Difficult to compare blood pressure distributions because of different sample sizes
- Divide by sample size to get *empirical relative frequency distribution*

ERFD – Example: Table 3.7

Blood Pressure	Native Japanese	Generation	
		1st	2nd
≤ 104	0.098	0.015	0.015
106-114	0.122	0.087	0.085
116-124	0.151	0.186	0.186
126-134	0.162	0.125	0.222
136-144	0.135	0.156	0.222
146-154	0.117	0.144	0.129
156-164	0.074	0.087	0.070
> 164	0.141	0.198	0.072
Total	2232	263	1561

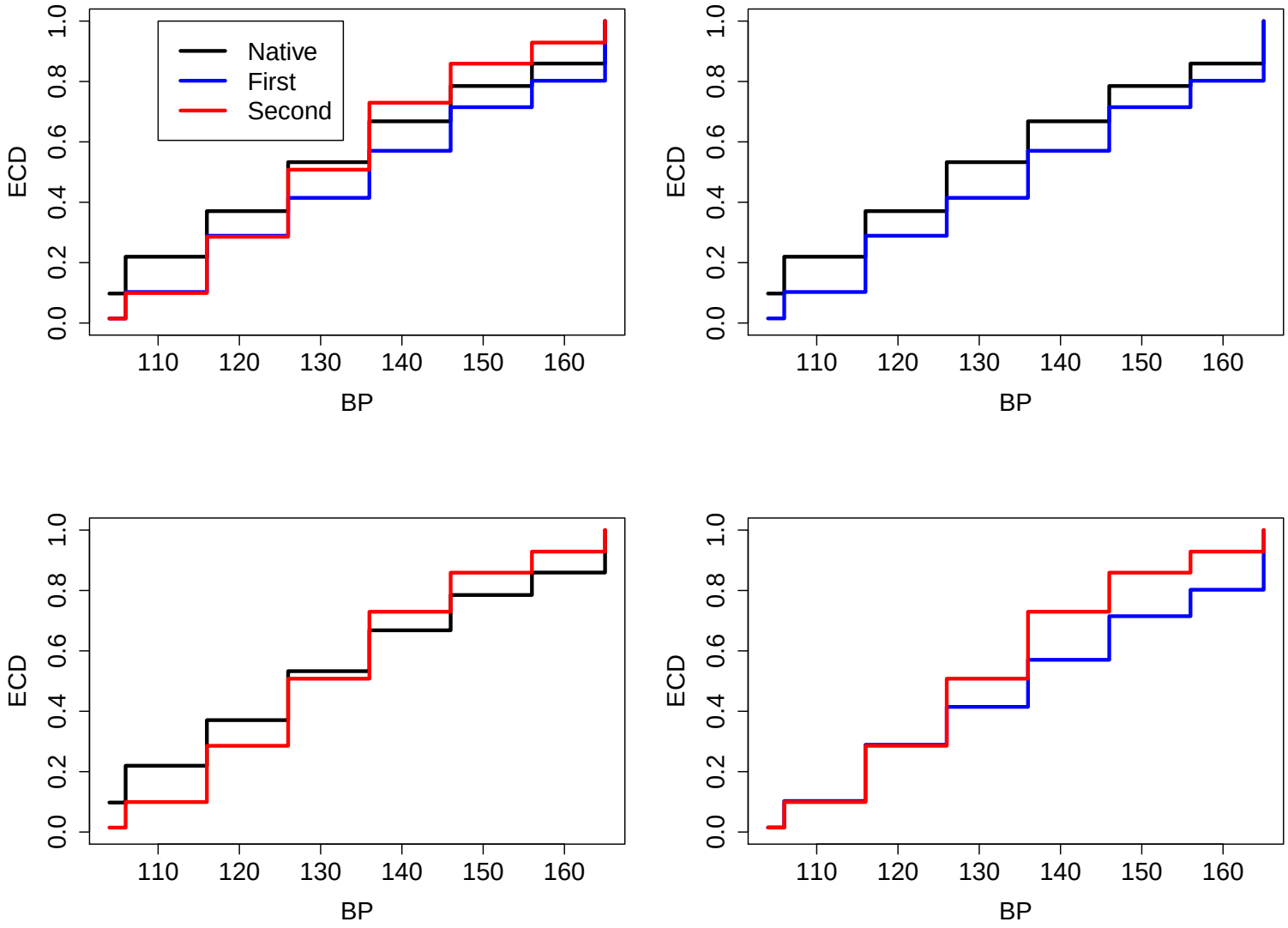
Empirical Distribution Function

- Definition 3.9. The *empirical cumulative distribution* of a variable is a listing of the values of the variable with the proportion of observations less than or equal to each value (cumulative proportion)
- Also known as the *empirical distribution function* (EDF)
- Does not necessarily entail binning (that is, grouping into intervals)

ECD – Example

Blood Pressure	Native Japanese	Generation	
		1st	2nd
≤ 104	0.098	0.015	0.015
≤ 114	0.220	0.103	0.100
≤ 124	0.371	0.289	0.285
≤ 134	0.533	0.414	0.507
≤ 144	0.668	0.570	0.729
≤ 154	0.785	0.715	0.858
≤ 164	0.859	0.802	0.928
$< \infty$	1.000	1.000	1.000
Total	2232	263	1561

ECD – Example


Native
First
Second
ECD
BP
0.0
0.2
0.4
0.6
0.8
1.0
110
120
130
140
150
160

Graphs

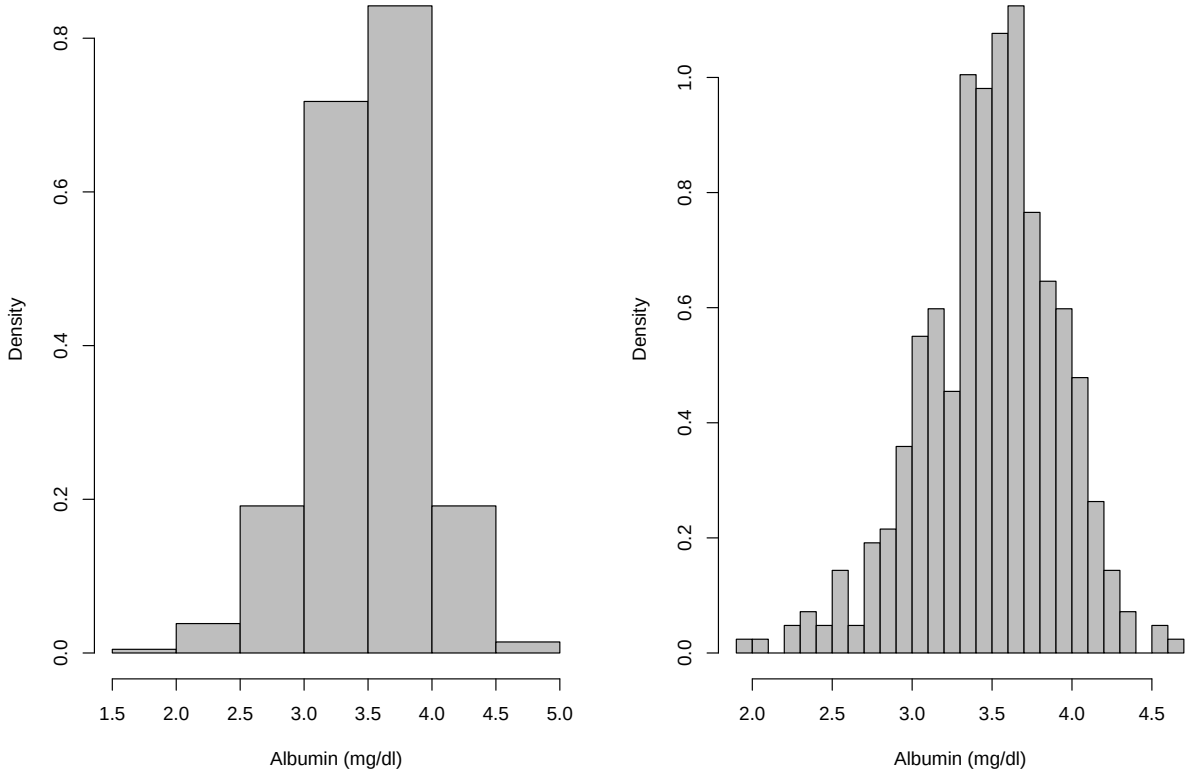
- ECD/EDF
- Histogram
- Stem and leaf plot
- Box plot
- Trellis/conditional plots

Histogram

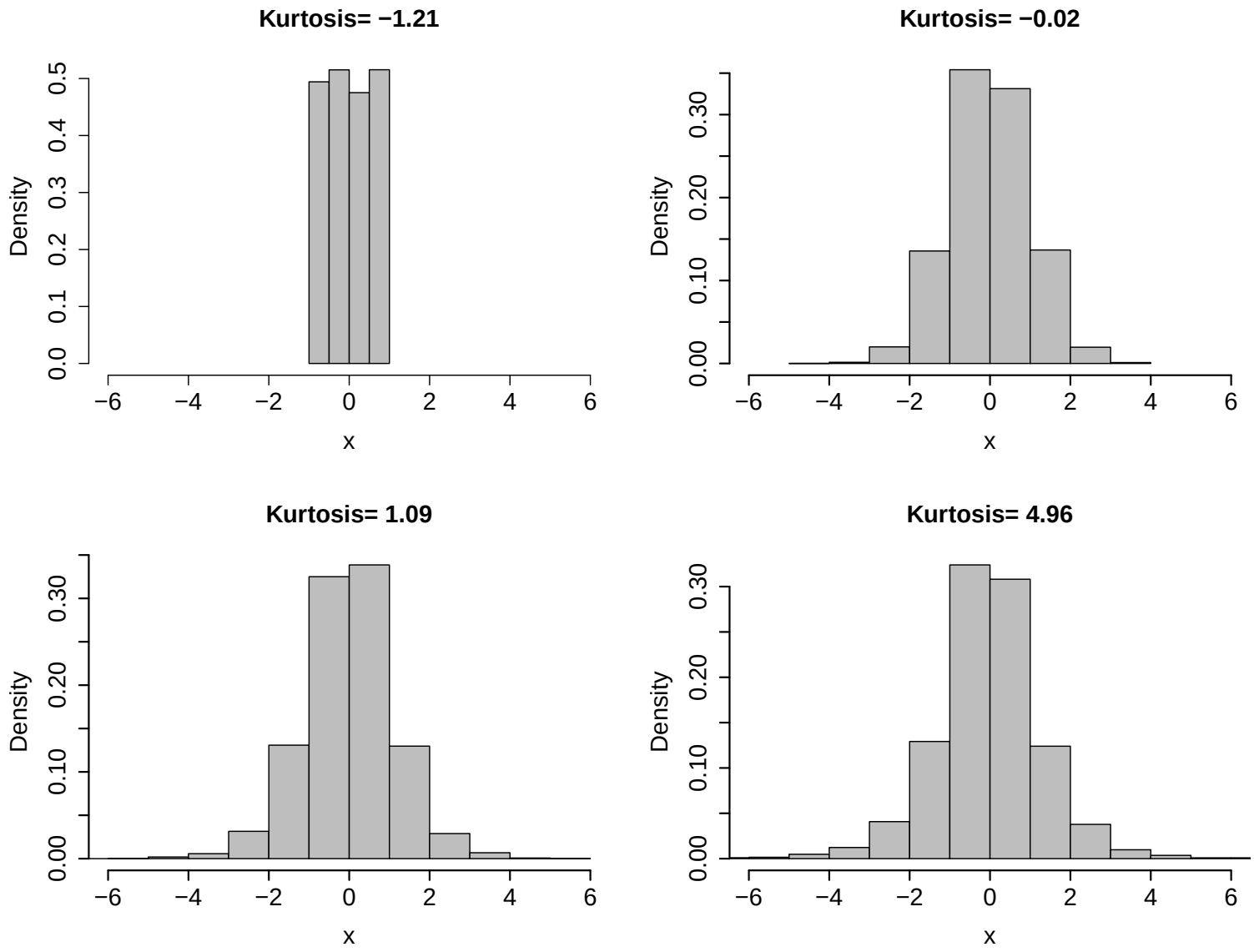
- Data are divided into intervals as in a frequency table
- A histogram is a bar graph with the area of each bar equal to the relative frequency in the interval.
- Can compare histograms from samples of different size
- Intervals need not be the same width
- Consider effect of choice of interval width (Figure 3.1 in Text)

Histogram: Example (Figure 3.1 in text)

```
> par(mfcol=c(1,2))

> hist(liver$albumin,col="gray",xlab="Albumin (mg/dl)",breaks=7,freq=F,main="")

> hist(liver$albumin,col="gray",xlab="Albumin (mg/dl)",breaks=30,freq=F,main="")
```



Histograms with Various Values for Kurtosis



Stem and Leaf Plot

- Stem consists of leading digits
- Leaves consist of last digit
- Example: for x = 496, stem = 49, leaf = 6
- Make a column of stems from smallest to largest
- To the right of each stem, list in a row the leaves, in ascending order.
- Note: there will be one leaf for each observation

Stem and Leaf Plot: Example

```
> stem(liver$albumin)

  The decimal point is 1 digit(s) to the left of the |

  18 | 6
  20 | 0
  22 | 37138
  24 | 3834468
  26 | 048345557
  28 | 0012344799033344566778
  30 | 000011112233344566677788800112222344555678899999
  32 | 000001223334456666669990111123344444455555555555666666777789
  34 | 000000011122222333334445566666788888899900000000111122223334444555555+5
  36 | 00000000111122233333344445555555666677778889999900000002233344445556+2
  38 | 000001112233333345555556779990001223333444556778888999 9
  40 | 001113344678888899990034566789 99
  42 | 022340088
  44 | 022
  46 | 4
```

Stem and Leaf Plot: Example

```
> stem(liver$albumin,width=100)

  The decimal point is 1 digit(s) to the left of the |

  18 | 6
  20 | 0
  22 | 37138
  24 | 3834468
  26 | 048345557
  28 | 0012344799033344566778
  30 | 000011112233344566677788800112222344555678899999
  32 | 000001223334456666669990111123344444455555555555666666777789
  34 | 00000001112222233333444556666678888889990000000011112222333444455555666666777777888889
  36 | 0000000011112223333334444555555566667777888999990000000223334444555666666777778999
  38 | 000001112233333345555556779990001223333444556778888999 9
  40 | 00111334467888889999003456678999
  42 | 022340088
  44 | 022
  46 | 4
```

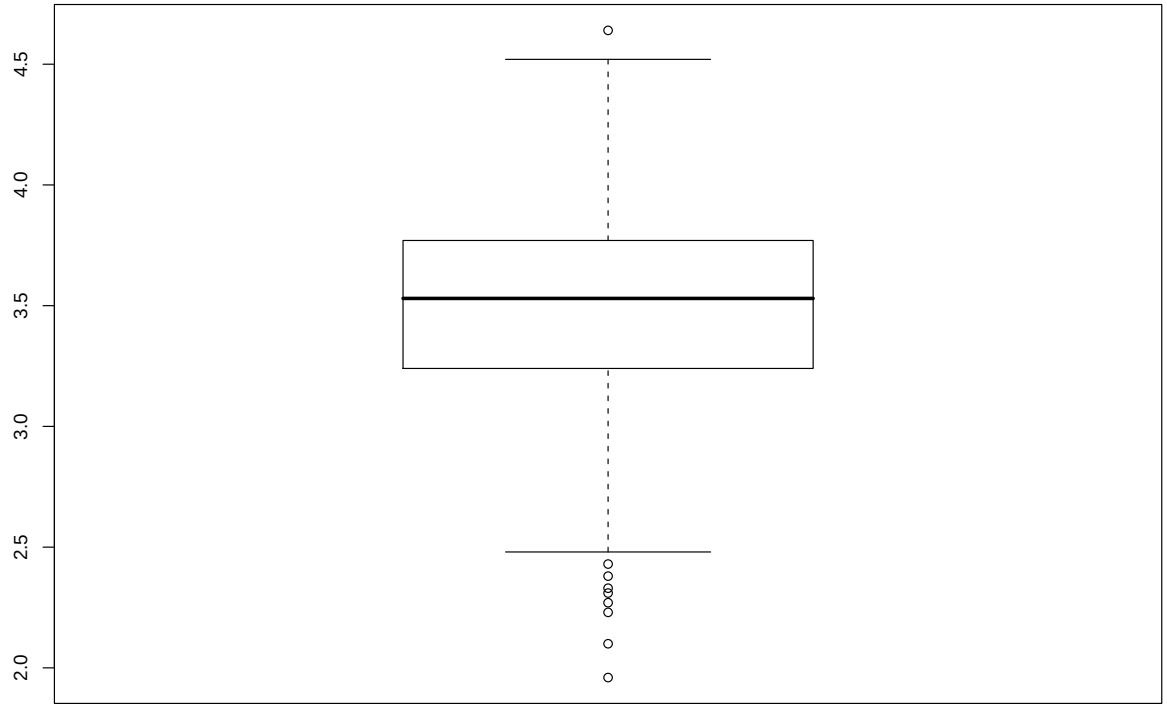
```
> stem(liver$albumin,scale=2)   # scale changes the (vertical) length of the display

  The decimal point is 1 digit(s) to the left of the |

  19 | 6
  20 |
  21 | 0
  22 | 37
  23 | 138
  24 | 38
  25 | 34468
  26 | 048
  27 | 345557
  28 | 0012344799
  29 | 033344566778
  30 | 000011112233445666777888
  31 | 0011222234455567889999
  32 | 000001223334566666999
  33 | 011112334444455555555555666666777789
  34 | 000000011122222333334445566667888888999
  35 | 0000000011112222333444455555666666777777788889
  36 | 00000000111122233333344445555556666777788899999
  37 | 000000022333444455566666777778999
  38 | 000001112233333334555555677999
  39 | 000122333344455677888899999
  40 | 00111334467888889999
  41 | 003456678999
  42 | 02234
  43 | 0088
  44 | 0
  45 | 22
  46 | 4
```

Box Plot

- The top of the box is the 75th percentile ($\hat{\xi}_{0.75}$); the bottom is the 25th percentile ($\hat{\xi}_{0.25}$)
- A line through the box is drawn at the median
- The lines extending out of the box (*whiskers*) may extend to
  - the 90th and 10th percentiles
  - the largest and smallest values
  - largest observation $\leq \hat{\xi}_{0.75}$ + 1.5 x IQR; smallest observation $\geq \hat{\xi}_{0.25} -$ 1.5 x IQR (Text is wrong! cf. Tukey 1977, Chambers *et al.* 1983)
- Data beyond whiskers may be plotted individually

Box Plot: Example Using R

```
> boxplot(liver$albumin)
```



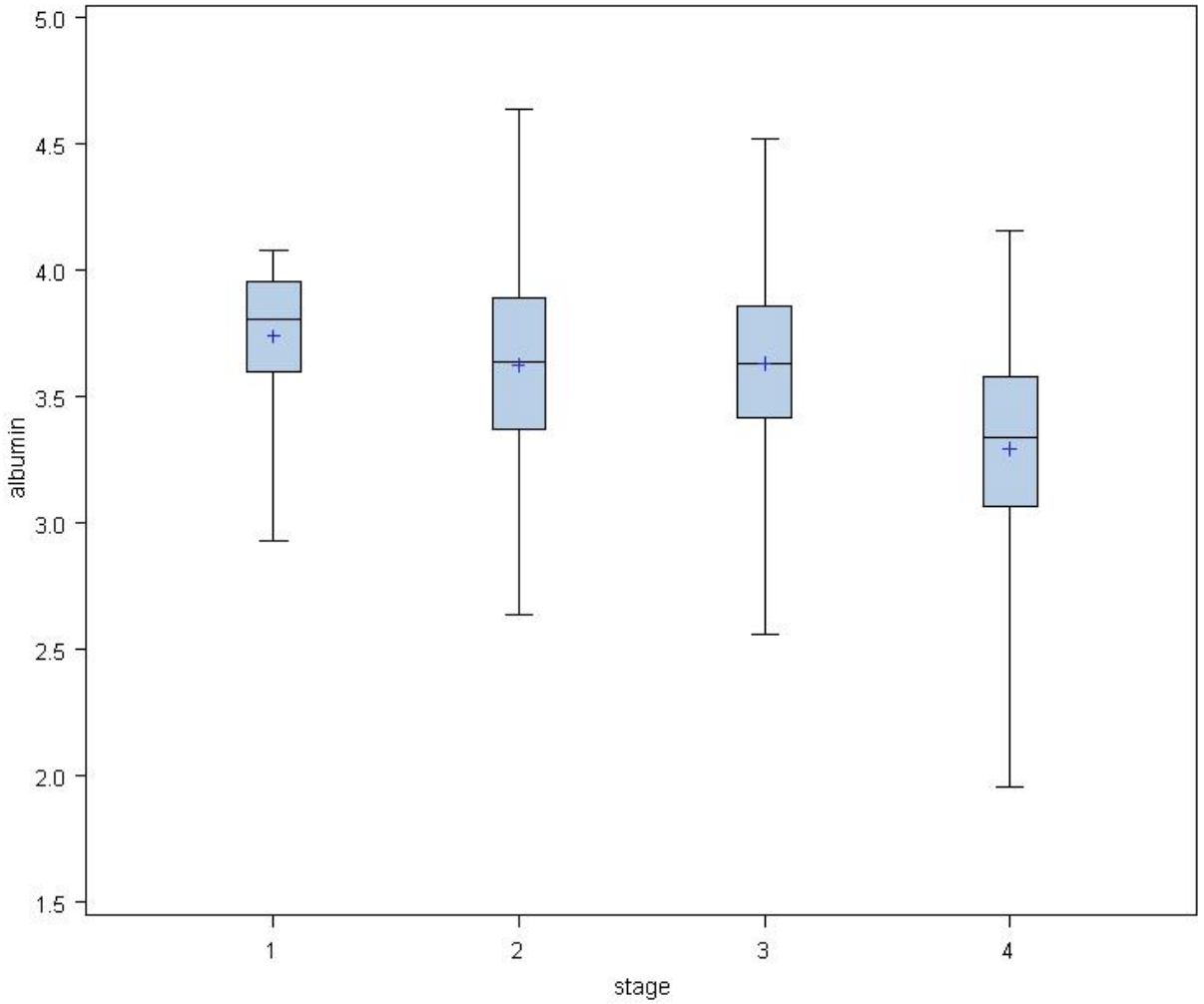
Histogram and Box Plot: Example Using SAS

```
proc univariate plot;
   var albumin;
```

```
                      Histogram                          #  Boxplot
    4.7+*                                                1     0
       .**                                               3     |
       .*****                                            9     |
       .****************                                32     |
       .****************************                    55     |
       .******************************************      82  +-----+
       .*******************************************    85  *--+--*
    3.3+*****************************                  59  +-----+
       .************************                        47     |
       .***********                                      22     |
       .*****                                            9     |
       .****                                             7     0
       .***                                              5     0
       .*                                                1     0
    1.9+*                                                1     0
        ----+----+----+----+----+----+----+----+---
        * may represent up to 2 counts
```

Box Plot: Example Using SAS

```
proc boxplot;
   plot albumin*stage;
```



Box Plot

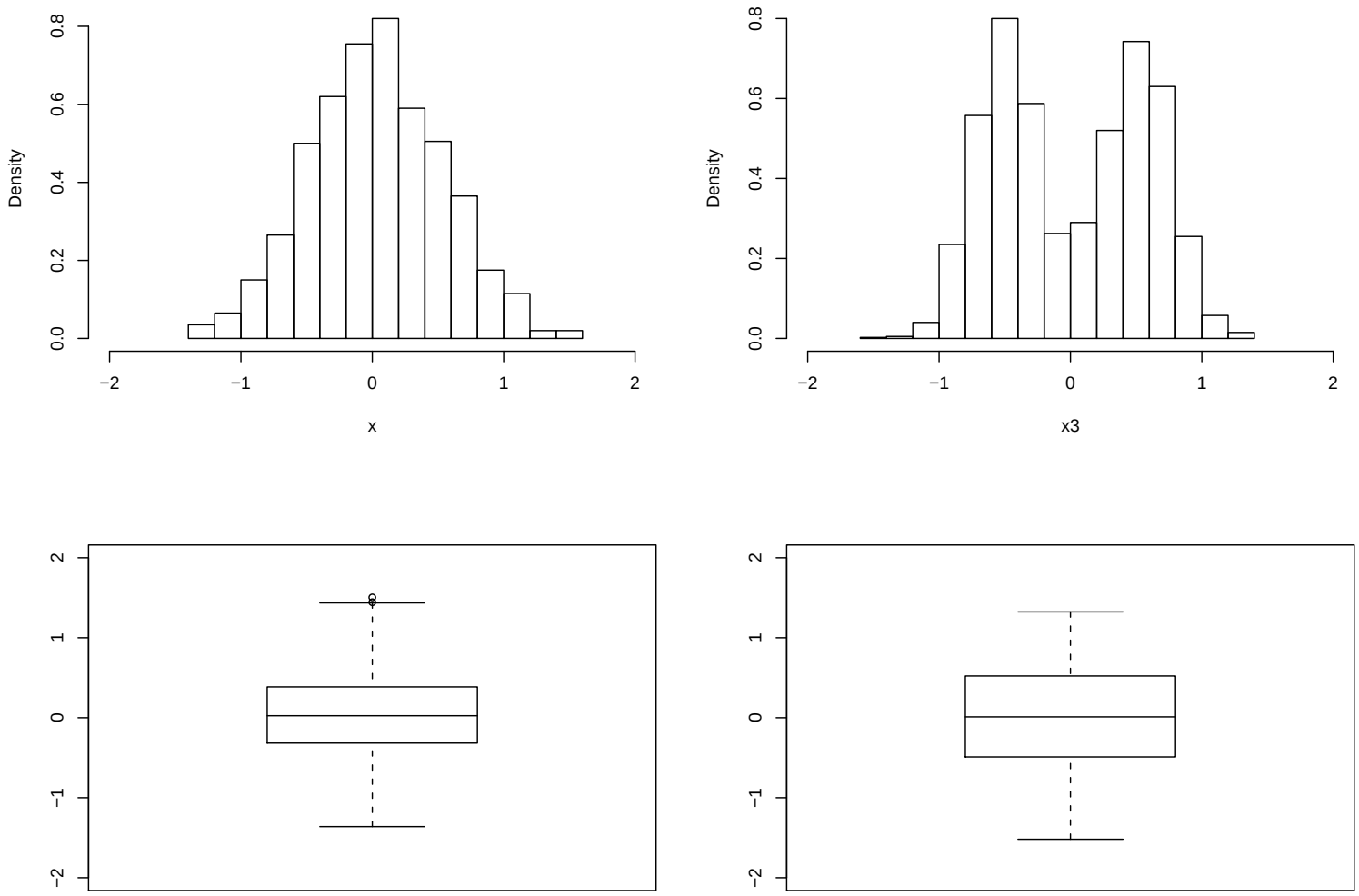
- What proportion of the data should we expect to be between the whiskers?
- If data normally distributed,
  - 95-98% for $6 \leq n \leq 20$,
  - 99% for $n > 20$
  - Ref: Hoaglin *et al.* (JASA 1986)
- Note

$$1.5 \times IQR \approx 1.5(1.35)s \approx 2s$$

so whiskers cover

$$\approx \hat{\xi}_{0.5} \pm 2.68s$$

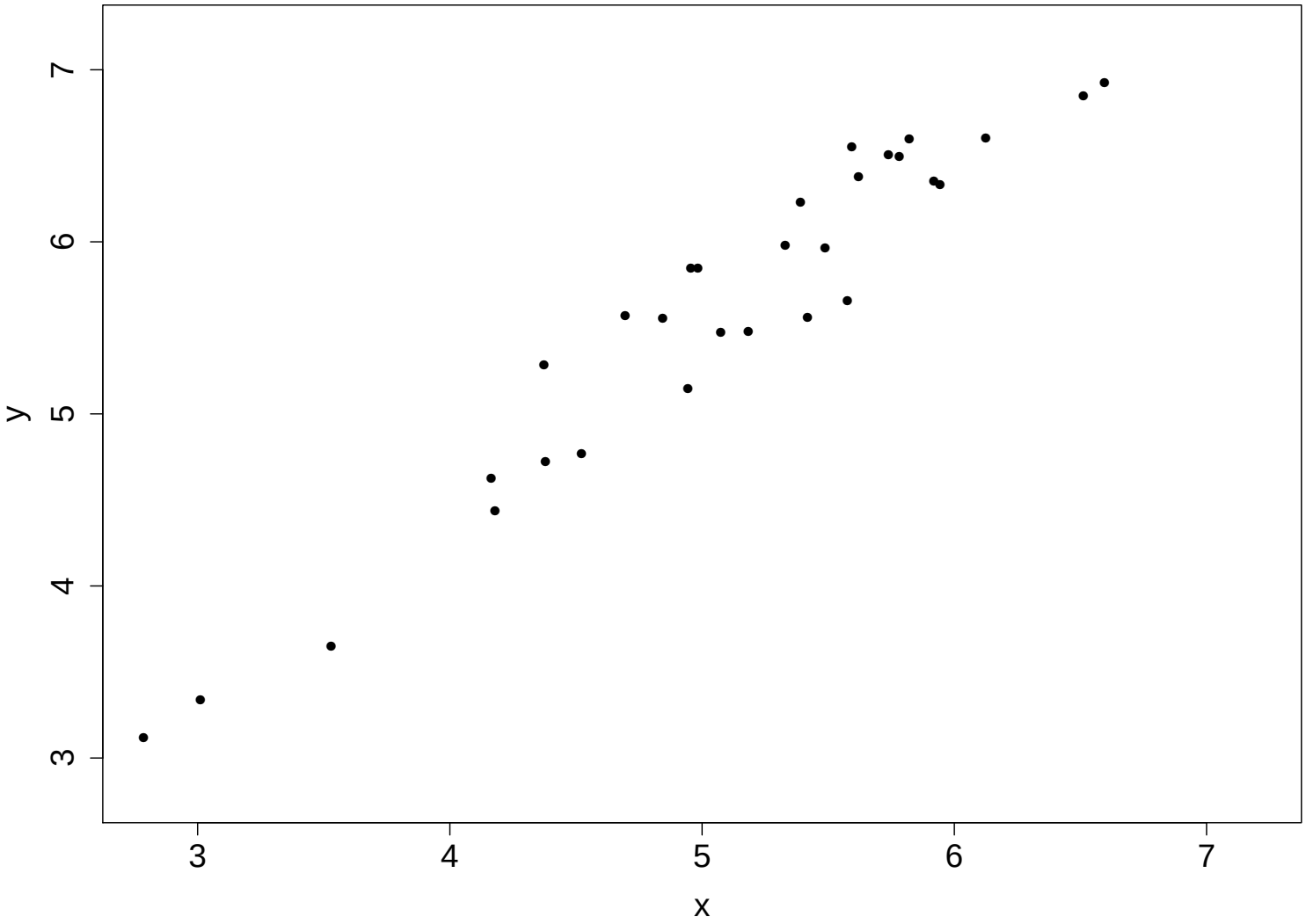
Box Plot and Histogram: Example



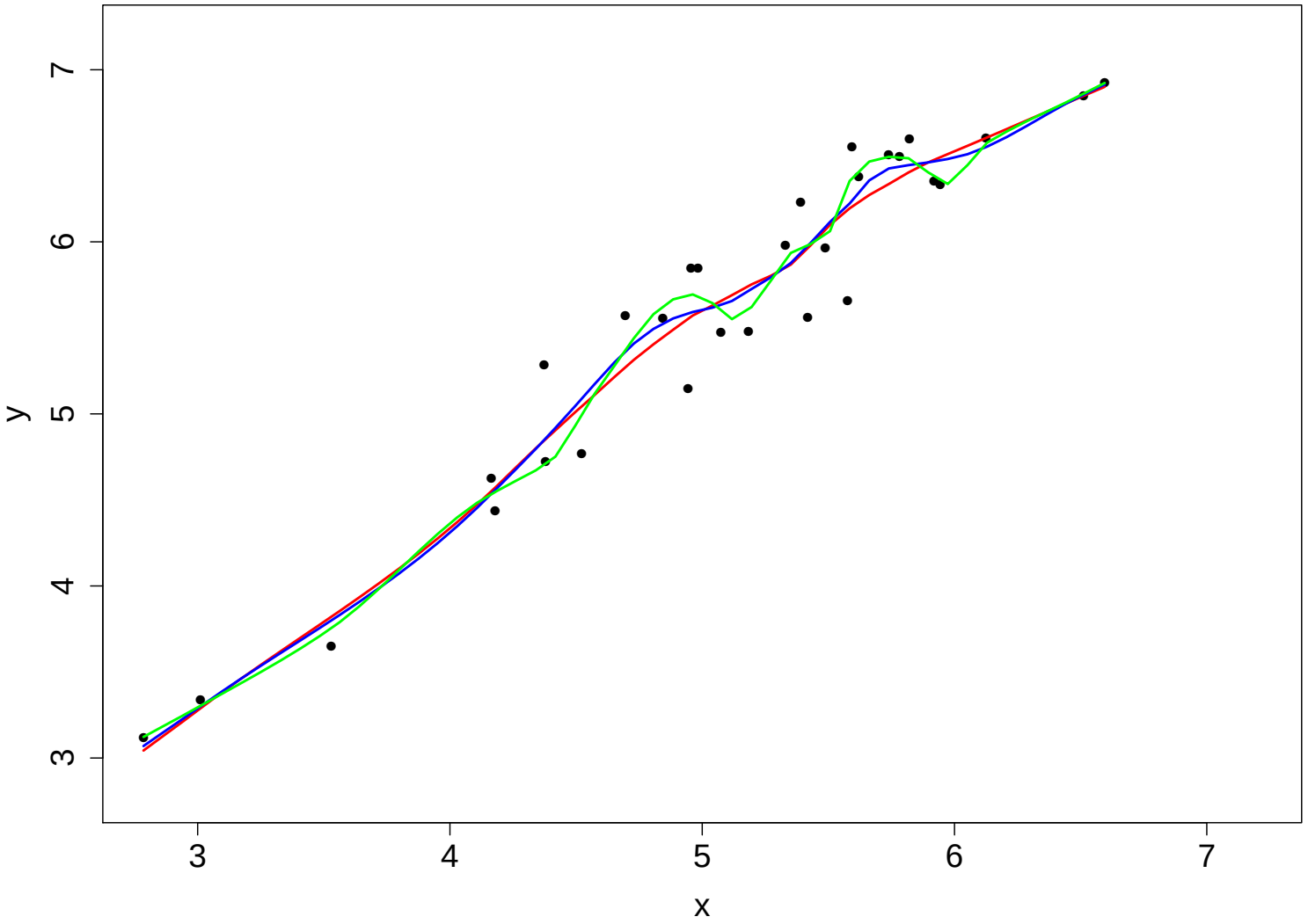
Multivariate Plots

- Describe relationships/associations between more than one variable
- Scatterplots
  - Simple for two variables
  - Add color, symbols for > 2 variables
- Trellis/conditional plots

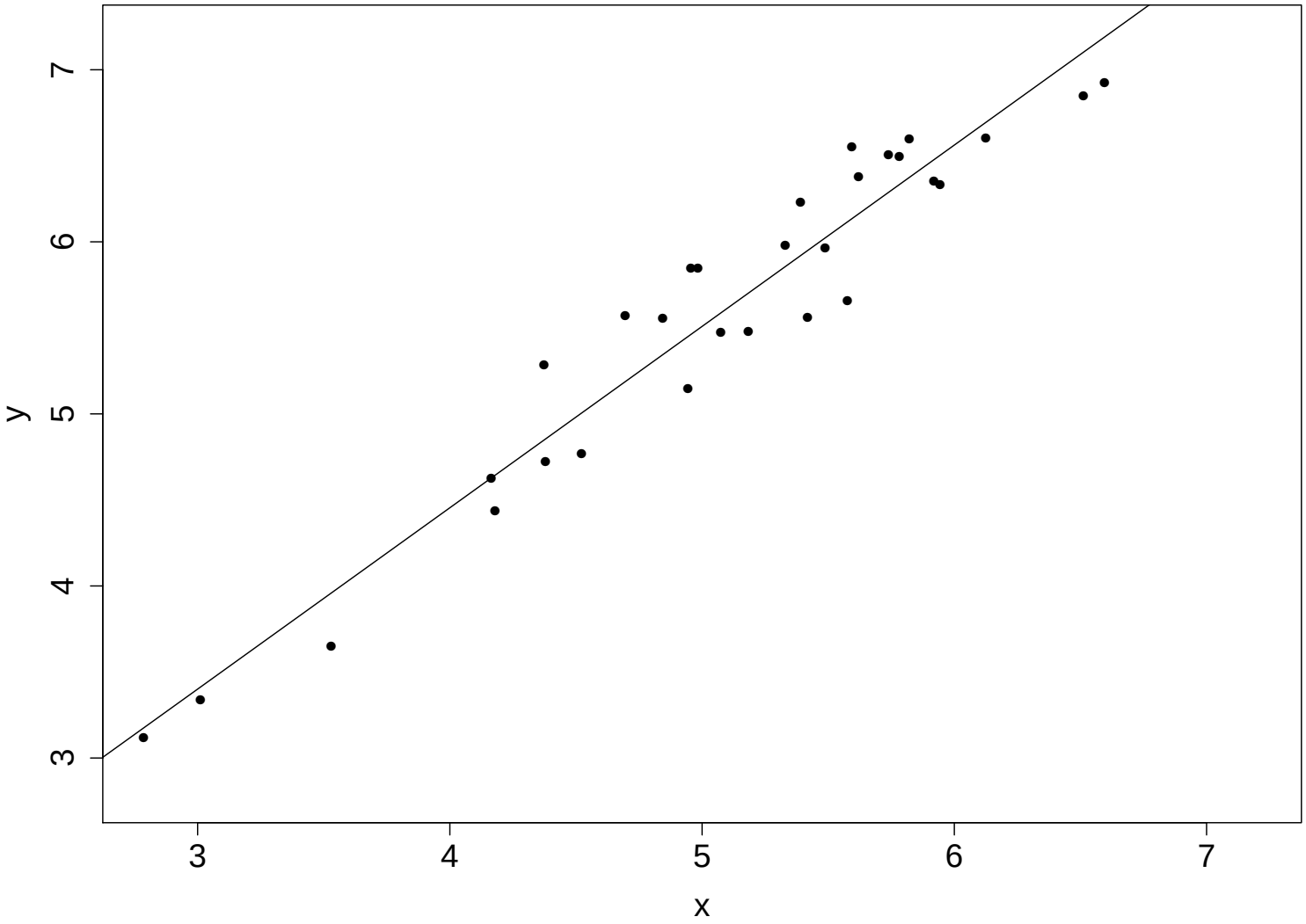
Scatterplot: Example



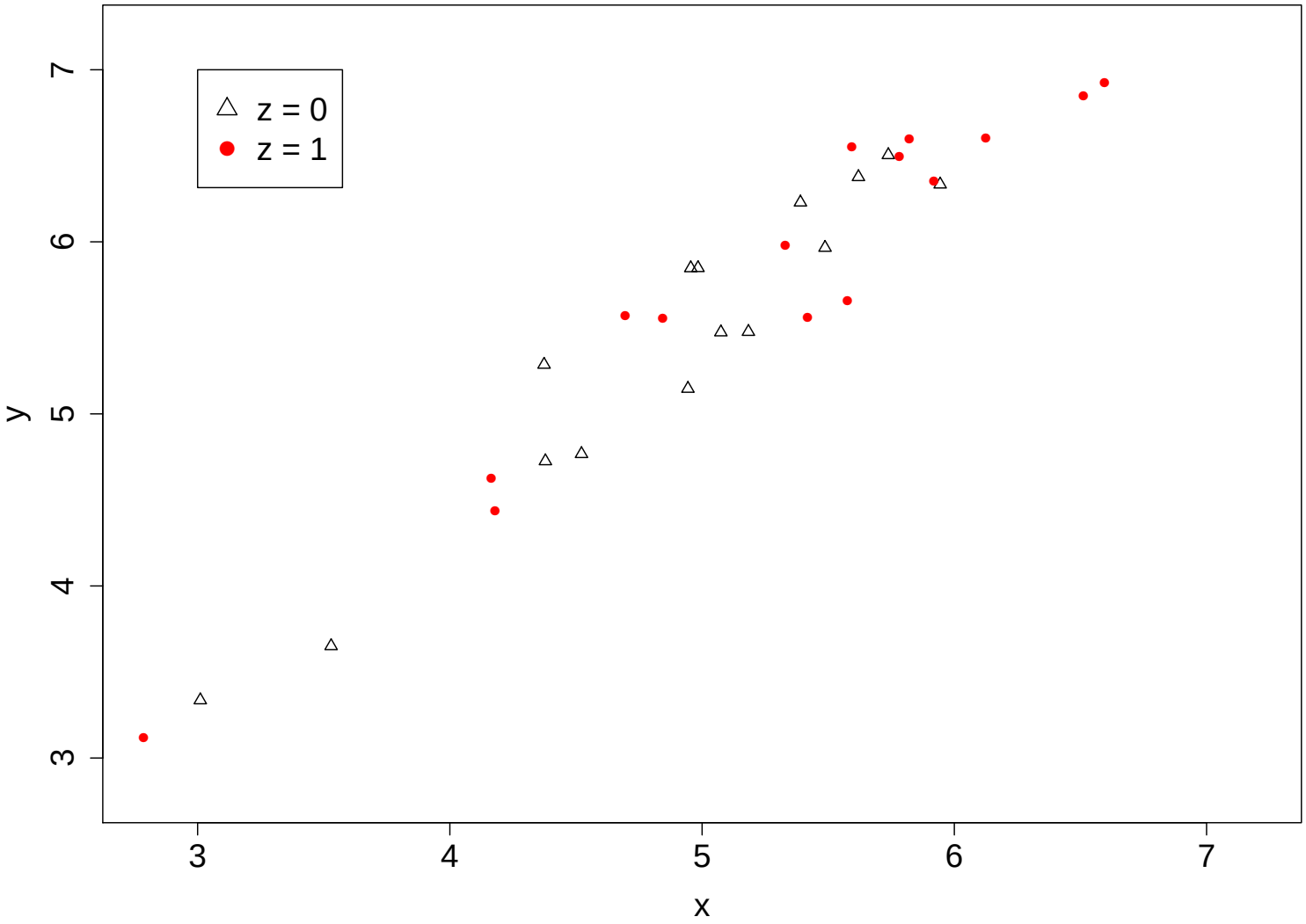
Scatterplot Example cont.



Scatterplot Example cont.



Scatterplot Example cont.



Trellis Plots

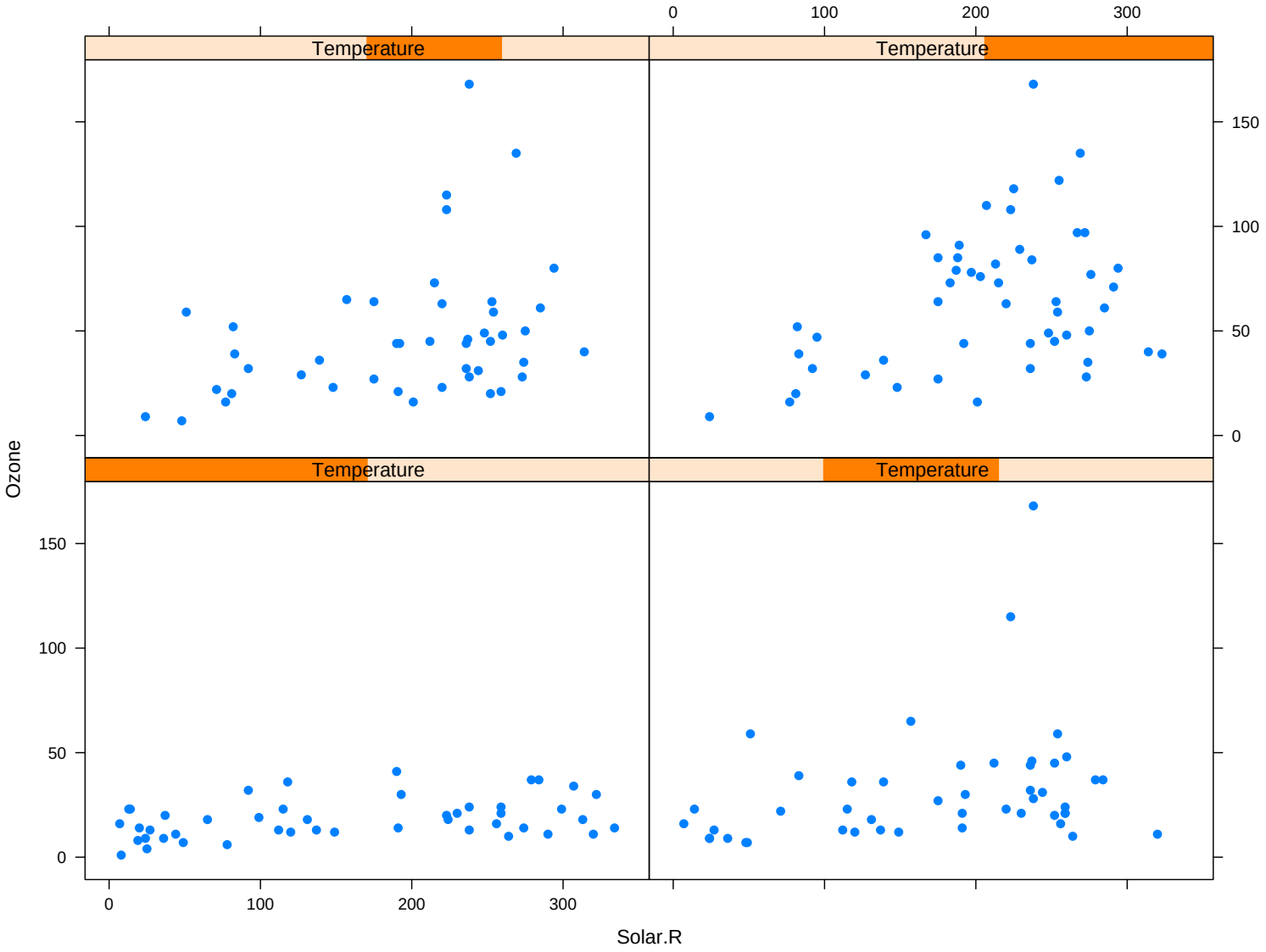


Table or Graph?

- Tables best suited for looking up specific information
- Graphs better for perceiving trends, making comparisons and predictions
- Ref. Gelman *et al.* (*Am Stat* 2002)